

systemd для администраторов

Lennart Poettering (автор)*

Сергей Пташник (русский перевод)†

Данный документ доступен на условиях лицензии [CC-BY-SA 3.0 Unported](#)

28 октября 2017 г.

Содержание

Предисловие автора	3
1 Контроль процесса загрузки	4
2 О службах и процессах	8
3 Преобразование SysV init-скрипта в systemd service-файл	20
4 Убить демона	24
5 Три уровня выключения	25
6 Смена корня	27
7 Поиск виновных	31
8 Новые конфигурационные файлы	35
9 О судьбе /etc/sysconfig и /etc/default	38
10 Экземпляры служб	42
11 Службы с активацией в стиле inetd	46
12 К вопросу о безопасности	50
12.1 Изолирование служб от сети	51
12.2 Предоставление службам независимых каталогов /tmp	52
12.3 Ограничение доступа служб к отдельным каталогам	52
12.4 Принудительное отключение полномочий (capabilities) для служб	53
12.5 Запрет форка, ограничение на создание файлов	54
12.6 Контроль доступа служб к файлам устройств	54
12.7 Прочие настройки	54
13 Отчет о состоянии службы и ее журнал	55
14 Самодокументированный процесс загрузки	56
15 Сторожевые таймеры	58

*Первоисточник (на английском языке) опубликован на сайте автора: <http://0pointer.de/blog/projects>

†Актуальная версия перевода доступна на личной странице переводчика: <http://www2.kangran.su/~nanz/pub/s4a/> и на GitHub: <https://github.com/nanz1024/s4a/>

16	Запуск getty на последовательных (и не только) консолях	61
16.1	Виртуальные консоли	61
16.2	Последовательные консоли	62
17	Работа с Journal	63
17.1	Сохранение логов на диск	64
17.2	Основы	64
17.3	Контроль доступа	65
17.4	Отслеживание логов в реальном времени	65
17.5	Простейшие методы выборки записей	65
17.6	Продвинутые методы выборки	66
17.7	И немного магии	67
18	Управление ресурсами с помощью cgroups	68
18.1	Процессор	69
18.2	Отслеживание использования ресурсов	70
18.3	Память	71
18.4	Ввод-вывод	71
18.5	Прочие параметры	72
19	Проверка на виртуальность	73
19.1	Условия на запуск юнитов	74
19.2	В скриптах	74
19.3	В программах	75
20	Сокет-активация служб и контейнеров	75
20.1	Сокет-активация сетевых служб	75
20.2	Сокет-активация контейнеров	76
21	Интеграция с контейнерами	79
	Защита от уязвимостей AF_PACKET при помощи systemd¹	86
	Динамические пользователи	87
	Что такое «динамический пользователь»?	89
	Долговременное хранение данных	90
	Примеры	92
	Практическое применение	94
	Рекомендации сопровождающим пакетов	95
	Дополнительные замечания	96
	Учет и фильтрация IP-трафика служб	97
	Учет IP-трафика	98
	Фильтрация IP-трафика	102
	Практическое применение	103
	Рекомендации сопровождающим пакетов	104
	Дополнительные замечания	105
A	FAQ (часто задаваемые вопросы)²	107

¹Прим. перев.: Эта и последующие статьи официально не входят в цикл «systemd для администраторов», и поэтому не имеют порядковых номеров. Однако они сходны по тематике с остальными статьями цикла и несут довольно много полезной информации, и поэтому их перевод все же присутствует в данном документе.

²Прим. перев.: Перевод статьи «Frequently Asked Questions» с официального сайта проекта, по состоянию на 2013-05-26 08:17:02 (коммит 8cf2b).

В Диагностика неполадок³	111
В.1 Диагностика проблем с загрузкой	111
В.1.1 Если у вас нет доступа к оболочке	111
В.1.2 Если у вас есть доступ к оболочке	113
В.2 Диагностика проблем с выключением системы	113
В.2.1 Система очень долго выключается	114
В.2.2 Система не может выключиться самостоятельно	114
В.3 Просмотр состояния службы и ее журнала	114
В.4 Подготовка сообщений об ошибках	115
В.4.1 Что нужно включить в сообщение об ошибке	115
С Совместимость с SysV⁴	115
Д Предсказуемые имена сетевых интерфейсов⁵	118
D.1 Зачем?	118
D.2 Что именно было изменено в версии 197?	119
D.3 Еще раз, что здесь хорошего?	120
D.4 Мне не нравится ваша схема. Как ее отключить?	120
D.5 Как именно работает новая схема?	121
Е Специальные файловые системы⁶	122
Ф Запуск служб после появления сети⁷	124
F.1 Как это реализовано в systemd	125
F.2 Короче, как заставить network.target работать?	127
F.3 А что делать нам, разработчикам?	128
Г Моя служба не может получить приоритет realtime⁸	128

Предисловие автора

Многие из вас, наверное, уже знают, что `systemd` — это новая система инициализации дистрибутива Fedora, начиная с Fedora 14⁹. Помимо Fedora, `systemd` также поддерживает и другие дистрибутивы, в частности, `OpenSUSE`¹⁰. `systemd` предоставляет администраторам целый ряд новых возможностей, значительно упрощающих процесс обслуживания системы. Эта статья является первой в серии публикаций, планируемых в ближайшие месяцы. В каждой из этих статей я попытаюсь рассказать об очередной новой возможности `systemd`. Большинство этих возможностей можно описать легко и просто, и подобные статьи должны быть интересны довольно широкой аудитории. Однако, время от времени мы будем рассматривать ключевые новшества `systemd`, что может потребовать несколько более подробного изложения.

Lennart Poettering, 23 августа 2010 г.

³Прим. перев.: Перевод статьи «[Debugging systemd Problems](#)» с официального сайта проекта, по состоянию на 2017-07-28 09:44:56 (коммит e8c76).

⁴Прим. перев.: Перевод статьи «[Compatibility with SysV](#)» с официального сайта проекта, по состоянию на 2017-05-23 08:04:14 (коммит 4c503).

⁵Прим. перев.: Перевод статьи «[Predictable Network Interface Names](#)» с официального сайта проекта, по состоянию на 2016-11-17 17:52:54 (коммит fe089).

⁶Прим. перев.: Перевод статьи «[API File Systems](#)» с официального сайта проекта, по состоянию на 2015-02-20 15:48:02 (коммит e2db9).

⁷Прим. перев.: Перевод статьи «[Running Services After the Network is up](#)» с официального сайта проекта, по состоянию на 2016-08-21 13:14:53 (коммит 4d5d5).

⁸Прим. перев.: Перевод статьи «[My Service Can't Get Realtime!](#)» с официального сайта проекта, по состоянию на 2013-05-18 08:20:36 (коммит 2f77b).

⁹Прим. перев.: к сожалению, разработчики Fedora приняли решение оставить в Fedora 14 в качестве системы инициализации по умолчанию `upstart`, однако `systemd` все равно включен в этот релиз и может быть использован в качестве альтернативной системы инициализации. Окончательный переход на `systemd` произошел лишь в Fedora 15.

¹⁰Прим. перев.: Сейчас `systemd` поддерживается практически во всех популярных дистрибутивах для настольных систем.

1 Контроль процесса загрузки

Как правило, во время загрузки Linux по экрану быстро пробегает огромное количество различных сообщений. Так как мы интенсивно работаем над параллелизацией и ускорением процесса загрузки, с каждой новой версией `systemd` эти сообщения будут пробегать все быстрее и быстрее, вследствие чего, читать их будет все труднее. К тому же, многие пользователи применяют графические оболочки загрузки (например, `Plymouth`), полностью скрывающие эти сообщения. Тем не менее, информация, которую они несут, была и остается чрезвычайно важной — они показывают, успешно ли запустилась каждая служба, или попытка ее запуска закончилась ошибкой (зеленое [`OK`] или красное [`FAILED`] соответственно). Итак, с ростом скорости загрузки систем, возникает неприятная ситуация: информация о результатах запуска служб бывает очень важна, а просматривать ее все тяжелее. `systemd` предлагает выход из этой ситуации: он отслеживает и запоминает факты успешного или неудачного запуска служб на этапе загрузки, а также сбой служб во время работы. К таким случаям относятся выходы с ненулевым кодом, ошибки сегментирования и т.п. Введя `systemctl status` в своей командной оболочке, вы можете ознакомиться с состоянием всех служб, как «родных» (native) для `systemd`, так и классических SysV/LSB служб, поддерживаемых в целях совместимости:

```
[root@lambda] ~# systemctl
```

```
UNIT                                LOAD    ACTIVE SUB    JOB    DESCRIPTION
dev-hugepages.automount             loaded active running Huge Pages File System Automount Point
dev-mqueue.automount                loaded active running POSIX Message Queue File System Automount Point
proc-sys-fs-binfmt_misc.automount   loaded active waiting Arbitrary Executable File Formats File System Automount Point
sys-kernel-debug.automount          loaded active waiting Debug File System Automount Point
sys-kernel-security.automount       loaded active waiting Security File System Automount Point
sys-devices-pc...0000:02:00.0-net-eth0.device loaded active plugged 82573L Gigabit Ethernet Controller
[...]
sys-devices-virtual-tty-tty9.device loaded active plugged /sys/devices/virtual/tty/tty9
-.mount                             loaded active mounted /
boot.mount                          loaded active mounted /boot
dev-hugepages.mount                 loaded active mounted Huge Pages File System
dev-mqueue.mount                    loaded active mounted POSIX Message Queue File System
home.mount                          loaded active mounted /home
proc-sys-fs-binfmt_misc.mount       loaded active mounted Arbitrary Executable File Formats File System
abrt.service                        loaded active running ABRT Automated Bug Reporting Tool
accounts-daemon.service             loaded active running Accounts Service
acpid.service                       loaded active running ACPI Event Daemon
atd.service                         loaded active running Execution Queue Daemon
auditd.service                      loaded active running Security Auditing Service
avahi-daemon.service                loaded active running Avahi mDNS/DNS-SD Stack
bluetooth.service                  loaded active running Bluetooth Manager
console-kit-daemon.service          loaded active running Console Manager
cpuspeed.service                   loaded active exited LSB: processor frequency scaling support
crond.service                       loaded active running Command Scheduler
cups.service                        loaded active running CUPS Printing Service
dbus.service                        loaded active running D-Bus System Message Bus
getty@tty2.service                  loaded active running Getty on tty2
getty@tty3.service                  loaded active running Getty on tty3
getty@tty4.service                  loaded active running Getty on tty4
getty@tty5.service                  loaded active running Getty on tty5
getty@tty6.service                  loaded active running Getty on tty6
haldaemon.service                   loaded active running Hardware Manager
hdapsd@sda.service                  loaded active running sda shock protection daemon
irqbalance.service                  loaded active running LSB: start and stop irqbalance daemon
```

iscsi.service	loaded active	exited	LSB: Starts and stops login and scanning of iSCSI devices.
iscsid.service	loaded active	exited	LSB: Starts and stops login iSCSI daemon.
livesys-late.service	loaded active	exited	LSB: Late init script for live image.
livesys.service	loaded active	exited	LSB: Init script for live image.
lvm2-monitor.service	loaded active	exited	LSB: Monitoring of LVM2 mirrors, snapshots etc. using dmeventd or progr
mdmonitor.service	loaded active	running	LSB: Start and stop the MD software RAID monitor
modem-manager.service	loaded active	running	Modem Manager
netfs.service	loaded active	exited	LSB: Mount and unmount network filesystems.
NetworkManager.service	loaded active	running	Network Manager
ntpd.service	loaded maintenance	maintenance	Network Time Service
polkitd.service	loaded active	running	Policy Manager
prefdm.service	loaded active	running	Display Manager
rc-local.service	loaded active	exited	/etc/rc.local Compatibility
rpcbind.service	loaded active	running	RPC Portmapper Service
rsyslog.service	loaded active	running	System Logging Service
rtkit-daemon.service	loaded active	running	RealtimeKit Scheduling Policy Service
sendmail.service	loaded active	running	LSB: start and stop sendmail
sshd@172.31.0.53:22-172.31.0.4:36368.service	loaded active	running	SSH Per-Connection Server
sysinit.service	loaded active	running	System Initialization
systemd-logger.service	loaded active	running	systemd Logging Daemon
udev-post.service	loaded active	exited	LSB: Moves the generated persistent udev rules to /etc/udev/rules.d
udisks.service	loaded active	running	Disk Manager
upowerd.service	loaded active	running	Power Manager
wpa_supplicant.service	loaded active	running	Wi-Fi Security Service
avahi-daemon.socket	loaded active	listening	Avahi mDNS/DNS-SD Stack Activation Socket
cups.socket	loaded active	listening	CUPS Printing Service Sockets
dbus.socket	loaded active	running	dbus.socket
rpcbind.socket	loaded active	listening	RPC Portmapper Socket
sshd.socket	loaded active	listening	sshd.socket
systemd-initctl.socket	loaded active	listening	systemd /dev/initctl Compatibility Socket
systemd-logger.socket	loaded active	running	systemd Logging Socket
systemd-shutdown.socket	loaded active	listening	systemd Delayed Shutdown Socket
dev-disk-by\x1...\x1db22a\x1d870f1adf2732.swap	loaded active	active	/dev/disk/by-uuid/fd626ef7-34a4-4958-b22a-870f1adf2732
basic.target	loaded active	active	Basic System
bluetooth.target	loaded active	active	Bluetooth
dbus.target	loaded active	active	D-Bus

getty.target	loaded active	active	Login Prompts
graphical.target	loaded active	active	Graphical Interface
local-fs.target	loaded active	active	Local File Systems
multi-user.target	loaded active	active	Multi-User
network.target	loaded active	active	Network
remote-fs.target	loaded active	active	Remote File Systems
sockets.target	loaded active	active	Sockets
swap.target	loaded active	active	Swap
sysinit.target	loaded active	active	System Initialization

LOAD = Reflects whether the unit definition was properly loaded.
ACTIVE = The high-level unit activation state, i.e. generalization of SUB.
SUB = The low-level unit activation state, values depend on unit type.
JOB = Pending job for the unit.

221 units listed. Pass --all to see inactive units, too.

[root@lambda] ~#

~

(Листинг был сокращен за счет удаления строк, не относящихся к теме статьи.)

Обратите внимание на графу ACTIVE, в которой отображается обобщенный статус службы (или любого другого юнита systemd: устройства, сокета, точки монтирования — их мы рассмотрим подробнее в последующих статьях). Основными значениями обобщенного статуса являются active (служба выполняется) и inactive (служба не была запущена). Также существуют и другие статусы. Например, внимательно посмотрев на листинг выше, вы можете заметить, что служба ntpd (сервер точного времени) находится в состоянии, обозначенном как maintenance. Чтобы узнать, что же произошло с ntpd, воспользуемся командой `systemctl status`¹¹:

```
[root@lambda] ~# systemctl status ntpd.service
ntp.service - Network Time Service
   Loaded: loaded (/etc/systemd/system/ntp.service)
   Active: maintenance
     Main: 953 (code=exited, status=255)
   CGroup: name=systemd:/systemd-1/ntp.service
[root@lambda] ~#
```

systemd сообщает нам, что ntpd был запущен (с идентификатором процесса 953) и аварийно завершил работу (с кодом выхода 255)¹².

В последующих версиях systemd, мы планируем добавить возможность вызова в таких ситуациях ABRT (Automated Bug Report Tool), но для этого необходима поддержка со стороны самого ABRT. Соответствующий запрос уже [направлен](#) его разработчикам, однако пока не встретил среди них поддержки.

Резюме: использование `systemctl` и `systemctl status` является современной, более удобной и эффективной альтернативой разглядыванию быстро пробегающих по экрану сообщений в классическом SysV. `systemctl status` дает возможность получить развернутую информацию о характере ошибки и, кроме того, в отличие от сообщений SysV, показывает не только ошибки при запуске, но и ошибки, возникшие во время исполнения службы.

2 О службах и процессах

В большинстве современных Linux-систем количество одновременно работающих процессов обычно весьма значительно. Понять, откуда взялся и что делает тот или иной процесс, становится все сложнее и сложнее. Многие службы используют сразу несколько рабочих процессов, и это отнюдь не всегда можно легко распознать по выводу команды `ps`. Встречаются еще более сложные ситуации, когда демон запускает сторонние процессы — например, веб-сервер выполняет CGI-программы, а демон `cron` — команды, предписанные ему в `crontab`.

Немного помочь в решении этой проблемы может древовидная иерархия процессов, отображаемая по команде `ps xaf`. Именно «немного помочь», а не решить полностью. В частности, процессы, родители которых умирают раньше их самих, становятся потомками PID 1 (процесса `init`), что сразу затрудняет процесс выяснения их происхождения. Кроме того, процесс может избавиться от связи с родителем через две последовательные операции `fork()` (в целом, эта возможность признается нужной и полезной, и является частью используемого в Unix подхода к разработке демонов). Также, не будем забывать, что процесс легко может изменить свое имя посредством `PR_SETNAME`, или задав значение `argv[0]`, что также усложняет процесс его опознания¹³.

¹¹Прим. перев.: Стоит заметить, что формат вывода данной команды менялся по мере развития systemd — появлялись дополнительные поля с информацией, был добавлен вывод журнала службы (см. главу 13) и т.д. Здесь приведен пример вывода этой команды на момент написания исходной статьи (лето 2010 года).

¹²Прим. перев.: Впоследствии, по просьбам пользователей, считавших, что слово «maintenance» недостаточно точно отражает ситуацию, оно было заменено на «failed».

¹³Прим. перев.: Стоит отметить, что перечисленные ситуации могут возникнуть не только вследствие ошибок в коде и/или конфигурации программ, но и в результате злого умысла. Например, очень часто встречается ситуация, когда установленный на взломанном сервере процесс-бэкдор маскируется под нормального демона, меняя себе имя, скажем, на `httpd`.

systemd предлагает простой путь для решения обсуждаемой задачи. Запуская новый процесс, systemd помещает его в отдельную контрольную группу с соответствующим именем. Контрольные группы Linux предоставляют очень удобный инструмент для иерархической структуризации процессов: когда какой-либо процесс порождает потомка, этот потомок автоматически включается в ту же группу, что и родитель. При этом, что очень важно, непривилегированные процессы не могут изменить свое положение в этой иерархии. Таким образом, контрольные группы позволяют точно установить происхождение конкретного процесса, вне зависимости от того, сколько раз он форкался и переименовывал себя — имя его контрольной группы невозможно спрятать или изменить. Кроме того, при штатном завершении родительской службы, будут завершены и все порожденные ею процессы, как бы они ни пытались сбежать. С systemd уже невозможна ситуация, когда после остановки web-сервера, некорректно форкнувшийся CGI-процесс продолжает исполняться вплоть до последних секунд работы системы.

В этой статье мы рассмотрим две простых команды, которые позволят вам наглядно оценить схему взаимоотношений systemd и порожденных им процессов. Первая из этих команд — все та же `ps`, однако на этот раз в ее параметры добавлено указание выводить сведения по контрольным группам, а также другую интересную информацию:

```

$ ps xawf -eo pid,user,cgroup,args
PID USER      CGROUP          COMMAND
   2 root        -                [kthreadd]
   3 root        -                \_ [ksoftirqd/0]
[...]
4281 root        -                \_ [flush-8:0]
   1 root        name=systemd:/systemd-1 /sbin/init
  455 root        name=systemd:/systemd-1/sysinit.service /sbin/udevd -d
28188 root        name=systemd:/systemd-1/sysinit.service \_ /sbin/udevd -d
28191 root        name=systemd:/systemd-1/sysinit.service \_ /sbin/udevd -d
 1096 dbus        name=systemd:/systemd-1/dbus.service /bin/dbus-daemon --system --address=systemd: --nofork --systemd-activation
 1131 root        name=systemd:/systemd-1/auditd.service auditd
 1133 root        name=systemd:/systemd-1/auditd.service \_ /sbin/auditd
 1135 root        name=systemd:/systemd-1/auditd.service \_ /usr/sbin/sedispatch
 1171 root        name=systemd:/systemd-1/NetworkManager.service /usr/sbin/NetworkManager --no-daemon
 4028 root        name=systemd:/systemd-1/NetworkManager.service \_ /sbin/dhclient -d -4 -sf /usr/libexec/nm-dhcp-client.action -pf /var/run/dhclient-wlan0.pid
 1175 avahi        name=systemd:/systemd-1/avahi-daemon.service avahi-daemon: running [epsilon.local]
 1194 avahi        name=systemd:/systemd-1/avahi-daemon.service \_ avahi-daemon: chroot helper
 1193 root        name=systemd:/systemd-1/rsyslog.service /sbin/rsyslogd -c 4
 1195 root        name=systemd:/systemd-1/cups.service cupsd -C /etc/cups/cupsd.conf
 1207 root        name=systemd:/systemd-1/mdmonitor.service mdadm --monitor --scan -f --pid-file=/var/run/mdadm/mdadm.pid
 1210 root        name=systemd:/systemd-1/irqbalance.service irqbalance
 1216 root        name=systemd:/systemd-1/dbus.service /usr/sbin/modem-manager
 1219 root        name=systemd:/systemd-1/dbus.service /usr/libexec/polkit-1/polkitd
 1242 root        name=systemd:/systemd-1/dbus.service /usr/sbin/wpa_supplicant -c /etc/wpa_supplicant/wpa_supplicant.conf -B -u -f /var/log/wpa_supplicant.log -I
 1249 68          name=systemd:/systemd-1/haldaemon.service hald
 1250 root        name=systemd:/systemd-1/haldaemon.service \_ hald-runner
 1273 root        name=systemd:/systemd-1/haldaemon.service \_ hald-addon-input: Listening on /dev/input/event3 /dev/input/event9 /dev/input/event1 /dev/inpu
 1275 root        name=systemd:/systemd-1/haldaemon.service \_ /usr/libexec/hald-addon-rfkill-killswitch
 1284 root        name=systemd:/systemd-1/haldaemon.service \_ /usr/libexec/hald-addon-leds
 1285 root        name=systemd:/systemd-1/haldaemon.service \_ /usr/libexec/hald-addon-generic-backlight
 1287 68          name=systemd:/systemd-1/haldaemon.service \_ /usr/libexec/hald-addon-acpi
 1317 root        name=systemd:/systemd-1/abrttd.service /usr/sbin/abrttd -d -s
 1332 root        name=systemd:/systemd-1/getty@.service/tty2 /sbin/mingetty tty2
 1339 root        name=systemd:/systemd-1/getty@.service/tty3 /sbin/mingetty tty3
 1342 root        name=systemd:/systemd-1/getty@.service/tty5 /sbin/mingetty tty5

```

```
1343 root    name=systemd:/systemd-1/getty@.service/tty4 /sbin/mingetty tty4
1344 root    name=systemd:/systemd-1/crond.service crond
1346 root    name=systemd:/systemd-1/getty@.service/tty6 /sbin/mingetty tty6
1362 root    name=systemd:/systemd-1/sshd.service /usr/sbin/sshd
1376 root    name=systemd:/systemd-1/prefdm.service /usr/sbin/gdm-binary -nodaemon
1391 root    name=systemd:/systemd-1/prefdm.service \_ /usr/libexec/gdm-simple-slave --display-id /org/gnome/DisplayManager/Display1 --force-active-vt
1394 root    name=systemd:/systemd-1/prefdm.service \_ /usr/bin/Xorg :0 -nr -verbose -auth /var/run/gdm/auth-for-gdm-f2KU0h/database -nolisten tcp vt1
1495 root    name=systemd:/user/lennart/1 \_ pam: gdm-password
1521 lennart  name=systemd:/user/lennart/1 \_ gnome-session
1621 lennart  name=systemd:/user/lennart/1 \_ metacity
1635 lennart  name=systemd:/user/lennart/1 \_ gnome-panel
1638 lennart  name=systemd:/user/lennart/1 \_ nautilus
1640 lennart  name=systemd:/user/lennart/1 \_ /usr/libexec/polkit-gnome-authentication-agent-1
1641 lennart  name=systemd:/user/lennart/1 \_ /usr/bin/seapplet
1644 lennart  name=systemd:/user/lennart/1 \_ gnome-volume-control-applet
1646 lennart  name=systemd:/user/lennart/1 \_ /usr/sbin/restorecond -u
1652 lennart  name=systemd:/user/lennart/1 \_ /usr/bin/devilspie
1662 lennart  name=systemd:/user/lennart/1 \_ nm-applet --sm-disable
1664 lennart  name=systemd:/user/lennart/1 \_ gnome-power-manager
1665 lennart  name=systemd:/user/lennart/1 \_ /usr/libexec/gdu-notification-daemon
1670 lennart  name=systemd:/user/lennart/1 \_ /usr/libexec/evolution/2.32/evolution-alarm-notify
1672 lennart  name=systemd:/user/lennart/1 \_ /usr/bin/python /usr/share/system-config-printer/applet.py
1674 lennart  name=systemd:/user/lennart/1 \_ /usr/lib64/deja-dup/deja-dup-monitor
1675 lennart  name=systemd:/user/lennart/1 \_ abrt-applet
1677 lennart  name=systemd:/user/lennart/1 \_ bluetooth-applet
1678 lennart  name=systemd:/user/lennart/1 \_ gpk-update-icon
1408 root    name=systemd:/systemd-1/console-kit-daemon.service /usr/sbin/console-kit-daemon --no-daemon
1419 gdm      name=systemd:/systemd-1/prefdm.service /usr/bin/dbus-launch --exit-with-session
1453 root    name=systemd:/systemd-1/dbus.service /usr/libexec/upowerd
1473 rtkit    name=systemd:/systemd-1/rtkit-daemon.service /usr/libexec/rtkit-daemon
1496 root    name=systemd:/systemd-1/accounts-daemon.service /usr/libexec/accounts-daemon
1499 root    name=systemd:/systemd-1/systemd-logger.service /lib/systemd/systemd-logger
1511 lennart  name=systemd:/systemd-1/prefdm.service /usr/bin/gnome-keyring-daemon --daemonize --login
1534 lennart  name=systemd:/user/lennart/1 dbus-launch --sh-syntax --exit-with-session
1535 lennart  name=systemd:/user/lennart/1 /bin/dbus-daemon --fork --print-pid 5 --print-address 7 --session
1603 lennart  name=systemd:/user/lennart/1 /usr/libexec/gconfd-2
```

```

1612 lennart name=systemd:/user/lennart/1 /usr/libexec/gnome-settings-daemon
1615 lennart name=systemd:/user/lennart/1 /usr/libexec/gvfsd
1626 lennart name=systemd:/user/lennart/1 /usr/libexec/gvfs-fuse-daemon /home/lennart/.gvfs
1634 lennart name=systemd:/user/lennart/1 /usr/bin/pulseaudio --start --log-target=syslog
1649 lennart name=systemd:/user/lennart/1 \_ /usr/libexec/pulse/gconf-helper
1645 lennart name=systemd:/user/lennart/1 /usr/libexec/bonobo-activation-server --ac-activate --ior-output-fd=24
1668 lennart name=systemd:/user/lennart/1 /usr/libexec/im-settings-daemon
1701 lennart name=systemd:/user/lennart/1 /usr/libexec/gvfs-gdu-volume-monitor
1707 lennart name=systemd:/user/lennart/1 /usr/bin/gnote --panel-applet --oaf-activate-iid=OAFIID:GnoteApplet_Factory --oaf-ior-fd=22
1725 lennart name=systemd:/user/lennart/1 /usr/libexec/clock-applet
1727 lennart name=systemd:/user/lennart/1 /usr/libexec/wnck-applet
1729 lennart name=systemd:/user/lennart/1 /usr/libexec/notification-area-applet
1733 root name=systemd:/systemd-1/dbus.service /usr/libexec/udisks-daemon
1747 root name=systemd:/systemd-1/dbus.service \_ udisks-daemon: polling /dev/sr0
1759 lennart name=systemd:/user/lennart/1 gnome-screensaver
1780 lennart name=systemd:/user/lennart/1 /usr/libexec/gvfsd-trash --spawner :1.9 /org/gtk/gvfs/exec_spaw/0
1864 lennart name=systemd:/user/lennart/1 /usr/libexec/gvfs-afc-volume-monitor
1874 lennart name=systemd:/user/lennart/1 /usr/libexec/gconf-im-settings-daemon
1903 lennart name=systemd:/user/lennart/1 /usr/libexec/gvfsd-burn --spawner :1.9 /org/gtk/gvfs/exec_spaw/1
1909 lennart name=systemd:/user/lennart/1 gnome-terminal
1913 lennart name=systemd:/user/lennart/1 \_ gnome-pty-helper
1914 lennart name=systemd:/user/lennart/1 \_ bash
29231 lennart name=systemd:/user/lennart/1 | \_ ssh tango
2221 lennart name=systemd:/user/lennart/1 \_ bash
4193 lennart name=systemd:/user/lennart/1 | \_ ssh tango
2461 lennart name=systemd:/user/lennart/1 \_ bash
29219 lennart name=systemd:/user/lennart/1 | \_ emacs systemd-for-admins-1.txt
15113 lennart name=systemd:/user/lennart/1 \_ bash
27251 lennart name=systemd:/user/lennart/1 \_ empathy
29504 lennart name=systemd:/user/lennart/1 \_ ps xawf -eo pid,user,cgroup,args
1968 lennart name=systemd:/user/lennart/1 ssh-agent
1994 lennart name=systemd:/user/lennart/1 gpg-agent --daemon --write-env-file
18679 lennart name=systemd:/user/lennart/1 /bin/sh /usr/lib64/firefox-3.6/run-mozilla.sh /usr/lib64/firefox-3.6/firefox
18741 lennart name=systemd:/user/lennart/1 \_ /usr/lib64/firefox-3.6/firefox
28900 lennart name=systemd:/user/lennart/1 \_ /usr/lib64/nspluginwrapper/npviewer.bin --plugin /usr/lib64/mozilla/plugins/libflashplayer.so --con
4016 root name=systemd:/systemd-1/sysinit.service /usr/sbin/bluetoothd --udev

```

```
4094 smmsp name=systemd:/systemd-1/sendmail.service sendmail: Queue runner@01:00:00 for /var/spool/clientmqueue
4096 root name=systemd:/systemd-1/sendmail.service sendmail: accepting connections
4112 ntp name=systemd:/systemd-1/ntpd.service /usr/sbin/ntpd -n -u ntp:ntp -g
27262 lennart name=systemd:/user/lennart/1 /usr/libexec/mission-control-5
27265 lennart name=systemd:/user/lennart/1 /usr/libexec/telepathy-haze
27268 lennart name=systemd:/user/lennart/1 /usr/libexec/telepathy-logger
27270 lennart name=systemd:/user/lennart/1 /usr/libexec/dconf-service
27280 lennart name=systemd:/user/lennart/1 /usr/libexec/notification-daemon
27284 lennart name=systemd:/user/lennart/1 /usr/libexec/telepathy-gabble
27285 lennart name=systemd:/user/lennart/1 /usr/libexec/telepathy-salut
27297 lennart name=systemd:/user/lennart/1 /usr/libexec/geoclue-yahoo
```

(Данный листинг был сокращен за счет удаления из него строк, описывающих потоки ядра, так как они никак не относятся к обсуждаемой нами теме.)

Обратите внимание на третий столбец, показывающий имя контрольной группы, в которую `systemd` поместил данный процесс. Например, процесс `udev` находится в группе `name=systemd:/systemd-1/sysinit.service`. В эту группу входят процессы, порожденные службой `sysinit.service`, которая запускается на ранней стадии загрузки.

Вы можете очень сильно упростить себе работу, если назначите для вышеприведенной команды какой-нибудь простой и короткий псевдоним, например

```
alias psc='ps xawf -eo pid,user,cgroup,args'
```

— теперь для получения исчерпывающей информации по процессам достаточно будет нажать всего четыре клавиши.

Альтернативный способ получить ту же информацию — воспользоваться утилитой `systemd-cgls`, входящей в комплект поставки `systemd`. Она отображает иерархию контрольных групп в виде псевдографической диаграммы-дерева¹⁴:

¹⁴Прим. перев.: Стоит заметить, что в нижеприведенном листинге используется ASCII-псевдографика. Между тем, уже довольно давно отображение иерархии процессов в `systemd-cgls` и `systemctl status` производится в более выразительной Юникодной псевдографике. ASCII-вариант выводится только в том случае, если консоль не поддерживает Unicode.

```
$ systemd-cgls
+ 2 [kthreadd]
[...]
+ 4281 [flush-8:0]
+ user
| \ lennart
| \ 1
| + 1495 pam: gdm-password
| + 1521 gnome-session
| + 1534 dbus-launch --sh-syntax --exit-with-session
| + 1535 /bin/dbus-daemon --fork --print-pid 5 --print-address 7 --session
| + 1603 /usr/libexec/gconfd-2
| + 1612 /usr/libexec/gnome-settings-daemon
| + 1615 /usr/libexec/gvfsd
| + 1621 metacity
| + 1626 /usr/libexec/gvfs-fuse-daemon /home/lennart/.gvfs
| + 1634 /usr/bin/pulseaudio --start --log-target=syslog
| + 1635 gnome-panel
| + 1638 nautilus
| + 1640 /usr/libexec/polkit-gnome-authentication-agent-1
| + 1641 /usr/bin/seapplet
| + 1644 gnome-volume-control-applet
| + 1645 /usr/libexec/bonobo-activation-server --ac-activate --ior-output-fd=24
| + 1646 /usr/sbin/restorecond -u
| + 1649 /usr/libexec/pulse/gconf-helper
| + 1652 /usr/bin/devilspie
| + 1662 nm-applet --sm-disable
| + 1664 gnome-power-manager
| + 1665 /usr/libexec/gdu-notification-daemon
| + 1668 /usr/libexec/im-settings-daemon
| + 1670 /usr/libexec/evolution/2.32/evolution-alarm-notify
| + 1672 /usr/bin/python /usr/share/system-config-printer/applet.py
| + 1674 /usr/lib64/deja-dup/deja-dup-monitor
| + 1675 abrt-applet
| + 1677 bluetooth-applet
| + 1678 gpk-update-icon
```

```
| + 1701 /usr/libexec/gvfs-gdu-volume-monitor
| + 1707 /usr/bin/gnote --panel-applet --oaf-activate-iid=OAFIID:GnoteApplet_Factory --oaf-ior-fd=22
| + 1725 /usr/libexec/clock-applet
| + 1727 /usr/libexec/wnck-applet
| + 1729 /usr/libexec/notification-area-applet
| + 1759 gnome-screensaver
| + 1780 /usr/libexec/gvfsd-trash --spawner :1.9 /org/gtk/gvfs/exec_spaw/0
| + 1864 /usr/libexec/gvfs-afc-volume-monitor
| + 1874 /usr/libexec/gconf-im-settings-daemon
| + 1882 /usr/libexec/gvfs-gphoto2-volume-monitor
| + 1903 /usr/libexec/gvfsd-burn --spawner :1.9 /org/gtk/gvfs/exec_spaw/1
| + 1909 gnome-terminal
| + 1913 gnome-pty-helper
| + 1914 bash
| + 1968 ssh-agent
| + 1994 gpg-agent --daemon --write-env-file
| + 2221 bash
| + 2461 bash
| + 4193 ssh tango
| + 15113 bash
| + 18679 /bin/sh /usr/lib64/firefox-3.6/run-mozilla.sh /usr/lib64/firefox-3.6/firefox
| + 18741 /usr/lib64/firefox-3.6/firefox
| + 27251 empathy
| + 27262 /usr/libexec/mission-control-5
| + 27265 /usr/libexec/telepathy-haze
| + 27268 /usr/libexec/telepathy-logger
| + 27270 /usr/libexec/dconf-service
| + 27280 /usr/libexec/notification-daemon
| + 27284 /usr/libexec/telepathy-gabble
| + 27285 /usr/libexec/telepathy-salut
| + 27297 /usr/libexec/geoclue-yahoo
| + 28900 /usr/lib64/nspluginwrapper/npviewer.bin --plugin /usr/lib64/mozilla/plugins/libflashplayer.so --connection /org/wrapper/NSPlugins/libflashplayer
| + 29219 emacs systemd-for-admins-1.txt
| + 29231 ssh tango
| \ 29519 systemd-cgls
\ systemd-1
```

```
+ 1 /sbin/init
+ ntpd.service
| \ 4112 /usr/sbin/ntpd -n -u ntp:ntp -g
+ systemd-logger.service
| \ 1499 /lib/systemd/systemd-logger
+ accounts-daemon.service
| \ 1496 /usr/libexec/accounts-daemon
+ rtkit-daemon.service
| \ 1473 /usr/libexec/rtkit-daemon
+ console-kit-daemon.service
| \ 1408 /usr/sbin/console-kit-daemon --no-daemon
+ prefdm.service
| + 1376 /usr/sbin/gdm-binary -nodaemon
| + 1391 /usr/libexec/gdm-simple-slave --display-id /org/gnome/DisplayManager/Display1 --force-active-vt
| + 1394 /usr/bin/Xorg :0 -nr -verbose -auth /var/run/gdm/auth-for-gdm-f2KU0h/database -nolisten tcp vt1
| + 1419 /usr/bin/dbus-launch --exit-with-session
| \ 1511 /usr/bin/gnome-keyring-daemon --daemonize --login
+ getty@.service
| + tty6
| | \ 1346 /sbin/mingetty tty6
| + tty4
| | \ 1343 /sbin/mingetty tty4
| + tty5
| | \ 1342 /sbin/mingetty tty5
| + tty3
| | \ 1339 /sbin/mingetty tty3
| \ tty2
| \ 1332 /sbin/mingetty tty2
+ abrt.service
| \ 1317 /usr/sbin/abrt -d -s
+ crond.service
| \ 1344 crond
+ sshd.service
| \ 1362 /usr/sbin/sshd
+ sendmail.service
| + 4094 sendmail: Queue runner@01:00:00 for /var/spool/clientmqueue
```

```
| \ 4096 sendmail: accepting connections
+ haldaemon.service
| + 1249 hald
| + 1250 hald-runner
| + 1273 hald-addon-input: Listening on /dev/input/event3 /dev/input/event9 /dev/input/event1 /dev/input/event7 /dev/input/event2 /dev/input/event0 /dev/inpu
| + 1275 /usr/libexec/hald-addon-rfkill-killswitch
| + 1284 /usr/libexec/hald-addon-leds
| + 1285 /usr/libexec/hald-addon-generic-backlight
| \ 1287 /usr/libexec/hald-addon-acpi
+ irqbalance.service
| \ 1210 irqbalance
+ avahi-daemon.service
| + 1175 avahi-daemon: running [epsilon.local]
+ NetworkManager.service
| + 1171 /usr/sbin/NetworkManager --no-daemon
| \ 4028 /sbin/dhclient -d -4 -sf /usr/libexec/nm-dhcp-client.action -pf /var/run/dhclient-wlan0.pid -lf /var/lib/dhclient/dhclient-7d32a784-ed9-4cf6-9ee3-0
+ rsyslog.service
| \ 1193 /sbin/rsyslogd -c 4
+ mdmonitor.service
| \ 1207 mdadm --monitor --scan -f --pid-file=/var/run/mdadm/mdadm.pid
+ cups.service
| \ 1195 cupsd -C /etc/cups/cupsd.conf
+ auditd.service
| + 1131 auditd
| + 1133 /sbin/audispd
| \ 1135 /usr/sbin/sedispach
+ dbus.service
| + 1096 /bin/dbus-daemon --system --address=systemd: --nofork --systemd-activation
| + 1216 /usr/sbin/modem-manager
| + 1219 /usr/libexec/polkit-1/polkitd
| + 1242 /usr/sbin/wpa_supplicant -c /etc/wpa_supplicant/wpa_supplicant.conf -B -u -f /var/log/wpa_supplicant.log -P /var/run/wpa_supplicant.pid
| + 1453 /usr/libexec/upowerd
| + 1733 /usr/libexec/udisks-daemon
| + 1747 udisks-daemon: polling /dev/sr0
| \ 29509 /usr/libexec/packagekitd
+ dev-mqueue.mount
```

```
+ dev-hugepages.mount
\ sysinit.service
+ 455 /sbin/udev -d
+ 4016 /usr/sbin/bluetoothd --udev
+ 28188 /sbin/udev -d
\ 28191 /sbin/udev -d
```

(Как и предыдущий, этот листинг был сокращен за счет удаления перечня потоков ядра.)

Как видно из листинга, данная команда наглядно показывает принадлежность процессов к их контрольным группам, а следовательно, и к службам, так как `systemd` имеет группы в соответствии с названиями служб. Например, из приведенного листинга нетрудно понять, что служба системного аудита `auditd.service` порождает три отдельных процесса: `auditd`, `auditpd` и `sedispatch`.

Наиболее внимательные читатели, вероятно, уже заметили, что некоторые процессы помещены в группу `/user/lennart/1`. Дело в том, что `systemd` занимается отслеживанием и группировкой не только процессов, относящихся к системным службам, но и процессов, запущенных в рамках пользовательских сеансов. В последующих статьях мы обсудим этот вопрос более подробно.

3 Преобразование SysV init-скрипта в systemd service-файл

Традиционно, службы Unix и Linux (демоны) запускаются через SysV init-скрипты. Эти скрипты пишутся на языке Bourne Shell (`/bin/sh`), располагаются в специальном каталоге (обычно `/etc/rc.d/init.d/`) и вызываются с одним из стандартных параметров (`start`, `stop`, `reload` и т.п.) — таким образом указывается действие, которое необходимо произвести над службой (запустить, остановить, заставить перечитать конфигурацию). При запуске службы такой скрипт, как правило, вызывает бинарник демона, который, в свою очередь, форкается, порождая фоновый процесс (т.е. демонизируется). Заметим, что shell-скрипты, как правило, отличаются низкой скоростью работы, излишней подробностью изложения и крайней хрупкостью. Читать их, из-за избытка всевозможного вспомогательного и дополнительного кода, чрезвычайно тяжело. Впрочем, нельзя не упомянуть, что эти скрипты являются очень гибким инструментом (ведь, по сути, это всего лишь код, который можно модифицировать как угодно). С другой стороны, многие задачи, возникающие при работе со службами, довольно тяжело решить средствами shell-скриптов. К таким задачам относятся: организация параллельного исполнения, корректное отслеживание процессов, конфигурирование различных параметров среды исполнения процесса. `systemd` обеспечивает совместимость с init-скриптами, однако, с учетом описанных выше их недостатков, более правильным решением будет использование штатных service-файлов `systemd` для всех установленных в системе служб. Стоит отметить что, в отличие от init-скриптов, которые часто приходится модифицировать при переносе из одного дистрибутива в другой, один и тот же service-файл будет работать в любом дистрибутиве, использующем `systemd` (а таких дистрибутивов с каждым днем становится все больше и больше). Далее мы вкратце рассмотрим процесс преобразования SysV init-скрипта в service-файл `systemd`. Вообще говоря, service-файл должен создаваться разработчиками каждого демона, и включаться в комплект его поставки. Если вам удалось успешно создать работоспособный service-файл для какого-либо демона, настоятельно рекомендуем вам отправить этот файл разработчикам. Вопросы по полноценной интеграции демонов с `systemd`, с максимальным использованием всех его возможностей, будут рассмотрены в последующих статьях этого цикла, пока же ограничимся ссылкой на [страницу](#) официальной документации.

Итак, приступим. В качестве примера возьмем init-скрипт демона ABRT (Automatic Bug Reporting Tool, службы, занимающейся сбором `crash dump`'ов). Исходный скрипт (в варианте для дистрибутива Fedora) можно загрузить [здесь](#).

Начнем с того, что прочитаем исходный скрипт (неожиданный ход, правда?) и выделим полезную информацию из груды хлама. Практически у всех init-скриптов большая часть кода является чисто вспомогательной, и мало чем отличается от одного скрипта к другому. Как правило, при создании новых скриптов этот код просто копируется из уже существующих (разработка в стиле `copy-paste`). Итак, в исследуемом скрипте нас интересует следующая информация:

- Строка описания службы: «Daemon to detect crashing apps». Как нетрудно заметить, комментарии в заголовке скрипта весьма пространны и описывают не сколь-

ко саму службу, сколько скрипт, ее запускающий. service-файлы systemd тоже содержат описание, но оно относится исключительно к службе, а не к service-файлу.

- LSB-заголовок¹⁵, содержащий информацию о зависимостях. systemd, базирующийся на идеях socket-активации, обычно не требует явного описания зависимостей (либо требует самого минимального описания). Заметим, что основополагающие принципы systemd, включая socket-активацию, рассмотрены в статье [Rethinking PID 1](#), в которой systemd был впервые представлен широкой публике¹⁶. Возвращаясь к нашему примеру: в данном случае ценной информацией о зависимостях является только строка `Required-Start: $syslog`, сообщающая, что для работы `abrttd` требуется демон системного лога. Информация о второй зависимости, `$local_fs`, является избыточной, так как systemd приступает к запуску служб уже после того, как все файловые системы готовы для работы.
- Также, LSB-заголовок сообщает, что данная служба должна быть запущена на уровнях исполнения (runlevels) 3 (консольный многопользовательский) и 5 (графический многопользовательский).
- Исполняемый бинарник демона называется `/usr/sbin/abrttd`.

Вот и вся полезная информация. Все остальное содержимое 115-строчного скрипта является чисто вспомогательным кодом: операции синхронизации и упорядочивания запуска (код, относящийся к lock-файлам), вывод информационных сообщений (команды `echo`), разбор входных параметров (монструозный блок `case`).

На основе приведенной выше информации, мы можем написать следующий service-файл:

```
[Unit]
Description=Daemon to detect crashing apps
After=syslog.target

[Service]
ExecStart=/usr/sbin/abrttd
Type=forking

[Install]
WantedBy=multi-user.target
```

Рассмотрим этот файл поподробнее.

Секция `[Unit]` содержит самую общую информацию о службе. Не будем забывать, что systemd управляет не только службами, но и многими другими объектами, в частности, устройствами, точками монтирования, таймерами и т.п. Общее наименование всех этих объектов — юнит (`unit`). Одноименная секция конфигурационного файла определяет наиболее общие свойства, которые могут быть присущи любому юниту. В нашем случае это, во-первых, строка описания, и во-вторых, указание, что данный юнит рекомендуется активировать после запуска демона системного лога¹⁷. Эта информация,

¹⁵ LSB-заголовок — определенная в [Linux Standard Base](#) схема записи метаданных о службах в блоках комментариев соответствующих `init`-скриптов. Изначально эта схема была введена именно для того, чтобы стандартизировать `init`-скрипты во всех дистрибутивах. Однако разработчики многих дистрибутивов не считают нужным точно исполнять требования LSB, и поэтому формы представления метаданных в различных дистрибутивах могут отличаться. Вследствие этого, при переносе `init`-скрипта из одного дистрибутива в другой, скрипт приходится модифицировать. Например, демон пересылки почты при описании зависимостей может именоваться `MTA` или `smtpdaemon` (Fedora), `smtp` (openSUSE), `mail-transport-agent` (Debian и Ubuntu), `mail-transfer-agent`. Таким образом, можно утверждать, что стандарт LSB не справляется с возложенной на него задачей.

¹⁶ Прим. перев.: Ее русский перевод (сделанный совершенно независимо от данного документа, совсем другим человеком) можно прочитать здесь: [часть 1](#), [часть 2](#).

¹⁷ Строго говоря, эту зависимость здесь указывать не нужно — в системах, в которых демон системного лога активируется через сокет, данная зависимость является избыточной. Современные реализации демона системного лога (например, `rsyslog` начиная с пятой версии) поддерживают активацию через сокет. В системах, использующих такие реализации, явное указание `After=syslog.target` будет из-

как мы помним, была указана в LSB-заголовке исходного `init`-скрипта. В нашем конфигурационном файле мы указываем зависимость от демона системного лога при помощи директивы `After`, указывающей на юнит `syslog.target`. Это специальный юнит, позволяющий ссылаться на любую реализацию демона системного лога, независимо от используемой программы (например, `rsyslog` или `syslog-ng`) и типа активации (как обычной службы или через `log-socket`). Подробнее о таких специальных юнитах можно почитать [страницу](#) официальной документации. Обратите внимание, что директива `After`, в отсутствие директивы `Requires`, задает лишь порядок загрузки, но не задает жесткой зависимости. То есть, если при загрузке конфигурация `systemd` будет предписывать запуск как демона системного лога, так и `abrtd`, то сначала будет запущен демон системного лога, и только потом `abrtd`. Если же конфигурация не будет содержать явного указания запустить демон системного лога, он не будет запущен даже при запуске `abrtd`. И это поведение нас полностью устраивает, так как `abrtd` прекрасно может обходиться и без демона системного лога. В противном случае, мы могли бы воспользоваться директивой `Requires`, задающей жесткую зависимость между юнитами.

Следующая секция, `[Service]`, содержит информацию о службе. Сюда включаются настройки, относящиеся именно к службам, но не к другим типам юнитов. В нашем случае, таких настроек две: `ExecStart`, определяющая расположение бинарника демона и аргументы, с которыми он будет вызван (в нашем случае они отсутствуют), и `Type`, позволяющая задать метод, по которому `systemd` определит окончание периода запуска службы. Традиционный для Unix метод демонизации процесса, когда исходный процесс форкается, порождая демона, после чего завершается, описывается типом `forking` (как в нашем случае). Таким образом, `systemd` считает службу запущенной с момента завершения работы исходного процесса, и рассматривает в качестве основного процесса этой службы порожденный им процесс-демон.

И наконец, третья секция, `[Install]`. Она содержит рекомендации по установке конкретного юнита, указывающие, в каких ситуациях он должен быть активирован. В нашем случае, служба `abrtd` запускается при активации юнита `multi-user.target`. Это специальный юнит, примерно соответствующий роли третьего уровня исполнения классического SysV¹⁸. Директива `WantedBy` никак не влияет на уже работающую службу, но она играет важную роль при выполнении команды `systemctl enable`, задавая, в каких условиях должен активироваться устанавливаемый юнит. В нашем примере, служба `abrtd` будет активироваться при переходе в состояние `multi-user.target`, т.е., при каждой нормальной¹⁹ загрузке²⁰.

Вот и все. Мы получили минимальный рабочий `service`-файл `systemd`. Чтобы проверить его работоспособность, скопируем его в `/etc/systemd/system/abrtd.service`, после чего командой `systemctl daemon-reload` уведомим `systemd` об изменении конфигурации. Теперь нам остается только запустить нашу службу: `systemctl start abrtd.service`. Проверить состояние службы можно командой `systemctl status abrtd.service`, а чтобы остановить ее, нужно скомпандовать `systemctl stop abrtd.service`. И наконец, команда `systemctl enable abrtd.service` выполнит установку `service`-файла, обеспечив его активацию при каждой загрузке (аналог `chkconfig abrtd on` в классическом SysV).

Приведенный выше `service`-файл является практически точным переводом исходного `init`-скрипта, и он никак не использует широкий спектр возможностей, предоставляемых `systemd`. Ниже приведен немного улучшенный вариант этого же файла:

быточным, так как соответствующая функциональность поддерживается автоматически. Однако, эту строчку все-таки стоит указать для обеспечения совместимости с системами, использующими устаревшие реализации демона системного лога.

¹⁸В том контексте, в котором он используется в большинстве дистрибутивов семейства Red Hat, а именно, многопользовательский режим без запуска графической оболочки.

¹⁹Прим. перев.: К «ненормальным» загрузкам можно отнести, например, загрузки в режимах `emergency.target` или `rescue.target` (является аналогом первого уровня исполнения в классической SysV).

²⁰Обратите внимание, что режим графической загрузки в `systemd` (`graphical.target`, аналог `runlevel 5` в SysV) является надстройкой над режимом многопользовательской консольной загрузки (`multi-user.target`, аналог `runlevel 3` в SysV). Таким образом, все службы, запускаемые в режиме `multi-user.target`, будут также запускаться и в режиме `graphical.target`.

```

[Unit]
Description=ABRT Automated Bug Reporting Tool
After=syslog.target

[Service]
Type=dbus
BusName=com.redhat.abrt
ExecStart=/usr/sbin/abrttd -d -s

[Install]
WantedBy=multi-user.target

```

Чем же новый вариант отличается от предыдущего? Ну, прежде всего, мы уточнили описание службы. Однако, ключевым изменением является замена значения `Type` с `forking` на `dbus` и связанные с ней изменения: добавление имени службы в шине D-Bus (директива `BusName`) и задание дополнительных аргументов `abrttd` «`-d -s`». Но зачем вообще нужна эта замена? Каков ее практический смысл? Чтобы ответить на этот вопрос, мы снова возвращаемся к демонизации. В ходе данной операции процесс дважды форкается и отключается от всех терминалов. Это очень удобно при запуске демона через скрипт, но в случае использования таких продвинутых систем инициализации, как `systemd`, подобное поведение не дает никаких преимуществ, но вызывает неоправданные задержки. Даже если мы оставим в стороне вопрос скорости загрузки, останется такой важный аспект, как отслеживание состояния служб. `systemd` решает и эту задачу, контролируя работу службы и при необходимости реагируя на различные события. Например, при неожиданном падении основного процесса службы, `systemd` должен зарегистрировать идентификатор и код выхода процесса, также, в зависимости от настроек, он может попытаться перезапустить службу, либо активировать какой-либо заранее заданный юнит. Операция демонизации несколько затрудняет решение этих задач, так как обычно довольно сложно найти связь демонизированного процесса с исходным (собственно, смысл демонизации как раз и сводится к уничтожению этой связи) и, соответственно, для `systemd` сложнее определить, какой из порожденных в рамках данной службы процессов является основным. Чтобы упростить для него решение этой задачи, мы и воспользовались типом запуска `dbus`. Он подходит для всех служб, которые в конце процесса инициализации регистрируют свое имя на шине D-Bus²¹. `ABRTd` относится к ним. С новыми настройками, `systemd` запустит процесс `abrttd`, который уже не будет форкаться (согласно указанным нами ключам «`-d -s`»), и в качестве момента окончания периода запуска данной службы `systemd` будет рассматривать момент регистрации имени `com.redhat.abrt` на шине D-Bus. В этом случае основным для данной службы будет считаться процесс, непосредственно порожденный `systemd`. Таким образом, `systemd` располагает удобным методом для определения момента окончания запуска службы, а также может легко отслеживать ее состояние.

Собственно, это все, что нужно было сделать. Мы получили простой конфигурационный файл, в 10 строчках которого содержится больше полезной информации, чем в 115 строках исходного `init`-скрипта. Добавляя в наш файл по одной строчке, мы можем использовать различные полезные функции `systemd`, создание аналога которых в традиционном `init`-скрипте потребовало бы значительных усилий. Например, добавив строку `Restart=restart-always`, мы приказываем `systemd` автоматически перезапускать службу после каждого ее падения. Или, например, добавив `OOMScoreAdjust=-500`, мы попросим ядро сбросить эту службу, даже если OOM Killer выйдет на тропу войны. А если мы добавим строку `CPU schedulingPolicy=idle`, процесс `abrttd` будет работать только в те моменты, когда система больше ничем не занята, что позволит не создавать помех для процессов, активно использующих CPU.

За более подробным описанием всех опций настройки, вы можете обратиться к страницам руководства [systemd.unit](#), [systemd.service](#), [systemd.exec](#). Полный список доступных страниц можно посмотреть [здесь](#).

²¹В настоящее время практически все службы дистрибутива Fedora после запуска регистрируются на шине D-Bus.

Конечно, отнюдь не все `init`-скрипты так же легко преобразовать в `service`-файлы. Но, к счастью, «проблемных» скриптов не так уж и много.

4 Убить демона

Убить системного демона нетрудно, правда? Или... все не так просто?

Если ваш демон функционирует как один процесс, все действительно очень просто. Вы командуете `killall rsyslogd`, и демон системного лога останавливается. Впрочем, этот метод не вполне корректен, так как он действует не только на самого демона, но и на другие процессы с тем же именем. Иногда подобное поведение может привести к неприятным последствиям. Более правильным будет использование `pid`-файла: `kill $(cat /var/run/syslogd.pid)`. Вот, вроде бы, и все, что вам нужно... Или мы упускаем еще что-то?

Действительно, мы забываем одну простую вещь: существуют службы, такие, как `Apache`, `crond`, `atd`, которые по роду служебной деятельности должны запускать дочерние процессы. Это могут быть совершенно посторонние, указанные пользователем программы (например, задачи `crond/at`, `CGI`-скрипты) или полноценные серверные процессы (например, `Apache workers`). Когда вы убиваете основной процесс, он может остановить все дочерние процессы. А может и не остановить. В самом деле, если служба функционирует в штатном режиме, ее обычно останавливают командой `stop`. К прямому вызову `kill` администратор, как правило, прибегает только в аварийной ситуации, когда служба работает неправильно и может не среагировать на стандартную команду остановки. Таким образом, убив, например, основной сервер `Apache`, вы можете получить от него в наследство работающие `CGI`-скрипты, причем их родителем автоматически станет `PID 1 (init)`, так что установить их происхождение будет не так-то просто.

`systemd` спешит к нам на помощь. Команда `systemctl kill` позволит отправить сигнал всем процессам, порожденным в рамках данной службы. Например:

```
# systemctl kill crond.service
```

Вы можете быть уверены, что всем процессам службы `crond` будет отправлен сигнал `SIGTERM`. Разумеется, можно отправить и любой другой сигнал. Скажем, если ваши дела совсем уж плохи, вы можете воспользоваться и `SIGKILL`:

```
# systemctl kill -s SIGKILL crond.service
```

После ввода этой команды, служба `crond` будет жестоко убита вместе со всеми ее дочерними процессами, вне зависимости от того, сколько раз она форкалась, и как бы она ни пыталась сбежать из-под нашего контроля при помощи двойного форка или [форк-бомбардировки](#)²².

В некоторых случаях возникает необходимость отправить сигнал именно основному процессу службы. Например, используя `SIGHUP`, мы можем заставить демона перечитать файлы конфигурации. Разумеется, передавать `HUP` вспомогательным процессам в этом случае совершенно необязательно. Для решения подобной задачи неплохо подойдет и классический метод с `pid`-файлом, однако у `systemd` и на этот случай есть простое решение, избавляющее вас от необходимости искать нужный файл:

```
# systemctl kill -s HUP --kill-who=main crond.service
```

²²Прим. перев.: Стоит особо отметить, что использование контрольных групп не только упрощает процесс уничтожения форк-бомб, но и значительно уменьшает ущерб от работающей форк-бомбы. Так как `systemd` автоматически помещает каждую службу и каждый пользовательский сеанс в свою контрольную группу по ресурсу процессорного времени, запуск форк-бомбы одним пользователем или службой не создаст значительных проблем с отзывчивостью системы у других пользователей и служб. Таким образом, в качестве основной угрозы форк-бомбардировки остаются лишь возможности исчерпания памяти и идентификаторов процессов (`PID`). Впрочем, и их тоже можно легко устранить: достаточно задать соответствующие лимиты в конфигурационном файле службы (см. [systemd.exec\(5\)](#) и [systemd.resource-control\(5\)](#)).

Итак, что же принципиально новое привносит `systemd` в рутинный процесс убийства демона? Прежде всего: впервые в истории Linux представлен способ принудительной остановки службы, не зависящий от того, насколько добросовестно основной процесс службы выполняет свои обязательства по остановке дочерних процессов. Как уже упоминалось выше, необходимость отправить процессу `SIGTERM` или `SIGKILL` обычно возникает именно в нештатной ситуации, когда вы уже не можете быть уверены, что демон корректно исполнит все свои обязанности.

После прочтения сказанного выше у вас может возникнуть вопрос: в чем разница между `systemctl kill` и `systemctl stop`? Отличие состоит в том, что `kill` просто отправляет сигнал заданному процессу, в то время как `stop` действует по «официально» определенному методу, вызывая команду, определенную в параметре `ExecStop` конфигурации службы. Обычно команды `stop` бывает вполне достаточно для остановки службы, и к `kill` приходится прибегать только в крайних случаях, например, когда служба «зависла» и не реагирует на команды.

Кстати говоря, при использовании параметра «-s», вы можете указывать названия сигналов как с префиксом `SIG`, так и без него — оба варианта будут работать.

В завершение стоит сказать, что для нас весьма интересным и неожиданным оказался тот факт, что до появления `systemd` в Linux просто не существовало инструментов, позволяющих корректно отправить сигнал службе в целом, а не отдельному процессу.

5 Три уровня выключения

В `systemd` существует три уровня (разновидности) действий, направленных на прекращение работы службы (или любого другого юнита):

- Вы можете *остановить* службу, то есть прекратить выполнение уже запущенных процессов службы. При этом сохраняется возможность ее последующего запуска, как ручного (через команду `systemctl start`), так и автоматического (при загрузке системы, при поступлении запроса через сокет или системную шину, при срабатывании таймера, при подключении соответствующего оборудования и т.д.). Таким образом, остановка службы является временной мерой, не дающей никаких гарантий на будущее.

В качестве примера рассмотрим остановку службы `NTPd` (отвечающей за синхронизацию времени по сети):

```
| systemctl stop ntpd.service
```

Аналогом этой команды в классическом SysV init является

```
| service ntpd stop
```

Заметим, что в Fedora 15, использующей в качестве системы инициализации `systemd`, в целях обеспечения обратной совместимости допускается использование классических SysV-команд, и `systemd` будет корректно воспринимать их. В частности, вторая приведенная здесь команда будет эквивалентна первой.

- Вы можете *отключить* службу, то есть отсоединить ее от всех триггеров активации. В результате служба уже не будет автоматически запускаться ни при загрузке системы, ни при обращении к сокету или адресу на шине, ни при подключении оборудования, и т.д. Но при этом сохраняется возможность «ручного» запуска службы (командой `systemctl start`). Обратите внимание, что при отключении уже запущенной службы, ее выполнение в текущем сеансе не останавливается — это нужно сделать отдельно, иначе процессы службы будут работать до момента выключения системы (но при следующем включении, разумеется, уже не запустятся).

Рассмотрим отключение службы на примере все того же `NTPd`:

```
| systemctl disable ntpd.service
```

В классических SysV-системах аналогичная команда будет иметь вид

```
chkconfig ntpd off
```

Как и в предыдущем случае, в Fedora 15 вторая из этих команд будет действовать аналогично первой.

Довольно часто приходится сочетать действия отключения и остановки службы — такая комбинированная операция гарантирует, что уже исполняющиеся процессы службы будут прекращены, и служба больше не будет запускаться автоматически (но может быть запущена вручную):

```
systemctl disable ntpd.service
systemctl stop ntpd.service
```

Подобное сочетание команд используется, например, при деинсталляции пакетов в Fedora.

Обратите внимание, что отключение службы является перманентной мерой, и действует вплоть до явной отмены соответствующей командой. Перезагрузка системы не отменяет отключения службы.

- Вы можете *заблокировать* (замаскировать) службу. Действие этой операции аналогично отключению, но дает более сильный эффект. Если при отключении отменяется только возможность автоматического запуска службы, но сохраняется возможность ручного запуска, то при блокировке исключаются обе эти возможности. Отметим, что использование данной опции при непонимании принципов ее работы может привести к трудно диагностируемым ошибкам.

Тем не менее, рассмотрим пример блокировки все той же службы NTPd:

```
ln -s /dev/null /etc/systemd/system/ntpd.service
systemctl daemon-reload
```

Итак, блокировка сводится к созданию символической ссылки с именем соответствующей службы, указывающей на `/dev/null`²³. После такой операции служба не может быть запущена ни вручную, ни автоматически. Символическая ссылка создается в каталоге `/etc/systemd/system/`, а ее имя должно соответствовать имени файла описания службы из каталога `/lib/systemd/system/` (в нашем случае `ntpd.service`).

Заметим, что `systemd` читает файлы конфигурации из обоих этих каталогов, но файлы из `/etc` (управляемые системным администратором) имеют приоритет над файлами из `/lib` (которые управляются пакетным менеджером). Таким образом, создание символической ссылки (или обычного файла) `/etc/systemd/system/ntpd.service` предотвращает чтение штатного файла конфигурации `/lib/systemd/system/ntpd.service`.

В выводе `systemctl status` заблокированные службы отмечаются словом `masked`. Попытка запустить такие службы командой `systemctl start` завершится ошибкой.

В рамках классического SysV `init`, штатная реализация такой возможности отсутствует. Похожий эффект может быть достигнут с помощью «костылей», например, путем добавления команды `exit 0` в начало `init`-скрипта. Однако, подобные решения имеют ряд недостатков, например, потенциальная возможность конфликтов с пакетным менеджером (при очередном обновлении исправленный скрипт может быть просто затерт соответствующим файлом из пакета).

²³Прим. перев.: Впоследствии в программу `systemctl` была добавлена поддержка команд `mask` и `unmask`, упрощающих процесс блокирования и разблокирования юнитов. Например, для блокирования службы `ntpd.service` теперь достаточно команды `systemctl mask ntpd.service`. Фактически она делает то же самое, что и приведенная выше команда `ln`.

Стоит отметить, что блокировка службы, как и ее отключение, является перманентной мерой²⁴.

После прочтения изложенного выше, у читателя может возникнуть вопрос: как отменить произведенные изменения? Что ж, ничего сложного тут нет: `systemctl start` отменяет действия `systemctl stop`, `systemctl enable` отменяет действие `systemctl disable`, а `rm` отменяет действие `ln`.

6 Смена корня

Практически все администраторы и разработчики рано или поздно встречаются с **chroot-окружениями**. Системный вызов `chroot()` позволяет задать для определенного процесса (и его потомков) каталог, который они будут рассматривать как корневой `/`, тем самым ограничивая для них область видимости иерархии файловой системы отдельной ветвью. Большинство применений `chroot`-окружений можно отнести к двум классам задач:

1. Обеспечение безопасности. Потенциально уязвимый демон `chroot`'ится в отдельный каталог и, даже в случае успешной атаки, взломщик увидит лишь содержимое этого каталога, а не всю файловую систему — он окажется в ловушке `chroot`'а.
2. Подготовка и управление образом операционной системы при отладке, тестировании, компиляции, установке или восстановлении. При этом вся иерархия файловых систем гостевой ОС монтируется или создается в каталоге системы-хоста, и при запуске оболочки (или любого другого приложения) внутри этой иерархии, в качестве корня используется данный каталог. Система, которую «видят» такие программы, может сильно отличаться от ОС хоста. Например, это может быть другой дистрибутив, или даже другая аппаратная архитектура (запуск `i386`-гостя на `x86_64`-хосте). Гостевая ОС не может увидеть полного дерева каталогов ОС хоста.

В системах, использующих классический SysV `init`, использовать `chroot`-окружения сравнительно несложно. Например, чтобы запустить выбранного демона внутри дерева каталогов гостевой ОС, достаточно смонтировать внутри этого дерева `/proc`, `/sys` и остальные API ФС, воспользоваться программой `chroot(1)` для входа в окружение, и выполнить соответствующий `init`-скрипт, запустив `/sbin/service` внутри окружения.

Но в системах, использующих `systemd`, уже не все так просто. Одно из важнейших достоинств `systemd` состоит в том, что параметры среды, в которой запускаются демоны, никак не зависят от метода их запуска. В системах, использующих SysV `init`, многие параметры среды выполнения (в частности, лимиты на системные ресурсы, переменные окружения, и т.п.) наследуются от оболочки, из которой был запущен `init`-скрипт. При использовании `systemd` ситуация меняется радикально: пользователь просто уведомляет процесс `init` о необходимости запустить ту или иную службу, и тот запускает демона в чистом, созданном «с нуля» и тщательно настроенном окружении, параметры которого никак не зависят от настроек среды, из которой была отдана команда. Такой подход полностью отменяет традиционный метод запуска демонов в `chroot`-окружениях: теперь демон порождается процессом `init` (PID 1) и наследует корневой каталог от него, вне зависимости от того, находился ли пользователь, отдавший команду на запуск, в `chroot`-окружении, или нет. Кроме того, стоит особо отметить, что взаимодействие управляющих программ с `systemd` происходит через сокеты, находящиеся в каталоге `/run/systemd`, так что программы, запущенные в `chroot`-окружении, просто не смогут

²⁴Прим. перев.: Подробно описав принцип работы блокировки службы (юнита), автор забывает привести практические примеры ситуаций, когда эта возможность оказывается полезной.

В частности, иногда бывает необходимо полностью предотвратить запуск службы в любой ситуации. При этом не стоит забывать, что в `post-install` скриптах пакетного менеджера или, скажем, в заданиях `cron`, вместо `systemctl try-restart (service condrestart)` может быть ошибочно указано `systemctl restart (service restart)`, что является прямым указанием на запуск службы, если она еще не запущена. Вследствие таких ошибок, отключенная служба может «ожить» в самый неподходящий момент.

взаимодействовать с `init`-подсистемой (и это, в общем, неплохо, а если такое ограничение будет создавать проблемы, его можно легко обойти, используя `bind`-монтирование).

В свете вышесказанного, возникает вопрос: как правильно использовать `chroot`-окружения в системах на основе `systemd`? Что ж, постараемся дать подробный и всесторонний ответ на этот вопрос.

Для начала, рассмотрим первое из перечисленных выше применений `chroot`: изоляция в целях безопасности. Прежде всего, стоит заметить, что защита, предоставляемая `chroot`'ом, весьма эфемерна и ненадежна, так как `chroot` не является «дорогой с односторонним движением». Выйти из `chroot`-окружения сравнительно несложно, и соответствующее предупреждение даже [присутствует на странице руководства](#). Действительно эффективной защиты можно достичь, только сочетая `chroot` с другими методиками. В большинстве случаев, это возможно только при наличии поддержки `chroot` в самой программе. Прежде всего, корректное конфигурирование `chroot`-окружения требует глубокого понимания принципов работы программы. Например, нужно точно знать, какие каталоги нужно `bind`-монтировать из основной системы, чтобы обеспечить все необходимые для работы программы каналы связи. С учетом вышесказанного, эффективная `chroot`-защита обеспечивается только в том случае, когда она реализована в коде самого демона. Именно разработчик лучше других знает (*обязан* знать), как правильно сконфигурировать `chroot`-окружение, и какой минимальный набор файлов, каталогов и файловых систем необходим внутри него для нормальной работы демона. Уже сейчас существуют демоны, имеющие встроенную поддержку `chroot`. К сожалению, в системе Fedora, установленной с параметрами по умолчанию, таких демонов всего два: [Avahi](#) и [RealtimeKit](#). Оба они написаны одним очень хитрым человеком ;-). (Вы можете собственноручно убедиться в этом, выполнив команду `ls -l /proc/*/root`.)

Возвращаясь к теме нашего обсуждения: разумеется, `systemd` позволяет помещать выбранных демонов в `chroot`, и управлять ими точно так же, как и остальными. Достаточно лишь указать параметр `RootDirectory=` в соответствующем `service`-файле. Например:

```
[Unit]
Description=A chroot()ed Service

[Service]
RootDirectory=/srv/chroot/foobar
ExecStartPre=/usr/local/bin/setup-foobar-chroot.sh
ExecStart=/usr/bin/foobard
RootDirectoryStartOnly=yes
```

Рассмотрим этот пример подробнее. Параметр `RootDirectory=` задает каталог, в который производится `chroot` перед запуском исполняемого файла, заданного параметром `ExecStart=`. Заметим, что путь к этому файлу должен быть указан относительно каталога `chroot` (так что, в нашем случае, с точки зрения основной системы, на исполнение будет запущен файл `/srv/chroot/foobar/usr/bin/foobard`). Перед запуском демона будет вызван сценарий оболочки `setup-foobar-chroot.sh`, который должен обеспечить подготовку `chroot`-окружения к запуску демона (например, смонтировать в нем `/proc` и/или другие файловые системы, необходимые для работы демона). Указав `RootDirectoryStartOnly=yes`, мы задаем, что `chroot()` будет выполняться только перед выполнением файла из `ExecStart=`, а команды из других директив, в частности, `ExecStartPre=`, будут иметь полный доступ к иерархии файловых систем ОС (иначе наш скрипт просто не сможет выполнить `bind`-монтирование нужных каталогов). Более подробную информацию по опциям конфигурации вы можете получить на [страницах руководства](#).

Поместив приведенный выше текст примера в файл `/etc/systemd/system/foobar.service`, вы сможете запустить `chroot`'нутого демона командой `systemctl start foobar.service`. Информацию о его текущем состоянии можно получить с помощью команды `systemctl status foobar.service`. Команды управления и мониторинга службы не зависят от того, запущена ли она в `chroot`'е, или нет. Этим `systemd` отличается от классического SysV `init`.

Новые ядра Linux поддерживают возможность создания независимых пространств имен файловых систем (в дальнейшем FSNS, от «file system namespaces»). По функциональности этот механизм аналогичен `chroot()`, однако предоставляет гораздо более широкие возможности, и в нем отсутствуют проблемы с безопасностью, характерные для `chroot`. `systemd` позволяет использовать при конфигурировании юнитов некоторые возможности, предоставляемые FSNS. В частности, использование FSNS часто является гораздо более простой и удобной альтернативой созданию полновесных `chroot`-окружений. Используя директивы `ReadOnlyDirectories=`, `InaccessibleDirectories=`, вы можете задать ограничения по использованию иерархии файловых систем для заданной службы: ее корнем будет системный корневой каталог, однако указанные в этих директивах подкаталоги будут доступны только для чтения или вообще недоступны для нее. Например:

```
[Unit]
Description=A Service With No Access to /home

[Service]
ExecStart=/usr/bin/foobard
InaccessibleDirectories=/home
```

Такая служба будет иметь доступ ко всей иерархии файловых систем ОС, с единственным исключением — она не будет видеть каталог `/home`, что позволит защитить данные пользователей от потенциальных хакеров. (Подробнее об этих опциях можно почитать на [странице руководства](#).)

Фактически, FSNS по множеству параметров превосходят `chroot()`. Скорее всего, Avahi и RealtimeKit в ближайшем будущем перейдут от `chroot()` к использованию FSNS.

Итак, мы рассмотрели вопросы использования `chroot` для обеспечения безопасности. Переходим ко второму пункту: подготовка и управление образом операционной системы при отладке, тестировании, компиляции, установке или восстановлении.

`chroot`-окружения, по сути, весьма примитивны: они изолируют только иерархии файловых систем. Даже после `chroot`'а в определенный подкаталог, процесс по-прежнему имеет полный доступ к системным вызовам, может убивать процессы, запущенные в основной системе, и т.п. Вследствие этого, запуск полноценной ОС (или ее части) внутри `chroot`'а несет угрозу для хост-системы: у гостя и хоста отличается лишь содержимое файловой системы, все остальное у них общее. Например, если вы обновляете дистрибутив, установленный в `chroot`-окружении, и пост-установочный скрипт пакета отправляет `SIGTERM` процессу `init` для его перезапуска²⁵, на него среагирует именно хост-система! Кроме того, хост и `chroot`'нутая система будут иметь общую разделяемую память SysV (SysV shared memory), общие сокеты из абстрактных пространств имен (abstract namespace sockets) и другие элементы ИРС. Для отладки, тестирования, компиляции, установки и восстановления ОС не требуется абсолютно неуязвимая изоляция — нужна лишь надежная защита от *случайного* воздействия на ОС хоста изнутри `chroot`-окружения, иначе вы можете получить целый букет проблем, как минимум, от пост-инсталляционных скриптов при установке пакетов в `chroot`-окружении.

`systemd` имеет целый ряд возможностей, полезных для работы с `chroot`-системами:

Прежде всего, управляющая программа `systemctl` автоматически определяет, что она запущена в `chroot`-системе. В такой ситуации будут работать только команды `systemctl enable` и `systemctl disable`, во всех остальных случаях `systemctl` просто не будет ничего делать, возвращая код успешного завершения операции. Таким образом, пакетные скрипты смогут включить/отключить запуск «своих» служб при загрузке (или в других ситуациях), однако команды наподобие `systemctl restart` (обычно выполняется при обновлении пакета) не дадут никакого эффекта внутри `chroot`-окружения²⁶.

²⁵Прим. перев.: Во избежание путаницы отметим, что перезапуск процесса `init` (PID 1) «на лету» при получении `SIGTERM` поддерживается только в `systemd`, в классическом SysV `init` такой возможности нет.

²⁶Прим. перев.: Автор забывает отметить не вполне очевидный момент: такое поведение `systemctl` проявляется только в «мертвых» окружениях, т.е. в тех, где не запущен процесс `init`, и соответственно

Однако, куда более интересные возможности предоставляет программа `systemd-nspawn`, входящая в стандартный комплект поставки `systemd`. По сути, это улучшенный аналог `chroot(1)` — она не только подменяет корневой каталог, но и создает отдельные пространства имен для дерева файловых систем (FSNS) и для идентификаторов процессов (PID NS), предоставляя легковесную реализацию системного контейнера²⁷. `systemd-nspawn` проста в использовании как `chroot(1)`, однако изоляция от хост-системы является более полной и безопасной. Всего одной командой вы можете загрузить внутри контейнера *полноценную* ОС (на базе `systemd` или `SysV init`). Благодаря использованию независимых пространств идентификаторов процессов, процесс `init` внутри контейнера получит PID 1, что позволит работать ему в штатном режиме. Также, в отличие от `chroot(1)`, внутри окружения будут автоматически смонтированы `/proc` и `/sys`.

Следующий пример иллюстрирует возможность запустить Debian на Fedora-хосте всего тремя командами:

```
# yum install debootstrap
# debootstrap --arch=amd64 unstable debian-tree/
# systemd-nspawn -D debian-tree/
```

Вторая из этих команд обеспечивает развертывание в подкаталоге `./debian-tree/` файловой структуры дистрибутива Debian, после чего третья команда запускает внутри полученной системы процесс командной оболочки. Если вы хотите запустить внутри контейнера полноценную ОС, воспользуйтесь командой

```
# systemd-nspawn -D debian-tree/ /sbin/init
```

После быстрой загрузки вы получите приглашение оболочки, запущенной внутри полноценной ОС, функционирующей в контейнере. Изнутри контейнера невозможно увидеть процессы, которые находятся вне его. Контейнер сможет пользоваться сетью хоста²⁸, однако не имеет возможности изменить ее настройки (это может привести к серии ошибок в процессе загрузки гостевой ОС, но ни одна из этих ошибок не должна быть критической). Контейнер получает доступ к `/sys` и `/proc/sys`, однако, во избежание вмешательства контейнера в конфигурацию ядра и аппаратного обеспечения хоста, эти каталоги будут смонтированы только для чтения. Обратите внимание, что эта защита блокирует лишь *случайные, непредвиденные* попытки изменения параметров. При необходимости, процесс внутри контейнера, обладающий достаточными полномочиями, сможет перемонтировать эти файловые системы в режиме чтения-записи.

Итак, что же такого хорошего в `systemd-nspawn`?

1. Использовать эту утилиту очень просто. Вам даже не нужно вручную монтировать внутри окружения `/proc` и `/sys` — она сделает это за вас, а ядро автоматически отмонтирует их, когда последний процесс контейнера завершится.
2. Обеспечивается надежная изоляция, предотвращающая случайные изменения параметров ОС хоста изнутри контейнера.
3. Теперь вы можете загрузить внутри контейнера полноценную ОС, а не единственную оболочку.

отсутствуют управляющие сокетты в `/run/systemd`. Такая ситуация возникает, например, при установке системы в `chroot` через `debootstrap/febootstrap`. В этом случае возможности `systemctl` ограничиваются операциями с символическими ссылками, определяющими триггеры активации юнитов, т.е. выполнением действий `enable` и `disable`, не требующих непосредственного взаимодействия с процессом `init`.

²⁷Прим. перев.: Используемые в `systemd-nspawn` механизмы ядра Linux, такие, как FS NS и PID NS, также лежат в основе LXC, системы контейнерной изоляции для Linux, которая позиционируется как современная альтернатива классическому OpenVZ. Стоит отметить, что LXC ориентирована прежде всего на создание независимых виртуальных окружений, с поддержкой отдельных сетевых стеков, ограничением на ресурсы, сохранением настроек и т.п., в то время как `systemd-nspawn` является лишь более удобной и эффективной заменой команды `chroot(1)`, предназначенной прежде всего для развертывания, восстановления, сборки и тестирования операционных систем. Далее автор разъясняет свою точку зрения на этот вопрос.

²⁸Прим. перев.: Впоследствии в `systemd-nspawn` была добавлена опция `--private-network`, позволяющая изолировать систему внутри контейнера от сети хоста: такая система будет видеть только свой собственный интерфейс обратной петли.

4. Эта утилита очень компактна и присутствует везде, где установлен `systemd`. Она не требует специальной установки и настройки.

`systemd` уже подготовлен для работы внутри таких контейнеров. Например, когда подается команда на выключение системы внутри контейнера, `systemd` на последнем шаге вызывает не `reboot()`, а просто `exit()`.

Стоит отметить, что `systemd-nspawn` все же не является полноценной системой контейнерной виртуализации/изоляции — если вам нужно такое решение, воспользуйтесь [LXC](#). Этот проект использует те же самые механизмы ядра, но предоставляет куда более широкие возможности, включая виртуализацию сети. Могу предложить такую аналогию: `systemd-nspawn` как реализация контейнера похожа на GNOME 3 — компактна и проста в использовании, опций для настройки очень мало. В то время как `LXC` больше похож на KDE: опций для настройки больше, чем строк кода. Я создал `systemd-nspawn` специально для тестирования, отладки, сборки, восстановления. Именно для этих задач вам стоит ее использовать — она неплохо с ними справляется, куда лучше, чем `chroot(1)`.

Что ж, пора заканчивать. Итак:

1. Использование `chroot()` для обеспечения безопасности дает наилучший результат, когда оно реализовано непосредственно в коде самой программы.
2. `ReadOnlyDirectories=` и `InaccessibleDirectories=` могут быть удобной альтернативой созданию полноценных `chroot`-окружений.
3. Если вам нужно поместить в `chroot`-окружение какую-либо службу, воспользуйтесь опцией `RootDirectory=`.
4. `systemd-nspawn` — очень неплохая штука.
5. `chroot`'ы убоги, FSNS — [1337](#).

И все это уже сейчас доступно в Fedora 15.

7 Поиск виновных

Fedora 15²⁹ является первым релизом Fedora, использующим `systemd` в качестве системы инициализации по умолчанию. Основной нашей целью при работе над выпуском F15 является обеспечение полной взаимной интеграции и корректной работы всех компонентов. При подготовке следующего релиза, F16, мы сконцентрируемся на дальнейшей полировке и ускорении системы. Для этого мы подготовили ряд инструментов (доступных уже в F15), которые должны помочь нам в поиске проблем, связанных с процессом загрузки. В этой статье я попытаюсь рассказать о том, как найти виновников медленной загрузки вашей системы, и о том, что с ними делать дальше.

Первый инструмент, который мы можем вам предложить, очень прост: по завершении загрузки, `systemd` регистрирует в системном журнале информацию о суммарном времени загрузки:

```
systemd[1]: Startup finished in 2s 65ms 924us (kernel) + 2s 828ms 195us (initrd)
+ 11s 900ms 471us (userspace) = 16s 794ms 590us.
```

Эта запись означает следующее: на инициализацию ядра (до момента запуска `initrd`, т.е. `dracut`) ушло 2 секунды. Далее, чуть менее трех секунд работал `initrd`. И наконец, почти 12 секунд было потрачено `systemd` на запуск программ из пространства пользователя. Итоговое время, начиная с того момента, как загрузчик передал управление коду ядра, до того момента, как `systemd` завершил все операции, связанные с загрузкой системы, составило почти 17 секунд. Казалось бы, смысл этого числа вполне очевиден. . . Однако не стоит делать поспешных выводов. Прежде всего, сюда не входит время, затраченное на инициализацию вашего сеанса в GNOME, так как эта задача уже выходит

²⁹Величайший в истории релиз свободной ОС.

за рамки задач процесса `init`. Кроме того, в этом показателе учитывается только время работы `systemd`, хотя часто бывает так, что некоторые демоны продолжают *свою* работу по инициализации уже после того, как секундомер остановлен. Проще говоря: приведенные числа позволяют лишь оценить общую скорость загрузки, однако они не являются точной характеристикой длительности процесса.

Кроме того, эта информация носит поверхностный характер: она не сообщает, какие именно системные компоненты заставляют `systemd` ждать так долго. Чтобы исправить это упущение, мы ввели команду `systemd-analyze blame`:

```
$ systemd-analyze blame
6207ms udev-settle.service
5228ms cryptsetup@luks\x2d9899b85d\x2df790\x2d4d2a\x2da650\x2d8b7d2fb92cc3.service
 735ms NetworkManager.service
 642ms avahi-daemon.service
 600ms abrt.service
 517ms rtkit-daemon.service
 478ms fedora-storage-init.service
 396ms dbus.service
 390ms rpcidmapd.service
 346ms systemd-tmpfiles-setup.service
 322ms fedora-sysinit-unhack.service
 316ms cups.service
 310ms console-kit-log-system-start.service
 309ms libvirt.service
 303ms rpcbind.service
 298ms ksmtuned.service
 288ms lvm2-monitor.service
 281ms rpcgssd.service
 277ms sshd.service
 276ms livesys.service
 267ms iscsid.service
 236ms mdmonitor.service
 234ms nfslock.service
 223ms ksm.service
 218ms mcelog.service
...
```

Она выводит список юнитов `systemd`, активированных при загрузке, с указанием времени инициализации для каждого из них. Список отсортирован по убыванию этого времени, поэтому наибольший интерес для нас представляют первые строчки. В нашем случае это `udev-settle.service` и `cryptsetup@luks\x2d9899b85d\x2df790\x2d4d2a\x2da650\x2d8b7d2fb92cc3.service`, инициализация которых занимает более одной секунды. Стоит отметить, что к анализу вывода команды `systemd-analyze blame` тоже следует подходить с осторожностью: она не поясняет, *почему* тот или иной юнит тратит столько-то времени, она лишь констатирует факт, что время было затрачено. Кроме того, не стоит забывать, что юниты могут запускаться параллельно. В частности, если две службы были запущены одновременно, то время их инициализации будет значительно меньше, чем сумма времен инициализации каждой из них.

Рассмотрим повнимательнее первого осквернителя нашей загрузки: службу `udev-settle.service`³⁰. Почему ей требуется так много времени для запуска, и что мы можем с этим сделать? Данная служба выполняет очень простую функцию: она ожидает, пока `udev` завершит опрос устройств, после чего завершается. Опрос же устройств может занимать довольно много времени. Например, в нашем случае опрос устройств длится более 6 секунд из-за подключенного к компьютеру 3G-модема, в котором отсутствует SIM-карта. Этот модем очень долго отвечает на запросы `udev`. Опрос

³⁰Прим. перев.: После объединения проектов `systemd` и `udev`, эта служба называется `systemd-udev-settle.service`.

устройств является частью схемы, обеспечивающей работу ModemManager'a и позволяющей NetworkManager'у упростить для вас настройку 3G. Казалось бы, очевидно, что виновником задержки является именно ModemManager, так как опрос устройств для него занимает слишком много времени. Но такое обвинение будет заведомо ошибочным. Дело в том, что опрос устройств очень часто оказывается довольно длительной процедурой. Медленный опрос 3G-устройств для ModemManager является частным случаем, отражающим это общее правило. Хорошая система опроса устройств обязательно должна учитывать тот факт, что операция опроса любого из устройств может затянуться надолго. Истинной причиной наших проблем является необходимость ожидать завершения опроса, т.е., наличие службы `udev-settle.service` как обязательной части нашего процесса загрузки.

Но почему эта служба вообще присутствует в нашем процессе загрузки? На самом деле, мы можем прекрасно обойтись и без нее. Она нужна лишь как часть используемой в Fedora схемы инициализации устройств хранения, а именно, набора скриптов, выполняющих настройку LVM, RAID и multipath-устройств. На сегодняшний день, реализация этих систем не поддерживает собственного механизма поиска и опроса устройств, и поэтому их запуск должен производиться только после того, как эта работа уже проделана — тогда они могут просто последовательно перебирать все диски. Однако, такой подход противоречит одному из фундаментальных требований к современным компьютерам: возможности подключать и отключать оборудование в любой момент, как при выключенном компьютере, так и при включенном. Для некоторых технологий невозможно точно определить момент завершения формирования списка устройств (например, это характерно для USB и iSCSI), и поэтому процесс опроса таких устройств обязательно должен включать некоторую фиксированную задержку, гарантирующую, что все подключенные устройства успеют отозваться. С точки зрения скорости загрузки это, безусловно, негативный эффект: соответствующие скрипты заставляют нас ожидать завершения опроса устройств, хотя большинство этих устройств не являются необходимыми на данной стадии загрузки. В частности, в рассматриваемой нами системе LVM, RAID и multipath вообще не используются!³¹

С учетом вышесказанного, мы можем спокойно исключить `udev-settle.service` из нашего процесса загрузки, так как мы не используем ни LVM, ни RAID, ни multipath. Замаскируем эти службы, чтобы увеличить скорость загрузки:

```
# ln -s /dev/null /etc/systemd/system/udev-settle.service
# ln -s /dev/null /etc/systemd/system/fedora-wait-storage.service
# ln -s /dev/null /etc/systemd/system/fedora-storage-init.service
# systemctl daemon-reload
```

После перезагрузки мы видим, что загрузка стала на одну секунду быстрее. Но почему выигрыш оказался таким маленьким? Из-за второго осквернителя нашей загрузки — `cryptsetup`. На рассматриваемой нами системе зашифрован раздел `/home`. Специально для нашего тестирования я записал пароль в файл, чтобы исключить влияние скорости ручного набора пароля. К сожалению, для подключения зашифрованного раздела `cryptsetup` требует более пяти секунд. Будем ленивы, и вместо того, чтобы исправлять

³¹ Наиболее правильным решением в данном случае будет ожидание событий подключения устройств (реализованное через `libudev` или аналогичную технологию) с соответствующей обработкой каждого такого события — подобный подход позволит нам продолжить загрузку, как только будут готовы все устройства, необходимые для ее продолжения. Для того, чтобы загрузка была быстрой, мы должны ожидать завершения инициализации только тех устройств, которые *действительно* необходимы для ее продолжения на данной стадии. Ожидать остальные устройства в данном случае смысла нет. Кроме того, стоит отметить, что в числе программ, не приспособленных для работы с динамически меняющимся оборудованием и построенных в предположении о неизменности списка устройств, кроме служб хранения, есть и другие подсистемы. Например, в нашем случае работа `initrd` занимает так много времени главным образом из-за того, что для запуска `Plymouth` необходимо дождаться завершения инициализации всех устройств видеовывода. По неизвестной причине (во всяком случае, неизвестной для меня) подгрузка модулей для моих видеокарт Intel занимает довольно продолжительное время, что приводит к беспричинной задержке процесса загрузки. (Нет, я возражаю не против опроса устройств как такового, но против необходимости ожидать его завершения, чтобы продолжить загрузку.)

ошибку в `cryptsetup`³², попробуем найти обходной путь³³. Во время загрузки `systemd` должен дожидаться, пока все файловые системы, перечисленные в `/etc/fstab` (кроме помеченных как `noauto`) будут обнаружены, проверены и смонтированы. Только после этого `systemd` может продолжить загрузку и приступить к запуску служб. Но первое обращение к `/home` (в отличие, например, от `/var`), происходит на поздней стадии процесса загрузки (когда пользователь входит в систему). Следовательно, нам нужно сделать так, чтобы этот каталог автоматически монтировался при загрузке, но процесс загрузки не ожидал завершения работы `cryptsetup`, `fsck` и `mount` для этого раздела. Как же сделать точку монтирования доступной, не ожидая, пока завершится процесс монтирования? Этого можно достичь, воспользовавшись магической силой `systemd` — просто добавим опцию монтирования `comment=systemd.automount` в `/etc/fstab`. После этого, `systemd` будет создавать в `/home` точку автоматического монтирования, и при первом же обращении к этому каталогу, если файловая система еще не будет готова к работе, `systemd` подготовит соответствующее устройство, проверит и смонтирует ее.

После внесения изменений в `/etc/fstab` и перезагрузки мы получаем:

```
systemd[1]: Startup finished in 2s 47ms 112us (kernel) + 2s 663ms 942us (initrd)
+ 5s 540ms 522us (userspace) = 10s 251ms 576us.
```

Прекрасно! Несколькими простыми действиями мы выиграли почти семь секунд. И эти два изменения исправляют только две наиболее очевидные проблемы. При более аккуратном и детальном исследовании, обнаружится еще множество моментов, которые можно улучшить. Например, на другом моем компьютере, ноутбуке X300 двухлетней давности (и даже два года назад он был не самым быстрым на Земле), после небольшой доработки, время загрузки до полноценной среды GNOME составило около четырех секунд, и это еще не предел совершенства.

`systemd-analyze blame` — простой и удобный инструмент для поиска медленно запускающихся служб, однако он обладает одним существенным недостатком: он не показывает, насколько эффективно параллельный запуск снижает потерю времени для медленно запускающихся служб. Чтобы вы могли наглядно оценить этот фактор, мы подготовили команду `systemd-analyze plot`. Использовать ее очень просто:

```
$ systemd-analyze plot > plot.svg
$ eog plot.svg
```

Она создает наглядные диаграммы, показывающие моменты запуска служб и время, затраченное на их запуск, по отношению к другим службам. На текущий момент, она не показывает явно, кто кого ожидает, но догадаться обычно несложно³⁴.

Чтобы продемонстрировать эффект, порожденный двумя нашими оптимизациями, приведем ссылки на соответствующие графики: [до](#) и [после](#) (для полноты описания приведем также и соответствующие данные `systemd-analyze blame`: [до](#) и [после](#)).

У наиболее эрудированных читателей может возникнуть вопрос: как это соотносится с программой `bootchart`? Действительно, `systemd-analyze plot` и `bootchart` рисуют похожие графики. Однако, `bootchart` является намного более мощным инструментом — он детально показывает, что именно происходило во время загрузки, и отображает соответствующие графики использования процессора и ввода-вывода. `systemd-analyze plot` оперирует более высокоуровневой информацией: сколько времени затратила та или иная служба во время запуска, и какие службы были вынуждены ее ожидать. Используя оба

³²На самом деле, я действительно пытался исправить эту ошибку. Задержки в `cryptsetup`, по моим наблюдениям, обусловлены, главным образом, слишком большим значением по умолчанию для опции `--iter-time`. Я попытался доказать сопровождающим пакета `cryptsetup`, что снижение этого времени с 1 секунды до 100 миллисекунд не приведет к трагическим последствиям для безопасности, однако они мне не поверили.

³³Вообще-то я предпочитаю решать проблемы, а не искать пути для их обхода, однако в данном конкретном случае возникает отличная возможность продемонстрировать одну интересную возможность `systemd`...

³⁴Прим. перев.: Начиная с `systemd 203`, добавлена поддержка специальной команды `systemd-analyze critical-chain`, которая выводит самую медленную цепочку юнитов (цепь в графе зависимостей, формируемом при загрузке). Время запуска всех юнитов в этой цепочке определяет итоговое время загрузки системы.

этих инструмента, вы значительно упростите себе поиск причин замедления вашей загрузки.

Но прежде, чем вы вооружитесь описанными здесь средствами и начнете отправлять баг-репорты авторам и сопровождающим программ, которые замедляют вашу загрузку, обдумайте все еще раз. Эти инструменты предоставляют вам «сырую» информацию. Постарайтесь не ошибиться, интерпретируя ее. Например, в при рассмотрении приведенного выше примера я показал, что проблема была вовсе не в `udev-settle.service`, и не в опросе устройств для `ModemManager`'а, а в неудачной реализации подсистемы хранения, требующей ожидать окончания опроса устройств для продолжения загрузки. Именно эту проблему и нужно исправлять. Поэтому постарайтесь правильно определить источник проблем. Возлагайте вину на тех, кто действительно виноват.

Как уже говорилось выше, все три описанных здесь инструмента доступны в Fedora 15 «из коробки».

Итак, какие же выводы мы можем сделать из этой истории?

- `systemd-analyze` — отличный инструмент. Фактически, это встроенный профилировщик `systemd`.
- Постарайтесь не ошибиться, интерпретируя вывод профилировщика!
- Всего два небольших изменения могут ускорить загрузку системы на семь секунд.
- Программное обеспечение, не способное работать с динамически меняющимся набором устройств, создает проблемы, и должно быть исправлено.
- Принудительное использование в стандартной установке Fedora 15 промышленной реализации подсистем хранения, возможно, является не самым правильным решением.

8 Новые конфигурационные файлы

Одно из ключевых достоинств `systemd` — наличие полного набора программ, необходимых на ранних стадиях загрузки, причем эти программы написаны на простом, быстром, надежном и легко поддающемся распараллеливанию языке C. Теперь можно отказаться от «простыней» shell-скриптов, разработанных для этих задач различными дистрибутивами. Наш «Проект нулевой оболочки»³⁵ увенчался полным успехом. Уже сейчас возможности предоставляемого нами инструментария покрывают практически все нужды настольных и встраиваемых систем, а также большую часть потребностей серверов:

- Проверка и монтирование всех файловых систем.
- Обновление и активация квот на всех файловых системах.
- Установка имени хоста.
- Настройка сетевого интерфейса обратной петли (10).
- Подгрузка правил SELinux, обновление меток безопасности в динамических каталогах `/run` и `/dev`.
- Регистрация в ядре дополнительных бинарных форматов (например, Java, Mono, WINE) через API-файловую систему `binfmt_misc`.
- Установка системной локали.
- Настройка шрифта и раскладки клавиатуры в консоли.
- Создание, очистка, удаление временных файлов и каталогов.

³⁵Наш девиз — «Первой оболочкой, запускающейся при старте системы, должна быть GNOME shell». Формулировка оставляет желать лучшего, но все же неплохо передает основную идею.

- Применение предписанных в `/etc/fstab` опций к смонтированным ранее API-файловым системам.
- Применение настроек `sysctl`.
- Поддержка технологии упреждающего чтения (`read ahead`), включая автоматический сбор информации.
- Обновление записей в `utmp` при включении и выключении системы.
- Сохранение и восстановление затравки для генерации случайных чисел (`random seed`).
- Принудительная загрузка указанных модулей ядра.
- Поддержка шифрованных дисков и разделов.
- Автоматический запуск `getty` на `serial`-консолях.
- Взаимодействие с `Plymouth`.
- Создание уникального идентификатора системы.
- Настройка часового пояса.

В стандартной установке Fedora 15 запуск shell-скриптов требуется только для некоторых устаревших служб, а также для подсистемы хранения данных (поддержка LVM, RAID и `multipath`). Если они вам не нужны, вы легко можете отключить их, и наслаждаться загрузкой, полностью очищенной от shell-костылей (лично я это сделал уже давно). Такая загрузка является уникальной возможностью Linux-систем.

Большинство перечисленных выше компонентов настраиваются через конфигурационные файлы в каталоге `/etc`. Некоторые из этих файлов стандартизированы для всех дистрибутивов, и поэтому реализация их поддержки в наших инструментах не представляла особого труда. Например, это относится к файлам `/etc/fstab`, `/etc/crypttab`, `/etc/sysctl.conf`. Однако множество других, нестандартно расположенных файлов и каталогов вынуждали нас добавлять в код огромное количество операторов `#ifdef`, чтобы обеспечить поддержку различных вариантов расположения конфигураций в разных дистрибутивах. Такое положение дел сильно усложняет жизнь нам всем, и при этом ничем не оправдано — все эти файлы решают одни и те же задачи, просто немного по-разному.

Чтобы улучшить ситуацию и установить единый стандарт расположения базовых конфигурационных файлов во всех дистрибутивах, мы заставили `systemd` пользоваться дистрибутивно-специфическими конфигурациями только в качестве *резервного* варианта — основным источником информации становится определенный нами стандартный набор конфигурационных файлов. Разумеется, там, где это возможно, мы старались не придумывать чего-то принципиально нового, а брали лучшее из решений, предложенных существующими дистрибутивами. Ниже приводится небольшой обзор этого нового набора конфигурационных файлов, поддерживаемых `systemd` во всех дистрибутивах:

- `/etc/hostname`: имя хоста для данной системы. Одна из наиболее простых и важных системных настроек. В разных дистрибутивах оно настраивалось по-разному: Fedora использовала `/etc/sysconfig/network`, OpenSUSE — `/etc/HOSTNAME`, Debian — `/etc/hostname`. Мы остановились на варианте, предложенном Debian.
- `/etc/vconsole.conf`: конфигурация раскладки клавиатуры и шрифта для консоли.
- `/etc/locale.conf`: конфигурация общесистемной локали.

- `/etc/modules-load.d/*.conf`: каталог³⁶ для перечисления модулей ядра, которые нужно принудительно подгрузить при загрузке (впрочем, необходимость в этом возникает достаточно редко).
- `/etc/sysctl.d/*.conf`: каталог для задания параметров ядра (`sysctl`). Дополняет классический конфигурационный файл `/etc/sysctl.conf`.
- `/etc/tmpfiles.d/*.conf`: каталог для управления настройками временных файлов (`systemd` обеспечивает создание, очистку и удаление временных файлов и каталогов, как во время загрузки, так и во время работы системы).
- `/etc/binfmt.d/*.conf`: каталог для регистрации дополнительных бинарных форматов (например, форматов Java, Mono, WINE).
- `/etc/os-release`: стандарт для файла, обеспечивающего идентификацию дистрибутива и его версии. Сейчас различные дистрибутивы используют для этого разные файлы (например, `/etc/fedora-release` в Fedora), и поэтому для решения такой простой задачи, как вывод имени дистрибутива, необходимо использовать базу данных, содержащую перечень возможных названий файлов. Проект LSB попытался создать такой инструмент — `lsb_release` — однако реализация столь простой функции через скрипт на Python'e является не самым оптимальным решением. Чтобы исправить сложившуюся ситуацию, мы решили перейти к единому простому формату представления этой информации.
- `/etc/machine-id`: файл с идентификатором данного компьютера (перекрывает аналогичный идентификатор D-Bus). Гарантируется, что в любой системе, использующей `systemd`, этот файл будет существовать и содержать корректную информацию (если его нет, он автоматически создается при загрузке). Мы вынесли этот файл из-под эгиды D-Bus, чтобы упростить решение множества задач, требующих наличия уникального и постоянного идентификатора компьютера.
- `/etc/machine-info`: новый конфигурационный файл, хранящий информации о полном (описательном) имени хоста (например, «Компьютер Леннарта») и значке, которым он будет обозначаться в графических оболочках, работающих с сетью (раньше этот значок мог определяться, например, файлом `/etc/favicon.png`). Данный конфигурационный файл обслуживается демоном `systemd-hostnamed`.

Одна из важнейших для нас задач — убедить *вас* использовать эти новые конфигурационные файлы в ваших инструментах для настройки системы. Если ваши конфигурационные фронтенды будут использовать новые файлы, а не их старые аналоги, это значительно облегчит портирование таких фронтендов между дистрибутивами, и вы внесете свой вклад в стандартизацию Linux. В конечном счете это упростит жизнь и администраторам, и пользователям. Разумеется, на текущий момент эти файлы полностью поддерживаются только дистрибутивами, основанными на `systemd`, но уже сейчас в их число входят практически все ключевые дистрибутивы, **за исключением одного**³⁷. В этом есть что-то от «проблемы курицы и яйца»: стандарт становится настоящим стандартом только тогда, когда ему начинают следовать. В будущем мы намерены аккуратно форсировать процесс перехода на новые конфигурационные файлы: поддержка старых файлов будет удалена из `systemd`. Разумеется, процесс будет идти медленно, шаг

³⁶Прим. перев.: Для описания этого и трех последующих каталогов автор пользуется термином «drop-in directory». Данный термин означает каталог, в который можно поместить множество независимых файлов настроек, и при чтении конфигурации все эти файлы будут обработаны (впрочем, часто накладывается ограничение — обрабатываются только файлы с именами, соответствующими маске, обычно `*.conf`). Такой подход позволяет значительно упростить процесс как ручного, так и автоматического конфигурирования различных компонентов — для внесения изменений в настройки уже не нужно редактировать основной конфигурационный файл, достаточно лишь скопировать/переместить в нужный каталог небольшой файл с указанием специфичных параметров.

³⁷Прим. перев.: В конце 2010 года энтузиаст Andrew Edmunds **добавил** в `systemd` базовую поддержку Ubuntu и **подготовил** соответствующие пакеты, однако его инициатива не встретила поддержки среди менеджеров Canonical. На момент написания этих строк проект остается заброшенным с декабря 2010 г.

за шагом. Но конечной его целью является переход всех дистрибутивов на единый набор базовых конфигурационных файлов.

Многие из этих файлов используются не только программами для настройки системы, но и апстримными проектами. Например, мы предлагаем проектам Mono, Java, WINE и другим помещать конфигурацию для регистрации своих бинарных форматов в `/etc/binfmt.d/` средствами их собственной сборочной системы. Специфичные для дистрибутивов механизмы поддержки бинарных форматов больше не нужны, и ваш проект будет работать одинаково хорошо во всех дистрибутивах. Аналогичное предложение мы обращаем и ко всем разработчикам программ, которым требуется автоматическое создание/очистка временных файлов и каталогов, например, в каталоге `/run` (ранее известном как `/var/run`). Таким проектам достаточно просто поместить соответствующий конфигурационный файл в `/etc/tmpfiles.d/`, тоже средствами собственной сборочной системы. Помимо прочего, подобный подход позволит увеличить скорость загрузки, так как, в отличие от SysV, не требует множества shell-скриптов, выполняющих тривиальные задачи (регистрация бинарных форматов, удаление/создание временных файлов/каталогов и т.п.). И пример того случая, когда апстримная поддержка стандартной конфигурации дала бы огромные преимущества — X11 (и его аналоги) могли бы устанавливать раскладку клавиатуры на основании данных из `/etc/vconsole.conf`.

Разумеется, я понимаю, что отнюдь не всех полностью устроят выбранные нами имена и форматы конфигурационных файлов. Но нам все же нужно было что-то выбрать, и мы выбрали то, что должно устроить большинство людей. Форматы конфигурационных файлов максимально просты, и их можно легко читать и записывать даже из shell-скриптов. Да, `/etc/bikeshed.conf` могло бы быть неплохим именем для файла конфигурации!³⁸

Помогите нам стандартизировать Linux! Используйте новые конфигурационные файлы! Поддерживайте их в апстриме, поддерживайте их во всех дистрибутивах!

И если у вас возникнет такой вопрос: да, все эти файлы так или иначе обсуждались с разными разработчиками из различных дистрибутивов. И некоторые из разработчиков планируют обеспечить поддержку новой конфигурации даже в системах без `systemd`.

9 О судьбе `/etc/sysconfig` и `/etc/default`

В дистрибутивах, основанных на Red Hat и SUSE, это каталог называется `/etc/sysconfig`. В дистрибутивах на базе Debian, его зовут `/etc/default`. Во многих других дистрибутивах также присутствуют каталоги похожего назначения. Связанные с ними вопросы неоднократно появляются в дискуссиях пользователей и разработчиков `systemd`. В этой статье мне хотелось бы рассказать, что я, как разработчик `systemd`, думаю об этих каталогах, и пояснить, почему, на мой взгляд, от них лучше отказаться. Стоит отметить, что это мое личное мнение, и оно может не совпадать с позицией проекта Fedora или моего работодателя.

Начнем с небольшого исторического экскурса. Каталог `/etc/sysconfig` появился в дистрибутивах Red Hat и SUSE задолго до того, как я присоединился к этим проектам — иными словами, это было очень давно. Некоторое время спустя, в Debian появился аналогичный по смыслу каталог `/etc/default`. Многие дистрибутивы используют такие каталоги, называя их по-разному. Они имеются даже в некоторых ОС семейства Unix. (Например, в SCO. Если эта тема вас заинтересовала — рекомендую обратиться к вашему знакомому ветерану Unix, он расскажет гораздо подробнее и интереснее, чем я.) Несмотря на то, что подобные каталоги широко используются в Linux и Unix, они совершенно не стандартизированы — ни в POSIX, ни в LSB/FHS, и результате мы имеем целый зоопарк их различных реализаций в разных дистрибутивах.

³⁸Прим. перев.: Здесь автор намекает на [Паркинсоновский Закон Тривиальности](#), который гласит, что самые жаркие споры возникают вокруг наиболее простых вопросов. В частности, в качестве примера Паркинсон приводит обсуждение строительства атомной электростанции и гаража для велосипедов (bike shed) — если первое из этих решений принимается довольно быстро, то вокруг второго разгорается множество дискуссий по самым разным аспектам.

Назначение этих каталогов определено весьма расплывчато. Абсолютное большинство находящихся в них файлов являются включаемыми³⁹ shell-скриптами, содержащими, главным образом, определения переменных. Большинство файлов из этих каталогов включаются в одноименные скрипты SysV init. Данный принцип отражен в [Debian Policy Manual \(раздел 9.3.2\)](#) и в [Fedora Packaging Guidelines](#), однако в обоих дистрибутивах иногда встречаются файлы, не соответствующие описанной схеме, например, не имеющие соответствующего init-скрипта, или даже сами не являющиеся скриптами.

Но почему вообще появились эти каталоги? Чтобы ответить на такой вопрос, обратимся к истории развития концепции SysV init-скриптов. Исторически, сложилось так, что они располагаются в каталоге под названием `/etc/rc.d/init.d` (или что-то похожее). Отметим, что каталог `/etc` вообще-то предназначен для хранения файлов конфигурации, а не исполняемого кода (в частности, скриптов). Однако, в начале своей истории, init-скрипты рассматривались именно как файлы конфигурации, и редактирование их администратором было общепринятой практикой. Но со временем, по мере роста и усложнения этих скриптов, их стали рассматривать уже не как файлы конфигурации, а как некие программы. Чтобы упростить их настройку и обеспечить безопасность процесса обновления, настройки были вынесены в отдельные файлы, загружаемые при работе init-скриптов.

Попробуем составить некоторое представление о настройках, которые можно сделать через эти файлы. Вот краткий и неполный список различных параметров, которые могут быть заданы через переменные окружения в таких файлах (составлен мною по результатам исследования соответствующих каталогов в Fedora и Debian):

- Дополнительные параметры командной строки для бинарника демона.
- Настройки локали для демона.
- Тайм-аут останова для демона.
- Режим останова для демона.
- Общесистемные настройки, например, системная локаль, часовой пояс, параметры клавиатуры для консоли.
- Избыточная информация о системных настройках, например, указание, установлены ли аппаратные часы по Гринвичу или по местному времени.
- Списки правил брандмауэра, не являются скриптами (!).
- Привязка к процессорным ядрам для демона.
- Настройки, не относящиеся к процессу загрузки, например, информация по установке пакетов с новыми ядрами, конфигурация `nspluginwrapper`, разрешение на выполнение предварительного связывания (`prelinking`) библиотек.
- Указание, нужно ли запускать данную службу или нет.
- Настройки сети.
- Перечень модулей ядра, которые должны быть подгружены принудительно.
- Нужно ли отключать питание компьютера при остановке системы (`poweroff`) или нет (`halt`).
- Права доступа для файлов устройств (!).
- Описание соответствующей SysV службы.

³⁹Прим. перев.: Здесь автор использует термин `sourcable`, происходящий от bash'овской директивы `source`, обеспечивающей включение в скрипт кода из внешнего файла. В классическом POSIX shell это соответствует оператору-точке `«.»`. В отличие от прямого запуска одного скрипта из другого, включаемый код исполняется той же самой оболочкой, что и основной код, и при возвращении в основной скрипт сохраняются переменные окружения, определенные во включаемом коде. Как правило, код для включения не содержит `shebang'a` (`#!/bin/sh` в начале файла).

- Идентификатор пользователя/группы, значение `umask` для демона.
- Ограничения по ресурсам для демона.
- Приоритет OOM killer'a для демона.

А теперь давайте ответим на вопрос: что же такого неправильного в `/etc/sysconfig` (`/etc/default`) и почему этим каталогам нет места в мире `systemd`?

- Прежде всего, утрачены основная цель и смысл существования таких каталогов: файлы конфигурации юнитов `systemd` не являются программами, в отличие от `init`-скриптов `SysV`. Эти файлы представляют собой простые, декларативные описания конкретных задач и функций, и обычно содержат не более шести строк. Они легко могут быть сгенерированы и проанализованы без использования `Bourne shell`. Их легко читать и понимать. Кроме того, их легко модифицировать: достаточно скопировать файл из `/lib/systemd/system` в `/etc/systemd/system`, после чего внести в созданную копию необходимые изменения (при этом можно быть уверенным, что изменения не будут затерты пакетным менеджером). Изначальная причина появления обсуждаемых каталогов — необходимость разделять код и параметры конфигурации — больше не существует, так как файлы описания юнитов не являются кодом. Проще говоря, обсуждаемые каталоги являются решением проблемы, которой уже не существует.
- Обсуждаемые каталоги и файлы в них очень сильно привязаны к специфике дистрибутивов. Мы же планируем, используя `systemd`, способствовать стандартизации и унификации дистрибутивов. В частности, одним из факторов такой стандартизации является рекомендация распространять соответствующие файлы конфигурации юнитов сразу с апстримным продуктом, а не возлагать эту работу на создателей пакетов, как это делалась во времена `SysV`. Так как расположение обсуждаемых каталогов и настраиваемые через них параметры сильно отличаются от дистрибутива к дистрибутиву, пытаться поддерживать их в апстримных файлах конфигурации юнитов просто бессмысленно. Хранение параметров конфигурации в этих каталогах — один из факторов, превращающих Linux в зоопарк несовместимых решений.
- Большинство настроек, задаваемых через эти каталоги, являются избыточными в мире `systemd`. Например, различные службы позволяют задать таким методом параметры исполнения процесса, в частности, идентификатор пользователя/группы, ограничения ресурсов, привязки к ядрам CPU, приоритет OOM killer'a. Однако, эти настройки поддерживаются лишь некоторыми `init`-скриптами, причем одна и та же настройка в различных скриптах может называться по-разному. С другой стороны, в мире `systemd`, все эти настройки доступны для всех служб без исключения, и всегда задаются одинаково, через одни и те же параметры конфигурационных файлов.
- Файлы конфигурации юнитов имеют множество удобных и простых в использовании настроек среды исполнения процесса, гораздо больше, чем могут предоставить файлы из `/etc/sysconfig`.
- Необходимость в некоторых из этих настроек весьма сомнительна. Например, возьмем вышеупомянутую возможность задавать идентификатор пользователя/группы для процесса. Эту задачу должен решать разработчик ПО или дистрибутива. Вмешательство администратора в данную настройку, как правило, лишено смысла — только разработчик располагает всей информацией, позволяющий предотвратить конфликты идентификаторов и имен пользователей и групп.
- Формат файлов, используемых для сохранения настроек, плохо подходит для данной задачи. Так как эти файлы, как правило, являются включаемыми `shell`-скриптами, ошибки при их чтении очень трудно отследить. Например, ошибка в имени переменной приведет к тому, что переменная не будет изменена, однако никакого предупреждения при этом не выводится.

- Кроме того, такая организация не исключает влияния конфигурационных параметров на среду исполнения: например, изменение переменных `IFS` и `LANG` может существенно повлиять на результат интерпретации `init`-скрипта.
- Интерпретация этих файлов требует запуска еще одного экземпляра оболочки, что приводит к задержкам при загрузке⁴⁰.
- Файлы из `/etc/sysconfig` часто пытаются использовать в качестве суррогатной замены файлов конфигурации для тех демонов, которые не имеют встроенной поддержки конфигурационных файлов. В частности, вводятся специальные переменные, позволяющие задать аргументы командной строки, используемые при запуске демона. Встроенная поддержка конфигурационных файлов является более удобной альтернативой такому подходу, ведь, глядя на ключи «-k», «-a», «-f», трудно догадаться об их назначении. Очень часто, из-за ограниченности словаря, на различных демонах одни и те же ключи действуют совершенно по-разному (для одного демона ключ «-f» содержит указание демонизироваться при запуске, в то время как для другого эта опция действует прямо противоположным образом.) В отличие от конфигурационных файлов, строка запуска не может включать полноценных комментариев.
- Некоторые из настроек, задаваемых в `/etc/sysconfig`, являются полностью избыточными. Например, во многие дистрибутивах подобным методом указывается, установлены ли аппаратные часы компьютера по Гринвичу, или по местному времени. Однако эта же настройка задается третьей строкой файла `/etc/adjtime`, поддерживаемого во всех дистрибутивах. Использование избыточного и не стандартизированного параметра конфигурации только добавляет путаницу и не несет никакой пользы.
- Многие файлы настроек из `/etc/sysconfig` позволяют отключать запуск соответствующей службы. Однако эта операция уже поддерживается штатно для всех служб, через команды `systemctl enable/disable` (или `chkconfig on/off`). Добавление дополнительного уровня настройки не приносит никакой пользы и лишь усложняет работу администратора.
- Что касается списка принудительно загружаемых модулей ядра — в настоящее время существуют куда более удобные пути для автоматической подгрузки модулей при загрузке системы. Например, многие модули автоматически подгружаются `udev`'ом при обнаружении соответствующего оборудования. Этот же принцип распространяется на ACPI и другие высокоуровневые технологии. Одно из немногих исключений из этого правила — к сожалению, в настоящее время не поддерживается автоматическая загрузка модулей на основании информации о возможностях процессора, однако это будет исправлено в ближайшем будущем⁴¹. В случае, если нужный вам модуль ядра все же не может быть подгружен автоматически, все равно существует гораздо более удобные методы указать его принудительную подгрузку — например, путем создания соответствующего файла в каталоге `/etc/modules-load.d/` (стандартный метод настройки принудительной подгрузки модулей).
- И наконец, хотелось бы отметить, что каталог `/etc` определен как место для хранения системных настроек («Host-specific system configuration», согласно FHS). Наличие внутри него подкаталога `sysconfig`, который тоже содержит системную конфигурацию, является очевидно избыточным.

Что же можно предложить в качестве современной, совместимой с `systemd` альтернативы настройке системы через файлы в этих каталогах? Ниже приведены несколько

⁴⁰Прим. перев.: Здесь автор несколько заблуждается. Скрипты, включенные через директиву `source`, исполняются тем же экземпляром оболочки, что и вызвавший их скрипт.

⁴¹Прим. перев.: Необходимые патчи уже приняты в ядро, и соответствующая функция поддерживается Linux, начиная с версии 3.3.

рекомендаций, как лучше поступить с задаваемыми таким образом параметрами конфигурации:

- Попробуйте просто отказаться от них. Если они полностью избыточны (например, настройка аппаратных часов на Гринвич/местное время), то убрать их будет довольно легко (если не рассматривать вопросы обеспечения совместимости). Если аналогичные по смыслу опции штатно поддерживаются `systemd`, нет никакого смысла дублировать их где-то еще (перечень опций, которые можно задать для любой службы, приведен на страницах справки [systemd.exec\(5\)](#) и [systemd.service\(5\)](#).) Если же ваша настройка просто добавляет еще один уровень отключения запуска службы — не плодите лишние сущности, откажитесь от нее.
- Найдите для них более подходящее место. Например, в случае с некоторыми общесистемными настройками (такими, как локаль или часовой пояс), мы надеемся аккуратно подтолкнуть дистрибутивы в правильном направлении (см. предыдущую главу).
- Добавьте их поддержку в штатную систему настройки демона через собственные файлы конфигурации. К счастью, большинство служб, работающих в Linux, являются свободным программным обеспечением, так что сделать это довольно просто.

Существует лишь одна причина поддерживать файлы `/etc/sysconfig` еще некоторое время: необходимо обеспечить совместимость при обновлении. Тем не менее, как минимум в новых пакетах, от таких файлов лучше отказаться.

В том случае, если требование совместимости критично, вы можете задействовать эти конфигурационные файлы даже в том случае, если настраиваете службы через штатные `unit`-файлы `systemd`. Если ваш файл из `sysconfig` содержит лишь определения переменных, можно воспользоваться опцией `EnvironmentFile=-/etc/sysconfig/foobar` (подробнее об этой опции см. [systemd.exec\(5\)](#)), позволяющей прочитать из файла набор переменных окружения, который будет установлен при запуске службы. Если же для задания настроек вам необходим полноценный язык программирования — ничто не мешает им воспользоваться. Например, вы можете создать в `/usr/lib/<your package>/` простой скрипт, который включает соответствующие файлы, а затем запускает бинарник демона через `exec`. После чего достаточно просто указать этот скрипт в опции `ExecStart=` вместо бинарника демона.

10 Экземпляры служб

Большинство служб в Linux/Unix являются одиночными (`singleton`): в каждый момент времени на данном хосте работает только один экземпляр службы. В качестве примера таких одиночных служб можно привести `Syslogd`, `Postfix`, `Apache`. Однако, существуют службы, запускающие по несколько экземпляров себя на одном хосте. Например, службы наподобие `Dovecot` ИМАР запускают по одному экземпляру на каждый локальный порт и/или IP-адрес. Другой пример, который можно встретить практически во всех системах — `getty`, небольшая служба, запускающаяся на каждом ТТУ (от `tty1` до `tty6`). На некоторых серверах, в зависимости от сделанных администратором настроек или параметров загрузки, могут запускаться дополнительные экземпляры `getty`, для подключаемых к СОМ-портам терминалов или для консоли системы виртуализации. Еще один пример службы, работающей в нескольких экземплярах (по крайней мере, в мире `systemd`) — `fsck`, программа проверки файловой системы, которая запускается по одному экземпляру на каждое блочное устройство, требующее такой проверки. И наконец, стоит упомянуть службы с активацией в стиле `inetd` — при обращении через сокет, по одному экземпляру на каждое соединение. В этой статье я попытаюсь рассказать, как в `systemd` реализовано управление «многоэкземплярными» службами, и какие выгоды системный администратор может извлечь из этой возможности.

Если вы читали предыдущие статьи из этого цикла, вы, скорее всего, уже знаете, что службы `systemd` именуются по схеме `foobar.service`, где `foobar` — строка, идентифици-

рующая службу (проще говоря, ее имя), а `.service` — суффикс, присутствующий в именах всех файлов конфигурации служб. Сами эти файлы могут находиться в каталогах `/etc/systemd/systemd` и `/lib/systemd/system` (а также, возможно, и в других⁴²). Для служб, работающих в нескольких экземплярах, эта схема становится немного сложнее: `foobar@quux.service`, где `foobar` — имя службы, общее для всех экземпляров, а `quux` — идентификатор конкретного экземпляра. Например, `serial-getty@ttyS2.service` — это служба `getty` для COM-порта, запущенная на `ttyS2`.

При необходимости, экземпляры служб можно легко создать динамически. Скажем, вы можете, безо всяких дополнительных настроек, запустить новый экземпляр `getty` на последовательном порту, просто выполнив `systemctl start` для нового экземпляра:

```
# systemctl start serial-getty@ttyUSB0.service
```

Получив такую команду, `systemd` сначала пытается найти файл конфигурации юнита с именем, точно соответствующим запрошенному. Если такой файл найти не удастся (при работе с экземплярами служб обычно так и происходит), из имени файла удаляется идентификатор экземпляра, и полученное имя используется при поиске *шаблона* конфигурации. В нашем случае, если отсутствует файл с именем `serial-getty@ttyUSB0.service`, используется файл-шаблон под названием `serial-getty@.service`. Таким образом, для всех экземпляров данной службы, используется один и тот же шаблон конфигурации. В случае с `getty` для COM-портов, этот шаблон, поставляемый в комплекте с `systemd` (файл `/lib/systemd/system/serial-getty@.service`) выглядит примерно так:

```
[Unit]
Description=Serial Getty on %I
BindTo=dev-%i.device
After=dev-%i.device systemd-user-sessions.service

[Service]
ExecStart=-/sbin/agetty -s %I 115200,38400,9600
Restart=always
RestartSec=0
```

(Заметим, что приведенная здесь версия немного сокращена, по сравнению с реально используемой в `systemd`. Удалены не относящиеся к теме нашего обсуждения параметры конфигурации, обеспечивающие совместимость с `SysV`, очистку экрана и удаление предыдущих пользователей с текущего ТТУ. Если вам интересно, можете посмотреть [полную версию](#).)

Этот файл похож на обычный файл конфигурации юнита, с единственным отличием: в нем используются спецификаторы `%I` и `%i`. В момент загрузки юнита, `systemd` заменяет эти спецификаторы на идентификатор экземпляра службы. В нашем случае, при обращении к экземпляру `serial-getty@ttyUSB0.service`, они заменяются на «`ttyUSB0`». Результат такой замены можно проверить, например, запросив состояние для нашей службы:

```
$ systemctl status serial-getty@ttyUSB0.service
serial-getty@ttyUSB0.service - Getty on ttyUSB0
    Loaded: loaded (/lib/systemd/system/serial-getty@.service; static)
    Active: active (running) since Mon, 26 Sep 2011 04:20:44 +0200; 2s ago
    Main PID: 5443 (agetty)
    CGroup: name=systemd:/system/getty@.service/ttyUSB0
            └─ 5443 /sbin/agetty -s ttyUSB0 115200,38400,9600
```

⁴²Прим. перев.: Перечень каталогов, в которых выполняется поиск общесистемных юнит-файлов, приведен на странице руководства `systemd(1)` (раздел «System unit directories»). Указанные выше каталоги `/etc/systemd/systemd` и `/lib/systemd/system` соответствуют значениям по умолчанию для упомянутых там переменных `pkg-config systemdssystemconfdir` и `systemdssystemunitdir` соответственно. Начиная с `systemd` версии 198, данный перечень, в более точной и развернутой форме, присутствует на странице `systemd.unit(5)`.

Собственно, это и есть ключевая идея организации экземпляров служб. Как видите, `systemd` предоставляет простой в использовании механизм шаблонов, позволяющих динамически создавать экземпляры служб. Добавим несколько дополнительных замечаний, позволяющих эффективно использовать этот механизм:

Вы можете создавать дополнительные экземпляры таких служб, просто добавляя символичные ссылки в каталоги `*.wants/`. Например, чтобы обеспечить запуск `getty` на `ttyUSB0` при каждой загрузке, достаточно создать такую ссылку:

```
# ln -s /lib/systemd/system/serial-getty@.service \
/etc/systemd/system/getty.target.wants/serial-getty@ttyUSB0.service
```

При этом файл конфигурации, на который указывает ссылка (в нашем случае `serial-getty@.service`), будет вызван с тем именем экземпляра, которое указано в названии этой ссылки (в нашем случае — `ttyUSB0`).

Вы не сможете обратиться к юниту-шаблону без указания идентификатора экземпляра. В частности, команда `systemctl start serial-getty@.service` завершится ошибкой.

Иногда возникает необходимость отказаться от использования общего шаблона для конкретного экземпляра (т.е. конфигурация данного экземпляра настолько сильно отличается от параметров остальных экземпляров данной службы, что механизм шаблонов оказывается неэффективен). Специально для таких случаев, в `systemd` и заложен предварительный поиск файла с именем, точно соответствующим указанному (прежде чем использовать общий шаблон). Таким образом, вы можете поместить файл с именем, точно соответствующим полному титулу экземпляра, в каталог `/etc/systemd/system/` — и содержимое этого файла, при обращении к выбранному экземпляру, полностью переопределяет все настройки, сделанные в общем шаблоне.

В приведенном выше файле, в некоторых местах используется спецификатор `%I`, а в других — `%i`. У вас может возникнуть закономерный вопрос — чем они отличаются? `%i` всегда точно соответствует идентификатору экземпляра, в то время, как `%I` соответствует неэкранированной (unescaped) версии этого идентификатора. Когда идентификатор не содержит спецсимволов (например, `ttyUSB0`), результат в обоих случаях одинаковый. Но имена устройств, например, содержат слэши (`«/»`), которые не могут присутствовать в имени юнита (и в имени файла на Unix). Поэтому, перед использованием такого имени в качестве идентификатора устройства, оно должно быть экранировано — `«/»` заменяются на `«-»`, а большинство других специальных символов (включая `«-»`) заменяются последовательностями вида `\xAB`, где `AB` — ASCII-код символа, записанный в шестнадцатеричной системе счисления⁴³. Например, чтобы обратиться к последовательному USB-порту по его адресу на шине, нам нужно использовать имя наподобие `serial/by-path/pci-0000:00:1d.0-usb-0:1.4:1.1-port0`. Экранированная версия этого имени — `serial-by\x2dpath-pci\x2d0000:00:1d.0\x2dusb\x2d0:1.4:1.1\x2dport0`. `%I` будет ссылаться на первую из этих строк, `%i` — на вторую. С практической точки зрения, это означает, что спецификатор `%i` можно использовать в том случае, когда надо сослаться на имена других юнитов, например, чтобы описать дополнительные зависимости (в случае с `serial-getty@.service`, этот спецификатор используется для ссылки на юнит `dev-%i.device`, соответствующий одноименному устройству). В то время как `%I` удобно использовать в командной строке (`ExecStart`) и для формирования читабельных строк описания. Рассмотрим работу этих принципов на примере нашего юнит-файла⁴⁴:

⁴³Согласен, этот алгоритм дает на выходе не очень читабельный результат. Но этим грешат практически все алгоритмы экранирования. В данном случае, был использован механизм экранирования из `udev`, с одним изменением. В конце концов, нам нужно было выбрать что-то. Если вы собираетесь комментировать наш алгоритм экранирования — пожалуйста, оставьте свой адрес, чтобы я мог приехать к вам и раскрасить ваш сарай для велосипедов в синий с желтыми полосами. Спасибо!

⁴⁴Прим. перев.: как видно из нижеприведенного примера, в командной строке `systemctl` используется экранированное имя юнита, что создает определенные трудности даже при наличии в оболочке «умного» автодополнения. Однако, начиная с `systemd v186`, при работе с `systemctl` можно указывать неэкранированные имена юнитов.

```
# systemctl start 'serial-getty@serial-by\x2dpath-pci\x2d0000:00:1d.0\x2dusb\x2d0:1.4:1.1\x2dport0.service'  
# systemctl status 'serial-getty@serial-by\x2dpath-pci\x2d0000:00:1d.0\x2dusb\x2d0:1.4:1.1\x2dport0.service'  
serial-getty@serial-by\x2dpath-pci\x2d0000:00:1d.0\x2dusb\x2d0:1.4:1.1\x2dport0.service - Serial Getty on serial/by-path/pci-0000:00:1d.0-usb-0:1.4:1.1-port0  
  Loaded: loaded (/lib/systemd/system/serial-getty@.service; static)  
  Active: active (running) since Mon, 26 Sep 2011 05:08:52 +0200; 1s ago  
    Main PID: 5788 (agetty)  
     CGroup: name=systemd:/system/serial-getty@.service/serial-by\x2dpath-pci\x2d0000:00:1d.0\x2dusb\x2d0:1.4:1.1\x2dport0  
             └─ 5788 /sbin/agetty -s serial/by-path/pci-0000:00:1d.0-usb-0:1.4:1.1-port0 115200 38400 9600
```

Как видите, в качестве идентификатора экземпляра используется экранированная версия, в то время как в строке описания и в командной строке `getty` фигурирует неэкранированный вариант, как и предполагалось.

(Небольшое замечание: помимо `%i` и `%I`, существует еще несколько спецификаторов, и большинство из них доступно и в обычных файлах конфигурации юнитов, а не только в шаблонах. Подробности можно посмотреть на [странице руководства](#), содержащей полный перечень этих спецификаторов с краткими пояснениями.)

11 Службы с активацией в стиле `inetd`

В одной из предыдущих глав (гл. 3) я рассказывал, как можно преобразовать SysV `init`-скрипт в юнит-файл `systemd`. В этой главе мы рассмотрим, как провести аналогичное преобразование для служб `inetd`.

Начнем с небольшого экскурса в историю вопроса. Уже многие годы `inetd` считается одной из базовых служб Unix-систем. Работая как суперсервер, он слушает сетевые сокеты от имени различных служб, и активирует соответствующие службы при поступлении на эти сокеты входящих соединений. Таким образом, он обеспечивает механизм активации по запросу (`on-demand activation`). Подобный подход избавляет от необходимости держать постоянно работающими все серверные процессы, что позволяет поддерживать множество служб даже на системах с очень ограниченными ресурсами. В дистрибутивах Linux можно встретить несколько различных реализаций `inetd`. Наиболее популярные из них — BSD `inetd` и `xinetd`. Хотя `inetd` во многих дистрибутивах до сих пор устанавливается по умолчанию, сейчас он уже редко используется для запуска сетевых служб — теперь большинство из них запускаются автономно при загрузке системы (основной аргумент в пользу такой схемы — она позволяет обеспечить более высокую производительность служб).

Одной из ключевых возможностей `systemd` (а также `launchd` от Apple) является сокет-активация — тот же самый механизм, давным-давно реализованный `inetd`, однако в данном случае ключевые цели немного другие. Сокет-активация в стиле `systemd` прежде всего ориентирована на локальные сокеты (`AF_UNIX`), хотя поддерживаются также и сетевые сокеты (`AF_INET`). И более важное отличие — сокет-активация в `systemd` обеспечивает не только экономию системных ресурсов, но также и эффективную параллелизацию работы (так как она позволяет запускать клиентские и серверные процессы одновременно, непосредственно после создания сокета), упрощение процесса конфигурации (отпадает необходимость явно указывать зависимости между службами) и увеличение надежности (перезапуск службы, служебный или экстренный — в случае падения — не приводит к недоступности сокета). Тем не менее, `systemd` ничуть не хуже `inetd` может запускать службы в ответ на входящие сетевые соединения.

Любая сокет-активация требует соответствующей поддержки со стороны самой запускаемой службы. `systemd` предоставляет очень простой интерфейс, который может быть использован службами для обеспечения сокет-активации. В основе этого [простого и минималистичного](#) механизма лежит функция `sd_listen_fds()`. Однако, интерфейс, традиционно используемый в `inetd`, еще проще. Он позволяет передавать запускаемой службе только один сокет, который формируется из потоков `STDIN` и `STDOUT` запущенного процесса. Поддержка этого механизма также присутствует в `systemd` — для обеспечения совместимости со множеством служб, у которых сокет-активация реализована только в стиле `inetd`.

Прежде, чем мы перейдем к примерам, рассмотрим три различных схемы, использующих сокет-активацию:

1. **Сокет-активация, ориентированная на параллелизацию, упрощение и надежность:** сокеты создаются на ранних стадиях загрузки, и единственный экземпляр службы тут же начинает обслуживать все поступающие запросы. Такая схема подходит для часто используемых системных служб — очевидно, что такие службы лучше запускать как можно раньше, и по возможности параллельно. В качестве примера можно привести D-Bus и Syslog.

2. **Сокет-активация для одиночных служб:** сокеты создаются на ранних стадиях загрузки, и единственный экземпляр службы запускается при получении входящих запросов. Такая схема больше подходит для редко используемых служб, позволяя обеспечить экономию системных ресурсов и уменьшить время загрузки, просто отложив активацию службы до того момента, когда она действительно понадобится. Пример — CUPS.
3. **Сокет-активация для служб, запускающих по экземпляру на каждое соединение:** сокеты создаются на ранней стадии загрузки, и при каждом входящем соединении запускается экземпляр службы, которому передается сокет соединения (слушающий сокет при этом остается у суперсервера, inetd или systemd). Эта схема подходит для редко используемых служб, не критичных по производительности (каждый новый процесс занимает сравнительно немного ресурсов). Пример: SSH.

Описанные схемы отнюдь не эквивалентны с точки зрения производительности. Первая и вторая схема, после завершения запуска службы, обеспечивают точно такую же производительность, как и в случае с независимой (stand-alone) службой (т.е. службой, запущенной без использования суперсервера и сокет-активации), так как слушающий сокет передается непосредственно процессу службы, и дальнейшая обработка входящих соединений происходит точно так же, как и в независимой службе. В то же время, производительность третьей из предложенных схем порою оставляет желать лучшего: каждое входящее соединение порождает еще один процесс службы, что, при большом количестве соединений, может привести к значительному потреблению системных ресурсов. Впрочем, эта схема имеет и свои достоинства. В частности, обеспечивается эффективная изоляция обработки клиентских запросов и, кроме того, выбор такой схемы активации значительно упрощает процесс разработки службы.

В systemd наибольшее внимание уделяется первой из этих схем, однако остальные две тоже прекрасно поддерживаются. В свое время, я [рассказывал](#), какие изменения в коде службы нужно сделать для обеспечения работы по второй схеме, на примере сервера CUPS. Что же касается inetd, то он предназначен прежде всего для работы по третьей схеме, хотя поддерживает и вторую (но не первую). Именно из-за специфики этой схемы inetd получил репутацию «медленного» (что, на самом деле, немного несправедливо).

Итак, изучив теоретическую сторону вопроса, перейдем к практике и рассмотрим, как inetd-служба может быть интегрирована с механизмами сокет-активации systemd. В качестве примера возьмем SSH, известную и широко распространенную службу. На большинстве систем, где она используется, частота обращений к ней не превышает одного раза в час (как правило, она *значительно* меньше этой величины). SSH уже очень давно поддерживает сокет-активацию в стиле inetd, согласно третьей схеме. Так как необходимость в данной службе возникает сравнительно редко, и число одновременно работающих процессов обычно невелико, она хорошо подходит для использования по этой схеме. Перерасход системных ресурсов должен быть незначителен: большую часть времени такая служба фактически не выполняется и не тратит ресурсы. Когда кто-либо начинает сеанс удаленной работы, она запускается, и останавливается немедленно по завершении сеанса, освобождая ресурсы. Что ж, посмотрим, как в systemd можно воспользоваться режимом совместимости с inetd и обеспечить сокет-активацию SSH.

Так выглядит строка с конфигурацией службы SSH для классического inetd:

```
ssh stream tcp nowait root /usr/sbin/sshd sshd -i
```

Аналогичный фрагмент конфигурации для xinetd:

```
service ssh {
    socket_type = stream
    protocol = tcp
    wait = no
    user = root
    server = /usr/sbin/sshd
    server_args = -i
}
```

Большая часть опций интуитивно понятна, кроме того, нетрудно заметить, что оба этих текста повествуют об одном и том же. Однако, некоторые моменты не вполне очевидны. Например, номер прослушиваемого порта (22) не указывается в явном виде. Он определяется путем поиска имени службы в базе данных `/etc/services`. Подобный подход был некогда весьма популярен в Unix, но сейчас он уже постепенно выходит из моды, и поэтому в современных версиях `xinetd` поддерживается возможность явного указания номера порта. Одна из наиболее интересных настроек названа не вполне очевидно — `nowait` (`wait=no`). Она определяет, будет ли служба работать по второй (`wait`) или третьей (`nowait`) схеме. И наконец, ключ `-i` активирует `inetd`-режим в SSH-сервере (без этого ключа он работает в независимом режиме).

На языке `systemd` эти фрагменты конфигурации превращаются в два файла. Первый из них, `sshd.socket`, содержит информацию о прослушиваемом сокете:

```
[Unit]
Description=SSH Socket for Per-Connection Servers

[Socket]
ListenStream=22
Accept=yes

[Install]
WantedBy=sockets.target
```

Смысл большинства опций вполне очевиден. Сделаю лишь пару замечаний. `Accept=yes` соответствует режиму `nowait`. Надеюсь, предложенное мною название более точно отражает смысл опции — в режиме `nowait` суперсервер сам вызывает `accept()` для слушающего сокета, в то время как в режиме `wait` эта работа ложится на процесс службы. Опция `WantedBy=sockets.target` обеспечивает активацию данного юнита в нужный момент при загрузке системы.

Второй из этих файлов — `sshd@.service`:

```
[Unit]
Description=SSH Per-Connection Server

[Service]
ExecStart=-/usr/sbin/sshd -i
StandardInput=socket
```

Большинство представленных здесь опций, как всегда, понятны интуитивно. Особое внимание стоит обратить лишь на строчку `StandardInput=socket`, которая, собственно, и включает для данной службы режим совместимости с `inetd`-активацией. Опция `StandardInput=` позволяет указать, куда будет подключен поток `STDIN` процесса данной службы (подробности см. [на странице руководства](#)). Задав для нее значение `socket`, мы обеспечиваем подключение этого потока к сокету соединения, как того и требует механизм `inetd`-активации. Заметим, что явно указывать опцию `StandardOutput=` в данном случае необязательно — она автоматически устанавливается в то же значение, что и `StandardInput`, если явно не указано что-то другое. Кроме того, можно отметить наличие «-» перед именем бинарного файла `sshd`. Таким образом мы указываем `systemd` игнорировать код выхода процесса `sshd`. По умолчанию, `systemd` сохраняет коды выхода для всех экземпляров службы, завершившихся с ошибкой (сбросить эту информацию можно командой `systemctl reset-failed`). SSH довольно часто завершается с ненулевым кодом выхода, и мы разрешаем `systemd` не регистрировать подобные «ошибки».

Служба `sshd@.service` предназначена для работы в виде независимых экземпляров (такие службы мы рассматривали в предыдущей главе [10](#)). Для каждого входящего соединения `systemd` будет создавать свой экземпляр этой службы, причем идентификатор экземпляра формируется на основе реквизитов соединения.

Быть может, вы спросите: почему для настройки `inetd`-службы в `systemd` требуется два файла конфигурации, а не один? Отвечаем: чтобы избежать излишнего усложнения, мы обеспечиваем максимально прозрачную связь между работающими юнитами и

соответствующими им юнит-файлами. Такой подход позволяет независимо оперировать юнитом сокета и юнитами соответствующих служб при формировании графа зависимостей и при управлении юнитами. В частности, вы можете остановить (удалить) сокет, не затрагивая уже работающие экземпляры соответствующей службы, или остановить любой из этих экземпляров, не трогая другие экземпляры и сам сокет.

Посмотрим, как наш пример будет работать. После того, как мы поместим оба предложенных выше файла в каталог `/etc/systemd/system`, мы сможем включить сокет (то есть, обеспечить его активацию при каждой нормальной загрузке) и запустить его (то есть активировать в текущем сеансе работы):

```
# systemctl enable sshd.socket
ln -s '/etc/systemd/system/sshd.socket' '/etc/systemd/system/sockets.target.wants/sshd.socket'
# systemctl start sshd.socket
# systemctl status sshd.socket
sshd.socket - SSH Socket for Per-Connection Servers
    Loaded: loaded (/etc/systemd/system/sshd.socket; enabled)
    Active: active (listening) since Mon, 26 Sep 2011 20:24:31 +0200; 14s ago
    Accepted: 0; Connected: 0
    CGroup: name=systemd:/system/sshd.socket
```

Итак, наш сокет уже прослушивается, но входящих соединений на него пока не поступало (счетчик `Accepted`: показывает количество принятых соединений с момента создания сокета, счетчик `Connected`: — количество активных соединений на текущий момент).

Подключимся к нашему серверу с двух разных хостов, и посмотрим на список служб:

```
$ systemctl --full | grep ssh
sshd@172.31.0.52:22-172.31.0.4:47779.service loaded active running      SSH Per-Connection Server
sshd@172.31.0.52:22-172.31.0.54:52985.service loaded active running  SSH Per-Connection Server
sshd.socket                          loaded active listening    SSH Socket for Per-Connection Serv
```

Как и следовало ожидать, работают два экземпляра нашей службы, по одному на соединение, и их в названиях указаны IP-адреса и TCP-порты источника и получателя. (Заметим, что в случае с локальными сокетами типа `AF_UNIX` там были бы указаны идентификаторы процесса и пользователя, соответствующие подключенному клиенту.) Таким образом, мы можем независимо отслеживать и убивать отдельные экземпляры `sshd` (например, если нам нужно прервать конкретный удаленный сеанс):

```
# systemctl kill sshd@172.31.0.52:22-172.31.0.4:47779.service
```

Вот, пожалуй, и все, что вам нужно знать о портировании `inetd`-служб в `systemd` и дальнейшем их использовании.

Применительно к SSH, в большинстве случаев схема с `inetd`-активацией позволяет сэкономить системные ресурсы и поэтому оказывается более эффективным решением, чем использование обычного `service`-файла `sshd`, обеспечивающего запуск одиночной службы без использования сокет-активации (поставляется в составе пакета и может быть включен при необходимости). В ближайшее время я собираюсь направить соответствующий запрос относительно нашего пакета SSH в багтрекер Fedora.

В завершение нашей дискуссии, сравним возможности `xinetd` и `systemd`, и выясним, может ли `systemd` полностью заменить `xinetd`, или нет. Вкратце: `systemd` поддерживает далеко не все возможности `xinetd` и поэтому отнюдь не является его полноценной заменой на все случаи жизни. В частности, если вы заглянете в [список параметров конфигурации xinetd](#), вы заметите, что далеко не все эти опции доступны в `systemd`. Например, в `systemd` нет и никогда не будет встроенных служб `echo`, `time`, `daytime`, `discard` и т.д. Кроме того, `systemd` не поддерживает `TCPMUX` и `RPC`. Впрочем, большая часть этих опций уже не актуальна в наше время. Подавляющее большинство `inetd`-служб не используют эти опции (в частности, ни одна из имеющихся в Fedora `xinetd`-служб не требует поддержки перечисленных опций). Стоит отметить, что `xinetd` имеет некоторые интересные возможности, отсутствующие в `systemd`, например, списки контроля

доступа для IP-адресов (IP ACL). Однако, большинство администраторов, скорее всего, согласятся, что настройка брандмауэра является более эффективным решением подобных задач⁴⁵, а для ценителей устаревших технологий `systemd` предлагает поддержку `tcpwrar`. С другой стороны, `systemd` тоже предоставляет ряд возможностей, отсутствующих в `xinetd`, в частности, индивидуальный контроль над каждым экземпляром службы (см. выше), и внушительный набор настроек для контроля окружения, в котором запускаются экземпляры. Я надеюсь, что возможностей `systemd` должно быть достаточно для решения большинства задач, а в тех редких случаях, когда вам потребуются специфические опции `xinetd` — ничто не мешает вам запустить его в дополнение к `systemd`. Таким образом, уже сейчас в большинстве случаев `xinetd` можно выкинуть из числа обязательных системных компонентов. Можно сказать, что `systemd` не просто возвращает функциональность классического юниксового `inetd`, но еще и восстанавливает ее ключевую роль в Linux-системах.

Теперь, вооруженные этими знаниями, вы можете портировать свои службы с `inetd` на `systemd`. Но, конечно, будет лучше, если этим займутся разработчики из апстрима приложения, или сопровождающие вашего дистрибутива.

12 К вопросу о безопасности

Одно из важнейших достоинств Unix-систем — концепция разделения привилегий между различными компонентами ОС. В частности, службы обычно работают от имени специальных системных пользователей, имеющих ограниченные полномочия, что позволяет уменьшить ущерб для системы в случае взлома этих служб.

Однако, такой подход предоставляет лишь самую минимальную защиту, так как системные службы, хотя уже и не получают полномочий администратора (`root`), все равно имеют практически те же права, что и обычные пользователи. Чтобы обеспечить более эффективную защиту, нужно поставить более жесткие ограничения, отняв у служб некоторые привилегии, присущие обычным пользователям.

Такая возможность предоставляется системами мандатного контроля доступа (далее MAC, от Mandatory Access Control), например, SELinux. Если вам нужно обеспечить высокий уровень безопасности на своем сервере, то вам определенно стоит обратить свое внимание на SELinux. Что же касается `systemd`, то он предоставляет разработчикам и администраторам целый арсенал возможностей по ограничению локальных служб, и эти механизмы работают независимо от систем MAC. Таким образом, вне зависимости от того, смогли ли вы разобраться с SELinux — у вас появляется еще несколько инструментов, позволяющих повысить уровень безопасности.

В этой главе мы рассмотрим несколько таких опций, предоставляемых `systemd`, и обсудим вопросы их практического применения. Реализация этих опций основана на использовании ряда уникальных технологий безопасности, интегрированных в ядро Linux уже очень давно, но при этом практически неизвестных для большинства разработчиков. Мы постарались сделать соответствующие опции `systemd` максимально простыми в использовании, чтобы заинтересовать администраторов и апстримных разработчиков. Вот краткий перечень наиболее интересных возможностей⁴⁶:

- Изолирование служб от сети

⁴⁵Прим. перев.: Стоит отметить, что приведенный пример является не единственным случаем, когда возможности брандмауэра Linux дублируются опциями `xinetd`. Например, количество соединений с каждого хоста может быть ограничено критерием `connlimit`, а скорость поступления входящих соединений можно контролировать сочетанием критериев `limit` и `conntrack (ctstate NEW)`. Критерий `recent` позволяет создать аналог простейшей IDS/IPS, реализованной механизмом `SENSORS` в `xinetd`. Кроме того, в ряде случаев возможности брандмауэра значительно превосходят функциональность `xinetd`. Например, критерий `hashlimit`, опять-таки в сочетании с `conntrack`, позволяет ограничить скорость поступления входящих соединений с каждого хоста (не путать с критерием `connlimit`, ограничивающим количество одновременно открытых соединений). Также стоит отметить, что интегрированная в Linux подсистема `ipset` гораздо лучше подходит для работы с большими списками разрешенных/запрещенных адресов, нежели встроенные механизмы `xinetd`.

⁴⁶Прим. перев.: В приведенном здесь списке не упомянута встроенная в `systemd` поддержка фильтров `seccomp` (опция `SystemCallFilter=`), так как она была добавлена уже после написания исходной статьи, в выпуске `systemd` 187.

- Предоставление службам независимых каталогов `/tmp`
- Ограничение доступа служб к отдельным каталогам
- Принудительное отключение полномочий (capabilities) для служб
- Запрет форка, ограничение на создание файлов
- Контроль доступа служб к файлам устройств

Все эти опции описаны в man-страницах `systemd`, главным образом, в [systemd.exec\(5\)](#)⁴⁷. Если вам потребуется дополнительная информация, вы можете обратиться к ним.

Все эти опции доступны на системах с `systemd`, вне зависимости от использования SELinux или любой другой реализации MAC.

Все эти опции не так уж и обременительны, и поэтому их разумнее будет использовать даже в тех случаях, когда явная необходимость в них, казалось бы, отсутствует. Например: даже если вы полагаете, что ваша служба никогда не пишет в каталог `/tmp`, и поэтому использование `PrivateTmp=yes` (см. ниже) вроде бы и не обязательно — лучше включить эту опцию, просто потому, что вы не можете знать наверняка, как будут вести себя используемые вами сторонние библиотеки (и плагины для них). В частности, вы никогда не узнаете, какие модули NSS могут быть включены в каждой конкретной системе, и пользуются ли они каталогом `/tmp`.

Мы надеемся, что перечисленные опции окажутся полезными как для администраторов, защищающих свои системы, так и для апстримных разработчиков, желающих сделать свои службы безопасными «из коробки». Мы настоятельно рекомендуем разработчикам использовать такие опции по умолчанию в апстримных `service`-файлах — это сравнительно несложно, но дает значительные преимущества в плане безопасности.

12.1 Изолирование служб от сети

Простая, но мощная опция, которой вы можете воспользоваться при настройке службы — `PrivateNetwork=`:

```
...
[Service]
ExecStart=...
PrivateNetwork=yes
...
```

Добавление этой строки обеспечивает полную изоляцию от сети всех процессов данной службы. Они будут видеть лишь интерфейс обратной петли (`lo`), причем полностью изолированный от обратной петли основной системы. Чрезвычайно эффективная защита против сетевых атак.

Предупреждение: Некоторым службам сеть необходима для нормальной работы. Разумеется, никто и не говорит о том, чтобы применять `PrivateNetwork=yes` к сетевым службам, таким, как Apache. Однако даже те службы, которые не ориентированы на сетевое взаимодействие, могут нуждаться в сети для нормального функционирования, и порой эта потребность не вполне очевидна. Например, если ваша система хранит учетные записи пользователей в базе LDAP, для выполнения системных вызовов наподобие `getpwnam()` может потребоваться обращение к сети. Впрочем, даже в такой ситуации, как правило, можно использовать `PrivateNetwork=yes`, за исключением случаев, когда службе может потребоваться информация о пользователях с `UID ≥ 1000`.

Если вас интересуют технические подробности: эта возможность реализована с использованием технологии сетевых пространств имен (`network namespaces`). При задействовании данной опции, для службы создается новое пространство имен, в котором настраивается только интерфейс обратной петли.

⁴⁷Прим. перев.: Начиная с `systemd` версии 206, значительная часть обсуждаемых здесь настроек вынесена на отдельную страницу [systemd.cgroup\(5\)](#). Начиная с `systemd` 208, эта страница переименована в [systemd.resource-control\(5\)](#).

12.2 Предоставление службам независимых каталогов /tmp

Еще одна простая, но мощная опция настройки служб — `PrivateTmp=`:

```
...
[Service]
ExecStart=...
PrivateTmp=yes
...
```

При задействовании этой опции, служба получит свой собственный каталог `/tmp`, полностью изолированный от общесистемного `/tmp`. По давно сложившейся традиции, в Unix-системах каталог `/tmp` является общедоступным для всех локальных служб и пользователей. За все эти годы, он стал источником огромного количества проблем безопасности. Чаще всего встречаются атаки с использованием символьных ссылок (`symlink attacks`) и атаки на отказ в обслуживании (`DoS attacks`). Использование независимой версии данного каталога для каждой службы делает подобные уязвимости практически бесполезными.

Для релиза Fedora 17 [утверждена](#) инициатива, направленная на включение этой опции по умолчанию для большинства системных служб.

Предупреждение: Некоторые службы используют `/tmp` не по назначению, помещая туда сокетные IPC и другие подобные элементы, что само по себе уже является уязвимостью (хотя бы потому, что используемые при передаче информации файлы и каталоги должны иметь предсказуемые имена, что делает подобные службы уязвимыми к атакам через символьные ссылки и атакам на отказ в обслуживании). Подобные объекты лучше помещать в каталог `/run`, так как в нем присутствует строгое разделение доступа. В качестве примера такого некорректного использования `/tmp` можно привести X11, размещающий там свои коммуникационные сокеты (впрочем, в данном конкретном случае некоторые меры безопасности все же присутствуют: сокеты размещаются в защищенном подкаталоге, который создается на ранних стадиях загрузки). Разумеется, для служб, использующих `/tmp` в целях коммуникации, включение опции `PrivateTmp=yes` недопустимо. К счастью, подобных служб сейчас уже не так уж и много⁴⁸.

Эта опция использует технологию пространств имен файловых систем (`filesystem namespaces`), реализованную в Linux. При включении данной опции, для службы создается новое пространство имен, отличающееся от иерархии каталогов основной системы только каталогом `/tmp`.

12.3 Ограничение доступа служб к отдельным каталогам

Используя опции `ReadOnlyDirectories=` и `InaccessibleDirectories=`, вы можете ограничить доступ службы к указанным каталогам только чтением, либо вообще запретить его:

```
...
[Service]
ExecStart=...
InaccessibleDirectories=/home
ReadOnlyDirectories=/var
...
```

Добавление этих двух строчек в файл конфигурации приводит к тому, что служба полностью утрачивает доступ к содержимому каталога `/home` (она видит лишь пустой каталог с правами доступа `000`), а также не может писать внутрь каталога `/var`.

⁴⁸Прим. перев.: Начиная с `systemd 209`, поддерживается опция `JoinsNamespaceOf=` (секция `[Unit]`), позволяющая поместить два и более юнитов в одну «песочницу» (пространство имен), в результате чего такие юниты смогут взаимодействовать между собой, как через сетевой интерфейс обратной петли, так и через файлы в каталоге `/tmp`, при этом оставаясь изолированными от остальной системы. Подробности можно уточнить на странице руководства [systemd.unit\(5\)](#).

Предупреждение: К сожалению, в настоящее время опция `ReadOnlyDirectories=` не применяется рекурсивно к точкам монтирования, расположенным в поддереве указанного каталога (т.е. файловые системы, смонтированные в подкаталогах `/var`, по-прежнему останутся доступными на запись, если, конечно, не перечислить их все поименно). Мы собираемся исправить это в ближайшее время.

Механизм работы этой опции тоже реализован с использованием пространств имен файловых систем.

12.4 Принудительное отключение полномочий (`capabilities`) для служб

Еще одна чрезвычайно эффективная опция — `CapabilityBoundingSet=`, позволяющая контролировать список `capabilities`, которые смогут получить процессы службы:

```
...
[Service]
ExecStart=...
CapabilityBoundingSet=CAP_CHOWN CAP_KILL
...
```

В нашем примере, служба может иметь лишь полномочия `CAP_CHOWN` и `CAP_KILL`. Попытки какого-либо из процессов службы получить любые другие полномочия, даже с использованием `suid`-бинарников, будут пресекаться. Список возможных полномочий приведен на странице документации [capabilities\(7\)](#). К сожалению, некоторые из них являются слишком общими (разрешают очень многое), например, `CAP_SYS_ADMIN`, однако выборочное делегирование полномочий все же является неплохой альтернативой запуску службы с полными административными (рутовыми) привилегиями.

Как правило, определение списка полномочий, необходимых для работы конкретной службы, является довольно трудоемким процессом, требующим ряда проверок. Чтобы немного упростить эту задачу, мы добавили возможность создания «черного списка» привилегий, как альтернативы описанному выше механизму «белого списка». Вместо того, чтобы указывать, какие привилегии можно оставить, вы можете перечислить, какие из них оставлять точно нельзя. Например: привилегия `CAP_SYS_PTRACE`, предназначенная для отладчиков, дает очень широкие полномочия в отношении всех процессов системы. Очевидно, что такие службы, как Apache, не занимаются отладкой чужих процессов, и поэтому мы можем спокойно отнять у них данную привилегию:

```
...
[Service]
ExecStart=...
CapabilityBoundingSet=~CAP_SYS_PTRACE
...
```

Наличие символа `~` после знака равенства инвертирует принцип работы опции: следующий за ним перечень привилегий рассматривается уже не как белый, а как черный список.

Предупреждение: Некоторые службы, при отсутствии определенных привилегий, могут вести себя некорректно. Как уже говорилось выше, формирование списка необходимых полномочий для каждой конкретной службы может оказаться довольно трудной задачей, и лучше всего будет обратиться с соответствующим запросом к разработчикам службы.

Предупреждение 2: *Привилегии — отнюдь не лекарство от всех бед.* Чтобы они работали действительно эффективно, возможно, придется дополнить их другими методами защиты.

Чтобы проверить, какие именно привилегии имеют сейчас ваши процессы, вы можете воспользоваться программой `pscap` из комплекта `libcap-ng-utils`.

Применение опции `CapabilityBoundingSet=` является простой, прозрачной и удобной альтернативой введению аналогичных ограничений через модификацию исходного кода всех системных служб.

12.5 Запрет форка, ограничение на создание файлов

Некоторые меры безопасности можно ввести при помощи механизма ограничения ресурсов. Как следует из его названия, этот механизм предназначен прежде всего для контроля использования ресурсов, а не для контроля доступа. Однако, две его настройки могут быть использованы для запрета определенных действий: с помощью `RLIMIT_NPROC` можно запретить службе форкаться (запускать дополнительные процессы), а `RLIMIT_FSIZE` позволяет блокировать запись в файлы ненулевого размера:

```
...
[Service]
ExecStart=...
LimitNPROC=1
LimitFSIZE=0
...
```

Обратите внимание, что эти ограничения будут эффективно работать только в том случае, если служба запускается от имени простого пользователя (не `root`) и без привилегии `CAP_SYS_RESOURCE` (блокирование этой привилегии можно обеспечить, например, описанной выше опцией `CapabilityBoundingSet=`). В противном случае, ничто не мешает процессу поменять соответствующие ограничения.

Предупреждение: `LimitFSIZE=` действует очень жестко. Если процесс пытается записать файл ненулевого размера, он немедленно получает сигнал `SIGXFSZ`, который, как правило, прекращает работу процесса (в случае, если не назначен обработчик сигнала). Кроме того, стоит отметить, что эта опция не запрещает создание файлов нулевого размера.

Подробности об этих и других опциях ограничения ресурсов можно уточнить на странице документации [setrlimit\(2\)](#).

12.6 Контроль доступа служб к файлам устройств

Файлы устройств предоставляют приложениям интерфейс для доступа к ядру и драйверам. Причем драйверы, как правило, менее тщательно тестируются и не так аккуратно проверяются на предмет наличия уязвимостей, по сравнению с основными компонентами ядра. Именно поэтому драйверы часто становятся объектами атаки злоумышленников. `systemd` позволяет контролировать доступ к устройствам индивидуально для каждой службы:

```
...
[Service]
ExecStart=...
DeviceAllow=/dev/null rw
...
```

Приведенная конфигурация разрешит службе доступ только к файлу `/dev/null`, запретив доступ ко всем остальным файлам устройств.

Реализация данной опции построена на использовании `sgroups`-контроллера `devices`.

12.7 Прочие настройки

Помимо приведенных выше, простых и удобных в использовании опций, существует и ряд других настроек, позволяющих повысить уровень безопасности. Но для их использования, как правило, уже требуются определенные изменения в исходном коде службы, и поэтому такие настройки представляют интерес прежде всего для апстримных разработчиков. В частности, сюда относится опция `RootDirectory=` (позволяющая

запустить службу `chroot()`-окружении), а также параметры `User=` и `Group=`, определяющие пользователя и группу, от имени которых работает служба. Использование этих опций значительно упрощает написание демонов, так как работа по понижению привилегий ложится на плечи `systemd`.

Возможно, у вас возникнет вопрос, почему описанные выше опции не включены по умолчанию. Ответаем: чтобы не нарушать совместимость. Многие из них несколько отступают от традиций, принятых в Unix. Например, исторически сложилось, так что `/tmp` является общим для всех процессов и пользователей. Существуют программы, использующие этот каталог для IPC, и включение по умолчанию режима изоляции для `/tmp` нарушит работу таких программ.

И несколько слов в заключение. Если вы занимаетесь сопровождением `unit`-файлов в апстримном проекте или в дистрибутиве, пожалуйста, подумайте о том, чтобы воспользоваться приведенными выше настройками. Если сопровождаемая вами служба станет более защищенной в конфигурации по умолчанию («из коробки»), от этого выиграют не только ваши пользователи, но и все мы, потому что Интернет станет чуть более безопасным.

13 Отчет о состоянии службы и ее журнал

При работе с системами, использующими `systemd`, одной из наиболее важных и часто используемых команд является `systemctl status`. Она выводит отчет о состоянии службы (или другого юнита). В таком отчете приводится статус службы (работает она или нет), список ее процессов и другая полезная информация.

В Fedora 17 мы ввели [Journal](#), новую реализацию системного журнала, обеспечивающую структурированное, индексированное и надежное журналирование, при сохранении совместимости с классическими реализациями `syslog`. Собственно, мы начали работу над `Journal` только потому, что хотели добавить одну, казалось бы, простую, возможность: включить в вывод `systemctl status` последние 10 сообщений, поступивших от службы в системный журнал. Но на практике оказалось, что в рамках классических механизмов `syslog`, реализация этой возможности чрезвычайно сложна и малоэффективна. С другой стороны, сообщения службы в системный журнал несут очень важную информацию о ее состоянии, и такая возможность действительно была бы очень полезной, особенно при диагностике нештатных ситуаций.

Итак, мы интегрировали `Journal` в `systemd`, и научили `systemctl` работать с ним. Результат выглядит примерно так:

```
$ systemctl status avahi-daemon.service
avahi-daemon.service - Avahi mDNS/DNS-SD Stack
   Loaded: loaded (/usr/lib/systemd/system/avahi-daemon.service; enabled)
   Active: active (running) since Fri, 18 May 2012 12:27:37 +0200; 14s ago
 Main PID: 8216 (avahi-daemon)
   Status: "avahi-daemon 0.6.30 starting up."
   CGroup: name=systemd:/system/avahi-daemon.service
           └─ 8216 avahi-daemon: running [omega.local]
              8217 avahi-daemon: chroot helper

May 18 12:27:37 omega avahi-daemon[8216]: Joining mDNS multicast group on interface eth1.IPv4 with address 1
May 18 12:27:37 omega avahi-daemon[8216]: New relevant interface eth1.IPv4 for mDNS.
May 18 12:27:37 omega avahi-daemon[8216]: Network interface enumeration completed.
May 18 12:27:37 omega avahi-daemon[8216]: Registering new address record for 192.168.122.1 on virbr0.IPv4.
May 18 12:27:37 omega avahi-daemon[8216]: Registering new address record for fd00::e269:95ff:fe87:e282 on et
May 18 12:27:37 omega avahi-daemon[8216]: Registering new address record for 172.31.0.52 on eth1.IPv4.
May 18 12:27:37 omega avahi-daemon[8216]: Registering HINFO record with values 'X86_64'/'LINUX'.
May 18 12:27:38 omega avahi-daemon[8216]: Server startup complete. Host name is omega.local. Local service c
May 18 12:27:38 omega avahi-daemon[8216]: Service "omega" (/services/ssh.service) successfully established.
May 18 12:27:38 omega avahi-daemon[8216]: Service "omega" (/services/sftp-ssh.service) successfully establis
```

Очевидно, это отчет о состоянии всеми любимого демона mDNS/DNS-SD, причем после списка процессов, как мы и обещали, приведены сообщения этого демона в системный журнал. Миссия выполнена!

Команда `systemctl status` поддерживает ряд опций, позволяющих настроить вывод информации в соответствии с вашими пожеланиями. Отметим две наиболее интересные из них: ключ `-f` позволяет в реальном времени отслеживать обновление сведений (по аналогии с `tail -f`), а ключ `-n` задает количество выводимых журнальных записей (как нетрудно догадаться, по аналогии с `tail -n`).

Журнальные записи собираются из трех различных источников. Во-первых, это сообщения службы в системный журнал, отправленные через функцию `libc syslog()`. Во-вторых, это сообщения, отправленные через штатный API системы Journal. И наконец, это весь текст, выводимый демоном в `STDOUT` и `STDERR`. Проще говоря, все, что демон считает нужным сказать администратору, собирается, упорядочивается и отображается в одинаковом формате.

Добавленная нами возможность, при всей своей простоте, может оказаться чрезвычайно полезной практически любому администратору. По-хорошему, ее стоило реализовать еще 15 лет назад.

14 Самодокументированный процесс загрузки

Нам часто приходится слышать жалобы, что процесс загрузки при использовании `systemd` очень сложен для понимания. Не могу с этим согласиться. Более того, я беру даже утверждать обратное: по сравнению с тем, что мы имели раньше (когда для того, чтобы разобраться в загрузке, нужно было иметь хорошие навыки программирования на языке Bourne Shell⁴⁹), процесс загрузки стал более простым и прозрачным. Но определенная доля истины в этом критическом замечании все же есть: даже опытному Unix-администратору при переходе на `systemd` нужно изучить некоторые новые для себя вещи. Мы, разработчики `systemd`, обязаны максимально упростить такое обучение. Решая поставленную задачу, мы подготовили для вас кое-какие приятные сюрпризы, и предоставили хорошую документацию даже там, где это казалось невозможным.

Уже сейчас `systemd` располагает довольно обширной документацией, включая [страницы руководства](#) (на данный момент, их около сотни), [wiki-сайт проекта](#), а также ряд статей в моем блоге. Тем не менее, большое количество документации еще не гарантирует простоты понимания. Огромные груды руководств выглядят пугающе, и неподготовленный читатель не может разобраться, с какого места ему начинать чтение, если его интересует просто общая концепция.

Чтобы решить данную проблему, мы добавили в `systemd` небольшую, но очень изящную возможность: самодокументированный⁵⁰ процесс загрузки. Что мы под этим подразумеваем? То, что для каждого элемента процесса загрузки теперь имеется документация, прозрачно привязанная к этому элементу, так что ее поиск не представляет трудности.

Иными словами, все штатные юниты `systemd` (которые, собственно, и формируют процесс загрузки, вызывая соответствующие программы) включают ссылки на страницы документации, описывающие назначение юнита/программы, используемые ими конфигурационные файлы и т.д. Если пользователь хочет узнать, для чего предназначен какой-либо юнит, какова его роль в процессе загрузки и как его можно настроить, может получить ссылки на соответствующую документацию, просто воспользовавшись давно известной командой `systemctl status`. Возьмем для примера `systemd-logind.service`:

```
$ systemctl status systemd-logind.service
systemd-logind.service - Login Service
   Loaded: loaded (/usr/lib/systemd/system/systemd-logind.service; static)
   Active: active (running) since Mon, 25 Jun 2012 22:39:24 -0200; 1 day and 18h ago
     Docs: man:systemd-logind.service\(7\)
```

⁴⁹Чья привлекательность очень коварна и обманчива.

⁵⁰Прим. перев.: В оригинале использован термин «self-explanatory», который также можно перевести как «само-объясняющий», «самоочевидный».

```

man:logind.conf(5)
http://www.freedesktop.org/wiki/Software/systemd/multiseat
Main PID: 562 (systemd-logind)
  CGroup: name=systemd:/system/systemd-logind.service
          └─ 562 /usr/lib/systemd/systemd-logind

Jun 25 22:39:24 epsilon systemd-logind[562]: Watching system buttons on /dev/input/event2 (Power Button)
Jun 25 22:39:24 epsilon systemd-logind[562]: Watching system buttons on /dev/input/event6 (Video Bus)
Jun 25 22:39:24 epsilon systemd-logind[562]: Watching system buttons on /dev/input/event0 (Lid Switch)
Jun 25 22:39:24 epsilon systemd-logind[562]: Watching system buttons on /dev/input/event1 (Sleep Button)
Jun 25 22:39:24 epsilon systemd-logind[562]: Watching system buttons on /dev/input/event7 (ThinkPad Extra Bu
Jun 25 22:39:25 epsilon systemd-logind[562]: New session 1 of user gdm.
Jun 25 22:39:25 epsilon systemd-logind[562]: Linked /tmp/.X11-unix/X0 to /run/user/42/X11-display.
Jun 25 22:39:32 epsilon systemd-logind[562]: New session 2 of user lennart.
Jun 25 22:39:32 epsilon systemd-logind[562]: Linked /tmp/.X11-unix/X0 to /run/user/500/X11-display.
Jun 25 22:39:54 epsilon systemd-logind[562]: Removed session 1.

```

На первый взгляд, в выводе этой команды почти ничего не изменилось. Но если вы взгляните повнимательнее, то заметите новое поле `Docs`, в котором приведены ссылки на документацию по данной службе. В нашем случае это два URI, ссылающихся на `man`-страницы, и один URL, указывающий на веб-страницу. `man`-страницы описывают соответственно предназначение службы и ее настройки, а веб-страница — базовые концепции, которые реализуются этой службой.

Тем, кто использует современные графические эмуляторы терминала, чтобы открыть соответствующую страницу документации, достаточно щелкнуть мышью по URI⁵¹. Иными словами, доступ к документации компонентов загрузки становится предельно простым: достаточно вызвать `systemctl status`, после чего щелкнуть по полученным ссылкам.

Если же вы используете не графический эмулятор терминала (где можно просто щелкнуть по URI), а настоящий терминал, наличие URI где-то в середине вывода `systemctl status` вряд ли покажется вам удобным. Чтобы упростить обращение к соответствующим страницам документации без использования мыши и копирования/вставки, мы ввели новую команду

```
systemctl help systemd-logind.service
```

которая просто откроет в вашем терминале соответствующие `man`-страницы.

URI, ссылающиеся на документацию, формируются в соответствии со страницей руководства `uri(7)`. Поддерживаются схемы `http/https` (URL для веб-страниц) и `man/info` (локальные страницы руководства).

Разумеется, добавленная нами возможность не обеспечивает полной очевидности абсолютно всего, хотя бы потому, что пользователь должен как минимум знать про команду `systemctl status` (а также вообще про программу `systemctl` и про то, что такое юнит). Но при наличии этих базовых знаний, пользователю уже не составит труда получить информацию по интересующим его вопросам.

Мы надеемся, что добавленный нами механизм для связи работающего кода и соответствующей документации станет значительным шагом к полностью прозрачному и предельно простому для понимания процессу загрузки.

Описанная функциональность частично присутствует в Fedora 17, а в полном объеме она будет представлена в Fedora 18.

В завершение стоит отметить, что использованная нами идея не является такой уж новой: в SMF, системе инициализации Solaris, уже используется практика указания ссылок на документацию в описании службы. Однако, в Linux такой подход является принципиально новым, и `systemd` сейчас является наиболее документированной и прозрачной системой загрузки для данной платформы.

Если вы занимаетесь разработкой или сопровождением пакетов и создаете файл конфигурации юнита, пожалуйста, включите в него ссылки на документацию. Это

⁵¹Для нормальной работы данной функции, в эмуляторе терминала должен быть исправлена [эта ошибка](#), а в программе просмотра страниц помощи — [вот эта](#).

очень просто: достаточно добавить в секции [Unit] поле `Documentation=` и перечислить в нем соответствующие URI. Подробнее об этом поле см. на странице руководства [systemd.unit\(5\)](#). Чем больше будет документированных юнитов, тем проще станет работа системного администратора. (В частности, я уже направил в ФРС [предложение](#) внести соответствующие пожелания в нормативные документы, регулирующие подготовку пакетов для Fedora.)

Да, кстати: если вас интересует общий обзор процесса загрузки `systemd`, то вам стоит обратить внимание на [новую страницу руководства](#), где представлена псевдографическая потоковая диаграмма, описывающая процесс загрузки и роль ключевых юнитов.

15 Сторожевые таймеры

Разрабатывая `systemd`, мы ориентируемся на три основных области его применения: мобильные/встраиваемые устройства, настольные системы и промышленные серверы. Мобильные и встраиваемые системы, как правило, располагают очень скромными ресурсами и вынуждены очень экономно расходовать энергию. Настольные системы уже не так сильно ограничены по мощности, хотя все же проигрывают в этом плане промышленным серверам. Но, как ни странно, существуют возможности, востребованные в обоих крайних случаях (встраиваемые системы и серверы), но не очень актуальные в промежуточной ситуации (десктопы). В частности, к ним относится поддержка [сторожевых таймеров](#) (`watchdogs`), как аппаратных, так и программных.

Во встраиваемых системах часто используются аппаратные сторожевые таймеры, выполняющие сброс системы, когда программа перестает отвечать (точнее, перестает периодически посылать таймеру сигнал «все в порядке»). Таким образом, обеспечивается повышение надежности системы: что бы ни случилось, система обязательно попытается привести себя в рабочее состояние. На десктопах такая функциональность практически не востребована⁵². С другой стороны, она весьма актуальна для высокодоступных серверных систем.

Начиная с версии 183, `systemd` полностью поддерживает аппаратные сторожевые таймеры (доступные через интерфейс `/dev/watchdog`), а также обеспечивает программный сторожевой контроль системных служб. В целом эта схема (если она задействована) работает так. `systemd` периодически посылает сигналы аппаратному таймеру. В том случае, если `systemd` или ядро зависают, аппаратный таймер, не получив очередного сигнала, автоматически перезапустит систему. Таким образом, ядро и `systemd` защищены от бесконечного зависания — на аппаратном уровне. В то же время, сам `systemd` предоставляет интерфейс, реализующий логику программных сторожевых таймеров для отдельных служб. Таким образом можно обеспечить, например, принудительный перезапуск службы в случае ее зависания. Для каждой службы можно независимо настроить частоту опроса и задать соответствующее действие. Соединив оба звена (аппаратный сторожевой таймер, контролирующий ядро и `systemd`, и программные сторожевые таймеры, посредством которых `systemd` контролирует отдельные службы), мы получаем надежную схему надзора за всеми системными компонентами.

Чтобы задействовать аппаратный таймер, достаточно задать ненулевое значение параметра `RuntimeWatchdogSec=` в файле `/etc/systemd/system.conf`. По умолчанию этот параметр равен нулю (т.е. аппаратный таймер не задействован). Установив его равным, например, «20s», мы включим аппаратный сторожевой таймер. Если в течение 20 секунд таймер не получит очередного сигнала «все в порядке», система автоматически перезагрузится. Отметим, что `systemd` отправляет такие сигналы с периодом, равным половине заданного интервала — в нашем случае, через каждые 10 секунд. Собственно, это все. Просто задав один параметр, вы обеспечите аппаратный контроль работоспособности `systemd` и ядра⁵³.

⁵²Тем не менее, сейчас аппаратные сторожевые таймеры все чаще появляются и в настольных системах, так как стоят они относительно дешево и доступны практически во всех современных чипсетах.

⁵³Небольшой совет: если вы занимаетесь отладкой базовых системных компонентов — не забудьте включить сторожевой таймер. Иначе ваша система может неожиданно перезагрузиться, когда отладчик остановит процесс `init` и тот не может вовремя отправить сообщение таймеру.

Заметим, что с аппаратным сторожевым таймером (`/dev/watchdog`) должна работать только одна программа. Этой программой может быть либо `systemd`, либо отдельный демон сторожевого таймера (например, `watchdogd`) — выбор остается за вами.

Стоит упомянуть здесь еще одну опцию из файла `/etc/systemd/system.conf` — `ShutdownWatchdogSec=`. Она позволяет настроить аппаратный сторожевой таймер для контроля процесса перезагрузки. По умолчанию она равна «10min». Если в процессе остановки системы перед перезагрузкой она зависнет, аппаратный таймер принудительно перезагрузит ее по истечении данного интервала.

Это все, что я хотел сказать об аппаратных таймерах. Двух вышеописанных опций должно быть вполне достаточно для полноценного использования их возможностей. А сейчас мы рассмотрим логику программных сторожевых таймеров, обеспечивающих контроль работоспособности отдельных служб.

Прежде всего отметим, что для полноценной поддержки сторожевого контроля, программа должна содержать специальный код, периодически отправляющий таймеру сигналы «все в порядке». Добавить поддержку такой функциональности довольно просто. Для начала, демон должен проверить переменную окружения `WATCHDOG_USEC=`. Если она определена, то ее значение задает контрольный интервал в микросекундах, сформатированный в виде текстовой (ASCII) строки. В этом случае демон должен регулярно выполнять вызов `sd_notify("WATCHDOG=1")` с периодом, равным половине указанного интервала. Таким образом, поддержка программного сторожевого контроля со стороны демона сводится к проверке значения переменной окружения, и выполнении определенных действий в соответствии с этим значением.

Если интересующая вас служба обеспечивает поддержку такой функциональности, вы можете включить для нее сторожевой контроль, задав опцию `WatchdogSec=` в ее юнит-файле. Эта опция задает период работы таймера (подробнее см. [systemd.service\(5\)](#)). Если вы зададите ее, то `systemd` при запуске службы передаст ей соответствующее значение `WATCHDOG_USEC=` и, если служба перестанет своевременно отправлять сигналы «все в порядке», присвоит ей статус сбойной (failure state).

Очевидно, что одного только присвоения статуса недостаточно для обеспечения надежной работы системы. Поэтому нам также пригодятся настройки, определяющие, нужно ли перезапускать зависшую службу, количество попыток перезапуска, и дальнейшие действия, если она все равно продолжает сбойить. Чтобы включить автоматический перезапуск службы при сбое, нужно задать опцию `Restart=on-failure` в ее юнит-файле. Чтобы настроить, сколько раз `systemd` будет пытаться перезапустить службу, воспользуйтесь настройками `StartLimitBurst=` `StartLimitInterval=` (первая из них определяет предельное количество попыток, вторая — интервал времени, в течение которого они подсчитываются). В том случае, если достигнут предел количества попыток за указанное время, выполняется действие, заданное параметром `StartLimitAction=`. По умолчанию он установлен в `none` (никаких дополнительных действий не будет, службу просто оставят в покое со статусом сбойной). В качестве альтернативы можно указать одно из трех специальных действий: `reboot`, `reboot-force` и `reboot-immediate`. `reboot` соответствует обычной перезагрузке системы, с выполнением всех сопутствующих процедур (корректное завершение всех служб, отмонтирование файловых систем и т.д.). `reboot-force` действует более грубо — даже не пытаясь корректно остановить службы, оно просто убивает все их процессы, отмонтирует файловые системы и выполняет принудительную перезагрузку (в результате, перезагрузка происходит быстрее, чем обычно, но файловые системы остаются неповрежденными). И наконец, `reboot-immediate` даже не пытается отдать дань вежливости (убить процессы и отмонтировать файловые системы) — оно немедленно выполняет жесткую перезагрузку системы (это поведение практически аналогично срабатыванию аппаратного сторожевого таймера). Все перечисленные настройки подробно описаны на странице руководства [systemd.service\(5\)](#)⁵⁴.

⁵⁴Прим. перев.: Автор упускает из виду одну полезную опцию, непосредственно относящуюся к обсуждаемому вопросу — `OnFailure=`, задающую юнит, который будет активирован в случае сбоя исходного юнита. Таким образом можно обеспечить, например, запуск скриптов, отправляющих администратору уведомление о сбое в сочетании с дополнительной информацией (для сбора которой целесообразно задействовать команды `systemctl status` и `journalctl`). Кроме того, комбинирование данной настрой-

Таким образом, мы получаем гибкий механизм для настройки сторожевого контроля служб, их перезапуска при зависании, и реагирования в ситуации, когда перезапуск не помогает.

Рассмотрим применение этих настроек на простом примере:

```
[Unit]
Description=My Little Daemon
Documentation=man:mylittled(8)

[Service]
ExecStart=/usr/bin/mylittled
WatchdogSec=30s
Restart=on-failure
StartLimitInterval=5min
StartLimitBurst=4
StartLimitAction=reboot-force
```

Данная служба будет автоматически перезапущена, если она не передаст системному менеджеру очередной сигнал «все в порядке» в течение 30 секунд после предыдущего (кроме того, перезапуск будет произведен и в случае любого другого сбоя, например, завершения основного процесса службы с ненулевым кодом выхода). Если потребуется более 4 перезапусков службы за 5 минут — будет предпринято специальное действие, в данном случае, быстрая перезагрузка системы с корректным отмонтированием файловых систем.

Это все, что я хотел рассказать о сторожевых таймерах в `systemd`. Мы надеемся, что поддержки аппаратного отслеживания работоспособности процесса `init`, в сочетании с контролем функционирования отдельных служб, должно быть достаточно для решения большинства задач, связанных со сторожевыми таймерами. Разрабатываете ли вы встраиваемую либо мобильную систему, или работаете с высокодоступными серверами — вам определенно стоит попробовать наши решения!

(И да, если у вас возникнет вопрос, почему с аппаратным таймером должен работать именно `init`, и что мешает вынести эту логику в отдельный демон — пожалуйста, перечитайте эту главу еще раз, и попытайтесь увидеть всю цепочку сторожевого контроля как единое целое: аппаратный таймер надзирает за работой `systemd`, а тот, в свою очередь, следит за отдельными службами. Кроме того, мы полагаем, что отсутствие своевременного ответа от службы нужно рассматривать и обрабатывать так же, как и любые другие ее сбои. И наконец, взаимодействие с `/dev/watchdog` — одна из самых тривиальных задач в работе ОС (обычно, она сводится к простому вызову `ioctl()`), и для ее решения достаточно нескольких строк кода. С другой стороны, вынос данной функции в отдельный демон потребует организации сложного межпроцессного взаимодействия между этим демоном и процессом `init` — очевидно, реализация такой схемы предоставит значительно больший простор для ошибок, не говоря уже о повышенном потреблении ресурсов.)

Отметим, что встроенная в `systemd` поддержка аппаратного сторожевого таймера в конфигурации по умолчанию отключена, и поэтому никак не мешает работе с этим таймером из других программ. Вы без лишних проблем можете выбрать внешний сторожевой демон, если он лучше подходит для вашей задачи.

Да, и еще: если у вас возникнет вопрос, имеется ли в вашей системе аппаратный таймер — скорее всего да, если ваш компьютер не очень старый. Чтобы получить точный ответ, вы можете воспользоваться утилитой `wdctl`, включенной в последний релиз `util-linux`⁵⁵. Эта программа выведет всю необходимую информацию о вашем аппаратном сторожевом таймере.

И в завершение, я хотел бы поблагодарить ребят из [Penguintronix](#), которым принадлежит основная заслуга реализации сторожевого контроля в `systemd`.

ки с опцией `OnFailureIsolate=` позволяет при сбое юнита перевести систему в определенное состояние (например, остановить некоторые службы, перейти в режим восстановления, выполнить перезагрузку).

⁵⁵Прим. перев.: Утилита `wdctl` присутствует в `util-linux`, начиная с версии 2.22, которая, на момент написания этих строк, еще не вышла.

16 Запуск `getty` на последовательных (и не только) консолях

Если вам лень читать всю статью целиком: для запуска `getty` на последовательной консоли⁵⁶ достаточно указать в загрузчике параметр ядра `console=ttyS0`, и `systemd` автоматически запустит `getty` на этом терминале.

Физический последовательный порт **RS-232**, хотя уже и стал редкостью на современных настольных компьютерах, тем не менее, продолжает играть важную роль как на серверах, так и во встраиваемых системах. Он предоставляет простой и надежный доступ к управлению системой, даже когда сеть упала, а основной интерфейс управления завис. Кроме того, эмуляция последовательной консоли часто используется при управлении виртуальными машинами.

Разумеется, в Linux уже давно реализована поддержка работы с последовательными консолями но, при разработке `systemd`, мы постарались сделать работу с ними еще проще. В этой статье я хочу рассказать о том, как в `systemd` реализован запуск `getty` на терминалах различных типов.

Для начала, хотелось бы отметить один важный момент: в большинстве случаев, чтобы получить приглашение к логину на последовательном терминале, вам не нужно совершать никаких дополнительных действий: `systemd` сам проверит настройки ядра, определит их и использует ядром консоль, и автоматически запустит на ней `getty`. Таким образом, вам достаточно лишь правильно указать ядру соответствующую консоль (например, добавив к параметрам ядра в загрузчик `console=ttyS0`).

Тем не менее, для общего образования все же стоит рассмотреть некоторые тонкости запуска `getty` в `systemd`. Эта задача решается двумя шаблонами юнитов⁵⁷:

- `getty@.service` отвечает за **виртуальные консоли** (virtual terminals, VT, известные в системе под именами `/dev/tty1`, `/dev/tty2` и т.д.) — те, которые вы можете увидеть безо всякого дополнительного оборудования, просто переключившись на них из графического сеанса.
- `serial-getty@.service` обеспечивает поддержку всех прочих разновидностей терминалов, в том числе, подключаемых к последовательным портам (`/dev/ttyS0` и т.д.). Этот шаблон имеет ряд отличий от `getty@.service`, в частности, переменная `$TERM` в нем устанавливается в значение `vt102` (должно хорошо работать на большинстве физических терминалов), а не `linux` (которое работает правильно только на виртуальных консолях), а также пропущены настройки, касающиеся очистки буфера прокрутки (и поэтому имеющие смысл только на VT).

16.1 Виртуальные консоли

Рассмотрим механизм запуска `getty@.service`, обеспечивающий появление приглашений логина на виртуальных консолях (последовательные терминалы пока оставим в покое). По устоявшейся традиции, `init`-системы Linux обычно настраивались на запуск фиксированного числа экземпляров `getty`, как правило, шести (на первых шести виртуальных консолях, с `tty1` по `tty6`).

В `systemd` мы сделали этот процесс более динамичным: чтобы добиться большей скорости и эффективности, мы запускаем дополнительные экземпляры `getty` только при необходимости. Например, `getty@tty2.service` стартует только после того, как вы переключитесь на вторую виртуальную консоль. Отказавшись от обязательного запуска нескольких экземпляров `getty`, мы сэкономили немного системных ресурсов, а также

⁵⁶Прим. перев.: Здесь и в дальнейшем автор использует термин «serial console». Точный перевод этого выражения на русский язык звучит как «консоль, подключаемая к последовательному порту». Однако, для краткости изложения, при переводе используется не вполне корректный, но хорошо знакомый администраторам жаргонизм «последовательная консоль». Также отметим, что в данном документе термины «консоль» и «терминал» используются как синонимы.

⁵⁷Прим. перев.: Принципы работы с шаблонами и экземплярами служб изложены в главе 10. Для лучшего понимания нижеприведенного материала, рекомендуется перечитать эту главу, если вы успели ее подзабыть.

сделали загрузку системы чуть-чуть быстрее. При этом, с точки зрения пользователя, все осталось так же просто: как только он переключается на виртуальную консоль, на ней запускается `getty`, которая выводит приглашение к логину. Пользователь может и не подозревать о том, что до момента переключения приглашения не было. Тем не менее, если он войдет в систему и выполнит команду `ps`, он увидит, что `getty` запущены только на тех консолях, на которых он уже побывал.

По умолчанию, автоматический запуск `getty` производится на виртуальных консолях с первой по шестую (чтобы свести к минимуму отличия от привычной конфигурации)⁵⁸. Отметим, что автоматический запуск `getty` на конкретной консоли производится только при условии, что эта консоль не занята другой программой. В частности, при интенсивном использовании механизма быстрого переключения пользователей графические сеансы могут занять первые несколько консолей (чтобы такое поведение не заблокировало возможность запуска `getty`, мы предусмотрели специальную защиту, см. чуть ниже).

Две консоли играют особую роль: `tty1` и `tty6`. `tty1`, при загрузке в графическом режиме, используется для запуска дисплейного менеджера, а при загрузке в многопользовательском текстовом режиме, `systemd` принудительно запускает на ней экземпляр `getty`, не дожидаясь переключений⁵⁹.

Что касается `tty6`, то она используется исключительно для автоматического запуска `getty`, и недоступна другим подсистемам, в частности, графическому серверу⁶⁰. Мы сделали так специально, чтобы гарантировать возможность входа в систему в текстовом режиме, даже если графический сервер займет более пяти консолей.

16.2 Последовательные консоли

Работа с последовательными консолями (и всеми остальными видами не-виртуальных консолей) реализована несколько иначе, чем с VT. По умолчанию, `systemd` запускает один экземпляр `serial-getty@.service` на основной консоли ядра⁶¹ (если она не является виртуальной). Консолью ядра — это та консоль, на которую выводятся сообщения ядра. Обычно она настраивается в загрузчике, путем добавления к параметрам ядра аргумента наподобие `console=ttyS0`⁶². Таким образом, если пользователь перенаправил вывод ядра на последовательную консоль, то по завершении загрузки он увидит на этой консоли приглашение для логина⁶³. Кроме того, `systemd` выполняет поиск консолей, предоставляемых системами виртуализации, и запускает `serial-getty@.service` на первой из этих консолей (`/dev/hvc0`, `/dev/xvc0` или `/dev/hvsi0`). Такое поведение реализовано специальной программой-генератором — `systemd-getty-generator`. Генераторы запускаются в самом начале загрузки и автоматически настраивают различные службы в зависимости от соответствующих факторов.

В большинстве случаев, вышеописанного механизма автоматической настройки должно быть достаточно, чтобы получить приглашение логина там, где нужно — без каких-либо дополнительных настроек `systemd`. Тем не менее, иногда возникает необходимость в ручной настройке — например, когда необходимо запустить `getty` сразу на нескольких последовательных консолях, или когда вывод сообщений ядра направляется на один терминал, а управление производится с другого. Для решения таких задач

⁵⁸ Тем не менее, это поведение можно легко изменить, задавая параметр `NAutoVTs=` в файле `logind.conf`.

⁵⁹ В данном случае нет принципиальной разницы между принудительным запуском и запуском по запросу: первая консоль используется по умолчанию, так что запрос на ее активацию обязательно поступит.

⁶⁰ При необходимости, вы можете легко поменять номер резервируемой консоли (или отключить резервирование), используя параметр `ReserveVT=` в файле `logind.conf`.

⁶¹ Если для ядра настроен вывод в несколько консолей, *основной* считается та, которая идет *первой* в `/sys/class/tty/console/active`, т.е. указана *последней* в строке параметров ядра.

⁶² Подробнее об этой опции см. в файле `kernel-parameters.txt`.

⁶³ Отметим, что `getty`, а точнее, `agetty` на такой консоли вызывается с параметром `-s`, и поэтому не изменяет настроек символической скорости (baud rate) — сохраняется то значение, которое было указано в строке параметров ядра.

достаточно определить по экземпляру `serial-getty@.service` для каждого последовательного порта, на котором вы хотите запустить `getty`⁶⁴⁶⁵:

```
# systemctl enable serial-getty@ttyS2.service
# systemctl start serial-getty@ttyS2.service
```

После выполнения этих команд, `getty` будет принудительно запускаться для указанных последовательных портов при всех последующих загрузках.

В некоторых ситуациях может возникнуть необходимость в тонкой настройке параметров `getty` (например, заданная для вывода сообщений ядра символьная скорость непригодна для интерактивного сеанса). Тогда просто скопируйте штатный шаблон юнита в каталог `/etc/systemd/system` и отредактируйте полученную копию:

```
# cp /usr/lib/systemd/system/serial-getty@.service /etc/systemd/system/serial-getty@ttyS2.service
# vi /etc/systemd/system/serial-getty@ttyS2.service
... редактируем параметры запуска agetty ...
# ln -s /etc/systemd/system/serial-getty@ttyS2.service /etc/systemd/system/getty.target.wants/
# systemctl daemon-reload
# systemctl start serial-getty@ttyS2.service
```

В приведенном примере создает файл настроек, определяющий запуск `getty` на порту `ttyS2` (это определяется именем, под которым мы скопировали файл — `serial-getty@ttyS2.service`). Все изменения настроек, сделанные в данном файле, будут распространяться только на этот порт.

Собственно, это все, что я хотел рассказать о последовательных портах, виртуальных консолях и запуске `getty` на них. Надеюсь, рассказ получился интересным.

17 Работа с Journal

В свое время, я уже рассказывал о некоторых возможностях `journal` (см. главу 13), доступных из утилиты `systemctl`. Сейчас я попробую рассказать о `journal` более подробно, с упором на практическое применение его возможностей.

Если вы еще не в курсе, что такое `journal`: это компонент `systemd`, регистрирующий сообщения из системного журнала (`syslog`), сообщения ядра (`kernel log`) и `initrd`, а также сообщения, которые процессы служб выводят на `STDOUT` и `STDERR`. Полученная информация индексируется и предоставляется пользователю по запросу. `Journal` может работать одновременно с традиционным демоном `syslog` (например, `rsyslog` или `syslog-ng`), либо полностью его заменять. За подробностями стоит обратиться к [первому анонсу](#).

`Journal` был включен в `Fedora` начиная с `F17`. В `Fedora 18` `journal` вырос в мощный и удобный механизм работы с системным журналом. Однако, и в `F17`, и в `F18` `journal` по умолчанию сохраняет информацию только в небольшой кольцевой буфер в каталоге `/run/log/journal`. Как и все содержимое каталога `/run`, эта информация теряется при перезагрузке⁶⁶. Такой подход сильно ограничивает использование полезных возможностей `journal`, однако вполне достаточен для вывода актуальных сообщений от служб в `systemctl status`. Начиная с `Fedora 19`, мы собираемся включить сохранение логов на

⁶⁴Отметим, что `systemctl enable` для экземпляров служб работает только начиная с `systemd` версии 188 и старше (например, в `Fedora 18`). В более ранних версиях придется напрямую манипулировать символьными ссылками: `ln -s /usr/lib/systemd/system/serial-getty@.service /etc/systemd/system/getty.target.wants/serial-getty@ttyS2.service ; systemctl daemon-reload`.

⁶⁵Прим. перев.: На самом деле, работать с символьными ссылками пришлось бы даже в более свежих версиях `systemd` (до 209 включительно), так как в файле `serial-getty@.service` отсутствовала секция `[Install]`, в результате чего попытка выполнения `systemctl enable` для экземпляра соответствующей службы приводила к закономерной ошибке. Данная проблема была устранена только в `systemd 210`. В случае использования более старых версий, исправить ее можно и самостоятельно: достаточно скопировать указанный файл из `/usr/lib/systemd/system/` в `/etc/systemd/system/`, после чего дописать в него секцию `[Install]`, содержащую параметр `WantedBy=getty.target`, и затем выполнить `systemctl daemon-reload`.

⁶⁶Прим. перев.: Разумеется, это никак не относится к традиционному демону системного лога, даже если он работает поверх `journal`.

диск, в каталог `/var/log/journal`. При этом, логи смогут занимать гораздо больше места⁶⁷, а значит, смогут вместить больше полезной информации. Таким образом, `journal` станет еще более удобным.

17.1 Сохранение логов на диск

В F17 и F18 вы можете включить сохранение логов на диск вручную:

```
# mkdir -p /var/log/journal
```

После этого рекомендуется перезагрузить систему, чтобы заполнить журнал новыми записями.

Так как теперь у вас есть `journal`, `syslog` вам больше не нужен (кроме ситуаций, когда вам совершенно необходимо иметь `/var/log/messages` в текстовом виде), и вы спокойно можете удалить его:

```
# yum remove rsyslog
```

17.2 Основы

Итак, приступим. Нижеприведенный текст демонстрирует возможности `systemd 195`, входящего в `Fedora 18`⁶⁸, так что, если некоторые из описанных трюков не сработают в F17 — пожалуйста, дождитесь F18.

Начнем с основ. Доступ к логам `journal` осуществляется через утилиту `journalctl(1)`. Чтобы просто взглянуть на лог, достаточно ввести

```
# journalctl
```

Если вы выполнили эту команду с полномочиями `root`, вы увидите все журнальные сообщения, включая исходящие как от системных компонентов, так и от залогиненных пользователей⁶⁹. Вывод этой команды форматируется в стиле `/var/log/messages`, но при этом добавлены кое-какие улучшения:

- Строки с приоритетом `error` и выше подсвечены красным.
- Строки с приоритетом `notice` и `warning` выделены жирным шрифтом.
- Все отметки времени сформированы с учетом вашего часового пояса.
- Для навигации по тексту используется просмотрщик (`pager`), по умолчанию `less`⁷⁰.
- Выводятся *все* доступные данные, включая информацию из файлов, прошедших ротацию (`rotated logs`).
- Загрузка системы отмечается специальной строкой, отделяющей записи, сгенерированные между (пере)загрузками.

Отметим, что в данной статье не приводятся примеры такого вывода — прежде всего, для краткости изложения, но также и для того, чтобы дать вам повод поскорее попробовать `Fedora 18` с `systemd 195`. Надеюсь, отсутствие таких примеров не мешает вам уловить суть.

⁶⁷Прим. перев.: В `journal` отдельно задаются ограничения на размер для логов во временном хранилище (`/run`) и в постоянном (`/var`). При превышении лимита старые журналы удаляются. Так как `/run` размещается на `tmpfs`, т.е. в оперативной памяти, для временного хранения по умолчанию установлены более жесткие ограничения. При необходимости, соответствующие настройки можно задать в файле `journal.conf`.

⁶⁸Обновление со 195-й версией `systemd` на момент написания этих строк находится [на тестировании](#) и вскоре будет включено в состав `Fedora 18`.

⁶⁹Прим. перев.: А если вы выполнили эту команду от имени непривилегированного пользователя, не входящего в группу `adm`, и при этом не включили сохранение логов на диск, то вы не увидите ничего — без специальных полномочий пользователь может просматривать только собственный лог, а он по умолчанию ведется только если логи записываются на диск.

⁷⁰Прим. перев.: В инструментах `systemd`, включая `journalctl`, просмотрщик включается только при прямом выводе на экран, и отключается при перенаправлении вывода в файл или передаче его по каналу (`shell pipe`).

17.3 Контроль доступа

Итак, мы получили удобный и эффективный метод просмотра логов. Но для полного доступа к системным сообщениям требуются привилегии `root`, что не всегда удобно — в наше время администраторы предпочитают работать от имени непривилегированного пользователя, переключаясь на `root` только при крайней необходимости. По умолчанию, непривилегированные пользователи могут просматривать в `journal` только свои собственные логи (сообщения, сгенерированные их процессами). Чтобы предоставить пользователю доступ ко всем системным логам, нужно включить его в группу `adm`:

```
# usermod -a -G adm lennart
```

Разлогинившись, а затем вновь залогинившись под именем `lennart`⁷¹, я могу просматривать сообщения от всех пользователей и системных компонентов⁷²:

```
$ journalctl
```

17.4 Отслеживание логов в реальном времени

Когда вы запускаете программу `journalctl` без параметров, она выводит все сообщения, сгенерированные на текущий момент, и возвращает управление оболочке. Однако, иногда бывает полезно отслеживать их появление в режиме реального времени. В классической реализации `syslog` это осуществлялось командой `tail -f /var/log/messages`. В `journal` ее аналог выглядит так:

```
$ journalctl -f
```

И работает он точно так же: выводит последние десять сообщений, после чего переходит в режим ожидания, и выводит новые сообщения по мере их появления.

17.5 Простейшие методы выборки записей

При вызове `journalctl` без параметров, она выводит все сообщения, начиная с самого первого из сохраненных. Разумеется, это огромный объем информации. На практике иногда бывает достаточно ограничиться сообщениями, сгенерированными с момента последней загрузки системы⁷³:

```
$ journalctl -b
```

Но часто даже после такой фильтрации записей остается довольно много. Что ж, мы можем ограничиться только наиболее важными. Итак, все сообщения с приоритетом `err` и выше, начиная с момента последней загрузки:

```
$ journalctl -b -p err
```

Если вы уже успели перезагрузить систему после того, как произошли интересные вас события, целесообразнее будет воспользоваться выборкой по времени:

```
$ journalctl --since=yesterday
```

⁷¹Прим. перев.: Для того, чтобы обновить групповые полномочия в уже запущенных сеансах, можно воспользоваться командой `newgrp(1)`: `newgrp adm`.

⁷²Прим. перев.: Группа `adm` была выбрана на основании опыта дистрибутива Debian, в котором она устанавливается в качестве группы-владельца большинства лог-файлов. При этом авторы четко разделяют полномочия групп `adm` и `wheel`: если последняя используется для предоставления прав *изменять* что-либо в системе, то первая дает возможность лишь *просматривать* системную информацию. Начиная с версии `systemd` 198, группа-владелец файлов журнала изменена с `adm` на `systemd-journal`, однако штатный алгоритм установки `systemd` все равно выдает права на их чтение группам `adm` и `wheel` (через ACL).

⁷³Прим. перев.: Начиная с `systemd` версии 206, синтаксис опции «-b» был расширен: теперь она позволяет просматривать логи не только за текущую, но и за предыдущие загрузки, например, «-b -1» — прошлая загрузка, «-b -2» — позапрошлая и т.д.

В результате мы увидим все сообщения, зарегистрированные начиная со вчерашнего дня вплоть до настоящего момента. Прекрасно! Разумеется, этот критерий отбора можно комбинировать с другими, например, с `-p err`. Но, допустим, нам нужно узнать о чем-то, что случилось либо 15-го октября, либо 16-го:

```
$ journalctl --since=2012-10-15 --until="2011-10-16 23:59:59"
```

Отлично, мы нашли то, что искали. Но вот вам сообщают, что сегодня ранним утром наблюдались проблемы с CGI-скриптами Apache. Ладно, послушаем, что нам скажет индеец:

```
$ journalctl -u httpd --since=00:00 --until=9:30
```

Да, мы нашли это. Хм, похоже, что причиной стала проблема с диском `/dev/sdc`. Посмотрим, что с ним случилось:

```
$ journalctl /dev/sdc
```

Кошмар, ошибка ввода-вывода⁷⁴! Нужно срочно заменить диск, пока не начались более серьезные проблемы. Ладно, пойдем дальше. Что у нас там случилось с процессом `vpnc`?

```
$ journalctl /usr/sbin/vpnc
```

Хм, ничего подозрительного. Но, кажется, проблема где-то во взаимодействии между `vpnc` и `dhclient`. Посмотрим объединенный и отсортированный по времени список сообщений от этих процессов:

```
$ journalctl /usr/sbin/vpnc /usr/sbin/dhclient
```

Отлично, мы нашли причину проблемы!

17.6 Продвинутые методы выборки

Да, это все, конечно, здорово, но попробуем подняться еще на ступеньку выше. Чтобы понять описанные ниже приемы, нужно знать, что `systemd` добавляет к каждой лог-записи ряд скрытых метаданных. Эти метаданные по структуре напоминают набор переменных окружения, хотя на самом деле дают даже больше возможностей: во-первых, метаданные могут включать большие бинарные блоки данных (впрочем, это скорее исключение — обычно они содержат текст в кодировке UTF-8), и во-вторых, в пределах одной записи поле метаданных может содержать сразу несколько значений (и это тоже встречается нечасто — обычно поля содержат по одному значению). Эти метаданные автоматически собираются и добавляются для каждой лог-записи, безо всякого участия пользователя. И вы легко можете их использовать для более тонкой выборки записей. Посмотрим, как они выглядят:

```
$ journalctl -o verbose -n1
Tue, 2012-10-23 23:51:38 CEST [s=ac9e9c423355411d87bf0ba1a9b424e8;i=4301;b=5335e9cf5d954633bb99ae
    PRIORITY=6
    SYSLOG_FACILITY=3
    _MACHINE_ID=a91663387a90b89f185d4e860000001a
    _HOSTNAME=epsilon
    _TRANSPORT=syslog
    SYSLOG_IDENTIFIER=avahi-daemon
    _COMM=avahi-daemon
    _EXE=/usr/sbin/avahi-daemon
    _SYSTEMD_CGROUP=/system/avahi-daemon.service
    _SYSTEMD_UNIT=avahi-daemon.service
    _SELINUX_CONTEXT=system_u:system_r:avahi_t:s0
```

⁷⁴Ну ладно, признаюсь, здесь я немножко считерил. Индексирование сообщений ядра по блочным устройствам пока что не принято в алстрим, но Ганс [проделал огромную работу](#), чтобы реализовать эту функциональность, и я надеюсь, что к релизу F18 все будет.

```

_UID=70
_GID=70
_CMDLINE=avahi-daemon: registering [epsilon.local]
MESSAGE=Joining mDNS multicast group on interface wlan0.IPv4 with address 172.31.0.53.
_BOOT_ID=5335e9cf5d954633bb99aefc0ec38c25
_PID=27937
SYSLOG_PID=27937
_SOURCE_REALTIME_TIMESTAMP=1351029098747042

```

(Чтобы не утомлять вас огромным количеством текста, ограничимся одной записью. Ключ `-n` позволяет задать число выводимых записей, в нашем случае 1. Если указать его без параметра, будут показаны 10 последних записей.)

Задав параметр `-o verbose`, мы переключили формат вывода — теперь, вместо скупых строчек, копирующих `/var/log/messages`, для каждой записи выводится полный перечень всех метаданных. В том числе, информация о пользователе и группе, контент SELinux, идентификатор компьютера и т.д. Полный список общеизвестных полей метаданных приведен на соответствующей [странице руководства](#).

База данных Journal индексируется по *всем* этим полям! И мы можем использовать любое из них в качестве критерия выборки:

```
$ journalctl _UID=70
```

Как нетрудно догадаться, в результате будут выведены все сообщения от процессов пользователя с UID 70. При необходимости, критерии можно комбинировать:

```
$ journalctl _UID=70 _UID=71
```

Указание нескольких значений для одного и того же поля эквивалентно логическому ИЛИ. Таким образом, будут выведены записи как от процессов с UID 70, так и от процессов с UID 71.

```
$ journalctl _HOSTNAME=epsilon _COMM=avahi-daemon
```

А вот указание нескольких *различных* полей дает эффект логического И. В результате, будут выведены записи только от процесса `avahi-daemon`, работающего на хосте с именем `epsilon`.

Но мы этим не ограничимся! Мы же суровые компьютерщики, нам нужны сложные логические выражения!

```
$ journalctl _HOSTNAME=theta _UID=70 + _HOSTNAME=epsilon _COMM=avahi-daemon
```

При помощи плюса мы можем явно задать логическое ИЛИ, применяя его к разным полям и даже И-выражениям. Поэтому наш пример выведет как записи с хоста `theta` от процессов с UID 70, так и с хоста `epsilon` от процесса `avahi-daemon`⁷⁵.

17.7 И немного магии

Уже неплохо, правда? Но есть одна проблема — мы же не сможем запомнить все возможные значения всех полей журнала! Для этого была бы нужна очень хорошая память. Но `journalctl` вновь приходит к нам на помощь:

```
$ journalctl -F _SYSTEMD_UNIT
```

Эта команда выведет все значения поля `_SYSTEMD_UNIT`, зарегистрированные в базе данных журнала на текущий момент. То есть, имена всех юнитов `systemd`, которые писали что-либо в журнал. Аналогичный запрос работает для всех полей, так что найти точное значение для выборки по нему — уже не проблема. Но тут самое время сообщить вам,

⁷⁵Прим. перев.: Стоит отметить, что приоритет логических операций стандартный: сначала выполняются операции И, и только потом — операции ИЛИ. Используемая в `journalctl` система записи выражений аналогична принятой в классической алгебре: умножение (имеющее более высокий приоритет) не указывается знаком операции, а обозначается просто последовательной записью величин.

что эта функциональность встроена в механизм автодополнения оболочки⁷⁶! Это же просто прекрасно — вы можете просмотреть перечень значений поля и выбрать из него нужно прямо при вводе выражения. Возьмем для примера метки SELinux. Помнится, имя поля начиналось с букв SE...

```
$ journalctl _SE<TAB>
```

и оболочка сразу же дополнит:

```
$ journalctl _SELINUX_CONTEXT=
```

Отлично, и какое там значение нам нужно?

```
$ journalctl _SELINUX_CONTEXT=<TAB><TAB>
kernel                               system_u:system_r:rtkit_daemon_t:s0
system_u:system_r:accounts_d_t:s0    system_u:system_r:syslogd_t:s0
system_u:system_r:avahi_t:s0         system_u:system_r:system_cronjob_t:s0-s0:c0.c1023
system_u:system_r:bluetooth_t:s0     system_u:system_r:system_dbusd_t:s0-s0:c0.c1023
system_u:system_r:chkpwd_t:s0-s0:c0.c1023 system_u:system_r:systemd_logind_t:s0
system_u:system_r:chronyd_t:s0       system_u:system_r:systemd_tmpfiles_t:s0
system_u:system_r:cron_d_t:s0-s0:c0.c1023 system_u:system_r:udev_t:s0-s0:c0.c1023
system_u:system_r:devicekit_disk_t:s0 system_u:system_r:virt_d_t:s0-s0:c0.c1023 c0.c1023
system_u:system_r:dhcpc_t:s0         system_u:system_r:vpnc_t:s0 sd_t:s0-s0:c0.c1023
system_u:system_r:dnsmasq_t:s0-s0:c0.c1023 system_u:system_r:xdm_t:s0-s0:c0.c1023
system_u:system_r:init_t:s0          unconfined_u:system_r:rpm_t:s0-s0:c0.c1023
system_u:system_r:local_login_t:s0-s0:c0.c1023 unconfined_u:system_r:unconfined_t:s0-s0:c0.c1023
system_u:system_r:lvm_t:s0           unconfined_u:system_r:useradd_t:s0-s0:c0.c1023
system_u:system_r:modemmanager_t:s0-s0:c0.c1023 unconfined_u:unconfined_r:unconfined_dbusd_t:s0-s0:c0.c1023
system_u:system_r:NetworkManager_t:s0 unconfined_u:unconfined_r:unconfined_t:s0-s0:c0.c1023
system_u:system_r:policykit_t:s0
```

Ага, нас интересуют записи с меткой PolicyKit! Пользуясь дополнением, вводим:

```
$ journalctl _SELINUX_CONTEXT=system_u:system_r:policykit_t:s0
```

Очень просто, не правда ли! Пожалуй, самое простое из действий, связанных с SELinux ;-) Разумеется, такое дополнение работает для всех полей Journal.

На сегодня все. Впрочем, на странице руководства [journalctl\(1\)](#) можно узнать и о многих других возможностях, не описанных здесь. Например, о том, что `journalctl` может выводить данные в формате JSON, или в формате `/var/log/messages`, но с относительными метками времени, как в `dmesg`.

18 Управление ресурсами с помощью cgroups

Важную роль в современных компьютерных системах играют механизмы управления использованием ресурсов: когда вы запускаете на одной системе несколько программ, возникает необходимость распределять между ними ресурсы системы, в соответствии с некоторыми правилами. В частности, это особенно актуально на маломощных встраиваемых и мобильных системах, обладающих очень скудными ресурсами. Но та же задача актуальна и для очень мощных вычислительных кластеров, которые располагают огромными ресурсами, но при этом несут и огромную вычислительную нагрузку.

Исторически, в Linux поддерживался только одна схема управления ресурсами: все процессы получают примерно равные доли процессорного времени или потока ввода-вывода. При необходимости соотношение этих долей можно изменить при помощи значения *nice*, задаваемого для каждого процесса. Такой подход очень прост, и на протяжении долгих лет покрывал все нужды пользователей Linux. Но у него есть существенный недостаток: он оперирует лишь отдельными процессами, но не их группами. В результате, например, веб-сервер Apache с множеством CGI-процессов при прочих

⁷⁶Прим. перев.: В исходной статье речь идет только о `bash`, однако, начиная с релиза `systemd 196`, аналогичная функциональность доступна и для `zsh`.

равных получает гораздо больше ресурсов, чем служба `syslog`, у которой не так много процессов.

В процессе проектирования архитектуры `systemd`, мы практически сразу поняли, что управление ресурсами должно быть одной из базовых функций, заложенных в основы его структуры. В современной системе — серверной или встраиваемой — контроль использования процессора, памяти и ввода-вывода для различных служб нельзя добавлять задним числом. Такая функциональность должна быть доступна изначально, через базовые настройки запуска служб. При этом, ресурсы должны распределяться на уровне служб, а не процессов, как это делалось при помощи значений `nice` или [POSIX Resource Limits](#).

В этой статье я попробую рассказать о методах управления механизмами распределения ресурсов между службами `systemd`. Эта функциональность присутствует в `systemd` уже долгое время, и давно пора рассказать о ней пользователям и администраторам⁷⁷.

В свое время я [пояснял](#)⁷⁸, что контрольные группы Linux (`cgroups`) могут работать и как механизм группировки и отслеживания процессов, и как инструмент управления использованием ресурсов. Для функционирования `systemd` необходим только первый из этих режимов, а второй опционален. И именно этот опциональный второй режим дает вам возможность распределять ресурсы между службами. (А сейчас очень рекомендую вам, прежде чем продолжать чтение этой статьи, ознакомиться с [базовой информацией о cgroups](#). Хотя дальнейшие рассуждения и не будут затрагивать низкоуровневые аспекты, все же будет лучше, если у вас сформируется некоторое представление о них.)

Основными контроллерами `cgroups`, отвечающими за управление ресурсами, являются `cpu`, `memory` и `blkio`. Для их использования необходимо, чтобы они были включены на этапе сборки ядра; большинство дистрибутивов (в том числе и Fedora), включают их в штатных ядрах. `systemd` предоставляет ряд высокоуровневых настроек, позволяющих использовать эти контроллеры, не вникая в технические детали их работы.

18.1 Процессор

Если в ядре включен контроллер `cpu`, `systemd` по умолчанию создает контрольную группу по этому ресурсу для каждой службы. Даже без каких-либо дополнительных настроек это дает положительный эффект: на системе под управлением `systemd` все службы получают равные доли процессорного времени, независимо от количества процессов, запущенных в рамках службы. Например, на вашем веб-сервере MySQL с несколькими рабочими процессами получит такую же долю процессорного времени, как и Apache, даже если тот запустил 1000 CGI-процессов. Разумеется, такое поведение

⁷⁷Прим. перев.: Дальнейшее содержимое этой главы устарело и сохраняется лишь из исторических соображений (ну и для тех, кто администрирует системы со старыми версиями ядра и `systemd`). Практически все описанные в ней настройки объявлены устаревшими в ходе миграции с `cgroup v1` на `cgroup v2`, произведенной в версиях 230–232. В частности, директива `CPUShares=` заменена на `CPUWeight=`, `MemoryLimit=` — на `MemoryMax=`, `BlockIOWeight=` — на `IOWeight=`/`IODeviceWeight=`, `BlockIOReadBandwidth=` — на `IOReadBandwidthMax=` и т.д. Смысл остается тем же, поменялись лишь некоторые тонкости, связанные с работой `cgroups`. Например, допустимые значения для `CPUShares` лежали в диапазоне от 2 до 262144, а для `BlockIOWeight` — от 10 до 1000, в то время как новые «весовые» настройки имеют стандартный диапазон 1–10000. Также, добавлены новые настройки, например, `CPUQuota=` для деления процессора по долям, `MemorySwapMax=` для ограничения использования подкачки. В любом случае, лучше внимательно ознакомиться с актуальной версией страницы руководства [systemd.resource-control\(5\)](#). Кстати, используемый в статье механизм директив `.include` тоже устарел — вместо них нужно использовать фрагменты конфигов (см. прим. 79).

⁷⁸Прим. перев.: В указанном документе автор рассказывает, что контрольные группы Linux состоят из двух сущностей: **(А)** механизма иерархической группировки и маркировки процессов, и **(В)** механизма, позволяющего распределять ресурсы между полученными группами. Для работы **(В)** необходимо **(А)**, но не наоборот — **(А)** может прекрасно работать без **(В)**. Для нормально функционирования `systemd` **(А)** *необходим*, а **(В)** опционален (он лишь обеспечивает работу некоторых настроек). Вы можете собрать ядро только с необходимой для **(А)** опцией `CONFIG_CGROUPS=y`, отключив все связанные с **(В)** опции (такие как `CONFIG_CGROUP_FREEZER=y`, `CONFIG_CGROUP_DEVICE=y`, `CONFIG_CGROUP_CPUACCT=y`, `CONFIG_CGROUP_MEM_RES_CTLR=y`, `CONFIG_CGROUP_MEM_RES_CTLR_SWAP=y`, `CONFIG_CGROUP_MEM_RES_CTLR_KMEM=y`, `CONFIG_CGROUP_PERF=y`, `CONFIG_CGROUP_SCHED=y`, `CONFIG_BLK_CGROUP=y`, `CONFIG_NET_CLS_CGROUP=y`, `CONFIG_NET_PRIO_CGROUP=y`), и `systemd` будет нормально работать на такой системе (за исключением того, что связанные с этими контроллерами настройки не будут срабатывать). Однако, если собрать ядро без `CONFIG_CGROUPS=y`, функциональность `systemd` будет сильно ограничена. При этом, автор особо подчеркивает, что все негативные эффекты влияния контрольных групп на производительность обусловлены именно **(В)**, в то время как **(А)** на производительность практически не влияет.

при необходимости можно легко отключить — см. опцию `DefaultControllers=` в файле `/etc/systemd/system.conf`.

Если *равномерное* распределение процессорного времени между службами вас не устраивает, и вы хотите выделить определенным службам больше или меньше времени — используйте опцию `CPUShares=` в конфигурационном файле службы. По умолчанию это значение равно 1024. Увеличивая это число, вы дадите службе больше процессорного времени, уменьшая — соответственно, меньше.

Рассмотрим небольшой практический пример. Допустим, вам нужно увеличить для службы Apache относительную долю потребления процессора до 1500. Для этого создаем файл «ручных» настроек `/etc/systemd/system/httpd.service`, который включает в себя все те же опции, что и файл настроек по умолчанию `/usr/lib/systemd/system/httpd.service`, отличаясь от него только значением `CPUShares=`:

```
.include /usr/lib/systemd/system/httpd.service
[Service]
CPUShares=1500
```

Первая строка обеспечивает включение в нашу конфигурацию файла с настройками по умолчанию, сделанными разработчиками Apache или его сопровождающими в вашем дистрибутиве (если это включение не указать явно, данный файл будет проигнорирован). Далее, мы указываем тот параметр, который хотим изменить⁷⁹. Сохраняем файл, приказываем `systemd` перечитать конфигурацию, и перезапускаем Apache, чтобы настройки вступили в силу⁸⁰:

```
systemctl daemon-reload
systemctl restart httpd.service
```

Готово!

Обратите внимание, что явное указание значения `CPUShares=` в конфигурации службы заставит `systemd` создать для нее контрольную группу в иерархии контроллера `cpu`, даже если этот контроллер не указан в `DefaultControllers=` (см. выше).

18.2 Отслеживание использования ресурсов

Для того, чтобы правильно распределять ресурсы между службами, неплохо бы знать реальные потребности этих служб. Чтобы упростить для вас отслеживание потребления ресурсов службами, мы подготовили утилиту `systemd-cgtop`, которая находит все имеющиеся в системе контрольные группы, определяет для каждой из них количество потребляемых ресурсов (процессорное время, память и ввод-вывод) и выводит эти данные в динамически обновляемой сводной таблице, по аналогии с программой `top`. Используя вводимое `systemd` распределение служб по контрольным группам, эта утилита выводит для служб те же сведения, которые `top` выводит для отдельных процессов.

⁷⁹Прим. перев.: В новых версиях `systemd`, начиная со 198, можно поступить проще: вместо того, чтобы создавать свой юнит-файл в каталоге `/etc` и включать в него содержимое штатного (из `/usr`), достаточно просто создать каталог `/etc/systemd/system/httpd.service.d/` и поместить в него файл `my_resource_limit.conf` (суффикс `.conf` обязателен), в котором указываются только те настройки, которые необходимо изменить (последние две строки вышеприведенного листинга). Это никак не отменяет основного конфигурационного файла юнита (который может находиться в `/usr` или `/etc`), однако настройки из `.d`-каталогов имеют более высокий приоритет и могут перекрывать настройки основного файла. Все вышесказанное относится и к другим примерам из этого раздела.

⁸⁰Прим. перев.: `systemd` версий до 197 (включительно) не имел штатного механизма для изменения параметров контрольных групп «на лету» (без перезапуска службы). Если у вас все же возникнет такая необходимость, вы можете узнать контрольную группу службы командой `systemctl show -p ControlGroup имя_службы.service`, и выполнить требуемые настройки напрямую через запись значений в псевдофайлы `sgroupfs` (разумеется, при следующем запуске службы к ней будут применены параметры, указанные в конфигурационном файле). В версиях `systemd` со 198 по 204, программа `systemctl` поддерживала команды `set-cgroup-attr`, `unset-cgroup-attr` и `get-cgroup-attr`, позволявшие манипулировать настройками контрольных групп. Начиная с `systemd` 205, они были заменены командой `set-property`, работающей с параметрами конфигурации юнитов. Установленные через `systemctl` настройки сохраняются во вспомогательных конфигурационных файлах, которые размещаются в `.d`-каталогах (см. примечание выше).

К сожалению, по умолчанию `cgtop` может отдельно отслеживать для каждой службы только потребление процессорного времени, а сведения по использованию памяти и ввода-вывода доступны только для всей системы в целом. Это ограничение возникает из-за того, что в конфигурации по умолчанию контрольные группы для служб создаются только в иерархии контроллера `cpu`, но не `memory` и `blkio`. Без создания групп в иерархии этих контроллеров невозможно отследить использование ресурса по службам. Самый простой способ обойти это ограничение — приказать `systemd` создавать соответствующие группы, добавив `memory` и `blkio` в перечень `DefaultControllers=` в файле `system.conf`.

18.3 Память

Используя опции `MemoryLimit=` и `MemorySoftLimit=`, вы можете ограничивать суммарное потребление оперативной памяти всеми процессами службы. В них указывается предел потребления памяти в байтах⁸¹. При этом поддерживаются суффиксы К, М, G и Т, обозначающие соответственно, килобайт, мегабайт, гигабайт и терабайт (по основанию 1024).

```
.include /usr/lib/systemd/system/httpd.service

[Service]
MemoryLimit=1G
```

По аналогии с `CPUShares=`, явное указание этих опций заставит `systemd` создать для службы контрольную группу в иерархии контроллера `memory`, даже если он не был указан в `DefaultControllers=`.

18.4 Ввод-вывод

Для контроля пропускной полосы ввода-вывода с блочных устройств, доступно несколько настроек. Первая из них — `BlockIOWeight=`, задающая *долю* полосы ввода-вывода для указанной службы. Принцип похож на `CPUShares=` (см. выше), однако здесь величина относительной доли ограничена значениями от 10 до 1000. По умолчанию, она равна 1000. Уменьшить долю для службы Apache можно так:

```
.include /usr/lib/systemd/system/httpd.service

[Service]
BlockIOWeight=500
```

При необходимости, вы можете задать такое значение отдельно для каждого устройства:

```
.include /usr/lib/systemd/system/httpd.service

[Service]
BlockIOWeight=/dev/disk/by-id/ata-SAMSUNG_MMCRE28G8MXP-0VBL1_DC06K01009SE009B5252 750
```

При этом, точное название устройства знать не обязательно — достаточно указать интересующий вас каталог:

```
.include /usr/lib/systemd/system/httpd.service

[Service]
BlockIOWeight=/home/lennart 750
```

⁸¹Прим. перев.: Разница между `MemorySoftLimit=` и `MemoryLimit=` состоит в том, что первый предел можно превышать, если в системе еще есть достаточное количество свободной памяти. Второй из этих пределов превышать нельзя, независимо от наличия свободной памяти. Подробнее см. раздел «Soft limits» в [файле документации](#). Отметим, что из-за планируемого прекращения поддержки параметра `memory.soft_limit_in_bytes` на уровне ядра, опция `MemorySoftLimit=` была [удалена](#) из `systemd`, начиная с версии 208.

Если заданный вами путь не указывает на файл устройства, `systemd` автоматически определит, на каком устройстве расположен указанный файл/каталог, и выставит для этого устройства соответствующую настройку.

Вы можете добавить несколько таких строк, задавая долю пропускной полосы отдельно для различных устройств, и при этом также допускается указать «общее» значение (как в первом примере), которое будет использовано для всех остальных устройств.

В качестве альтернативы относительной доле пропускной полосы, вы также можете ограничивать абсолютную долю, используя настройки `BlockIOReadBandwidth=` и `BlockIOWriteBandwidth=`. В них нужно указать устройство или любой находящийся на нем файл/каталог, а также предельную скорость чтения/записи в байтах в секунду:

```
.include /usr/lib/systemd/system/httpd.service

[Service]
BlockIOReadBandwidth=/var/log 5M
```

В результате, для данной службы скорость чтения с устройства, содержащего каталог `/var/log`, будет ограничена величиной 5 мегабайт в секунду.

По аналогии с вышеописанными `CPUShares=` и `MemoryLimit=`, явное указание любой из приведенных настроек пропускной полосы заставит `systemd` создать для службы контрольную группу в иерархии контроллера `blkio`.

18.5 Прочие параметры

Вышеописанные опции покрывают лишь малую толику настроек, поддерживаемых различными контроллерами Linux `sgroups`. Мы добавили высокоуровневый интерфейс только к тем настройкам, которые кажутся нам наиболее важным для большинства пользователей. Из соображений удобства мы добавили механизмы, обеспечивающие поддержку крупных единиц измерения (килобайты, мегабайты и т.д.) и автоматическое определение блочных устройств по указанному файлу/каталогу.

В некоторых случаях описанных высокоуровневых настроек может оказаться недостаточно — допустим, вам нужно задать низкоуровневую настройку `sgroups`, для которой мы (пока) не добавили высокоуровневого аналога. На этот случай мы предусмотрели универсальный механизм задания таких опций в конфигурационных файлах юнитов⁸². Рассмотрим, например, задание для службы параметра `swappiness` (относительная интенсивность использования подкачки для процессов службы). В `systemd` нет высокоуровневой настройки для этого значения. Однако вы можете задать его, используя низкоуровневую настройку `ControlGroupAttribute=`:

```
.include /usr/lib/systemd/system/httpd.service

[Service]
ControlGroupAttribute=memory.swappiness 70
```

Как обычно, явное указание настройки, относящейся к какому-либо контроллеру (в нашем случае `memory`) приведет к автоматическому созданию группы в иерархии данного контроллера.

В дальнейшем, возможно, мы расширим возможности высокоуровневой настройки различных параметров контрольных групп. Если вы часто пользуетесь какими-то из

⁸²Прим. перев.: Описываемый здесь механизм управления низкоуровневыми настройками контрольных групп (`ControlGroupAttribute=`) присутствовал в `systemd` до 204 версии включительно. Начиная с версии 205, он был удален по требованию разработчиков `sgroups`. Также были удалены возможности напрямую задавать для юнита контрольную группу (`ControlGroup=`) и устанавливать права доступа для управления группами (`ControlGroupModify=`). Взамен, разработчики `systemd` планируют значительно расширить поддержку высокоуровневых настроек для параметров `sgroups` — сразу после того, как разработчики `sgroups` определятся, какие именно параметры они будут поддерживать. Наличие некоторого хаоса в этой области обусловлено происходящей сейчас полной переработкой реализации контрольных групп на уровне ядра и механизмов работы с ними из пространства пользователя. Техническая сторона данного вопроса раскрыта в статье [The New Control Group Interfaces](#) (пока не переведена).

них и полагаете, что для них можно добавить соответствующие опции — не стесняйтесь обращаться к нам. А лучше всего — присылайте сразу патч!

Предупреждение: Обратите внимание, что использование некоторых контроллеров может сильно сказаться на производительности системы. Это та цена, которую приходится платить за контроль над ресурсами. Использование таких контроллеров может ощутимо замедлить некоторые операции. В частности, весьма нежелательная в этом плане репутация закрепилась за контроллером `memory` (хотя, не исключено, что эта проблема уже исправлена в свежих выпусках ядра).

Для углубленного изучения темы, затронутой в этой статье, вы можете обратиться к документации по [поддерживаемым настройкам юнитов](#), а также по контроллерам [cpu](#), [memory](#) и [blkio](#).

Стоит подчеркнуть, что мы сейчас обсуждали распределение ресурсов *между службами*. В дополнение к этим современным механизмам, `systemd` также поддерживает и традиционные настройки, касающиеся распределения ресурсов *между отдельными процессами*. Хотя такие настройки обычно наследуются порожденными процессами, они, тем не менее, все равно ограничивают ресурсы на уровне отдельных процессов. В частности, к ним относятся `IOSchedulingClass=`, `IOSchedulingPriority=`, `CPUSchedulingPolicy=`, `CPUSchedulingPriority=`, `CPUAffinity=`, `LimitCPU=` и т.п. Для их работы не требуют контроллеры `srgroups`, и они не так сильно ухудшают производительность. Возможно, мы рассмотрим их в последующих статьях.

19 Проверка на виртуальность

Еще в начале разработки `systemd`, мы внимательно изучали существовавшие на тот момент `init`-скрипты, выделяя наиболее типичные для них операции. Среди прочих, в составленный нами список попала и такая функция, как определение виртуализации: некоторые скрипты проверяли, запускаются они в виртуальном окружении (например, KVM, VMWare, LXC и т.д.) или на полноценной, физической системе. Часть этих скриптов отказывалась работать на виртуальных системах (например, службы управления устройствами совершенно излишни в виртуальных контейнерах, не имеющих доступа к устройствам), другие же, наоборот, запускались только в определенных виртуальных окружениях (например, всевозможные «`guest additions`», рекомендуемые к запуску на гостевых системах VMWare и VirtualBox). По-хорошему, в некоторых ситуациях было бы более правильно проверять некоторые другие условия, а не пытаться явно определить наличие виртуализации. Тем не менее, всесторонне изучив вопрос, мы пришли к выводу, что во многих случаях возможность явной проверки такого условия при запуске служб была бы очень кстати. В результате, мы добавили поддержку соответствующей опции настройки юнитов — `ConditionVirtualization`; кроме того, мы создали небольшую утилиту, которую можно вызывать из скриптов — `systemd-detect-virt(1)`; и наконец, мы предоставили простой интерфейс для шины D-Bus, позволяющий получить информацию о виртуализации даже непривилегированным программам.

Определить, запущен код на виртуальной системе, или на физической, на самом деле **не так уж и сложно**. В зависимости от того, какие именно механизмы виртуализации вы хотите определить, основная работа сводится к выполнению инструкции `CPUID` и, возможно, проверке некоторых файлов в `/sys` и `/proc`. Основная трудность здесь — точно знать строки, которые нужно искать. Список таких строк необходимо поддерживать в актуальном состоянии. В настоящий момент, `systemd` определяет следующие механизмы виртуализации:

- Полная виртуализация (т.е. виртуальные машины):
 - `qemu`
 - `kvm`
 - `vmware`
 - `microsoft`

- oracle
- xen
- bochs
- Виртуализация на уровне ОС (т.е. контейнеры):
 - chroot
 - openvz
 - lxc
 - lxc-libvirt
 - **systemd-nspawn**

Рассмотрим, как можно использовать эту функциональность.

19.1 Условия на запуск юнитов

При помощи опции `ConditionVirtualization`, добавленной в секцию `[Unit]` файла конфигурации юнита, вы можете обеспечить запуск (или наоборот, отмену запуска) данного юнита в зависимости от того, работает ли он на виртуальной системе, или нет. В случае утвердительного ответа, также можно уточнить, какая система виртуализации при этом используется. Например:

```
[Unit]
Name=My Foobar Service (runs only only on guests)
ConditionVirtualization=yes

[Service]
ExecStart=/usr/bin/foobard
```

Помимо «yes» или «no», вы также можете указать идентификатор конкретной системы виртуализации (согласно списку выше, например, «kvm», «vmware» и т.д.), либо «container» или «vm» (что позволит отличить виртуализацию на уровне ОС от полной виртуализации). Кроме того, вы можете добавить перед значением восклицательный знак, и результат проверки будет инвертирован (юнит запустится только в том случае, если указанная технология *не* используется). Подробности вы можете узнать на [странице руководства](#).

19.2 В скриптах

В скриптах оболочки вы можете выполнить аналогичные проверки при помощи утилиты `systemd-detect-virt(1)`. Например:

```
if systemd-detect-virt -q ; then
    echo "Virtualization is used:" `systemd-detect-virt`
else
    echo "No virtualization is used."
fi
```

Эта утилита возвращает код 0 (успех), обнаружив виртуализацию, или ненулевое значение, если виртуализация не выявлена. Также она выводит идентификатор обнаруженной системы виртуализации (согласно списку выше), если это не было запрещено опцией `-q`. Кроме того, опции `-s` и `-v` позволяют ограничить проверки только механизмами виртуализации на уровне ОС, либо полной виртуализации, соответственно. Подробности см. на [странице руководства](#).

19.3 В программах

Информация о виртуализации также представлена на системной шине:

```
$ gdbus call --system --dest org.freedesktop.systemd1 --object-path /org/freedesktop/systemd1 \
> --method org.freedesktop.DBus.Properties.Get org.freedesktop.systemd1.Manager Virtualization
(<'systemd-nsspawn'>,< >)
```

Если виртуализация не выявлена, это свойство содержит пустую строку. Обратите внимание, что некоторые контейнерные системы не могут быть обнаружены напрямую из непривилегированного кода. Именно поэтому мы не стали создавать библиотеку, а воспользовались шиной D-Bus, которая позволяет корректно решить проблему привилегий.

Стоит отметить, что все эти инструменты определяют только «самый внутренний» из задействованных механизмов виртуализации. Если вы используете несколько систем, вложенных друг в друга, вышеописанные инструменты обнаружат только ту, в которой они непосредственно запущены. В частности, если они работают в контейнере, находящемся внутри виртуальной машины, они увидят только контейнер.

20 Сокет-активация служб и контейнеров

Сокет-активация — это одна из наиболее интересных возможностей systemd. Еще в [первом анонсе](#) нашего проекта мы рассказывали о том, как этот механизм улучшает параллелизацию и отказоустойчивость использующих сокеты служб, а также упрощает формирование зависимостей между службами при загрузке. В данной статье я продолжу рассказ о его возможностях — на этот раз речь пойдет об увеличении числа служб и контейнеров, работающих на одной и той же системе, без повышения потребления ресурсов. Другими словами: как можно увеличить количество клиентских сайтов на хостинговых серверах без затрат на новое оборудование.

20.1 Сокет-активация сетевых служб

Начнем с небольшого отступления. Итак, что же такое сокет-активация, и как она работает? На самом деле все довольно просто. systemd создает «слушающие» сокеты (не обязательно IP) от имени вашей службы (которая пока не запущена) и ожидает входящие соединения. Как только в сокет поступает первый запрос, systemd запускает службу и передает ей полученные данные. После обработки запроса, в зависимости от реализации службы, она может продолжать работу, ожидая новых соединений, или завершиться, переложив эту задачу обратно на systemd (который вновь запустит ее при поступлении очередного запроса). При этом, со стороны клиента невозможно отличить, когда служба запущена, а когда — нет. Сокет постоянно остается открытым для входящих соединений, и все они обрабатываются быстро и корректно.

Такая конфигурация позволяет снизить потребление системных ресурсов: службы работают и потребляют ресурсы только тогда, когда это действительно необходимо. Многие интернет-сайты и службы могут использовать это с выгодой для себя. Например, хостеры веб-сайтов знают, что из огромного количества существующих в Интернете сайтов лишь малая часть получает непрерывный поток запросов. Большинство же сайтов, хотя и должны постоянно оставаться доступными, получают запросы очень редко. Используя сокет-активацию, вы можете воспользоваться этим: разместив множество таких сайтов на одной системе и активируя их службы только при необходимости, вы получаете возможность «оверкоммита»: ваша система будет обслуживать сайтов больше, чем формально позволяют ее ресурсы⁸³. Разумеется, увлекаться оверкоммитом не стоит, иначе в моменты пиковой нагрузки ресурсов может действительно не хватить.

⁸³Прим. перев.: Стоит отметить, что подобная схема работает только при условии, что для каждого клиентского сайта запускаются отдельные процессы служб, хотя это и происходит в рамках одного хоста. Не очень распространенный в отечественном хостинге вариант: обычно следующей опцией после shared-хостинга (одна служба на всех клиентов) идет VPS (каждому клиенту по виртуальному хосту).

С помощью `systemd` вы без труда можете организовать такую схему. Множество современных сетевых служб уже поддерживают сокет-активацию «из коробки» (а в те, которые пока не поддерживают, ее **не так уж и сложно** добавить⁸⁴). Реализованный в `systemd` механизм управления **экземплярами служб** позволяет подготовить универсальные шаблоны конфигурации служб, и в соответствии с ними для каждого сайта будет запускаться свой экземпляр службы. Кроме того, не стоит забывать, что `systemd` предоставляет **внушительный арсенал** механизмов обеспечения безопасности и разграничения доступа, который позволит изолировать клиентские сайты друг от друга (например, службы каждого клиента будут видеть только его собственный домашний каталог, в то время как каталоги всех остальных пользователей будут им недоступны). Итак, в конечном итоге вы получаете надежную и масштабируемую серверную систему, на сравнительно небольших ресурсах которой функционирует множество безопасно изолированных друг от друга служб — и все это реализовано штатными возможностями вашей ОС⁸⁵.

Подобные конфигурации уже используются на рабочих серверах ряда компаний. В частности, специалисты из [Panttheon](#) используют такую схему для обслуживания масштабируемой инфраструктуры множества сайтов на базе Drupal. (Стоит упомянуть, что заслуга ее внедрения в компании Panttheon принадлежит Дэвиду Штрауссу. Дэвид, ты крут!)

20.2 Сокет-активация контейнеров

Все вышеописанные технологии уже реализованы в ранее вышедших версиях `systemd`. Если ваш дистрибутив поддерживает `systemd`, вы можете воспользоваться этими механизмами прямо сейчас. А теперь сделаем шаг вперед: начиная с `systemd 197` (и, соответственно, с Fedora 19), мы добавили поддержку сокет-активации *виртуальных контейнеров* с полноценными ОС внутри. И я считаю это действительно важным достижением.

Сокет-активация контейнеров работает следующим образом. Изначально, `systemd` хост-системы слушает порты от имени контейнера (например, порт SSH, порт веб-сервера и порт сервера СУБД). При поступлении на любой из этих портов входящего запроса, `systemd` запускает контейнер и передает ему все его сокеты. Внутри контейнера, еще один `systemd` (init гостевой системы) принимает эти сокеты, после чего вступает в работу вышеописанная схема обычной сокет-активации служб. При этом, SSH-сервер, веб-сервер и СУБД-сервер «видят» только ОС контейнера — хотя они были активированы сокетами, созданными на хосте! Для клиента все эти тонкости скрыты. Таким образом, мы получаем виртуальный контейнер с собственной ОС, активируемый при поступлении входящего сетевого соединения, причем совершенно прозрачно для клиента⁸⁶.

Внутри контейнера функционирует полноценная ОС, причем ее дистрибутив не обязательно совпадает с дистрибутивом хост-системы. Например, вы можете установить на хосте Fedora, и запускать на нем несколько контейнеров с Debian. Контейнеры имеют собственные `init`-системы, собственные SSH-серверы, собственные списки процессов и т.д., но при этом пользуются некоторыми механизмами ОС хоста (например, управлением памятью).

К настоящему моменту сокет-активация контейнеров поддерживается лишь встроенным в `systemd` простейшим контейнерным менеджером — `systemd-nspawn`. Мы надеемся, что соответствующая возможность вскоре появится и в `libvirt-lxc`⁸⁷. А пока,

⁸⁴Прим. перев.: Начиная с версии 209, в состав `systemd` входит утилита `systemd-socket-proxyd`, позволяющая использовать сокет-активацию даже для служб, которые ее не поддерживают.

⁸⁵Прим. перев.: В качестве практического примера использования сокет-активации служб `systemd` в промышленных серверных платформах, можно предложить [вот эту статью](#), наглядно описывающую применение этой и других технологий (мониторинг, использование Journal, ограничение ресурсов и доступа) на примере Node.js.

⁸⁶Кстати говоря, [это еще один аргумент](#) в пользу важности быстрой загрузки для серверных систем.

⁸⁷Прим. перев.: Патчи, реализующие эту возможность, были подготовлены и приняты разработчиками `libvirt` во второй декаде июля 2013 г., и входят в `libvirt` начиная с версии 1.1.1.

за отсутствием альтернатив, рассмотрим использование этого механизма на примере `systemd-nspawn`.

Начнем с установки файлов ОС контейнера в выбранный каталог. Детальное рассмотрение этого вопроса выходит далеко за рамки нашего обсуждения, и при том детально рассмотрено во многих статьях и руководствах. Поэтому ограничусь лишь несколькими наиболее важными замечаниями. В частности, команда для установки Fedora будет выглядеть следующим образом:

```
$ yum --releasever=19 --nogpg --installroot=/srv/mycontainer/ --disablerepo='*' \  
> --enablerepo=fedora install systemd passwd yum fedora-release vim-minimal
```

а для Debian —

```
$ debootstrap --arch=amd64 unstable /srv/mycontainer/
```

Также см. последние абзацы страницы руководства [systemd-nspawn\(1\)](#). Стоит отметить, что в настоящее время реализация системного аудита в Linux не поддерживает виртуальные контейнеры, и ее включение в ядре хоста вызовет множество ошибок при попытке запустить контейнер. Отключить ее можно добавлением `audit=0` в строку параметров загрузки ядра хоста.

Разумеется, внутри контейнера должен быть установлен `systemd` версии не ниже 197. Установить его и произвести другие необходимые настройки можно при помощи того же `systemd-nspawn` (пока что в режиме аналога `chroot`, т.е. без параметра `-b`). После этого можно попробовать загрузить ОС контейнера, используя `systemd-nspawn` уже с параметром `-b`.

Итак, ваш контейнер нормально загружается и работает. Подготовим для него `service`-файл, при помощи которого `systemd` сможет запускать и останавливать виртуальное окружение. Для этого, создадим на хост-системе файл `/etc/systemd/system/mycontainer.service` со следующим содержанием:

```
[Unit]  
Description=My little container  
  
[Service]  
ExecStart=/usr/bin/systemd-nspawn -jbd /srv/mycontainer 3  
KillMode=process
```

Теперь мы можем запускать и останавливать эту службу командами `systemctl start` и `stop`. Однако, пока что мы не можем войти в эту систему⁸⁸. Чтобы исправить это упущение, настроим на контейнере SSH-сервер, причем таким образом, что подключение к его порту активировало весь контейнер, а затем активировало сервер, работающий внутри. Начнем с того, что прикажем хосту слушать порт SSH для контейнера. Для этого создадим на хосте файл `/etc/systemd/system/mycontainer.socket`:

```
[Unit]  
Description=The SSH socket of my little container  
  
[Socket]  
ListenStream=23
```

После того, как мы запустим этот юнит командой `systemctl start`, `systemd` будет слушать 23-й TCP-порт хоста. В примере выбран именно 23-й, потому что 22-й скорее всего окажется занят SSH-сервером самого хоста. `nspawn` виртуализует списки процессов

⁸⁸Прим. перев.: Ручной запуск на хосте соответствующей команды `systemd-nspawn -b` во время работы такой службы просто создаст *еще один контейнер*. Хотя корневой каталог у них один и тот же, пространства имен (например, списки процессов) у каждого будут свои. Подключиться к работающему контейнеру можно при помощи утилиты `nsenter`, которая включена в состав `util-linux` начиная с версии 2.23 (на момент написания исходной статьи этой утилиты еще не существовало).

и точек монтирования, но не сетевые стеки, поэтому порты хоста и гостей не должны конфликтовать⁸⁹.

Пока что `systemd`, работающий внутри контейнера, не знает, что делать с тем сокетом, которые ему передает `systemd` хоста. Если вы попытаетесь подключиться к порту 23, контейнер запустится, но сетевое соединение немедленно будет закрыто, так как данному сокету не соответствует пока никакой сервер. Давайте это исправим!

Настройка сокет-активации службы SSH была детально рассмотрена в одной из предыдущих статей, поэтому ограничусь приведением содержимого необходимых для этого конфигурационных файлов. Файл настроек для сокета (`/srv/mycontainer/etc/systemd/system/sshd.socket`)⁹⁰:

```
[Unit]
Description=SSH Socket for Per-Connection Servers

[Socket]
ListenStream=23
Accept=yes
```

Соответствующий ему файл конфигурации службы (`/srv/mycontainer/etc/systemd/system/sshd@.service`):

```
[Unit]
Description=SSH Per-Connection Server for %I

[Service]
ExecStart=-/usr/sbin/sshd -i
StandardInput=socket
```

После чего, добавим наш сокет в зависимости к цели `sockets.target`, чтобы `systemd` создавал его автоматически при загрузке контейнера⁹¹:

```
$ chroot /srv/mycontainer ln -s /etc/systemd/system/sshd.socket \
> /etc/systemd/system/sockets.target.wants/
```

Собственно, все. После того, как мы запустим на хосте юнит `mycontainer.socket`, `systemd` начнет прослушивать TCP-порт 23. При подключении к этому порту, `systemd` запустит контейнер и передаст сокет ему. Внутри контейнера свой `systemd`, в соответствии с файлом `sshd.socket`, примет этот сокет и запустит для нашего соединения экземпляр `sshd@.service`, что позволит нам залогиниться в контейнер по SSH.

⁸⁹Прим. перев.: До выхода 209 версии `systemd`, возможности по виртуализации сети в `systemd-nspawn` были весьма ограниченными (поддерживалась только полная изоляция сети от хоста), что весьма затрудняло использование описываемой технологии. Впрочем, опытные администраторы легко могли найти обходные пути, например: присваивание хосту дополнительного IP-адреса (безо всякой виртуализации, командой `ip addr add`) и привязка слушающих сокетов к конкретным адресам (в параметре `ListenStream=` сокет-файлов и/или в директиве `ListenAddress` файла `sshd_config` хоста). Начиная с `systemd` версии 209, `systemd-nspawn` предоставляет довольно широкие возможности по виртуализации сети: проброс сетевого интерфейса хоста в контейнер (`--network-interface=`), а также создание в контейнере виртуального сетевого интерфейса (`--network-veth`), связывающего его с хостом. Такой виртуальный интерфейс при необходимости может быть объединен в мост с интерфейсами хоста (`--network-bridge=`), что обеспечит доступ из контейнера к соответствующим сегментам сети. Подробности см. на странице руководства [systemd-nspawn\(1\)](#).

⁹⁰Прим. перев.: Обратите внимание на разницу между файлами конфигурации сокетов на хосте и в контейнере. Внутри контейнера используются параметры `Accept=yes` и `StandardInput=socket`, которые не задействованы на хосте. Это обусловлено тем фактом, что внутри контейнера сокет-активация производится в стиле `inetd` (служба получает только один сокет, причем замкнутый на ее потоки `STDIN` и `STDOUT`), в то время как на стороне хоста используется штатный стиль активации `systemd`, при котором запущенному процессу (в данном случае `systemd-nspawn`) просто передаются открытые файловые дескрипторы (`fd`) всех сокетов. Одно из достоинств такого подхода — возможность передавать сразу несколько сокетов, что позволяет активировать внутри контейнера несколько серверов.

⁹¹Прим. перев.: Возиться вручную с командой `ln` здесь совершенно необязательно. Можно просто добавить в файл `sshd.socket` секцию `[Install]`, содержащую параметр `WantedBy=sockets.target`, после чего добавление правильного симлинка будет выполняться куда более очевидной командой `systemctl --root=/srv/mycontainer enable sshd.socket`.

Если нам нужно запустить внутри контейнера другие службы с сокет-активацией, мы можем добавить в `myscontainer.socket` дополнительные сокеты. Все они будут прослушиваться, обращение к любому из них приведет к активации контейнера, и все эти сокеты будут переданы контейнеру при его активации. Внутри контейнера они будут обработаны соответствии с настройками имеющихся там сокет-юнитов. Те сокеты, для которых соответствующих юнитов не найдется, будут закрыты⁹², а те сокеты, которые будут настроены для прослушивания внутри контейнера, но не получены от хоста, будут активированы и доступны изнутри контейнера (а если это сетевые или файловые `unix`-сокеты, то и извне).

Итак, давайте отступим чуть назад и полюбуемся на результаты наших трудов. Что мы получили в итоге? Возможность настраивать на одном хосте множество контейнеров с полноценными ОС внутри, причем контейнеры запускаются только по запросу, что позволяет снизить потребление системных ресурсов и, соответственно, увеличить количество контейнеров (по сравнению с принудительной их активацией при загрузке хоста).

Разумеется, описанный подход работает только для контейнерной виртуализации, и неприменим к полной, т.е. может быть использован только с технологиями наподобие `libvirt-lxc` или `nspawn`, но не с `qemu/kvm` или `hpx`.

Если вы будете администрировать несколько таких контейнеров, вас наверняка порадует одна из возможностей `journal`: при запуске на хосте утилиты `journalctl` с ключом `-n`, она автоматически обнаружит журналы гостевых контейнеров и объединит их вывод с выводом журнала хоста⁹³. Ловко, не правда ли?

Необходимый минимум технологий для сокет-активации контейнеров, присутствует в `systemd`, начиная с версии 197. Тем не менее, наша работа в этой области еще не закончена, и в ближайшее время мы планируем доработать некоторые моменты. Например, сейчас, даже если все серверные службы внутри контейнера закончили обработку запросов и завершились, контейнер все равно продолжает функционировать и потреблять ресурсы хоста. Мы просто обязаны реализовать возможность автоматического завершения работы гостевой системы в такой ситуации. Причем у нас уже есть готовые наработки в этой области — мы можем задействовать уже существующую инфраструктуру, обеспечивающую автоматическое засыпание/выключение ноутбука при отсутствии активных задач и пользователей.

Впрочем, пора закругляться, а то статья получается чересчур длинной. Надеюсь, что вы смогли прорваться через все эти длинные и скучные рассуждения о виртуализации, сокетах, службах, различных ОС и прочем колдунстве. Также надеюсь, что эта статья станет хорошей отправной точкой при конфигурировании мощных и хорошо масштабируемых серверных систем. За дополнительной информацией обращайтесь к документации или приходите на наш IRC-канал. Спасибо за внимание!

21 Интеграция с контейнерами

В последнее время, все большую популярность набирает тема использования Linux-контейнеров, а вместе с ней на первый план выходят различные реализации систем управления контейнерами, такие, как `libvirt-lxc`, `LXC` и `Docker`. В данной статье я попробую рассказать о некоторых механизмах интеграции между `systemd` и подобными системами, обеспечивающих унифицированное, «бесшовное» управление службами, работающими на хосте и в его контейнерах.

⁹²Прим. перев.: Стоит особо отметить, что описанная технология работает только для служб, поддерживающих сокет-активацию в режимах `inetd` (все классические `inetd`-службы, кроме встроенных) или `systemd` (обычно зависят от библиотеки `libsystemd-daemon.so`, либо содержат в исходниках заголовочный файл `sd-daemon.h`). Службы, которые сами открывают себе слушающий сокет и не содержат кода для приема уже открытого сокета, так активировать нельзя.

⁹³Прим. перев.: Этот трюк работает благодаря опции `-j`, которую мы передаем программе `systemd-nspawn` при запуске контейнера (см. файл юнита выше). В соответствии с ней, в каталоге `/var/log/journal` хоста создается символическая ссылка на соответствующий каталог гостя. При этом, так как сообщения от гостей имеют другие `machine ID`, `journalctl` хоста не выводит их, если явно не указать `-n`. Подробности см. на страницах руководства [systemd-nspawn\(1\)](#) и [journalctl\(1\)](#).


```

[ OK ] Created slice system-getty.slice.
[ OK ] Reached target Slices.
[ OK ] Listening on Delayed Shutdown Socket.
[ OK ] Listening on /dev/initctl Compatibility Named Pipe.
[ OK ] Listening on Journal Socket.
      Starting Journal Service...
[ OK ] Started Journal Service.
[ OK ] Reached target Paths.
      Mounting Debug File System...
      Mounting Configuration File System...
      Mounting FUSE Control File System...
      Starting Create static device nodes in /dev...
      Mounting POSIX Message Queue File System...
      Mounting Huge Pages File System...
[ OK ] Reached target Encrypted Volumes.
[ OK ] Reached target Swap.
      Mounting Temporary Directory...
      Starting Load/Save Random Seed...
[ OK ] Mounted Configuration File System.
[ OK ] Mounted FUSE Control File System.
[ OK ] Mounted Temporary Directory.
[ OK ] Mounted POSIX Message Queue File System.
[ OK ] Mounted Debug File System.
[ OK ] Mounted Huge Pages File System.
[ OK ] Started Load/Save Random Seed.
[ OK ] Started Create static device nodes in /dev.
[ OK ] Reached target Local File Systems (Pre).
[ OK ] Reached target Local File Systems.
      Starting Trigger Flushing of Journal to Persistent Storage...
      Starting Recreate Volatile Files and Directories...
[ OK ] Started Recreate Volatile Files and Directories.
      Starting Update UTMP about System Reboot/Shutdown...
[ OK ] Started Trigger Flushing of Journal to Persistent Storage.
[ OK ] Started Update UTMP about System Reboot/Shutdown.
[ OK ] Reached target System Initialization.
[ OK ] Reached target Timers.
[ OK ] Listening on D-Bus System Message Bus Socket.
[ OK ] Reached target Sockets.
[ OK ] Reached target Basic System.
      Starting Login Service...
      Starting Permit User Sessions...
      Starting D-Bus System Message Bus...
[ OK ] Started D-Bus System Message Bus.
      Starting Cleanup of Temporary Directories...
[ OK ] Started Cleanup of Temporary Directories.
[ OK ] Started Permit User Sessions.
      Starting Console Getty...
[ OK ] Started Console Getty.
[ OK ] Reached target Login Prompts.
[ OK ] Started Login Service.
[ OK ] Reached target Multi-User System.
[ OK ] Reached target Graphical Interface.

```

```

Fedora release 20 (Heisenbug)
Kernel 3.18.0-0.rc4.git0.1.fc22.x86_64 on an x86_64 (console)

```

```

mycontainer login: root
Password:
-bash-4.2#

```

Наш испытательный стенд готов. Начнем рассмотрение механизмов интеграции systemd и контейнеров с утилиты machinectl. Запустив ее без параметров, мы получим список работающих в данный момент контейнеров:

```
$ machinectl
MACHINE                CONTAINER SERVICE
mycontainer             container nspawn
```

```
1 machines listed.
```

Команда `machinectl status` позволяет получить детальную информацию о выбранном контейнере:

```
$ machinectl status mycontainer
mycontainer:
  Since: Mi 2014-11-12 16:47:19 CET; 51s ago
  Leader: 5374 (systemd)
  Service: nspawn; class container
  Root: /srv/mycontainer
  Address: 192.168.178.38
          10.36.6.162
          fd00::523f:56ff:fe00:4994
          fe80::523f:56ff:fe00:4994
  OS: Fedora 20 (Heisenbug)
  Unit: machine-mycontainer.scope
      └─5374 /usr/lib/systemd/systemd
         └─system.slice
            └─dbus.service
               └─5414 /bin/dbus-daemon --system --address=systemd: --nofork --nopidfile --systemd-act...
            └─systemd-journald.service
               └─5383 /usr/lib/systemd/systemd-journald
            └─systemd-logind.service
               └─5411 /usr/lib/systemd/systemd-logind
            └─console-getty.service
               └─5416 /sbin/agetty --noclear -s console 115200 38400 9600
```

В частности, она отображает дерево контрольных групп и процессов контейнера, его IP-адреса и корневой каталог.

Команда `machinectl login` позволяет войти внутрь работающего контейнера, запустив в нем еще одну командную оболочку:

```
# machinectl login mycontainer
Connected to container mycontainer. Press ^] three times within 1s to exit session.

Fedora release 20 (Heisenbug)
Kernel 3.18.0-0.rc4.git0.1.fc22.x86_64 on an x86_64 (pts/0)

mycontainer login:
```

Команда `machinectl reboot` перезагружает контейнер:

```
# machinectl reboot mycontainer
```

Команда `machinectl poweroff` выключает контейнер:

```
# machinectl poweroff mycontainer
```

Пожалуй, этих минимальных сведений о `machinectl` должно быть достаточно. Стоит заметить, что ее функциональность не ограничивается приведенными здесь командами. За подробностями рекомендуем обратиться к [странице руководства](#). Еще раз отметим, что данные команды применимы не только к контейнерам на базе `systemd-nspawn`, но и к другим системам управления контейнерами, реализованным с учетом [наших рекомендаций](#), в частности, `libvirt-lxc`.

Помимо `machinectl`, в `systemd` существует и ряд других инструментов, упрощающих работу с контейнерами. В частности, большинство управляющих утилит `systemd` имеют встроенную поддержку контейнеров: при помощи ключа `-M` вы можете указать имя нужного вам контейнера, и соответствующая утилита будет работать так, как будто запущена внутри этого контейнера⁹⁵. Например (не забудьте снова запустить наш тестовый контейнер, если вы его до этого выключили):

```
# hostnamectl -M mycontainer set-hostname "wuff"
```

Приведенная здесь команда `hostnamectl(1)` устанавливает в локальном контейнере `mycontainer` имя хоста «wuff».

А вот пример использования того же ключа `-M` в программе `systemctl(1)`:

```
# systemctl -M mycontainer
UNIT                                LOAD   ACTIVE SUB    DESCRIPTION
-.mount                             loaded active mounted /
dev-hugepages.mount                 loaded active mounted Huge Pages File System
dev-mqueue.mount                    loaded active mounted POSIX Message Queue File System
proc-sys-kernel-random-boot_id.mount loaded active mounted /proc/sys/kernel/random/boot_id
[...]
time-sync.target                    loaded active active  System Time Synchronized
timers.target                        loaded active active  Timers
systemd-tmpfiles-clean.timer        loaded active waiting Daily Cleanup of Temporary Directories

LOAD   = Reflects whether the unit definition was properly loaded.
ACTIVE = The high-level unit activation state, i.e. generalization of SUB.
SUB     = The low-level unit activation state, values depend on unit type.

49 loaded units listed. Pass --all to see loaded but inactive units, too.
To show all installed unit files use 'systemctl list-unit-files'.
```

Как нетрудно догадаться, эта команда выводит список активных юнитов гостевой системы контейнера. (Чтобы не загромождать нашу статью лишним текстом, часть листинга пропущена.)

Перезапустить службу внутри гостевой системы также не представляет проблемы:

```
# systemctl -M mycontainer restart systemd-resolved.service
```

Поддержка контейнеров в утилите `systemctl` не ограничивается ключом `-M`. Она имеет еще и ключ `-r` (рекурсивный режим)⁹⁶, позволяющий выводить объединенный список юнитов для хоста и всех работающих гостевых систем:

```
# systemctl -r
UNIT                                LOAD   ACTIVE SUB    DESCRIPTION
boot.automount                       loaded active waiting EFI System Partition Automount
proc-sys-fs-binfmt_misc.automount     loaded active waiting Arbitrary Executable File Formats File S
sys-devices-pci0000:00-0000:00:02.0-drm-card0-card0\x2dLVDS\x2d1-intel_backlight.device loaded active plugg
[...]
timers.target                         loaded active active  Timers
mandb.timer                           loaded active waiting Daily man-db cache update
systemd-tmpfiles-clean.timer          loaded active waiting Daily Cleanup of Temporary Directories
mycontainer:-.mount                   loaded active mounted /
mycontainer:dev-hugepages.mount       loaded active mounted Huge Pages File System
mycontainer:dev-mqueue.mount          loaded active mounted POSIX Message Queue File System
[...]
mycontainer:time-sync.target          loaded active active  System Time Synchronized
mycontainer:timers.target             loaded active active  Timers
```

⁹⁵Прим. перев.: Поддержка данной опции добавлена в `systemctl`, `journalctl`, `loginctl`, `machinectl`, `hostnamectl`, `timedatectl`, `localectl`, `busctl`, `systemd-analyze`, `systemd-run` начиная с `systemd` 209, в `systemd-cgls` — с `systemd` 203.

⁹⁶Прим. перев.: Поддержка этой опции при выводе списка юнитов появилась начиная с `systemd` 212, при выводе списков таймеров и сокетов (`systemctl list-timers` и `systemctl list-sockets`) — начиная с `systemd` 213.

```
mycontainer:systemd-tmpfiles-clean.timer    loaded active waiting    Daily Cleanup of Temporary Directories

LOAD    = Reflects whether the unit definition was properly loaded.
ACTIVE  = The high-level unit activation state, i.e. generalization of SUB.
SUB     = The low-level unit activation state, values depend on unit type.

191 loaded units listed. Pass --all to see loaded but inactive units, too.
To show all installed unit files use 'systemctl list-unit-files'.
```

Как видно из листинга, сначала выводится список юнитов хоста, после которого идет список юнитов нашего контейнера. Перед именами юнитов гостевой системы, через двоеточие, указывается имя контейнера. (По традиции, листинг сокращен.)

Команда `systemctl list-machines` опрашивает процессы `init` всех работающих контейнеров, и по результатам выводит список работающих систем (хостовой и всех гостевых), с указанием их текущего состояния и наличия ошибок загрузки.

```
# systemctl list-machines
NAME          STATE   FAILED JOBS
delta (host)  running  0     0
mycontainer   running  0     0
miau          degraded 1     0
waldi         running  0     0

4 machines listed.
```

Для большей наглядности мы запустили еще два контейнера. В одном из них произошла ошибка запуска юнита, поэтому его состояние характеризуется как «`degraded`».

А теперь обсудим поддержку контейнеров в утилите [journalctl\(1\)](#). Во-первых, она позволяет использовать описанную выше опцию `-M`:

```
# journalctl -M mycontainer -n 8
Nov 12 16:51:13 wuff systemd[1]: Starting Graphical Interface.
Nov 12 16:51:13 wuff systemd[1]: Reached target Graphical Interface.
Nov 12 16:51:13 wuff systemd[1]: Starting Update UTMP about System Runlevel Changes...
Nov 12 16:51:13 wuff systemd[1]: Started Stop Read-Ahead Data Collection 10s After Completed Startup.
Nov 12 16:51:13 wuff systemd[1]: Started Update UTMP about System Runlevel Changes.
Nov 12 16:51:13 wuff systemd[1]: Startup finished in 399ms.
Nov 12 16:51:13 wuff sshd[35]: Server listening on 0.0.0.0 port 24.
Nov 12 16:51:13 wuff sshd[35]: Server listening on :: port 24.
```

Во-вторых, стоит упомянуть опцию `-m` (режим слияния), выводящую объединенный поток журнальных записей хоста и всех его гостевых систем⁹⁷:

```
# journalctl -m -e
```

(Здесь вывод команды опущен полностью. Думаю, вам будет несложно представить себе, как он выглядит.)

Поддержка контейнеров присутствует не только в утилитах `systemd`, но и в программах из комплекта `procs`⁹⁸:

```
# ps -eo pid,machine,args
PID MACHINE          COMMAND
  1 -                /usr/lib/systemd/systemd --switched-root --system --deserialize 20
[...]
2915 -                emacs contents/projects/containers.md
3403 -                [kworker/u16:7]
```

⁹⁷Прим. перев.: Чтобы данная опция действительно работала именно так, необходимо, чтобы в каталоге `/var/log/journal` хоста присутствовали символичные ссылки на каталоги с логами гостевых систем, либо было выполнено их `bind`-монтирование. См. примечание [93](#). Стоит заметить что, помимо логов контейнеров, в этом каталоге могут оказаться и логи других компьютеров, если у вас настроен их сбор с помощью [systemd-journal-remote\(8\)](#) или иных инструментов.

⁹⁸Прим. перев.: Начиная с выпуска `procs` 3.3.8.

```

3415 -          [kworker/u16:9]
4501 -          /usr/libexec/nm-vpnc-service
4519 -          /usr/sbin/vpnc --non-inter --no-detach --pid-file /var/run/NetworkManager/nm-vpnc-bfda86
4749 -          /usr/libexec/dconf-service
4980 -          /usr/lib/systemd/systemd-resolved
5006 -          /usr/lib64/firefox/firefox
5168 -          [kworker/u16:0]
5192 -          [kworker/u16:4]
5193 -          [kworker/u16:5]
5497 -          [kworker/u16:1]
5591 -          [kworker/u16:8]
5711 -          sudo -s
5715 -          /bin/bash
5749 -          /home/lennart/projects/systemd/systemd-nspawn -D /srv/mycontainer -b
5750 mycontainer /usr/lib/systemd/systemd
5799 mycontainer /usr/lib/systemd/systemd-journald
5862 mycontainer /usr/lib/systemd/systemd-logind
5863 mycontainer /bin/dbus-daemon --system --address=systemd: --nofork --nopidfile --systemd-activation
5868 mycontainer /sbin/agetty --noclear --keep-baud console 115200 38400 9600 vt102
5871 mycontainer /usr/sbin/sshd -D
6527 mycontainer /usr/lib/systemd/systemd-resolved
[...]

```

Здесь приводится (сокращенный) список процессов хоста и всех гостевых систем. Во втором столбце указывается имя контейнера («-» соответствует системе хоста).

Но наша работа по поддержке взаимодействия с контейнерами этим не ограничивается. В частности, мы добавили такую поддержку в «sd-bus» — разрабатываемую нами клиентскую библиотеку D-Bus/kdbus. Если вызов `sd_bus_open_system()` позволяет подключаться к шине вашей локальной системы, то `sd_bus_open_system_container()` открывает соединение к шине выбранного гостевого контейнера, обеспечивая прямой доступ к ее интерфейсам и методам.

`sd-login.h` и [D-Bus интерфейс демона machined](#) предоставляют API для реализации поддержки контейнеров в сторонних программах. В частности, они содержат функции для определения имени контейнера по PID процесса, получения списка работающих контейнеров и т.д.

`systemd-networkd` также имеет встроенную поддержку контейнеров. Когда он запускается внутри контейнера, он автоматически задействует DHCP-клиент и выполняет настройку IPv4LL-адреса⁹⁹ на виртуальном сетевом интерфейсе с именем `host0` (имя и назначение этого интерфейса оговаривается в [наших рекомендациях](#), которые должны поддерживаться вашей системой управления контейнерами). Когда `networkd` работает на хост-системе, он предоставляет функции DHCP-сервера и настраивает IPv4LL-адреса на всех виртуальных интерфейсах с именами вида `ve-имя_контейнера`.

И наконец, последний из рассматриваемых в этой статье аспектов интеграции `systemd` и контейнеров: взаимодействие с NSS. В составе свежих версий `systemd`¹⁰⁰ поставляется NSS-модуль `nss-mymachines`, позволяющий преобразовывать имена контейнеров в их IP-адреса при помощи вызовов `gethostbyname()` и `getaddrinfo()`. Разумеется, это применимо лишь к контейнерам, которые имеют собственное сетевое пространство имен (`netns`), т.е. свой сетевой стек. Команда `systemd-nspawn`, приведенная выше, запускает контейнер в сетевом стеке хоста, поэтому, чтобы продемонстрировать работу обсуждаемых механизмов, нам придется перезапустить этот контейнер с отдельным сетевым стеком, в котором `systemd-nspawn` создаст виртуальный сетевой интерфейс (`veth`), связывающий контейнер с хостом:

```

# machinectl poweroff mycontainer
# systemd-nspawn -D /srv/mycontainer --network-veth -b

```

⁹⁹Прим. перев.: Механизм IPv4LL предоставляет простой метод автоматической настройки сетевых адресов в пределах одного сегмента. Сетевому интерфейсу присваивается случайно выбранный незанятый IP-адрес из специально зарезервированного диапазона 162.254.0.0/16.

¹⁰⁰Прим. перев.: Начиная с версии 216.

Теперь, если в контейнере и на хосте работает `systemd-networkd`, мы можем обращаться к контейнеру через сеть просто по его имени, которое при помощи модуля `nss-myhostname` будет автоматически преобразовано в его IP-адрес¹⁰¹:

```
# ping mycontainer
PING mycontainer (10.0.0.2) 56(84) bytes of data.
64 bytes from mycontainer (10.0.0.2): icmp_seq=1 ttl=64 time=0.124 ms
64 bytes from mycontainer (10.0.0.2): icmp_seq=2 ttl=64 time=0.078 ms
```

Разумеется, преобразование имен контейнеров в IP-адреса и обратно работает не только для утилиты `ping`, но и для всех остальных программ, использующих функции `libc` `gethostbyname()` и `getaddrinfo()`, в том числе и для нашей любимой и незаменимой `ssh`.

Вот и все, о чем я хотел рассказать в данной статье. Мы кратко прошлись по основным моментам интеграции `systemd` и контейнеров. Изложение не претендует на полноту, и очень многие детали оказались «за бортом» — о них вы можете узнать из документации. Кроме того, в настоящее время мы продолжаем работу над более тесной интеграцией с контейнерами, так что ожидайте новых возможностей в этой области с ближайшими выпусками `systemd`.

В завершение стоит отметить, что концепция машины (`machine`) применима не только к контейнерам, но и, в некоторой степени, к полноценным виртуальным машинами. Однако, в случае виртуальных машин, доступ к гостевым системам с хоста реализуется несколько сложнее, чем в контейнерах — вместо прямого доступа к системным вызовам, приходится использовать взаимодействие через сеть¹⁰².

В любом случае, надеюсь, что статья для кого-то оказалась полезной. За подробностями обращайтесь к документации (начать изучение темы можно со ссылок, приведенных в этой статье).

Защита от уязвимостей `AF_PACKET` при помощи `systemd`¹⁰³

Небольшая заметка о том, как ограничительные механизмы, встроенные в последние версии `systemd`, позволяют защитить ваши службы от эксплуатации уязвимости [CVE-2016-8655](#)¹⁰⁴.

Начиная с версии 211, в `systemd` добавлена поддержка новой директивы для `service`-файлов — `RestrictAddressFamilies=`, которая отбирает у службы возможность создавать сокет определенного типа¹⁰⁵. Достаточно добавить в секцию `[Service]` юнит-файла вашей службы строчку `RestrictAddressFamilies=~AF_PACKET`, и служба утратит

¹⁰¹Прим. перев.: Для корректной работы модуля `nss-myhostname`, он должен быть указан в конфигурационном файле `/etc/nsswitch.conf`, в строке `hosts`. В частности, в поставляемом с `systemd` варианте этого файла, данная строчка выглядит так: `hosts: files myhostname resolve myhostname`. Помимо `myhostname`, в ней фигурируют и другие `NSS`-модули `systemd: resolve` отвечает за взаимодействие с кэширующим `DNS`-сервером `systemd-resolved`, а `myhostname` обеспечивает корректное преобразование системного имени хоста в IP-адрес `127.0.0.2`, даже при отсутствии соответствующей записи в `/etc/hosts`.

¹⁰²Прим. перев.: Можно предположить, что здесь автор пытается намекнуть на опцию `-H [user@]host[:container]`, позволяющую обращаться к другим хостам и их контейнерам через сеть при помощи `SSH`-транспорта, реализованного в шине `D-Bus`. Эта опция поддерживается в большинстве управляющих утилит `systemd` (`systemctl`, `loginctl`, `machinectl`, `hostnamectl`, `timedatectl`, `localectl`, `busctl`, `systemd-analyze`, `systemd-run`).

¹⁰³Прим. перев.: Эта и последующие статьи официально не входят в цикл «`systemd` для администраторов», и поэтому не имеют порядковых номеров. Однако они сходны по тематике с остальными статьями цикла и несут довольно много полезной информации, и поэтому их перевод все же присутствует в данном документе.

¹⁰⁴Прим. перев.: Решение перевести эту заметку связано с появлением информации о целом ряде уязвимостей, связанных с `AF_PACKET`. Помимо уже упомянутой автором [CVE-2016-8655](#), это [CVE-2017-7308](#), [CVE-2017-10001111](#), [CVE-2017-14497](#), и еще одна, на момент написания этих строк не имеющая идентификатора `CVE`. Все они имеют схожую природу (повышение привилегий до `root` путем манипуляций с сокетами `AF_PACKET`) и успешно блокируются описанным в статье методом. Соответственно, был изменен и заголовок статьи (в оригинале речь шла только о [CVE-2016-8655](#)).

¹⁰⁵Прим. перев.: Под «типом сокета» здесь подразумевается `address family`. Не вполне корректно, зато кратко.

возможность создавать сокеты типа `AF_PACKET`, в результате чего эксплуатация соответствующих уязвимостей будет исключена. Приведенный пример демонстрирует режим «черного списка», когда указываются *запрещенные* типы сокетов. Однако, более безопасным подходом является «белый список» *разрешенных* типов. В отличие от черного, белый список указывается без символа `~`. Простой пример:

```
...
[Service]
ExecStart=/usr/bin/mydaemon
RestrictAddressFamilies=AF_INET AF_INET6 AF_UNIX
...
```

Служба, настроенная таким образом, может использовать только сокеты типа `AF_INET`, `AF_INET6` и `AF_UNIX`, что должно быть достаточно для большинства системных демонов (`AF_INET` — низкоуровневое название для IPv4, `AF_INET6` — для IPv6, а `AF_UNIX` соответствует локальным UNIX-сокетам).

Начиная с [systemd 232](#), мы добавили директивы `RestrictAddressFamilies=` в юнит-файлы всех служб, входящих в состав `systemd`, указав минимально необходимый набор типов сокетов для каждой службы.

С выходом `systemd 233` мы [добавили](#) еще один метод блокировки подобных уязвимостей. Директива `RestrictNamespaces=` позволяет ограничить перечень пространств имен (`namespaces`), которые может использовать служба¹⁰⁶. В частности, `RestrictNamespaces=yes` полностью отключает доступ к пространствам имен. Перечисление, например `RestrictNamespaces=net ipc`, позволяет задать список разрешенных к использованию пространств имен (в нашем случае это пространства имен для сети и IPC)¹⁰⁷. Учитывая, что механизм пространства имен пользователей (`user namespace`) уже давно является перманентным источником опасных уязвимостей¹⁰⁸, возможно, будет правильным заблокировать использование пространств имен для тех служб, которым они не требуются (а таких служб большинство).

Хочется надеяться, что в будущем сопровождающие дистрибутивов и апстримные разработчики программ будут сами добавлять соответствующие ограничительные директивы в поставляемые по умолчанию юнит-файлы, так как эти люди лучше знают, какие именно привилегии минимально необходимы для функционирования их демонов.

Динамические пользователи

Коротко о главном: теперь вы можете настроить systemd таким образом, чтобы он автоматически выделял новый идентификатор пользователя (UID) для службы при ее запуске, и автоматически освобождал его при остановке службы. Такой подход абсолютно безопасен и может также применяться в сочетании с другими технологиями systemd, в частности, одноразовыми службами (transient services), службами с активацией через сокет и экземплярами служб.

¹⁰⁶Прим. перев.: Пространства имен — специфичная для ядра Linux технология контейнерной изоляции, позволяющая группе процессов иметь свой собственный сетевой стек (`network namespace`), иерархию монтирования (`mount namespace`, также использовался термин `filesystem namespace`), иерархию контрольных групп (`cgroup namespace`), очереди сообщений POSIX и интерфейсы SysV IPC (`IPC namespace`), имя хоста (`UTS namespace`), независимые списки идентификаторов процессов (`PID namespace`) и пользователей/групп (`user namespace`). Лежат в основе систем контейнерной изоляции, таких как LXC и Docker. Также широко используются в `systemd` для обеспечения безопасности — см. главы 6 и 12.

¹⁰⁷Прим. перев.: Как и в предыдущем случае, можно поменять белый список на черный, добавив перед ним символ `~`.

¹⁰⁸Прим. перев.: В частности, для создания сокетов `AF_PACKET` требуется наличие полномочий `CAP_NET_RAW` (или прав `рута`), отсутствующих у большинства программ. Однако, если система разрешает непривилегированным процессам создавать новое пространство имен пользователей (`sysctl kernel.unprivileged_userns_clone=1`), то злоумышленник может создать новое `user namespace`, в котором его процесс будет иметь полномочия `рута` и, соответственно, сможет создать новое сетевое пространство имен, и в нем сокет типа `AF_PACKET`, с помощью которого он попытается получить полномочия `рута` уже в основной системе.

В выпуске [systemd 235](#), помимо прочих улучшений, значительно расширена поддержка *динамических пользователей* (базовая функциональность для которой появилась еще в версии 232). Это мощная и удобная, но пока малоизвестная технология, и в данной статье я попытаюсь исправить последний недостаток, рассказав о ней более подробно.

Понятие *пользователя* является одной из ключевых концепций безопасности POSIX-совместимых операционных систем. Большинство созданных впоследствии механизмов обеспечения безопасности (SELinux и другие системы мандатного контроля доступа, *sarabilities*, пространства имен пользователей и т.д.) так или иначе связаны или взаимодействуют с концепцией пользователя. Даже если вы хотите пересобрать ядро Linux, полностью отключив все возможные механизмы безопасности, от пользователей вы, скорее всего, не избавитесь.

Изначально, понятие пользователя было введено при создании многопользовательских систем, рассчитанных на одновременную работу нескольких пользователей-*людей*, обеспечивая взаимную изоляцию и разделение ресурсов. Однако, большинство современных UNIX-систем используют концепцию пользователей иначе: несмотря на то, что реальный пользователь у них может быть только один (или вообще ни одного!), в их списках пользователей (то есть файлах `/etc/passwd`), записей намного больше. В наше время, большинство пользователей типичной UNIX-системы являются *системными* — они соответствуют не сидящим перед компьютером живым людям, а наборам полномочий (*security identity*), с которыми запускаются системные службы (т.е. программы). Хотя традиционные многопользовательские системы постепенно теряют свою актуальность, лежащее в их основе понятие «пользователь» стало одним из краеугольных камней в технологиях защиты UNIX-систем. Большинство служб в современной системе работает от имени своего пользователем, которому даны минимально возможные полномочия.

Создатели ОС Android хорошо понимали значимость концепции пользователя для безопасности системы, и пошли еще дальше: в Android пользователи создаются не только для системных служб, но и для каждого графического приложения, что позволяет обеспечить разделение ресурсов и защиту процессов разных приложений друг от друга.

Тем не менее, в традиционных Linux-системах концепции пользователей пока уделяется меньше внимания. Несмотря на то, что она является ключевой технологией обеспечения безопасности UNIX-системы, механизмы создания пользователей и управления ими пока не обладают достаточной гибкостью. В большинстве случаев, если вы устанавливаете службу, которая выполняется от своего пользователя, работу по созданию учетных записей берут на себя установочные скрипты RPM или DEB-пакетов. После этого, созданный пользователь останется в системе уже навсегда, даже если вы удалите соответствующий пакет. Большинство дистрибутивов Linux используют для идентификаторов системных пользователей диапазон от 0 до 1000 — не так уж и много. Следовательно, создание пользователей является «дорогой» операцией: количество идентификаторов ограничено, и не существует механизма для их автоматического освобождения после использования. Если вы будете активно использовать концепцию системных пользователей, вы рано или поздно исчерпаете доступный лимит.

У вас может возникнуть вопрос: почему системные пользователи не удаляются при удалении создавшего их пакета (как минимум, в большинстве дистрибутивов)? Причиной этому является одно из важных свойств концепции пользователя (его можно даже назвать *ошибкой дизайна*): идентификаторы пользователей сохраняются в атрибутах файлов (и некоторых других объектов, в частности, элементов IPC). Если служба, работающая из-под отдельного системного пользователя, создаст где-либо новый файл, то даже после остановки службы и удаления ее пакета, числовой идентификатор ее пользователя (UID) останется в метаданных этого файла. Если бы наша система удаляла системных пользователей вместе с их пакетами, то рано или поздно этот идентификатор заняла бы какая-нибудь другая служба. И это уже можно считать уязвимостью, так как файл с данными одной службы становится доступен для другой. Как следствие, дистрибутивы склонны избегать повторного использования UID, и каждый созданный системный пользователь остается в системе навсегда.

Выше мы рассматривали сложившуюся к настоящему моменту ситуацию. А теперь посмотрим, что нового вносит наша концепция динамических пользователей, и какие проблемы она может решить.

Что такое «динамический пользователь»?

Реализованный в `systemd` механизм динамических пользователей нацелен на упрощение и удешевление операций создания пользователей «на лету», и в перспективе позволит значительно расширить область применения концепции пользователей в целом.

При создании или редактировании `service`-файла вашей службы, вы можете включить для нее механизм динамических пользователей, добавив в секцию `[Service]` директиву `DynamicUser=yes`. После этого, при каждом запуске данной службы, для нее будет на лету создаваться системный пользователь. При остановке службы он будет автоматически удаляться. `UID` такого пользователя выбирается из диапазона 61184–65519 (при этом производится автоматическая проверка, исключающая использование `UID`, занятых кем-либо еще).

Вы спросите, как же в таком случае решается вышеописанная проблема привязки идентификаторов пользователей к файлам? Существуют как минимум два очевидных подхода к ее решению:

1. Запретить службе создавать файлы, каталоги и объекты `IPC`.
2. Автоматически удалять созданные службой файлы, каталоги и объекты `IPC` при ее остановке.

`systemd` реализует одновременно обе стратегии, но применяет их к различным областям рабочего окружения службы. А именно:

1. Установка `DynamicUser=yes` также применяет директивы `ProtectSystem=strict` и `ProtectHome=read-only`. В результате служба теряет возможность что-либо писать практически во все каталоги системы, за исключением специальных файловых систем (`/dev`, `/proc` и `/sys`) и временных каталогов (`/tmp` и `/var/tmp`). (Кстати, включать эти опции имеет смысл даже для служб, которые не используют режим `DynamicUser=yes`, так как это сильно снижает возможный ущерб при взломе службы.)
2. Установка `DynamicUser=yes` автоматически применяет директиву `PrivateTmp=yes`. В результате для этой службы создаются свои собственные, изолированные от других служб экземпляры каталогов `/tmp` и `/var/tmp`¹⁰⁹, причем их жизненный цикл привязан к жизненному циклу службы: при остановке службы удаляется не только ее пользователь, но и ее временные каталоги. (Опять же замечу, что эту директиву имеет смысл применять и без `DynamicUser=yes`, так как она тоже повышает безопасность вашей системы.)
3. Установка `DynamicUser=yes` автоматически применяет директиву `RemoveIPC=yes`, которая обеспечивает автоматическое удаление объектов межпроцессного взаимодействия (`IPC`) `SysV` и `POSIX` (общая память, очереди сообщений, семафоры), принадлежащих службе, при ее остановке. Как следствие, жизненный цикл `IPC`-объектов также привязан к жизненному циклу службы и ее пользователя. (И снова: эту директиву имеет смысл применять и для обычных служб!)

Использование этих четырех опций позволяет практически полностью изолировать службу с динамическим пользователем от основной системы. Она не может создавать

¹⁰⁹Прим. перев.: Технически это реализовано следующим образом: в системных `/tmp` и `/var/tmp` создаются подкаталоги со специальными именами (построенными на основе имени службы и (псевдо)случайных последовательностей символов), принадлежащие руту и имеющие права доступа 0700 (это граничные каталоги — принцип их работы описан в данной статье ниже), а внутри них создаются подкаталоги `tmp` с правами 0777. Для самой службы создается пространство имен монтирования (`mount namespace`), в котором эти подкаталоги `bind`-монтируются в `/tmp` и `/var/tmp` соответственно. См. также раздел 12.2.

файлы или каталоги где-либо, кроме `/tmp` и `/var/tmp`, где они будут удалены при ее остановке, как и созданные ею объекты ИРС. Таким образом, проблема владения файлами, каталогами и объектами ИРС, успешно решена.

Если вам нужно немного приоткрыть «железный занавес» и дать службе возможность взаимодействовать с другими программами, это можно сделать при помощи параметра `RuntimeDirectory=`. Подкаталог с указанным в этом параметре именем создается в каталоге `/run` при запуске службы, принадлежит ее пользователю, и тоже автоматически удаляется при ее остановке. Однако, при этом он открыт для доступа других программ, так что служба может размещать в нем различные интерфейсные объекты (например, UNIX-сокеты), жизненный цикл которых также будет привязан к жизненному циклу службы. Например, если вы зададите для своей службы `RuntimeDirectory=foobar`, то увидите, что при ее запуске создается каталог `/run/foobar`, а при ее остановке он удаляется. (Как и другие описанные здесь директивы, `RuntimeDirectory=` прекрасно может работать и без `DynamicUser=`, предоставляя для вашей службы автоматически создаваемый и удаляемый каталог с правильным владельцем.)

Долговременное хранение данных

Вышеописанный способ изоляции службы, хотя и может оказаться полезен в некоторых случаях, имеет одно существенное ограничение: служба не может сохранить какие-либо данные так, чтобы получить к ним доступ при последующих запусках. Почти все системное дерево каталогов доступно ей только для чтения, а все доступные на запись каталоги удаляются при остановке и перезапуске.

В выпуске `systemd` 235 это ограничение было снято: мы добавили три новых опции — `StateDirectory=`, `LogsDirectory=` и `CacheDirectory=`. По большей части они действуют аналогично вышеописанной опции `RuntimeDirectory=`, но создают подкаталоги не в `/run`, а в `/var/lib`, `/var/log` и `/var/cache` соответственно. Но самое главное их отличие состоит в том, что эти подкаталоги *не удаляются* при остановке службы, что позволяет использовать их для долговременного хранения данных.

Очевидный вопрос: как при этом решается вышеописанная проблема с динамическими идентификаторами пользователей?

Для решения этой задачи мы обратились к опыту систем контейнерной изоляции. При работе с контейнерами возникает схожая проблема: контейнеры и хост используют частично или полностью перекрывающиеся наборы числовых идентификаторов пользователей, а значит, пользователи хоста могут получить доступ к файлам контейнера, которые принадлежат пользователям с таким же UID — а ведь это может быть совсем другой пользователь, никак не связанный со своим хостовым «двойником» (кроме случаев, когда используются пространства имен пользователей¹¹⁰). Особенно неприятно, если это происходит с исполняемым файлом, имеющим `suid`-бит — тогда речь идет уже не просто о доступе к чужим данным, а о потенциальной возможности повышения привилегий. Чтобы предотвратить подобные ситуации, системы управления контейнерами обычно помещают корневые системы гостей в специальный *граничный* каталог, имеющих ограниченные права доступа (обычно: права `0700`, владелец `root:root`). При этом, непривилегированные пользователи хоста уже не могут добраться до файлов, принадлежащих их контейнерным «двойникам», просто потому, что не имеют доступа ко всему, что лежит в граничном каталоге. В UNIX, чтобы получить доступ к файлу, вы должны иметь доступ ко всем каталогам по пути к нему, начиная с корневого.

Как это работает для наших служб с динамическими пользователями? Предположим, что вы указали `StateDirectory=foobar`, но *не включали* режим `DynamicUser=`. В момент запуска такой службы, для нее будет создан каталог `/var/lib/foobar`,

¹¹⁰Прим. перев.: В таких случаях используется механизм отображения UID (`user-remap`), когда пользователям, работающим с `user namespaces` (например, `dockermap` в `Docker`), выдаются диапазоны UID хоста (см. `/etc/subuid`), на которые отображаются UID гостевой системы. Это решает проблему пересечения UID хоста и гостя. Без присвоения такого диапазона гость будет иметь только один UID (в пространстве гостя он может быть любым, даже нулевым), который для хоста (и его файловой системы) будет виден как UID пользователя контейнера.

принадлежащий пользователю, от которого запускается служба¹¹¹. Этот каталог и его содержимое *не удаляются* после завершения службы. Если же мы включим для нашей службы режим `DynamicUser=`, алгоритм станет немного сложнее — теперь `/var/lib/foobar` уже не каталог, а принадлежащая руту символьная ссылка на каталог `/var/lib/private/foobar`, который принадлежит динамическому пользователю, выделенному для службы. Каталог `/var/lib/private` играет роль граничного: он принадлежит `root:root`, и имеет права `0700`. При остановке службы, и символьная ссылка, и каталог, на который она указывает, не удаляются. Каталог продолжает принадлежать теперь уже освобожденному идентификатору пользователя, однако граничный каталог не позволит получить к нему доступ ни пользователям хоста, ни другим службам, получившим тот же UID.

Следующий вопрос: если граничный каталог столь успешно защищает подкаталог с данными службы, то как сама служба сможет получить к ним доступ? Для этого, служба запускается в специально модифицированном пространстве имен точек монтирования (mount namespace): помимо уже упомянутых хитростей с монтированием `/tmp` и `/var/tmp`¹¹², добавляется `tmpfs`, смонтированная в `/var/lib/private` и доступная только на чтение, с правами, разрешающими чтение из этого каталога (`0755`). Внутри этой файловой системы создан каталог `foobar`, в который `bind`-монтируется каталог `/var/lib/private/foobar` с хоста. В результате и для хоста, и для службы содержимое этого каталога находится по одному и тому же пути, но при этом «с точки зрения» службы каталог `/var/lib/private` уже не является ограничительным, и в нем отсутствуют подкаталоги, принадлежащие другим службам, что обеспечивает полную изоляцию служб друг от друга. Символьная ссылка `/var/lib/foobar` позволяет не задумываться о том, используется граничный каталог или нет: и при включенном, и при выключенном режиме динамических пользователей, каталог с данными программы доступен по указанному пути.

Назревает очередной вопрос. Предположим, что мы запустили службу от динамического пользователя и с настроенным каталогом для хранения данных. Она получила некоторый UID (назовем его X), которому принадлежит этот каталог. Потом мы перезапустили службу, и она получила новый UID $Y \neq X$. Что тогда произойдет? Неужели каталог и его содержимое будут все еще принадлежать UID X , и наша служба уже не сможет получить к ним доступ? Конечно же, нет — `systemd` рекурсивно меняет владельца для каталога и его содержимого.

Разумеется, операция рекурсивной смены владельца (`chown()`) для дерева каталогов может оказаться весьма затратной (хотя, по моему личному опыту, для большинства служб это не так критично, как кажется на первый взгляд), и поэтому мы ввели две оптимизации, сводящие к минимуму вероятность такой операции. Во-первых, `systemd` начинает подбор подходящего UID с некоторого значения, полученного путем хеширования имени службы. Таким образом, если имя службы не поменялось, то при следующем запуске она, скорее всего, получит тот же UID. В результате, необходимость в `chown()` отпадает (разумеется, после соответствующих проверок). Во-вторых, если указанный каталог уже существует, и его владельцем является неиспользуемый UID из динамического диапазона, то службе присваивается именно этот UID, что также увеличивает шансы избежать `chown()`. (На самом деле, выделенный сейчас диапазон на четыре с лишним тысячи UID не так уж и велик, и при активном применении динамических пользователей рано или поздно появятся ситуации, когда обойтись без `chown()` уже не удастся.)

Директивы `CacheDirectory=` и `LogsDirectory=` работают по аналогии со `StateDirectory=`. Единственное их отличие состоит в том, что они управляют подкаталогами в `/var/cache` и `/var/log`, и используют граничные каталоги `/var/cache/private` и `/var/log/private` соответственно.

¹¹¹Прим. перев.: Если режим `DynamicUser=` отключен, пользователь службы определяется директивой `User=`. Более подробно ее роль при включенном и выключенном режиме динамических пользователей рассмотрена [ниже](#).

¹¹²Прим. перев.: Строго говоря, «хитростей» там гораздо больше — упомянутые выше директивы `ProtectSystem=strict` и `ProtectHome=read-only` предполагают перемонтирование целого ряда каталогов, включая корневой. Но к теме это действительно не относится.

Примеры

Итак, мы ознакомились с теорией. Попробуем теперь посмотреть, как все это работает на практике. Простой пример:

```
# cat > /etc/systemd/system/dynamic-user-test.service <<EOF
[Service]
ExecStart=/usr/bin/sleep 4711
DynamicUser=yes
EOF
# systemctl daemon-reload
# systemctl start dynamic-user-test
# systemctl status dynamic-user-test
dynamic-user-test.service
   Loaded: loaded (/etc/systemd/system/dynamic-user-test.service; static; vendor preset: disabled)
   Active: active (running) since Fri 2017-10-06 13:12:25 CEST; 3s ago
 Main PID: 2967 (sleep)
   Tasks: 1 (limit: 4915)
   CGroup: /system.slice/dynamic-user-test.service
           └─2967 /usr/bin/sleep 4711

Okt 06 13:12:25 sigma systemd[1]: Started dynamic-user-test.service.
# ps -e -o pid,comm,user | grep 2967
 2967 sleep          dynamic-user-test
# id dynamic-user-test
uid=64642(dynamic-user-test) gid=64642(dynamic-user-test) groups=64642(dynamic-user-test)
# systemctl stop dynamic-user-test
# id dynamic-user-test
id: 'dynamic-user-test': no such user
```

Мы создали юнит-файл службы с включенным режимом `DynamicUser=`, запустили эту службу, убедились, что она работает нормально, запросили данные о пользователе, от которого она запущена (в нашем случае его имя совпадает с именем службы — `systemd` использует его автоматически, если оно соответствует синтаксису имени пользователя, и вы не указали явно другого имени), остановили службу и убедились, что такого пользователя больше не существует.

Уже неплохо, правда? Следующий шаг — повторим то же самое с интерактивной *одноразовой* службой. Для тех, кто не очень хорошо разбирается в тонкостях `systemd`: одноразовая служба (transient service) создается и запускается «на лету», без каких-либо конфигурационных файлов — например, с помощью программы `systemd-run`, которую можно запустить непосредственно из командной оболочки. Короче: запуск службы без предварительного создания юнит-файла.

```
# systemd-run --pty --property=DynamicUser=yes --property=StateDirectory=wuff /bin/sh
Running as unit: run-u15750.service
Press ~] three times within 1s to disconnect TTY.
sh-4.4$ id
uid=63122(run-u15750) gid=63122(run-u15750) groups=63122(run-u15750) context=system_u:system_r:in
sh-4.4$ ls -al /var/lib/private/
total 0
drwxr-xr-x. 3 root      root          60  6. Okt 13:21 .
drwxr-xr-x. 1 root      root          852 6. Okt 13:21 ..
drwxr-xr-x. 1 run-u15750 run-u15750    8  6. Okt 13:22 wuff
sh-4.4$ ls -ld /var/lib/wuff
lrwxrwxrwx. 1 root root 12  6. Okt 13:21 /var/lib/wuff -> private/wuff
sh-4.4$ ls -ld /var/lib/wuff/
drwxr-xr-x. 1 run-u15750 run-u15750 0  6. Okt 13:21 /var/lib/wuff/
sh-4.4$ echo hello > /var/lib/wuff/test
sh-4.4$ exit
```

```

exit
# id run-u15750
id: 'run-u15750': no such user
# ls -al /var/lib/private
total 0
drwx-----. 1 root  root   66  6. Okt 13:21 .
drwxr-xr-x.  1 root  root  852  6. Okt 13:21 ..
drwxr-xr-x.  1 63122 63122   8  6. Okt 13:22 wuff
# ls -ld /var/lib/wuff
lrwxrwxrwx. 1 root root 12  6. Okt 13:21 /var/lib/wuff -> private/wuff
# ls -ld /var/lib/wuff/
drwxr-xr-x. 1 63122 63122 8  6. Okt 13:22 /var/lib/wuff/
# cat /var/lib/wuff/test
hello

```

В приведенном примере, мы запускаем интерактивную оболочку `/bin/sh` как однократную службу `run-u15750.service` (`systemd-run` выбрал это имя автоматически, так как мы не указали имя службы явно¹¹³) под динамическим пользователем, имя которого, как и в предыдущем примере, унаследовано от имени службы. Так как мы задали `StateDirectory=wuff`, то каталог для долговременного хранения данных нашей службы должен быть доступен под именем `/var/lib/wuff`. В интерактивной оболочке, запущенной в рамках службы, команда `ls` показывает граничный каталог `/var/lib/private` и его содержимое, а также символическую ссылку `/var/lib/wuff`, указывающую на его подкаталог `wuff`. Наконец, перед завершением оболочки, мы создаем там тестовый файл. Вернувшись в нашу исходную оболочку, мы проверяем, существует ли еще пользователь, выделенный для нашей службы — нет, он автоматически удален в момент завершения службы (оболочки). При помощи аналогичных команд `ls` мы снова проверяем каталог долговременного хранения данных службы, на этот раз с хоста. Видим мы почти то же самое, с двумя исключениями: во-первых, пользователь и группа, владеющие каталогом и его содержимым, отображаются уже в виде числовых идентификаторов, а не имен. Это обусловлено тем, что пользователь (то есть ассоциация числового идентификатора с некоторым именем) был удален в момент завершения службы. Во-вторых, отличаются права доступа к граничному каталогу: `0755` (читать могут все) изнутри службы, и `0700` (читать может только владелец, т.е. `root`) — с хоста.

А теперь попробуем запустить еще одну однократную службу, указав ей тот же каталог с данными:

```

# systemd-run --pty --property=DynamicUser=yes --property=StateDirectory=wuff /bin/sh
Running as unit: run-u16087.service
Press ~] three times within 1s to disconnect TTY.
sh-4.4$ cat /var/lib/wuff/test
hello
sh-4.4$ ls -al /var/lib/wuff/
total 4
drwxr-xr-x. 1 run-u16087 run-u16087  8  6. Okt 13:22 .
drwxr-xr-x. 3 root          root          60  6. Okt 15:42 ..
-rw-r--r--. 1 run-u16087 run-u16087  6  6. Okt 13:22 test
sh-4.4$ id
uid=63122(run-u16087) gid=63122(run-u16087) groups=63122(run-u16087) context=system_u:system_r:in
sh-4.4$ exit
exit

```

Как видим, `systemd-run` сгенерировал для этой службы уже другое имя, которое перешло и к ее пользователю, однако числовой идентификатор пользователя остался прежним — `systemd` подхватил UID владельца каталога с данными (предварительно убедившись, что он больше никем не используется). Этим иллюстрируется вышеописанная оптимизация алгоритма выбора UID (цикл выбора идентификатора начинается с

¹¹³Прим. перев.: Это можно было бы сделать параметром `systemd-run --unit=`.

UID владельца существующего каталога с данными): рекурсивного выполнения `chown()` удалось избежать.

Надеюсь, вышеприведенные примеры помогли вам разобраться и в самой идее, и в особенностях ее реализации.

Практическое применение

Итак, мы рассмотрели, как включить механизм динамических пользователей для юнита, и разобрались, как он реализован. Самое время попытаться понять, где и для чего он может применяться.

- Одно из главных достоинств технологии динамического выделения UID — возможность запускать службы с урезанными привилегиями (то есть, от отдельного пользователя), не оставляя в системе никаких артефактов. Системный пользователь создается и используется, но после применения автоматически удаляется, и повторное применение его UID не несет никаких рисков для безопасности. Мы можем одной командой запускать одноразовые службы для выполнения каких-либо операций, изолируя их под отдельным идентификатором пользователя, без необходимости вручную создавать и удалять учетную запись, и не расходуя доступные UID попусту.
- Во многих случаях, запуск службы уже не требует предварительной подготовки со стороны пакетного менеджера. Другими словами, большинство операций `useradd/mkdir/chown/chmod`, выполняемых пост-инсталляционными скриптами пакетов, а также дополнительные конфигурационные файлы в `sysusers.d` и `tmpfiles.d` становятся необязательными, так как эти операции выполняются автоматически благодаря директивам `DynamicUser=` и `StateDirectory=/CacheDirectory=/LogsDirectory=`, причем не при установке/удалении пакета, а непосредственно при запуске/остановке службы.
- Сочетание технологий динамических пользователей и одноразовых служб предоставляет простой механизм изоляции приложений. Например, предположим, что мы не доверяем бинарнику `sort`. Мы можем поместить его в простую и надежную песочницу динамического пользователя при помощи `systemd-run`, и при этом сохранить возможность подключать ее через каналы (pipelines) к другим программам. Простенький пример конвейера, второй элемент которого запущен от динамического пользователя (который уничтожается после завершения конвейера):

```
# cat some-file.txt | systemd-run --pipe --property=DynamicUser=1 sort -u | \
    grep -i foobar > some-other-file.txt
```

- Сочетая технологию динамических пользователей и экземпляров служб¹¹⁴, можно получить гибкой и полностью автоматический механизм управления идентификаторами пользователей. Допустим, мы создаем шаблон службы `/etc/systemd/system/foobard@.service`:

```
[Service]
ExecStart=/usr/bin/myfoobarserviced
DynamicUser=1
StateDirectory=foobar/%i
```

Теперь предположим, что вы запускаете экземпляр этой службы для одного из своих клиентов:

```
# systemctl enable foobard@customerxyz.service --now
```

Готово! (Надеюсь, понятно, что эту операцию можно повторять многократно, подставляя каждый раз вместо `customerxyz` идентификаторы различных клиентов.)

¹¹⁴Прим перев.: Более подробно работа с экземплярами служб рассмотрена в главе 10.

- Сочетая технологии динамических пользователей и сокет-активации, вы легко можете получить систему, где каждое входящее соединение обслуживается экземпляром процесса, работающим в песочнице динамически выделенного UID¹¹⁵. Пример конфигурации сокета `waldo.socket`:

```
[Socket]
ListenStream=2048
Accept=yes
```

И соответствующего ему шаблона службы `waldo@.service`:

```
[Service]
ExecStart=-/usr/bin/myservicebinary
DynamicUser=yes
```

В результате, `systemd` будет слушать TCP-порт 2048, и на каждое входящее соединение создавать новый экземпляр `waldo@.service`, каждый раз с новым идентификатором пользователя, обеспечивающим его изоляцию от остальных экземпляров.

- Динамическое выделение пользователей хорошо сочетается с «системами без состояния» (*state-less systems*), то есть системами, которые запускаются с пустыми каталогами `/etc` и `/var`. Динамическим выделение UID и директивы `StateDirectory=`, `CacheDirectory=`, `LogsDirectory=` и `RuntimeDirectory=` позволяет автоматически создавать пользователя и необходимые службе каталоги непосредственно перед ее запуском.

Динамическое выделение пользователей — масштабная и глобальная концепция, и ее применение, разумеется, не ограничивается приведенным списком. Этот список — всего лишь попытка разбудить ваше воображение, дать начальный импульс к размышлениям.

Рекомендации сопровождающим пакетов

Я уверен, что для значительной доли служб, поставляемых в современных дистрибутивах, опции `DynamicUser=`, `StateDirectory=` и т.д., могут оказаться весьма полезны. Во многих случаях они позволят вообще отказаться от `post-inst` скриптов, а также конфигурационных файлов в `sysusers.d` и `tmpfiles.d`, объединив все необходимые настройки непосредственно в юнит-файле. Так что, если вы сопровождаете какой-либо пакет со службой, пожалуйста, рассмотрите возможность использования этих директив. Тем не менее, существует ряд ситуаций, когда данные директивы неэффективны или неприменимы:

1. Службы, которым нужно писать куда-либо за пределами разрешенного списка каталогов (`/run/<package>`, `/var/lib/<package>`, `/var/cache/<package>`, `/var/log/<package>`, `/var/tmp`, `/tmp`, `/dev/shm`) не совместимы с описанным подходом. Например, демон, обновляющий систему — ему, как минимум, необходим доступ на запись в `/usr`.
2. Службы, которые управляют набором процессов, запущенных от различных пользователей, например, некоторые SMTP-серверы. Если ваша служба построена по принципу *суперсервера*, то управление идентификаторами пользователей для своих процессов она должна осуществлять сама — `systemd` не должен в это вмешиваться.
3. Службы, запускаемые от рута, и вообще требующие расширенных привилегий.
4. Службы, которые должны запускаться в пространстве имен монтирования хоста (например, если служба должна создавать точки монтирования, видимые для

¹¹⁵Прим перев.: Здесь идет речь о сокет-активации «в стиле `inetd`», когда на каждое соединение создается экземпляр службы. Более подробно она обсуждается в главе 11.

всей системы). Как уже упоминалось выше, `DynamicUser=` задействует механизмы `ProtectSystem=`, `PrivateTmp=` и т.д., которые основаны на запуске службы в отдельном пространстве имен монтирования.

5. В вашем дистрибутиве пока нет свежих версий `systemd`: 232 (поддержка `DynamicUser=`) или 235 (поддержка `StateDirectory=` и аналогичных ей опций).
6. Правила создания пакетов для вашего дистрибутива не разрешают подобный подход. Уточните эти моменты в правилах и, при необходимости, обсудите данный вопрос в рассылке вашего дистрибутива.

Дополнительные замечания

Еще несколько замечаний, непосредственно относящихся к обсуждаемой теме:

1. Обратите внимание, что процесс выделения и удаления динамических пользователей никак не затрагивает `/etc/passwd`. Добавление пользователя в базу данных осуществляется при помощи NSS-модуля `glibc nss-systemd`¹¹⁶, и эта информация никогда не попадает на диск.
2. В традиционных UNIX-системах, демоны сбрасывают привилегии с рута до обычного пользователя самостоятельно, в то время как механизм динамических пользователей предполагает, что этим должен заниматься `systemd`. В версии `systemd` 235 добавлена возможность совместить механизм динамических пользователей с самостоятельным сбросом привилегий процессом службы. Для этого, включите опцию `DynamicUser=`, а в опции `User=` укажите имя пользователя, в которого ваша служба перевоплотится (`setuid()`) после инициализации. В результате, в момент запуска службы `systemd` создаст динамического пользователя с указанным именем. Далее, в директиве `ExecStart=` непосредственно перед командой запуска укажите символ «!». После этого, пользователь для службы будет создаваться, но запускаться она будет от рута — `systemd` будет считать, что служба сама сбросит полномочия в ходе инициализации. Например: `ExecStart=!/usr/bin/mydaemond`. При этом, регистрация соответствия имени и идентификатора пользователя в базе данных производится, как и прежде, при запуске службы, и поэтому процесс демона сможет без проблем преобразовать имя пользователя в UID.
3. У вас может возникнуть вопрос: почему для динамического выделения UID выбран именно диапазон 61184–65519 (в шестнадцатеричной записи 0xEF00–0xFFEF)? Он был выбран потому, что большинство дистрибутивов (например, Fedora) используют для обычных пользователей идентификаторы ниже 60000, и мы не хотим переступать эту границу. Также мы делаем небольшой отступ от 65535, так как некоторые UID вблизи этого значения имеют специальное значение (65535 часто трактуется как «некорректный» или «отсутствующий» UID, так как является представлением числа -1 в 16-битном целом типе; 65534 обычно соответствует пользователю `nobody`, и некоторые подсистемы ядра отображают в это значение «посторонние» идентификаторы¹¹⁷). И наконец, мы не хотим выходить за пределы 16-битного целого типа. Даже с распространением технологии пространств имен пользователей, контейнерам все равно не нужен весь диапазон значений, предоставляемый 32-битным целым типом, который используется в ядре Linux для UID. Там не менее, очевидно, что контейнеры должны поддерживать весь 16-битный диапазон — как минимум, из-за `nobody`. (Если честно, я считаю выделение 64 тысяч идентификаторов на контейнер оптимальным вариантом: верхние 16 бит из 32-битного UID можно использовать как идентификатор контейнера, в том время как нижние

¹¹⁶Прим. перев.: Разумеется, чтобы преобразование «имя-идентификатор» для пользователя и его группы работало корректно, это модуль должен быть указан в строках `passwd:` и `group:` файла `/etc/nsswitch.conf`. Примеры приведены на странице руководства модуля.

¹¹⁷Прим. перев.: Например, внутри пространства имен пользователей, все пользователи хоста, кроме создателя пространства, выглядят как `nobody`.

будут соответствовать идентификатору пользователя в этом контейнере... Надеюсь, вы не потеряли нить рассуждений.) И, не дожидаясь вашего вопроса: пока нет никакого способа изменить этот диапазон — его границы заданы в исходном коде. Но когда-нибудь мы обязательно добавим соответствующие настройки.

4. Вы можете поинтересоваться, что произойдет, если вы уже используете идентификаторы из диапазона 61184–65519 для других целей? `systemd` должен обработать такую ситуацию корректно, если эти идентификаторы зарегистрированы в базе данных пользователей: выбрав UID, `systemd` проверят, не используется ли он кем-то, и если он занят, выбирает другой — до тех пор, пока не найдет свободный. Проверка производится прежде всего при помощи функций NSS. Также просматриваются списки объектов IPC, и их владельцы проверяются на совпадение с нашим кандидатом. Таким образом, `systemd` избегает использования UID, занятых кем-то еще. Тем не менее, это сокращает набор доступных идентификаторов, и в худшем случае выделение пользователя может завершиться ошибкой из-за отсутствия свободных UID в рабочем диапазоне.
5. Если имя для выделяемого пользователя не указано явно, `systemd` пытается вывести его из имени службы. Однако, далеко не каждое корректное имя службы является также корректным именем пользователя, и чтобы обойти это, используется случайное имя. Возможно, вам будет удобнее задать имя пользователя вручную — используйте для этого директиву `User=`.
6. Если вы используете `User=` в сочетании с `DynamicUser=on`, но пользователь с указанным именем уже существует, то для службы будет использован именно он, а механизм динамического выделения пользователя для этой службы автоматически отключится. Таким образом, упрощается переход между статическими и динамическими пользователями: вы указываете нужное вам имя в `User=`, и пока этот пользователь существует в системной базе, система будет использовать его, и лишь при отсутствии такого пользователя он будет создаваться в динамическом режиме. Также это может быть полезно в других ситуациях, например, чтобы подготовить службы, использующие динамических пользователей, к возможности перехода на статические UID, скажем, чтобы применить к ним квоты файловой системы.
7. `systemd` всегда выделяет вместе с пользователем еще и группу, с тем же самым значением идентификатора (UID = GID).
8. Если бы ядро Linux имело механизм наподобие `shiftfs`, то есть способ смонтировать существующий каталог куда-либо с подменой UID/GID по некоторому правилу, задаваемому при монтировании, это значительно упростило бы реализацию работы `StateDirectory=` в сочетании с `DynamicUser=`, в частности, позволив отказаться от рекурсивной смены владельца, и просто монтировать каталог с хоста в пространство имен гостя, подменив владельца каталога на UID/GID службы. Однако я не питаю больших надежд на подобный вариант, так как все работы в этой области сейчас завязаны на пространство имен пользователей — механизм, который *никак не используется* в обсуждаемой технологии (есть мнение, что он создает гораздо больше проблем, чем решает, хотя вы можете с этим и не согласиться).

На сегодня все!

Учет и фильтрация IP-трафика служб

Коротко о главном: теперь `systemd` может подсчитывать и фильтровать по подсетям IP-трафик любой службы.

В недавнем вышедшем выпуске [systemd 235](#), помимо расширения поддержки динамических пользователей (рассмотренной в предыдущей статье), добавлена еще одна важная возможность: учет и фильтрация IP-трафика.

systemd и раньше поддерживал механизмы управления ресурсами, доступными юнитам: процессорным временем, дисковым вводом-выводом, потреблением памяти, количеством запускаемых процессов. А в версии 235 был добавлен контроль над еще одним видом ресурсов — IP-трафиком.

Для этого были введены три новых директивы конфигурации юнитов:

1. `IPAccounting=` — булева переменная, позволяющая включить подсчет IP-трафика, принятого и полученного службой (как количества пакетов, так и их суммарного объема в байтах).
2. `IPAddressDeny=` указывает «черный» список адресов подсетей или хостов для службы. Весь трафик, отправляемый на эти адреса процессами службы, а также полученный ими с этих адресов, будет заблокирован.
3. `IPAddressAllow=` имеет противоположный смысл — процессы службы могут обмениваться трафиком с перечисленными в этой директиве адресами, даже если они пересекаются с тем, что задано в `IPAddressDeny=` (т.е. белый список имеет приоритет над черным).

Эти три опции предоставляют интерфейс к новой функциональности ядра Linux, добавленной в выпуске 4.11, а именно, eBPF-хукам контрольных групп. Основную работу берет на себя ядро, а systemd лишь обеспечивает возможность настройки этих механизмов. Обратите внимание, что `sgroup/eBPF` никак не относятся к традиционному брандмауэру Linux — `NetFilter/iptables`. Вы можете использовать любую из этих технологий, или обе сразу, или вообще не пользоваться ни одной из них.

Учет IP-трафика

Давайте посмотрим, как работает учет трафика. Создадим тестовую службу `/etc/systemd/system/ip-accounting-test.service`:

```
[Service]
ExecStart=/usr/bin/ping 8.8.8.8
IPAccounting=yes
```

Этот простой юнит вызывает команду `ping(8)`, которая отправляет серию ICMP эхо-запросов на IP-адрес 8.8.8.8 (это адрес DNS-сервера Google; он используется в нашем примере потому, что его легко запомнить, он отовсюду доступен и отвечает на эхо-запросы; вы можете использовать любой другой адрес, отвечающий на пинги). Опция `IPAccounting=yes` включает учет IP-трафика для нашей службы.

Запустим нашу службу и посмотрим ее состояние:

```
# systemctl daemon-reload
# systemctl start ip-accounting-test
# systemctl status ip-accounting-test
ip-accounting-test.service
  Loaded: loaded (/etc/systemd/system/ip-accounting-test.service; static; vendor preset: disabled)
  Active: active (running) since Mon 2017-10-09 18:05:47 CEST; 1s ago
  Main PID: 32152 (ping)
  IP: 168B in, 168B out
  Tasks: 1 (limit: 4915)
  CGroup: /system.slice/ip-accounting-test.service
          └─32152 /usr/bin/ping 8.8.8.8

Okt 09 18:05:47 sigma systemd[1]: Started ip-accounting-test.service.
Okt 09 18:05:47 sigma ping[32152]: PING 8.8.8.8 (8.8.8.8) 56(84) bytes of data.
Okt 09 18:05:47 sigma ping[32152]: 64 bytes from 8.8.8.8: icmp_seq=1 ttl=59 time=29.2 ms
Okt 09 18:05:48 sigma ping[32152]: 64 bytes from 8.8.8.8: icmp_seq=2 ttl=59 time=28.0 ms
```

Как видим, программа `ping` сейчас работает. Судя по логу в конце вывода `systemctl status`, она как раз закончила второй цикл пинга. Однако, для нас сейчас более интересна строка выше, начинающаяся с `IP:` и содержащая счетчики принятых и полученных байт IP-трафика. Сейчас она показывает, что было отправлено 168 байт, и ровно столько же принято. Это вполне предсказуемо: ICMP эхо-запросы и эхо-ответы должны иметь одинаковый размер¹¹⁸. Обратите внимание, что эта строка появляется только для юнитов с включенной опцией `IPAccounting=`. Если она выключена, трафик юнита не подсчитывается¹¹⁹.

Подождем немного, и вызовем `systemctl status` снова:

```
# systemctl status ip-accounting-test
ip-accounting-test.service
  Loaded: loaded (/etc/systemd/system/ip-accounting-test.service; static; vendor preset: disabled)
  Active: active (running) since Mon 2017-10-09 18:05:47 CEST; 4min 28s ago
  Main PID: 32152 (ping)
    IP: 22.2K in, 22.2K out
    Tasks: 1 (limit: 4915)
  CGroup: /system.slice/ip-accounting-test.service
          └─32152 /usr/bin/ping 8.8.8.8

Okt 09 18:10:07 sigma ping[32152]: 64 bytes from 8.8.8.8: icmp_seq=260 ttl=59 time=27.7 ms
Okt 09 18:10:08 sigma ping[32152]: 64 bytes from 8.8.8.8: icmp_seq=261 ttl=59 time=28.0 ms
Okt 09 18:10:09 sigma ping[32152]: 64 bytes from 8.8.8.8: icmp_seq=262 ttl=59 time=33.8 ms
Okt 09 18:10:10 sigma ping[32152]: 64 bytes from 8.8.8.8: icmp_seq=263 ttl=59 time=48.9 ms
Okt 09 18:10:11 sigma ping[32152]: 64 bytes from 8.8.8.8: icmp_seq=264 ttl=59 time=27.2 ms
Okt 09 18:10:12 sigma ping[32152]: 64 bytes from 8.8.8.8: icmp_seq=265 ttl=59 time=27.0 ms
Okt 09 18:10:13 sigma ping[32152]: 64 bytes from 8.8.8.8: icmp_seq=266 ttl=59 time=26.8 ms
Okt 09 18:10:14 sigma ping[32152]: 64 bytes from 8.8.8.8: icmp_seq=267 ttl=59 time=27.4 ms
Okt 09 18:10:15 sigma ping[32152]: 64 bytes from 8.8.8.8: icmp_seq=268 ttl=59 time=29.7 ms
Okt 09 18:10:16 sigma ping[32152]: 64 bytes from 8.8.8.8: icmp_seq=269 ttl=59 time=27.6 ms
```

Как видно из вывода, после 269 пингов счетчики достигли значения 22 килобайт.

Несмотря на то, что `systemctl status` показывает только счетчики байт, ведется также учет и количества пакетов. Для просмотра этих значений можно использовать низкоуровневую команду `systemctl show`:

```
# systemctl show ip-accounting-test -p IPIngressBytes -p IPIngressPackets \
    -p IPEgressBytes -p IPEgressPackets
IPIngressBytes=37776
IPIngressPackets=449
IPEgressBytes=37776
IPEgressPackets=449
```

Разумеется, эта информация доступна также и через API D-Bus. Если вы хотите автоматизировать обработку таких данных, использование вызовов D-Bus будет гораздо удобнее, чем разбор вывода `systemctl show`.

Теперь остановим нашу службу:

```
# systemctl stop ip-accounting-test
```

После остановки службы, для которой включен учет потребления ресурсов, в системный журнал добавляется запись с итоговой суммой потребленных службой ресурсов. Просмотреть ее можно командой `journalctl`:

¹¹⁸Прим. перев.: Если не считать заголовков IP и ICMP, имеющих фиксированную длину, размер пакета определяется только телом сообщения (payload), которое должно полностью копироваться из запроса в ответ.

¹¹⁹Прим. перев.: Можно заметить, что для приведенного в примере юнита подсчитывается не только трафик, но и количество процессов/поток (строка `Tasks:`). Это связано с введенным лимитом на их количество `TasksMax=4915` (`TasksMax=` автоматически включает `TasksAccounting=`).

```
# journalctl -u ip-accounting-test -n 5
-- Logs begin at Thu 2016-08-18 23:09:37 CEST, end at Mon 2017-10-09 18:17:02 CEST. --
Okt 09 18:15:50 sigma ping[32152]: 64 bytes from 8.8.8.8: icmp_seq=603 ttl=59 time=26.9 ms
Okt 09 18:15:51 sigma ping[32152]: 64 bytes from 8.8.8.8: icmp_seq=604 ttl=59 time=27.2 ms
Okt 09 18:15:52 sigma systemd[1]: Stopping ip-accounting-test.service...
Okt 09 18:15:52 sigma systemd[1]: Stopped ip-accounting-test.service.
Okt 09 18:15:52 sigma systemd[1]: ip-accounting-test.service: Received 49.5K IP traffic, sent 49.5K IP traff
```

Последняя строка — и есть та самая запись о потребленных ресурсах. На самом деле, эта запись является структурной¹²⁰, и содержит поля метаданных, в которых приводится более полная информация:

```
# journalctl -u ip-accounting-test -n 1 -o verbose
-- Logs begin at Thu 2016-08-18 23:09:37 CEST, end at Mon 2017-10-09 18:18:50 CEST. --
Mon 2017-10-09 18:15:52.649028 CEST [s=89a2cc877fdf4dafb2269a7631afedad;i=14d7;b=4c7e7adcba0c45b69d612857270
  PRIORITY=6
  _BOOT_ID=4c7e7adcba0c45b69d612857270716d3
  _MACHINE_ID=e87bfd866aea4ae4b761aff06c9c3cb3
  _HOSTNAME=sigma
  SYSLOG_FACILITY=3
  SYSLOG_IDENTIFIER=systemd
  _UID=0
  _GID=0
  _TRANSPORT=journal
  _PID=1
  _COMM=systemd
  _EXE=/usr/lib/systemd/systemd
  _CAP_EFFECTIVE=3ffffffff
  _SYSTEMD_CGROUP=/init.scope
  _SYSTEMD_UNIT=init.scope
  _SYSTEMD_SLICE=-.slice
  CODE_FILE=../src/core/unit.c
  _CMDLINE=/usr/lib/systemd/systemd --switched-root --system --deserialize 25
  _SELINUX_CONTEXT=system_u:system_r:init_t:s0
  UNIT=ip-accounting-test.service
  CODE_LINE=2115
  CODE_FUNC=unit_log_resources
  MESSAGE_ID=ae8f7b866b0347b9af31fe1c80b127c0
  INVOCATION_ID=98a6e756fa9d421d8dfc82b6df06a9c3
  IP_METRIC_INGRESS_BYTES=50880
  IP_METRIC_INGRESS_PACKETS=605
  IP_METRIC_EGRESS_BYTES=50880
  IP_METRIC_EGRESS_PACKETS=605
  MESSAGE=ip-accounting-test.service: Received 49.6K IP traffic, sent 49.6K IP traffic
  _SOURCE_REALTIME_TIMESTAMP=1507565752649028
```

Нас интересуют поля `IP_METRIC_INGRESS_BYTES=`, `IP_METRIC_INGRESS_PACKETS=`, `IP_METRIC_EGRESS_BYTES=`, `IP_METRIC_EGRESS_PACKETS=`, отображающие значения соответствующих счетчиков.

Все подобные записи имеют один и тот же идентификатор типа сообщения (`message ID`), при помощи которого их можно легко найти в журнале (`ae8f7b866b0347b9af31fe1c80b127c0`). Вызовем `journalctl` с указанием этого идентификатора, добавив также ключ `-u`, который ограничит выборку только записями нашей службы:

```
# journalctl -u ip-accounting-test MESSAGE_ID=ae8f7b866b0347b9af31fe1c80b127c0
-- Logs begin at Thu 2016-08-18 23:09:37 CEST, end at Mon 2017-10-09 18:25:27 CEST. --
Okt 09 18:15:52 sigma systemd[1]: ip-accounting-test.service: Received 49.6K IP traffic, sent 49.6K IP traff
```

Приведенный вывод содержит пока только одну запись, так как мы запускали службу всего один раз. Каждый последующий запуск будет добавлять новую запись.

¹²⁰Прим. перев.: Подробнее о структуре записей Journal можно почитать в разделе 17.6.

Учет IP-трафика доступен также и для одноразовых служб, запускаемых через `systemd-run`, что позволяет прозрачно выполнять различные команды как службы `systemd`, и подсчитывать их трафик:

```
# systemd-run -p IPAccounting=yes --wait \  
    wget https://cfp.all-systems-go.io/en/ASG2017/public/schedule/2.pdf  
Running as unit: run-u2761.service  
Finished with result: success  
Main processes terminated with: code=exited/status=0  
Service runtime: 878ms  
IP traffic received: 231.0K  
IP traffic sent: 3.7K
```

Мы использовали `wget`, чтобы загрузить [PDF с расписанием второго дня](#) нашей любимой конференции [All Systems Go! 2017](#). Эта операция потребовала 231 килобайт входящего трафика и 4 килобайта исходящего. Особого внимания заслуживают параметры командной строки, с которыми мы вызвали `systemd-run`. Первый, `-p IPAccounting=yes`, включает режим учета IP-трафика для нашей одноразовой службы (аналогично тому, как та же строка в юнит-файле делает это для обычной службы). Второй параметр, `--wait`, приказывает `systemd-run` дожидаться завершения работы созданной службы, после чего вывести информацию о результатах ее работы, включая интересующую нас статистику по IP-трафику (разумеется, только при условии, что мы включили его подсчет).

Еще интереснее получается, если применять учет IP-трафика для *интерактивных* одноразовых служб:

```
# systemd-run -p IPAccounting=1 -t /bin/sh  
Running as unit: run-u2779.service  
Press ~] three times within 1s to disconnect TTY.  
sh-4.4# dnf update  
...  
sh-4.4# dnf install firefox  
...  
sh-4.4# exit  
Finished with result: success  
Main processes terminated with: code=exited/status=0  
Service runtime: 5.297s  
IP traffic received: ...B  
IP traffic sent: ...B
```

Мы использовали ключ `systemd-run --pty` (в краткой форме `-t`), который открывает интерактивное псевдо-терминальное подключение к запускаемой службе — в нашем случае это оболочка `/bin/sh`. Таким образом, мы получаем полноценную оболочку, с управлением фоновыми заданиями и прочими возможностями, запущенную в рамках службы `systemd`. Так как для этой службы включен учет IP-трафика, при ее завершении мы получаем суммарную статистику по входящим и исходящим данным. (Для краткости я не стал показывать вывод целиком, а оставил только ключевые моменты. Если вам хочется посмотреть на полный вывод — попробуйте запустить что-нибудь сами.)

Иногда бывает нужно включить учет IP-трафика для юнита, который уже запущен. Это можно сделать командой

```
systemctl set-property foobar.service IPAccounting=yes
```

Обратите внимание, что она не имеет обратной силы: трафик учитывается только с момента выполнения этой команды. Аналогичной командой (`IPAccounting=no`) учет трафика можно отключить.

Если вы хотите подсчитывать IP-трафик сразу для всех служб, вам вовсе не обязательно добавлять `IPAccounting=yes` во все юнит-файлы. Достаточно задействовать глобальную опцию `DefaultIPAccounting=` в файле `/etc/systemd/system.conf`.

Фильтрация IP-трафика

От учета трафика переходим к его фильтрации. Рассмотрим поподробнее механизм контрольных списков IP-адресов (IP ACL), добавленный в systemd 235. Как уже упоминалось выше, за него отвечают две директивы: `IPAddressAllow=` и `IPAddressDeny=`. Работают они следующим образом:

1. Если адрес источника входящего пакета, либо адрес назначения исходящего пакета соответствует какому-либо из адресов хостов или попадает в одну из подсетей, указанных в `IPAddressAllow=`, пакет проходит свободно.
2. Если пакет не соответствует `IPAddressAllow=`, но подпадает под `IPAddressDeny=` (аналогично, проверяются исходные адреса входящих пакетов и адреса назначения исходящих), он блокируется.
3. Если пакет не подпадает ни под одну из этих директив, он тоже проходит свободно.

Иными словами, `IPAddressDeny=` является черным списком, но белый список `IPAddressAllow=` имеет над ним приоритет.

Посмотрим, как это работает. Поменяем предыдущий пример, чтобы получить интерактивную оболочку, работающую в рамках одноразовой службы с настроенными контрольными списками IP-адресов:

```
# systemd-run -p IPAddressDeny=any -p IPAddressAllow=8.8.8.8 \  
-p IPAddressAllow=127.0.0.0/8 -t /bin/sh  
Running as unit: run-u2850.service  
Press ^] three times within 1s to disconnect TTY.  
sh-4.4# ping 8.8.8.8 -c1  
PING 8.8.8.8 (8.8.8.8) 56(84) bytes of data.  
64 bytes from 8.8.8.8: icmp_seq=1 ttl=59 time=27.9 ms  
  
--- 8.8.8.8 ping statistics ---  
1 packets transmitted, 1 received, 0% packet loss, time 0ms  
rtt min/avg/max/mdev = 27.957/27.957/27.957/0.000 ms  
sh-4.4# ping 8.8.4.4 -c1  
PING 8.8.4.4 (8.8.4.4) 56(84) bytes of data.  
ping: sendmsg: Operation not permitted  
^C  
--- 8.8.4.4 ping statistics ---  
1 packets transmitted, 0 received, 100% packet loss, time 0ms  
sh-4.4# ping 127.0.0.2 -c1  
PING 127.0.0.1 (127.0.0.2) 56(84) bytes of data.  
64 bytes from 127.0.0.2: icmp_seq=1 ttl=64 time=0.116 ms  
  
--- 127.0.0.2 ping statistics ---  
1 packets transmitted, 1 received, 0% packet loss, time 0ms  
rtt min/avg/max/mdev = 0.116/0.116/0.116/0.000 ms  
sh-4.4# exit
```

Мы задали `IPAddressDeny=any`, чтобы работать по схеме белого списка: для нашей службы разрешен обмен трафика только с теми адресами, которые перечислены в `IPAddressAllow=`. А это, в данном примере, уже знакомый нам адрес 8.8.8.8 (указанный без маски подсети, что соответствует адресу хоста, т.е. маске /32), и подсеть 127.0.0.0/8. Таким образом, служба может взаимодействовать только с одним из DNS-серверов Google и адресами обратной петли, и больше ни с кем. Команды, запущенные в оболочке, иллюстрируют это. Сначала мы пытаемся пинговать 8.8.8.8 — успешно. Затем, мы пробуем пропинговать 8.8.4.4 (это другой DNS-сервер Google, не входящий в наш белый список), и видим ошибку «Operation not permitted». Наконец, мы пингуем адрес 127.0.0.2 (принадлежащий подсети обратной петли), и снова успешно.

Обратите внимание на специальное значение `any`, которое мы использовали в примере выше. Оно является сокращением для «`0.0.0.0/0 ::/0`», то есть *всех* возможных адресов IPv4 и IPv6. Есть и другие сокращения. Например, вместо `127.0.0.0/8` мы могли бы указать более понятное обозначение `localhost`, соответствующее «`127.0.0.0/8 ::1/128`», то есть IPv4 и IPv6 подсетям обратной петли.

Возможность настраивать контрольные списки IP-адресов независимо для каждого юнита — это уже неплохо. Однако, для большего удобства, существует возможность задавать их для целых групп юнитов, или даже для всей системы. Это можно сделать при помощи юнитов `.slice` (для тех, кто не очень хорошо разбирается в `systemd`: slice-юниты предоставляют возможность организовать юниты в группы для управления системными ресурсами¹²¹): контрольные списки адресов юнита формируются путем объединения списков самого юнита, а также всех slice-юнитов, в которые он вложен.

По умолчанию, все системные службы помещаются в группу `system.slice`, которая, в свою очередь, входит в корневую группу `-.slice`. Соответственно, при помощи любой из этих групп можно заблокировать *все* системные службы разом. Разница состоит в том, что ограничения `system.slice` будут применяться только к системным службам, а `-.slice` действует сразу на все процессы системы, включая не только службы, но и процессы пользовательских сеансов (объединенные в группу `user.slice`, которая находится в корневой группе).

Воспользуемся этим:

```
# systemctl set-property system.slice IPAddressDeny=any IPAddressAllow=localhost
# systemctl set-property apache.service IPAddressAllow=10.0.0.0/8
```

Две приведенные команды имеют очень мощный эффект: сначала мы отключаем взаимодействие через IP (кроме обратной петли) для всех служб системы, а затем добавляем белый список `10.0.0.0/8` (скажем, это локальная сеть вашей компании) только для службы веб-сервера Apache.

Практическое применение

Вот несколько идей по практическому применению рассмотренных возможностей:

1. Механизм контрольных списков IP адресов можно рассматривать как современную альтернативу классической технологии `TCP Wrapper`. Однако, в отличие от нее, контрольные списки применяются сразу ко всем IP-сокетам службы, и не требуют никакой поддержки со стороны приложения. С другой стороны, `TCP Wrapper` предоставляет ряд опций, которые отсутствуют в нашей схеме, в частности, возможность указания DNS-имен вместо IP-адресов (лично я считаю это весьма сомнительной опцией — выполнять сетевые операции (разрешение имен), причем незащищенные, для того, чтобы ограничить работу сети).
2. В некоторых аспектах наши механизмы могут заменить или хотя бы дополнить классический брандмауэр Linux — `NetFilter/iptables`. На текущий момент, контрольные списки IP-адресов `systemd` предоставляют гораздо меньше возможностей, чем `NetFilter`, однако имеют перед ним серьезное преимущество: они работают на уровне *приложений*, а не абстрактных TCP/UDP-портов. Классические брандмауэры, в частности, `NetFilter`, вынуждены делать предположение о принадлежности пакетов тем или иным службам, основываясь только на номерах

¹²¹Прим. перев.: Чуть более развернутое определение: slice-юниты соответствуют «промежуточным» уровням иерархии контрольных групп, которые не содержат непосредственно процессов, а объединяют другие `.slice`, `.service` и `.scope`-юниты. Последние два типа юнитов уже соответствуют контрольным группам процессов, причем группы `.service` создаются автоматически при запуске служб `systemd`, а `.scope` — это специально сформированные контрольные группы процессов, запущенных другими программами, что позволяет «перекидывать» отдельные процессы в другие точки иерархии контрольных групп. В частности, процессы, созданные в рамках пользовательских сессий, а также деревья процессов контейнеров, автоматически перемещаются из контрольных групп породивших их служб (например, `kdm.service`) в специальные `scope`-юниты (например, `session-1.scope`).

портов¹²², однако на практике порты часто выбираются динамически. Например, клиенты BitTorrent для передачи данных могут использовать любые удобные им порты, что крайне затрудняет корректную выборку таких пакетов. В случае с контрольными списками все просто: достаточно настроить разрешения для юнита службы BitTorrent, и дело в шляпе.

Замечу что, по большей части, сравнение NetFilter и контрольных списков systemd — это сравнение теплого с мягким. Контрольные списки ориентированы исключительно на конечные машины (клиенты и серверы), так как работают только с локальными службами. В то время как NetFilter отлично работает на промежуточных маршрутизаторах, передающих чистый IP-трафик, никак не привязанный к их локальным процессам.

3. Контрольные списки предоставляют простой способ обеспечить безопасность служб «из коробки». Например, если вы сопровождаете пакет службы, которая не должна требовать доступа к сети, добавьте в ее юнит-файл `IPAddressDeny=any` (и, при необходимости, `IPAddressAllow=localhost`), и она будет помещена в песочницу, из которой не сможет сбежать. В systemd подобные ограничения уже введены по умолчанию для целого ряда служб, например, для системного журнала `systemd-journald.service`¹²³, менеджера логинов `systemd-logind` и обработчика дампов памяти `systemd-coredump@.service`, так как мы знаем, что этим службам ни при каких условиях не нужна сеть.
4. Механизм контрольных списков IP-адресов можно сочетать с одноразовыми службами, что позволяет быстро и эффективно изолировать произвольные команды, и даже включать их в конвейеры. Предположим, что мы не доверяем нашему бинарному `curl` (может быть, он модифицирован хакером, и обращается на подконтрольные ему серверы?), но все равно хотим использовать его для загрузки [презентации с моего последнего доклада по casync](#), чтобы распечатать их, и при этом быть уверенными, что он не будет связываться ни с кем, кроме нужного нам сервера (а чтобы было еще интереснее и безопаснее, включим описанный в предыдущей главе механизм динамических пользователей):

```
# systemd-resolve 0pointer.de
0pointer.de: 85.214.157.71
                2a01:238:43ed:c300:10c3:bcf3:3266:da74
-- Information acquired via protocol DNS in 2.8ms.
-- Data is authenticated: no
# systemd-run --pipe -p IPAddressDeny=any \
                -p IPAddressAllow=85.214.157.71 \
                -p IPAddressAllow=2a01:238:43ed:c300:10c3:bcf3:3266:da74 \
                -p DynamicUser=yes \
                curl http://0pointer.de/public/casync-kinvolk2017.pdf | lp
```

Как и в прошлой главе, в силу масштабности обсуждаемой концепции, приведенный список применений не претендует на полноту, а лишь является затравкой для вашего воображения.

Рекомендации сопровождающим пакетов

Механизмы учета и контроля IP-трафика ориентированы прежде всего на системных администраторов, а не разработчиков. Тем не менее, как я уже заметил выше, для

¹²²Прим. перев.: В ядрах Linux 2.4-2.6 существовала теоретическая возможность выборки пакетов по идентификатору процесса (`-m owner --pid-owner PID`), однако она была удалена в выпуске 2.6.14. Много лет спустя, в ядре 3.14 был добавлен модуль `netfilter xt_cgroup`, позволяющий выбирать пакеты по индексу класса, присвоенного `sgroup`-контроллером `net_class`. Кстати, поддержка соответствующей опции `NetClass=` существовала в `systemd` с 227 по 229 версию, но была удалена при миграции на `cgroup v2`, где такого контроллера уже нет, зато есть вышеописанный механизм контрольных списков.

¹²³Прим. перев.: Знаменитый HTTP-сервер вынесен в отдельную службу `systemd-journal-gatewayd(8)`, которая **отключена** по умолчанию, а в ряде дистрибутивов вообще поставляется в отдельном опциональном пакете (в частности, RHEL/CentOS, Debian/Ubuntu).

служб, которые никак и никогда не требуют сети, целесообразно добавлять настройку `IPAddressDeny=any` (и опционально `IPAddressAllow=localhost`), чтобы повысить безопасность системы сразу «из коробки».

Для специализированных дистрибутивов, ориентированных на максимальную безопасность, можно предложить более радикальный подход: добавить `IPAddressDeny=any` сразу в `-.slice` или `system.slice`, чтобы администратор, когда ему нужно выпустить в сеть какую-либо службу, вручную командовал `systemctl set-property ... IPAddressAllow=...`. Разумеется, это вариант для тех дистрибутивов, которые не боятся ломать обратную совместимость.

Дополнительные замечания

И еще несколько замечаний:

1. Описанные механизмы учета и фильтрации IP-трафика можно совмещать с сокет-активацией. При этом, целесообразно настраивать их и для сокета, и для активируемой им службы, так как это разные юниты с независимыми настройками. Обратите внимание, что учет и фильтрация трафика, настроенные для сокет-юнита, применяются для всех сокетов, созданных в рамках этого юнита, включая переданные активированным службам. Как следствие, трафик учитывается именно для юнита сокета, а не службы. Тот факт, что для сокетов, созданных под эгидой сокет-юнита (т.е. при сокет-активации), и для сокетов, созданных из кода службы, используются *разные* контрольные списки, открывает весьма интересные возможности. Например, можно настроить относительно свободный доступ для сокета, при этом полностью запретив IP-трафик для активируемых им служб — в результате эти службы смогут взаимодействовать с внешним миром только через переданный им при активации сокет.
2. Учет и фильтрация IP-трафика работают только с IP-сокетами. В частности, сокеты типа `AF_PACKET` (так называемые «сырые сокеты») под эти ограничения не подпадают¹²⁴. Поэтому целесообразно дополнять контрольные списки опцией `RestrictAddressFamilies=AF_UNIX AF_INET AF_INET6`.
3. Вы также можете поинтересоваться, может ли журнальная запись о потребленных ресурсах, а также `systemd-run` с ключом `--wait` показывать статистику по другим видам ресурсов? Ответ утвердительный: например, если вы зададите для юнита `CPUAccounting=yes`, то журнальная запись и `systemd-run` будут показывать статистику по потребленному процессорному времени. В ближайшее время мы планируем добавить аналогичную поддержку для `IOAccounting=`.
4. Обратите внимание, что учет и фильтрация IP-трафика сами могут потреблять определенные ресурсы. Чтобы эти функции работали, `systemd` вставляет в путь прохождения IP-трафика специальную eBPF-программу. Однако, в последних версиях ядра выполнение eBPF очень неплохо оптимизировано, и работа по оптимизации продолжается, так что серьезных проблем с производительностью ожидать не стоит.
5. Учет трафика не является иерархическим, то есть, счетчики slice-юнита не суммируют результаты вложенных в него юнитов. Это одна из задач, которую мы намерены решить, однако она требует доработки кода ядра.
6. У вас может возникнуть вопрос — как соотносятся механизмы `PrivateNetwork=yes` и `IPAccessDeny=any`? На первый взгляд, они работают одинаково: блокируют доступ к сети для службы. Однако, при более пристальном рассмотрении обнаруживается ряд отличий. Механизм `PrivateNetwork=` реализован на основе сетевых

¹²⁴Прим. перев.: Они также «обходят» и `netfilter` — именно поэтому бесполезно пытаться фильтровать трафик DHCP-клиента с той же машины.

пространств имен ядра Linux. Он полностью изолирует от хоста все сетевые операции службы — не только обычные IP-сокеты, но и другие виды сетевого взаимодействия (в частности, сырые сокеты). Для этого он создает службе отдельное сетевое пространство имен, в котором она может общаться только сама с собой (через свой личный интерфейс обратной петли). Однако, при использовании опции `JoinsNamespaceOf=` в ту же песочницу могут быть помещены другие службы — в результате, службы из одной песочницы смогут взаимодействовать между собой, но при этом будут изолированы от остального мира. С другой стороны, `IPAddressAllow=` и `IPAccessDeny=` действуют не столь жестко. Прежде всего, они работают только с IP-сокетами и IP-адресами. Кроме того, служба с выключенным режимом `PrivateNetwork=`, но включенным `IPAccessDeny=any`, все равно способна прочитать список сетевых интерфейсов хоста и узнать присвоенные им адреса, хотя и не сможет пересылать данные по протоколу IP. В то же время, с включенным режимом `PrivateNetwork=` служба может видеть только свой личный интерфейс `lo`. Короче: в зависимости от конкретной ситуации, для изоляции службы можно использовать тот или иной вариант, или оба сразу, или вообще ни одного. Когда это возможно, для максимальной безопасности лучше использовать обе эти защиты, как это делаем мы для всех основных служб из состава `systemd`.

На этом пока все.

А FAQ (часто задаваемые вопросы)¹²⁵

Также смотрите статью [Tips & Tricks](#)¹²⁶.

Вопрос: Как изменить текущий уровень выполнения (runlevel)?

Ответ: В `systemd` уровни выполнения оформлены в виде целевых состояний (target units). Команда перехода в заданное состояние выглядит так:

```
# systemctl isolate runlevel5.target
```

Отметим, что концепция уровней исполнения уже порядком устарела, и вместо номеров уровней удобнее оперировать более выразительными именами целевых состояний:

```
# systemctl isolate graphical.target
```

Подобные команды изменят только текущий уровень выполнения. Их действие никак не повлияет на последующие загрузки системы.

Вопрос: Как изменить уровень выполнения, в который система грузится по умолчанию?

Ответ: Соответствующее целевое состояние задается символьной ссылкой `/etc/systemd/system/default.target`. Для его смены достаточно перезаписать эту ссылку¹²⁷. Например, так:

```
# ln -sf /usr/lib/systemd/system/multi-user.target /etc/systemd/system/default.target
```

или так

```
# ln -sf /usr/lib/systemd/system/graphical.target /etc/systemd/system/default.target
```

Вопрос: Как узнать текущий уровень выполнения?

Ответ: В один и тот же момент времени может быть активно несколько целевых состояний, поэтому понятие текущего уровня выполнения (единственного и однозначно определенного) применимо уже далеко не всегда. Узнать, какие состояния сейчас активны, вы можете при помощи команды

```
$ systemctl list-units --type=target
```

Если вам нужна именно одна цифра, вы можете воспользоваться классической командой `runlevel`, однако, по изложенным выше причинам, она далеко не всегда способна адекватно оценить текущую ситуацию.

Вопрос: Я изменил service-файл, но пакетный менеджер перезаписывает мои изменения при каждом обновлении. Что делать?

Ответ: Наиболее разумным решением будет скопировать исходный файл из `/usr/lib/systemd/system` в `/etc/systemd/system` и внести изменения в полученную копию. Юнит-файлы из `/etc` защищены от посягательств пакетного менеджера, и при этом имеют приоритет над файлами из `/usr`. Если же вы захотите вернуться к настройкам по умолчанию, достаточно будет просто удалить или переименовать созданный вами файл¹²⁸.

¹²⁵Прим. перев.: Перевод статьи «[Frequently Asked Questions](#)» с официального сайта проекта, по состоянию на 2013-05-26 08:17:02 (коммит 8cf2b).

¹²⁶Прим. перев.: На самом деле, смотреть там особо не на что. Большая часть приведенного там материала вполне соответствует формату FAQ, и поэтому была перенесена в текущую главу. Исключением являются только «sgroup tree» и «ps with sgroups», но они подробно рассмотрены в главе 2.

¹²⁷Прим. перев.: Начиная с `systemd 205`, в программу `systemctl` были добавлены команды `set-default` и `get-default`, упрощающие подобные действия, например, `systemctl set-default multi-user.target`.

¹²⁸Прим. перев.: Начиная с версии `systemd 198`, необязательно копировать файл целиком, если вы хотите добавить/изменить отдельные настройки. Можно просто создать в `/etc/systemd/system/` подкаталог `имя_службы.service.d`, а в нем — файл с именем, заканчивающимся на `.conf`, например, `mysettings.conf`, после чего вписать в этот файл необходимые настройки. Они будут применены «поверх» настроек из штатного файла (расположенного в `/usr`).

Вопрос: Служба `foo.service` по умолчанию запускается только при поступлении входящего соединения (или подключении определенного оборудования, или при появлении заданного файла). Как сделать так, чтобы она запускалась сразу при загрузке?

Ответ: Достаточно поместить символическую ссылку на соответствующий `service`-файл в каталог `multi-user.target.wants/`.

```
# ln -sf /usr/lib/systemd/system/foobar.service \
> /etc/systemd/system/multi-user.target.wants/foobar.service
# systemctl daemon-reload
```

В это каталоге содержатся ссылки на все юниты, которые нужно запустить в состоянии `multi-user` (эквивалент уровня выполнения 3, т.е. загрузка с запуском всех служб, но не графического интерфейса). Так как цель `graphical.target` тянет по зависимости и `multi-user.target`, то ваша служба будет запущена также и при «графических» загрузках.

Вопрос: Я хочу запустить еще один экземпляр `getty`, как это сделать?

Ответ: Для запуска `getty` на последовательном порту, достаточно запустить соответствующий экземпляр службы `serial-getty@.service`:

```
# systemctl start serial-getty@ttyS2.service
```

Чтобы обеспечить автоматически запуск `getty` на этом порту при каждой загрузке, нужно поместить соответствующую символическую ссылку в каталог `getty.target.wants/`¹²⁹:

```
# ln -s /usr/lib/systemd/system/serial-getty@.service \
> /etc/systemd/system/getty.target.wants/serial-getty@ttyS2.service
# systemctl daemon-reload
```

Обратите внимание, что на виртуальных терминалах `getty` запускаются автоматически, как только пользователь переключается на данный терминал. Максимальное количество таких терминалов задается параметром `NAutoVTs=` в файле `logind.conf(7)`. Также см. главу 16.

Вопрос: Как узнать, какой службе принадлежат вот эти процессы?

Ответ: Для этого можно воспользоваться командой `ps`

```
$ alias psc='ps xawf -eo pid,user,cgroup,args'
$ psc
PID USER      CGROUP                COMMAND
...
   1 root      name=systemd:/systemd-1 /bin/systemd systemd.log_target=kmsg systemd.log_level=de
 415 root      name=systemd:/systemd-1/sysinit.service /sbin/udevd -d
 928 root      name=systemd:/systemd-1/atd.service /usr/sbin/atd -f
 930 root      name=systemd:/systemd-1/ntpd.service /usr/sbin/ntpd -n
 932 root      name=systemd:/systemd-1/crond.service /usr/sbin/crond -n
 935 root      name=systemd:/systemd-1/auditd.service /sbin/auditd -n
 943 root      name=systemd:/systemd-1/auditd.service \_ /sbin/audispd
 964 root      name=systemd:/systemd-1/auditd.service \_ /usr/sbin/sedispach
 937 root      name=systemd:/systemd-1/acpid.service /usr/sbin/acpid -f
 941 rpc      name=systemd:/systemd-1/rpcbind.service /sbin/rpcbind -f
 944 root      name=systemd:/systemd-1/rsyslog.service /sbin/rsyslogd -n -c 4
 947 root      name=systemd:/systemd-1/systemd-logger.service /lib/systemd/systemd-logger
 950 root      name=systemd:/systemd-1/cups.service /usr/sbin/cupsd -f
 955 dbus     name=systemd:/systemd-1/messagebus.service /bin/dbus-daemon --system --address=systemd: --nof
 969 root      name=systemd:/systemd-1/getty@.service/tty6 /sbin/mingetty tty6
 970 root      name=systemd:/systemd-1/getty@.service/tty5 /sbin/mingetty tty5
 971 root      name=systemd:/systemd-1/getty@.service/tty1 /sbin/mingetty tty1
```

¹²⁹Прим. перев.: Приведенная в оригинале команда `systemctl enable serial-getty@ttyS2.service` в `systemd` версии 209 и ниже работать не будет. Подробнее см. примечание 65.

```

973 root    name=systemd:/systemd-1/getty@.service/tty4 /sbin/mingetty tty4
974 root    name=systemd:/user/lennart/2          login -- lennart
1824 lennart name=systemd:/user/lennart/2          \_ -bash
975 root    name=systemd:/systemd-1/getty@.service/tty3 /sbin/mingetty tty3
988 root    name=systemd:/systemd-1/polkitd.service /usr/libexec/polkit-1/polkitd
994 rtkit   name=systemd:/systemd-1/rtkit-daemon.service /usr/libexec/rtkit-daemon
...

```

или просто посмотреть `/proc/$PID/cgroup`. Также см. главу 2.

Вопрос: Почему вы не используете `inotify` для отслеживания изменений в настройках?

Ответ: К сожалению, реализовать это весьма проблематично. За подробностями вы можете обратиться к обсуждению в [bugzilla](#)¹³⁰.

Вопрос: У моей службы есть как штатный `service`-файл, так и `init`-скрипт (с тем же именем). Какой из этих файлов главнее — `/usr/lib/systemd/system/foobar.service` или `/etc/init.d/foobar`?

Ответ: При наличии обоих этих файлов, приоритет всегда отдается штатному `service`-файлу, независимо от того, включен ли запуск соответствующего скрипта (например, через `chkconfig`) или нет. Обратите внимание, что отключение юнит-файла (в частности, через `systemctl disable`) приведет к полному отключению службы, даже если `init`-скрипт при этом будет включен.

Вопрос: Как заставить `journalctl` выводить полные (не сокращенные) строки?

Ответ: Используйте

```
# journalctl --full
```

Вопрос: Моя служба пытается получить приоритет реального времени, но получает ошибку `EPERM`, хотя обладает всеми необходимыми привилегиями. А без вашего `systemd` все работало!

Ответ: По умолчанию, `systemd` помещает каждую системную службу в собственную контрольную группу контроллера `cpu`. К сожалению, из-за существующего в реализации данного контроллера ограничения, это приводит к невозможности получения приоритета реального времени для служб. Подробное обсуждение этой проблемы и методы ее решения смотрите в приложении G.

Вопрос: Моя служба настроена на запуск после `network.target`, однако она все равно запускается раньше, чем появляется сеть. Почему?

Ответ: Это длинная история. См. приложение F.

Вопрос: `systemd` монтирует в `/tmp tmpfs`. Как это отключить?

Ответ: Это тоже долгая история. См. приложение E.

Вопрос: Как просмотреть список работающих служб?

Ответ: Запустите команду `systemctl` без параметров¹³¹:

```

$ systemctl
UNIT                                LOAD    ACTIVE SUB    JOB DESCRIPTION
accounts-daemon.service             loaded active running    Accounts Service
atd.service                          loaded active running    Job spooling tools
avahi-daemon.service                loaded active running    Avahi mDNS/DNS-SD Stack
bluetooth.service                   loaded active running    Bluetooth Manager

```

¹³⁰Прим. перев.: Вкратце, суть проблемы состоит в том, что перемещение/переименование файла не является атомарной операцией, а состоит из удаления одного файла и создания другого. Эти операции могут следовать в произвольном порядке с некоторым интервалом, что может привести к временному исчезновению службы, либо к появлению двух конфликтующих служб.

¹³¹Прим. перев.: `systemctl` без параметров выведет состояние всех активных юнитов (в т.ч. сокетов, устройств, точек монтирования, целевых состояний, таймеров, и т.д.). Чтобы ограничиться только службами, добавьте параметр `-t service`. Чтобы вывести все службы/юниты, включая неактивные, добавьте параметр `-a`.

```
colord-sane.service    loaded active running    Daemon for monitoring attached scanners and registering
colord.service        loaded active running    Manage, Install and Generate Color Profiles
crond.service         loaded active running    Command Scheduler
cups.service          loaded active running    CUPS Printing Service
dbus.service          loaded active running    D-Bus System Message Bus
...
```

Вопрос: Как узнать текущее состояние службы?

Ответ: Воспользуйтесь командой `systemctl status`:

```
$ systemctl status udisks2.service
udisks2.service - Storage Daemon
   Loaded: loaded (/usr/lib/systemd/system/udisks2.service; static)
   Active: active (running) since Wed, 27 Jun 2012 20:49:25 +0200; 1 day and 1h ago
 Main PID: 615 (udisksd)
   CGroup: name=systemd:/system/udisks2.service
           └─ 615 /usr/lib/udisks2/udisksd --no-debug

Jun 27 20:49:25 epsilon udisksd[615]: udisks daemon version 1.94.0 starting
Jun 27 20:49:25 epsilon udisksd[615]: Acquired the name org.freedesktop.UDisks2 on the system message bus
```

Вопрос: Как отследить зависимости между юнитами?

Ответ: Допустим, вы хотите узнать, какие юниты будут запущены при активации `multi-user.target`:

```
$ systemctl show -p "Wants" multi-user.target
Wants=rc-local.service avahi-daemon.service rpcbind.service
NetworkManager.service acpid.service dbus.service atd.service crond.service
auditd.service ntpd.service udisks.service bluetooth.service cups.service
wpa_supplicant.service getty.target modem-manager.service portreserve.service
abrt.service yum-updatesd.service upowerd.service test-first.service
pcscd.service rsyslog.service haldaemon.service remote-fs.target
plymouth-quit.service systemd-update-utmp-runlevel.service sendmail.service
lvm2-monitor.service cpuspeed.service udev-post.service mdmonitor.service
iscsid.service livesys.service livesys-late.service irqbalance.service
iscsi.service netfs.service
```

Вместо `Wants` вы можете использовать другие типы зависимостей: `WantedBy`, `Requires`, `RequiredBy`, `Conflicts`, `ConflictedBy`, `Before`, `After`¹³²¹³³.

Вопрос: Как узнать, что будет запущено при загрузке системы в заданное целевое состояние?

Ответ: Вы можете заставить `systemd` рассчитать «начальную» транзакцию, определяющую алгоритм загрузки в заданное состояние `foobar.target`:

```
$ systemd --test --system --unit=foobar.target
```

Обратите внимание, что эта команда предназначена исключительно для отладки, и ее работа на самом деле не ограничивается расчетом транзакции, поэтому не стоит вызывать ее в скриптах.

¹³²Прим. перев.: Разница между `*s` и `*edBy` — в направлении зависимости. Например, если цель `multi-user.target` запрашивает запуск (`Wants`) службы `rc-local.service`, то она будет указана в свойстве `WantedBy` этой службы. Интересно, что свойство `ConflictedBy` существует, однако задать его напрямую нельзя — оно формируется только на основании `Conflicts` других служб. Что касается разницы между `Wants` и `Requires`, то первое является пожеланием, а второе требованием. Если требуемый (`required`) юнит не сможет запуститься, то не запустится и сам зависящий от него юнит. Пожелания (`wants`) не накладывают такого ограничения — будет сделана попытка запуска зависимого юнита, но результат ее никак не отразится на зависящем юните. А указания *порядка* запуска (`Before`, `After`) вообще не формируют зависимости и работают только тогда, когда такие зависимости, прямо или косвенно, уже заданы.

¹³³Прим. перев. Отметим, что начиная с `systemd 198`, `systemctl` поддерживает специализированную команду `list-dependencies`, позволяющую отследить прямые и косвенные («рекурсивные») зависимости заданного юнита.

В Диагностика неполадок¹³⁴

В.1 Диагностика проблем с загрузкой

Если система зависает во время загрузки, прежде всего нужно разобраться, на каком этапе возникает проблема — до запуска `systemd`, или после.

Для этого надо удалить из командной строки ядра параметры `quiet` и `rhgb`. При работе `systemd` на экран выводятся примерно такие сообщения:

```
Welcome to Fedora ВЕРСИЯ (имя релиза)!  
Starting название...  
[ OK ] Stared название...
```

(Пример можно посмотреть на [скриншоте](#).)

Если у вас есть доступ к оболочке, это значительно упрощает диагностику и решение проблем. В том случае, когда до приглашения входа в систему дело так и не доходит, попробуйте переключиться на другую виртуальную консоль, нажав `CTRL-ALT-F<цифра>`. При проблемах, связанных с запуском графического сервера, может возникнуть ситуация, когда на первой консоли (`tty1`) приглашение ко входу отсутствует, но все остальные консоли при этом работают нормально.

Если ни на одной из виртуальных консолей приглашение так и не появилось — попробуйте выждать еще *не менее 5 минут*. Только после этого можно будет считать процесс загрузки окончательно зависшим. Если подвисание обусловлено всего лишь сбоем при запуске какой-то службы, то после истечения тайм-аута проблемная служба будет убита, и загрузка продолжится. Другой вариант — отсутствует устройство, которое должно быть смонтировано для нормальной работы системы. В этом случае загрузка будет остановлена, и система перейдет в *аварийный режим (emergency mode)*.

В.1.1 Если у вас нет доступа к оболочке

Если система не предоставила вам ни нормального приглашения, ни аварийной оболочки, то для диагностики проблемы нужно выполнить следующие действия:

- Попробуйте перезагрузить систему, нажав `CTRL-ALT-DEL`. Если после этого перезагрузки не произойдет — обязательно укажите данное обстоятельство, когда будете писать отчет об ошибке (`bugreport`). Чтобы перезагрузить систему, вы можете воспользоваться `SysRq` или «аппаратным» методом.
- При следующей загрузке попробуйте воспользоваться некоторыми нижеописанными стратегиями.

Вывод диагностических сообщений на последовательную консоль Если у вас под рукой есть терминал последовательной консоли, либо дело происходит в виртуальной машине (в частности, `virt-manager` позволяет просматривать вывод виртуальной машины на последовательную консоль: меню Вид (View) ⇒ Текстовые консоли (Text Consoles); также можно воспользоваться командой `virsh console ИМЯ`), вы можете попросить `systemd` выводить на эту консоль подробную отладочную информацию о ходе загрузки, добавив к параметрам ядра следующие аргументы¹³⁵:

```
systemd.log_level=debug systemd.log_target=console console=ttyS0,38400 console=tty1
```

¹³⁴Прим. перев.: Перевод статьи «[Debugging systemd Problems](#)» с официального сайта проекта, по состоянию на 2017-07-28 09:44:56 (коммит `e8c76`).

¹³⁵Прим. перев.: Стоит упомянуть еще об одном отладочном инструменте — `netconsole`. Это механизм ядра, позволяющий отправлять содержимое буфера `kmsg` (см. `dmesg(8)`) по сети на заданный компьютер. Обратите внимание, что его настройка через командную строку ядра (параметр `netconsole=...`) работает только если он вкомпилирован в ядро монолитно. Если же он собран модулем, его необходимо настраивать через `/etc/modprobe.d/*.conf` или `configfs` (впрочем, некоторые версии `modprobe` поддерживают чтение параметров модулей из командной строки ядра, так что можно попробовать добавить туда `netconsole.netconsole=...`), а также задавать его принудительную подгрузку через параметр ядра `modules-load=netconsole`. Кроме того, при этом надо обеспечить перенаправление логов `systemd` в буфер ядра — соответствующие параметры см. в разделе [В.1.2](#).

Вышеприведенный способ удобен для диагностики проблем, непосредственно связанных с `systemd`. Если же сбой происходит при запуске системных служб (например, сети), целесообразно перенаправить на консоль вывод системного журнала:

```
| systemd.journald.forward_to_console=1 console=ttyS0,38400 console=tty1
```

Вы можете указывать `console=` несколько раз — `systemd` будет выводить информацию во все перечисленные консоли.

Загрузка в восстановительном (`rescue`) или аварийном (`emergency`) режимах

Чтобы загрузиться в восстановительном режиме, добавьте к параметрам ядра `systemd.unit=rescue.target`, или просто `1`. Это режим эффективен для решения проблем, возникающих на этапе запуска обычных служб, когда ключевые компоненты системы уже инициализированы. В такой ситуации, вы можете просто отключить проблемную службу. Если же загрузка не доходит даже до восстановительного режима — попробуйте менее требовательный, аварийный режим.

Для загрузки напрямую в режим аварийной оболочки, добавьте к параметрам ядра `systemd.unit=emergency.target`, или просто `emergency`. Обратите внимание, что в аварийном режиме корневая система по умолчанию монтируется в режиме «только для чтения», поэтому перед восстановительными работами, связанными с записью на диск, необходимо перемонтировать ее в режиме «для чтения и записи»:

```
| mount -o remount,rw /
```

Как правило, аварийная оболочка используется для исправления некорректных записей в `/etc/fstab`. После внесения необходимых изменений, скомануйте `systemctl daemon-reload`, чтобы `systemd` увидел ваши исправления.

Если не работает даже аварийный режим, попробуйте загрузиться напрямую в оболочку, добавив к параметрам ядра `init=/bin/sh`. Такая ситуация может возникнуть вследствие повреждения бинарного файла `systemd`, либо библиотек, которые он использует. В этом случае может помочь переустановка соответствующих пакетов.

Если не срабатывает даже `init=/bin/sh`, остается лишь попробовать загрузиться с другого носителя.

Отладочная оболочка Вы можете включить специальную отладочную оболочку, которая запускается в отдельной консоли на раннем этапе загрузки и позволяет собрать необходимую диагностическую информацию, а также провести восстановительные операции. Для включения отладочной оболочки скомануйте

```
| systemctl enable debug-shell.service
```

или укажите

```
| systemd.debug-shell=1
```

в строке параметров ядра при загрузке.

Совет: Если вышеприведенная команда `systemctl enable` сообщает об ошибке соединения с системным менеджером (такое возможно, например, если вы загрузились с помощью другой операционной системы), вы можете запустить ее в «автономном режиме», явно указав параметр `--root=`:

```
| systemctl --root=/ enable debug-shell.service
```

Отладочная оболочка будет запущена с правами `root` на консоли `tty9` при следующей загрузке системы. Чтобы переключиться на нее, нажмите `CTRL-ALT-F9`. Оболочка запускается на самом раннем этапе загрузки и позволяет вам проверять состояние служб, читать системные журналы, выявлять зависшие задачи (командой `systemctl list-jobs`) и т.д.

Предупреждение: Используйте эту оболочку только для отладки! Не забудьте отключить ее после того, как разберетесь с проблемами. Оставлять доступную всем и каждому оболочку с правами `root`, мягко говоря, небезопасно.

Также можно настроить `kbrequest.target` как псевдоним для `debug-shell.service`¹³⁶ — тогда отладочная оболочка будет запускаться по запросу (`ALT+↑` из `tty`). Указанные проблемы с безопасностью при этом остаются, но оболочка хотя бы не будет работать постоянно.

Проверка параметров ядра Для корректной загрузки системы необходимо, чтобы каталог `/dev` был заполнен, хотя бы частично. Убедитесь, что ядро Linux собрано с опциями `CONFIG_DEVTMPFS` и `CONFIG_DEVTMPFS_MOUNT`. Кроме того, для корректной работы `systemd` рекомендуется включить поддержку контрольных групп и `fanotify` (опции `CONFIG_CGROUPS` и `CONFIG_FANOTIFY` соответственно). Отключение этих опций может привести к появлению сообщений об ошибках вида «Failed to get D-Bus connection: No connection to service manager.» при попытке запуска `systemctl`.

В.1.2 Если у вас есть доступ к оболочке

Если вам все-таки удалось получить доступ к оболочке системы, вы можете воспользоваться ею для сбора диагностической информации. Загрузите систему со следующими параметрами ядра:

```
systemd.log_level=debug systemd.log_target=kmsg log_buf_len=1M printk.devkmsg=on
```

В соответствии с ними, `systemd` будет выводить максимально подробные сообщения о процессе загрузки, и направлять их в кольцевой буфер ядра (предпоследний параметр обеспечивает соответствующее увеличение размера буфера, последний параметр снимает ограничение на скорость вывода сообщений в буфер). Дождавшись запуска оболочки, вы можете просмотреть полученный журнал:

```
journalctl -b
```

Отправляя отчет об ошибке, перенаправьте вывод этой команды в файл и присоедините его к отчету.

Также, вы можете просмотреть список операций, чтобы выявить зависшие задачи:

```
systemctl list-jobs
```

Операции, находящиеся в состоянии «waiting», будут запущены на исполнение только после того, как завершатся операции, выполняемые в данный момент (состояние «running»).

В.2 Диагностика проблем с выключением системы

При зависании системы во время выключения, как и в случае с загрузкой, рекомендуется подождать *минимум 5 минут*, чтобы отличить полное зависание системы от временного подвисания из-за проблем с отдельными службами. Также стоит проверить, реагирует ли система на нажатие `CTRL+ALT+DEL`.

Если процесс останова системы (при выключении или перезагрузке) зависает полностью, прежде всего нужно убедиться, способно ли ядро Linux выключить или перезагрузить систему. Для этого воспользуйтесь одной из команд:

```
reboot -f
poweroff -f
```

Если хотя бы одна из этих команд не сработает — значит, проблема скорее всего не в `systemd`, а в ядре.

¹³⁶Прим. перев.: Надо полагать, речь идет о чем-то вроде «`systemctl add-wants kbrequest.target debug-shell.service`».

В.2.1 Система очень долго выключается

Если ваша система все же может выключиться/перезагрузиться, но этот процесс длится подозрительно долго, выполните нижеописанные операции:

- Загрузите систему со следующими параметрами ядра:

```
systemd.log_level=debug systemd.log_target=kmsg log_buf_len=1M printk.devkmsg=on enforcing=0
```

- Создайте файл `/usr/lib/systemd/system-shutdown/debug.sh`, добавьте ему право на запуск и запишите в него следующие строки:

```
#!/bin/sh
mount -o remount,rw /
dmesg > /shutdown-log.txt
mount -o remount,ro /
```

- Перезагрузите систему.

После этого вы можете самостоятельно проанализировать файл `/shutdown-log.txt`, и/или присоединить его к вашему сообщению об ошибке.

В.2.2 Система не может выключиться самостоятельно

Если процесс выключения или перезагрузки вашей системы не завершается даже через несколько минут, и вышеописанный метод с `shutdown-log` не сработал, вы можете собрать диагностическую информацию другими методами (которые мы уже рассматривали применительно к проблемам загрузки):

- Используйте [последовательную консоль](#)¹³⁷.
- Воспользуйтесь [отладочной оболочкой](#) — она полезна не только на ранних стадиях загрузки, но и на поздних стадиях остановки системы.

В.3 Просмотр состояния службы и ее журнала

Когда при запуске службы происходит сбой, `systemd` выводит весьма абстрактное сообщение об ошибке:

```
# systemctl start foo.service
Job failed. See system journal and 'systemctl status' for details.
```

При этом сама служба может выводить собственное сообщение, но вы (пока что) его не видите. Дело в том, что запуск служб происходит не из вашей оболочки, а из процесса `systemd`, и поэтому вывод программы не привязан к вашей консоли. Тем не менее, это вовсе не означает, что выводимые сообщения теряются. По умолчанию, потоки `STDOUT` и `STDERR`, принадлежащие запускаемым службам, перенаправляются в системный журнал (`journal`). Туда же попадают и сообщения, отправляемые с помощью функции `syslog(3)`. Кроме того, `systemd` записывает код выхода сбойных процессов. Посмотреть собранные данные можно, например, так:

```
# systemctl status foo.service
foo.service - mmm service
   Loaded: loaded (/etc/systemd/system/foo.service; static)
   Active: failed (Result: exit-code) since Fri, 11 May 2012 20:26:23 +0200; 4s ago
   Process: 1329 ExecStart=/usr/local/bin/foo (code=exited, status=1/FAILURE)
   CGroup: name=systemd:/system/foo.service

May 11 20:26:23 scratch foo[1329]: Failed to parse config
```

¹³⁷Прим. перев.: Здесь опять же стоит напомнить про `netconsole` (см. примечание [135](#)).

В нашем примере, служба запустилась как процесс с идентификатором (PID) 1329, после чего завершилась с кодом выхода 1. Если вы запустили команду `systemctl status` от имени пользователя `root`, либо от имени пользователя, входящего в группу `adm`, вы также увидите последние несколько строчек, записанные службой в журнал. В нашем примере служба выдала всего одно сообщение («Failed to parse config»).

Чтобы просмотреть весь журнал целиком, воспользуйтесь командой `journalctl`.

Если одновременно с `journal` вы используете и классический демон системного лога (например, `rsyslog`), то все сообщения из журнала будут переданы также и этому демону, который запишет их в традиционные лог-файлы (в какие именно — зависит от его настроек, обычно `/var/log/messages`).

В.4 Подготовка сообщений об ошибках

Если вы собираетесь отправить сообщение об ошибке `systemd`, пожалуйста, включите в него диагностическую информацию, в частности, содержимое системных журналов. Журналы должны быть полными (без вырезок) и не заархивированными.

Прежде всего, отправьте сообщение в багтрекер своего дистрибутива. Если же вы твердо уверены, что проблема именно в апстримном `systemd`, проверьте сначала список [уже известных ошибок](#). Если вы не найдете в нем своего случая — [заводите новую запись](#).

В.4.1 Что нужно включить в сообщение об ошибке

По возможности, пожалуйста, укажите в самом сообщении, либо присоедините к нему, следующую информацию:

- Строку параметров ядра. Ее можно найти в файле конфигурации загрузчика (например, `/boot/grub2/grub.cfg`) или в специальном файле `/proc/cmdline`.

- Файл `journal.txt`, полученный после выполнения команды

```
journalctl -b > journal.txt
```

Перед ее выполнением желательно загрузиться с параметрами ядра

```
systemd.log_level=debug systemd.log_target=kmsg log_buf_len=1M printk.devkmsg=on
```

- Файл `systemd-dump.txt`, полученный в результате выполнения команды¹³⁸

```
systemd-analyze dump > systemd-dump.txt
```

- Файл `systemd-test.txt`, полученный при помощи команды

```
/usr/bin/systemd --test --system --log-level=debug > systemd-test.txt 2>&1
```

С Совместимость с SysV¹³⁹

`systemd` обеспечивает высокий уровень совместимости с поведением классической системы инициализации SysV `init`, реализованной во многих дистрибутивах. Это касается как непосредственного взаимодействия с пользователем, так и поддержки API для скриптов. Тем не менее, существует ряд ограничений, обусловленных техническими соображениями, а также особенностями дизайна `systemd` и/или дистрибутивов. Ниже приведен список таких ограничений. Большая их часть затрагивает дистрибутивно-специфичные расширения SysV `init`, и поэтому будет ощутима только для пользователей отдельных дистрибутивов.

¹³⁸Прим. перев.: В версиях `systemd` до 206 включительно, вместо `systemd-analyze dump` нужно было использовать `systemctl dump`.

¹³⁹Прим. перев.: Перевод статьи «[Compatibility with SysV](#)» с официального сайта проекта, по состоянию на 2017-05-23 08:04:14 (коммит 4c503).

- Если разработчики вашего дистрибутива отказались от SysV init-скрипта в пользу штатного юнит-файла, прямой вызов этого скрипта (например, `/etc/init.d/foobar start`), очевидно, работать не будет. Лучше воспользоваться стандартной и универсальной командой `/sbin/service foobar start`, которая одинаково хорошо работает как с `systemd`, так и с SysV `init`. Отметим, что прямое обращение к скрипту в любом случае не вполне корректно, так как запущенная служба унаследует часть контекста выполнения (переменные окружения, `umask`, ограничения по ресурсам, идентификаторы аудита, и т.д.) от пользовательской оболочки, из которой был запущен скрипт. Вызов скрипта через `/sbin/service` исправляет этот недостаток, пусть и частично. Кроме того, стоит заметить, что стандарт LSB предписывает обращаться к службам только через `/sbin/service`. (Вышеуказанное ограничение *не* распространяется на дистрибутивы, которые одновременно поддерживают и `init`-скрипты, и юнит-файлы. В таких дистрибутивах даже прямое обращение к скрипту при необходимости будет преобразовано в соответствующий запрос к `systemd`¹⁴⁰.)
- Информация о зависимостях служб, прописанная в LSB-заголовке скрипта, играет существенную роль. Большинство реализаций SysV в различных дистрибутивах практически не используют эту информацию, либо используют ее очень ограниченно. Вследствие этого, информацию о зависимостях часто указывали некорректно или не полностью. В отличие от подобных реализаций, `systemd` интерпретирует эти заголовки корректно, и использует содержащуюся в них информацию при выполнении различных операций со службами (а не только в момент их установки, как это делают некоторые другие реализации SysV¹⁴¹).
- Все операции со скриптами ограничены по времени при помощи тайм-аутов. В отличие от классических SysV-систем, где зависший `init`-скрипт полностью останавливает загрузку, `systemd` ограничивает максимальную длительность работы скрипта пятью минутами.
- Службы запускаются в совершенно чистом контексте исполнения. Они уже не наследуют контекст у оболочки, из которой был вызван их скрипт. В частности, это касается специфических переменных, таких как `$HOME`. Как следствие, скрипты, их использующие, не смогут работать корректно.
- Службы и их скрипты не могут читать с `STDIN` — для них этот поток подключен к `/dev/null`. Следовательно, интерактивные `init`-скрипты не поддерживаются (в частности, игнорируется используемый в LSB-заголовках скриптов Debian флаг `X-Interactive`). К счастью, большинство дистрибутивов все равно не поддерживало интерактивные `init`-скрипты. Если вашему скрипту нужно запросить пароль для зашифрованного диска или SSL-ключа, вы можете воспользоваться для этого специальным механизмом, предоставляемым `systemd` (см. [описание](#) и [страницу руководства](#)).
- Дополнительные команды для `init`-скриптов (кроме стандартных, таких, как `start`, `stop`, `status`, `restart`, `reload`, `try-restart`, `force-reload`¹⁴²) не поддерживаются¹⁴³.

¹⁴⁰Прим. перев.: Такое поведение задают разработчики дистрибутива, при помощи корректировки файла, содержащего стандартные функции для `init`-скриптов (например, `/etc/rc.d/init.d/functions` в Fedora или `/etc/rc.status` в openSUSE). Этот файл вызывается практически из всех `init`-скриптов в самом начале их работы.

¹⁴¹Прим. перев.: Например, `insserv`, используемый как дополнение к SysV `init` в Debian (а ранее и в openSUSE).

¹⁴²Прим. перев.: Точный список «стандартных» команд для SysV `init`-скриптов зависит от используемого дистрибутива. Так уж исторически сложилось, что практически каждый дистрибутив использовал собственные стандарты интерфейсов SysV `init`. Это отразилось и на «заглушках», используемых для обратной совместимости. Выше приведен список команд, которые поддерживаются как в Fedora, так и в openSUSE.

¹⁴³Прим. перев.: В частности, это ограничение касается специфичных для `init`-скрипта `/etc/init.d/iptables` команд `save` и `panic`.

- Также не поддерживается возможность указывать после стандартных команд скрипту дополнительные параметры. Такое расширение SysV не прописано ни в каких стандартах, и реализация его поддержки была бы весьма проблематичной.
- Переопределение команды `restart` не поддерживается. Она реализована непосредственно в `systemd` как последовательное выполнение команд `stop` и `start`.
- `systemd` останавливает только те службы, которые выполняются. Традиционный SysV `init` при завершении работы системы запускает все скрипты, предписанные K*-ссылками для текущего уровня исполнения. `systemd` в аналогичной ситуации действует более последовательно, и не пытается остановить те службы, которые не были запущены.
- Для уровней исполнения 0 и 6 все S*- и K*-ссылки игнорируются. При завершении работы системы останавливаются все работающие службы, и никаких новых SysV-служб не запускается.
- Если `systemd` не знает PID главного процесса службы, он не сможет отслеживать ее работу. В частности, если этот процесс «самостоятельно» (без сигнала от системного менеджера) завершит работу, `systemd` не будет знать о том, что служба остановилась. Поэтому настоятельно рекомендуется указывать в `init`-скрипте предложенный Red Hat заголовок-комментарий `pidfile:`, позволяющий `systemd` найти PID-файл вашей службы и получить из него PID главного процесса. Из-за того, что в некоторых `init`-скриптах запускаются дополнительные процессы, выполняющие предварительную настройку для нужд службы, в общем случае `systemd` не должен рассматривать завершение процесса, запущенного из `init`-скрипта, как завершение работы всей службы. Именно поэтому для корректного отслеживания ее статуса `systemd` требуется явное указание PID-файла. (Обратите внимание, что заголовок `pidfile:` можно указывать в `init`-скрипте не более одного раза.)
- Поддержка уровней исполнения (`runlevels`) в `systemd` имеет некоторые ограничения. Хотя все уровни исполнения SysV имеют соответствующие им целевые состояния (`target units`), далеко не все целевые состояния имеют соответствующие им уровни исполнения. Это обусловлено тем, что механизм целевых состояний отличается гораздо большей гибкостью и эффективностью, чем система уровней исполнения. Как следствие, в некоторых случаях попытка узнать текущий уровень исполнения (например, при помощи `/sbin/runlevel`) может вернуть просто «N» (т.е. «уровень выполнения неизвестен»), что приведет к нарушению работы скриптов, использующих явную проверку текущего уровня исполнения. Избегайте подобных проверок.
- `/sbin/chkconfig` и аналогичные программы могут выводить некорректную информацию о состоянии службы (включена/отключена). Они оперируют только `init`-скриптами, игнорируя штатные юнит-файлы. Кроме того, они опираются на механизм уровней выполнения, который поддерживается не полностью (см. выше).
- По умолчанию, уровни исполнения 2, 3 и 4 являются эквивалентами целевого состояния `multi-user.target`. Если включить службу на одном из них, то она будет включена и на остальных. Впрочем, это лишь настройка по умолчанию, и ничто не мешает вам определить их независимо. Тем не менее, мы рекомендуем отказаться от уровней исполнения в пользу более гибкого механизма целевых состояний `systemd`.
- Дистрибутивно-специфичные квази-уровни исполнения, используемые для организации запуска скриптов на ранних стадиях загрузки, больше не поддерживаются¹⁴⁴. Примерами таких уровней являются S в Debian и b в openSUSE. Разумеется, это никак не затрагивает поддержку обычных уровней исполнения, которую мы намерены сохранять еще очень долго.

¹⁴⁴Прим. перев.: Начиная с версии 196, см. коммит [3cdeb](#) от 16 ноября 2012 г.

- В системах с SysV, изменение `init`-скриптов и любых других файлов, отвечающих за настройку процесса загрузки (например, `/etc/fstab`), обычно вступают в силу немедленно. В то же время, в системах с `systemd`, `init`-скрипты и другие конфигурационные файлы, влияющие на процесс загрузки, перечитываются только после выполнения команды `systemctl daemon-reload` (обратите внимание, что некоторые команды, например, `systemctl enable` и `systemctl disable`, выполняют данную операцию автоматически)¹⁴⁵. Это сделано специально, для обеспечения консистентности вносимых изменений, и позволяет избежать ситуации, когда считывание происходит непосредственно в момент редактирования конфигурации администратором¹⁴⁶.
- Не поддерживается задание двух и более записей для одного и того же пути монтирования в `/etc/fstab`. В `systemd` каждой точке монтирования в каждый момент времени может соответствовать только один `mount`-юнит. Также стоит отметить, что порядок перечисления записей в `/etc/fstab` не играет никакой роли: монтирование всех точек выполняется параллельно, с учетом зависимостей, автоматически рассчитанных на основе расположения монтируемого объекта и точки монтирования¹⁴⁷.

D Предсказуемые имена сетевых интерфейсов¹⁴⁸

Начиная с версии 197, `systemd/udev` присваивает сетевым интерфейсам (Ethernet, WLAN, WWAN¹⁴⁹) стабильные, предсказуемые имена. Новая схема именования несколько отличается от классической (`eth0`, `eth1`, `wlan0`, ...), однако при этом лишена и многих ее недостатков.

D.1 Зачем?

Классическая схема именования сетевых интерфейсов, используемая ядром, предполагает последовательное присвоение интерфейсам имен `eth0`, `eth1` и т.д., по мере опроса оборудования соответствующими драйверами (`device probing`). На современных системах такой опрос не обеспечивает гарантированного соблюдения одного и того же порядка поступления ответов. Вследствие этого, соответствие имен реальным интерфейсам не является чем-то фиксированным, и очень может быть, что интерфейс, который сейчас называется `eth0`, при следующей загрузке превратится в `eth1`. Такая ситуация приводит к целому ряду проблем, в частности, к проблемам с безопасностью: в настройках брандмауэра интерфейсы идентифицируются как раз по именам.

Существует несколько подходов к решению этой проблемы. В течение многих лет `udev` поддерживал механизм постоянной привязки имен к интерфейсам на основе MAC-адресов. Такой подход имел множество недостатков, в том числе: требование доступности корневого каталога на запись (что возможно далеко не всегда); необходимость

¹⁴⁵Прим. перев.: Стоит отметить, что на самом деле все немного сложнее. В частности, даже в системе с `systemd` изменения в исполняемом коде SysV `init`-скриптов должны вступать в действие немедленно. В то же время, изменения в их *заголовках* (оформленных в виде комментариев и содержащих служебную информацию, в частности, о зависимостях между службами), а также сделанные «вручную» (командами `ln` и `rm`) изменения символьных ссылок в `/etc/rc?.d/` действительно срабатывают только после `systemctl daemon-reload`. Это обусловлено тем, что симлинки и информация из заголовков используются при генерации `unit`-файлов на основе `init`-скриптов, которая осуществляется программой `systemd-sysv-generator`(8) в начале процесса загрузки, а также каждый раз после выполнения `systemctl daemon-reload`. Аналогично и с `/etc/fstab`, на основе которого `systemd-fstab-generator`(8) генерирует набор `mount`-юнитов.

¹⁴⁶Прим. перев.: Пример подобной проблемы рассмотрен в примечании 130.

¹⁴⁷Прим. перев.: Например, если указаны точки монтирования `/var/tmp` и `/var`, сначала монтируется `/var`, и только потом `/var/tmp`. Более подробно автоматические зависимости `mount`-юнитов рассмотрены в *руководстве*.

¹⁴⁸Прим. перев.: Перевод статьи «[Predictable Network Interface Names](#)» с официального сайта проекта, по состоянию на 2016-11-17 17:52:54 (коммит fe089).

¹⁴⁹Прим. перев.: WWAN (Wireless Wide Area Network) — относительно новый термин, обозначающий технологии глобальных беспроводных компьютерных сетей, такие, как GPRS, 3G, WiMAX и т.д.

внесения изменений в образ системы после загрузки на новом оборудовании; привязка к MAC-адресам, которые далеко не всегда являются фиксированными (в частности, это касается многих встраиваемых систем и механизмов виртуализации). Тем не менее, основная проблема такого подхода состояла в том, что он использовал имена из того же адресного пространства (`ethX`), что и ядро. Вследствие этого, возникали ситуации «гонки» между `udev` и ядром, когда `udev` пытался присвоить интерфейсу имя, которое ядро уже выдало другому интерфейсу, что приводило к сбою операции переименования. Вследствие перечисленных недостатков, данный механизм был удален из `systemd/udev`¹⁵⁰.

Другая попытка решения обсуждаемой проблемы — `biosdevname`, программа, формирующая имена интерфейсов на основании их физического расположения на материнской плате. Соответствующая информация запрашивается у BIOS. В чем-то такая схема похожа на ту, которую `udev` уже давно использует для формирования стабильных символьных ссылок на различные устройства (`/dev*/by-path/*`), но, тем не менее, в ее основе лежат несколько иные алгоритмы (`udev`, формируя свои символьные ссылки, опирается на идентификационные данные, предоставленные ядром, в то время как `biosdevname` использует свои собственные схемы).

И наконец, многие дистрибутивы в своих сетевых скриптах поддерживают механизмы, позволяющие присваивать интерфейсам имена, выбранные пользователем (например, `internet0`, `dmz0`, ...). Если бы не необходимость явного вмешательства пользователя, это было бы замечательным решением.

Мы пришли к выводу, что наилучшим вариантом для настроек по умолчанию будет схема, являющаяся обобщением и развитием механизма `biosdevname`. Присвоение интерфейсам имен на основании информации об их физическом расположении имеет ряд существенных достоинств: такие имена формируются автоматически, всегда предсказуемы, а также остаются неизменными даже при добавлении и удалении оборудования, что позволяет без лишних проблем заменять сломанные сетевые устройства. Тем не менее, такие выглядят немного сложнее, чем привычные `eth0` и `wlan0`, например: `enp5s0`.

D.2 Что именно было изменено в версии 197?

В версии 197 мы добавили в `systemd/udev` встроенную поддержку нескольких различных механизмов именования сетевых интерфейсов, получив в итоге схему, похожую на `biosdevname`, но отличающуюся большей гибкостью и максимально приближенную к алгоритмам идентификации устройств, используемым в ядре. В частности, `udev` теперь штатно поддерживает следующие схемы именования сетевых интерфейсов:

1. Имена устройств, встроенных в материнскую плату, формируются на основании информации от прошивки¹⁵¹, например: `enp01`.
2. Имена устройств, подключенных в слоты PCI Express, формируются также на основании информации от прошивки, например: `ens1`.
3. Имена устройств формируются исходя из физического расположения точки подключения оборудования, например, `enp2s0`.
4. Имена устройств формируются на основе их MAC-адресов, например, `enx78e7d1ea46da`.
5. Используются классические, непредсказуемые имена, присвоенные ядром, например, `eth0`.

По умолчанию, `systemd 197` при именовании сетевых интерфейсов последовательно пытается применить схемы 1–3 (для первых двух требуется информация от прошивки). Если ни одна из них не подходит, используется схема 5. Что касается схемы 4 — она не задействована по умолчанию, однако ее можно включить вручную.

¹⁵⁰Прим. перев.: См. коммит [3e214](#) от 3 апреля 2012 года, в котором, среди прочего, был удален каталог `src/udev/src/rule_generator`.

¹⁵¹Прим. перев.: BIOS, (U)EFI, SPARC Boot PROM, ...

Вышеописанная комбинированная схема используется лишь *в последнюю очередь*. Если у вас установлена программа `biosdevname`, для именованя сетевых устройств будет использоваться именно она. Кроме того, приоритет предоставляется любым правилам, добавленным пользователем или разработчиками дистрибутива.

D.3 Еще раз, что здесь хорошего?

Достоинства нашей новой схемы:

- Имена интерфейсов остаются неизменными после перезагрузок.
- Имена интерфейсов остаются неизменными при добавлении или удалении устройств (если ваша прошивка это позволяет).
- Имена интерфейсов остаются неизменными при обновлении/изменении ядра и драйверов¹⁵².
- Имена интерфейсов остаются неизменными при замене сломанных сетевых карт новыми.
- Имена формируются автоматически, безо всякого вмешательства пользователя.
- Имена являются предсказуемыми: достаточно просто взглянуть на вывод `lspci`, чтобы узнать, как будет называться конкретное устройство.
- Изменения в аппаратной конфигурации не приводят к необходимости записи в каталог `/etc`.
- Полная поддержка систем, корень которых доступен только для чтения.
- Логика именованя аналогична той, которая используется `udev` для формирования стабильных символьных ссылок в каталоге `/dev (by-path)`.
- Работает как на x86, так и на других архитектурах.
- Работает одинаково во всех дистрибутивах, использующих `systemd/udev`.
- От этой схемы очень легко отказаться (см. ниже).

Есть ли у нее недостатки? Да. Раньше, если на компьютере имелся всего один сетевой интерфейс, можно было с высокой долей вероятности утверждать, что он называется `eth0`. Теперь же, прежде чем администратор начнет настраивать сеть, он должен как минимум просмотреть список сетевых интерфейсов.

D.4 Мне не нравится ваша схема. Как ее отключить?

У вас есть три варианта:

1. Вы можете полностью отключить новую схему, вернувшись к классическим непредсказуемым именам. Для этого достаточно заблокировать (замаскировать) link-файл `udev`, отвечающий за именованя интерфейсов

```
ln -s /dev/null /etc/systemd/network/99-default.link
```

2. Вы можете вручную назначить интерфейсам наиболее понятные для вас имена (например, `internet0`, `dmz0`, `lan0`). Для этого создайте соответствующие link-файлы в каталоге `/etc/systemd/network` и укажите в них нужные вам имена (или схемы именованя) для одного, нескольких или сразу всех интерфейсов¹⁵³. За подробностями вы можете обратиться к странице руководства [systemd.link\(5\)](#).

¹⁵²Прим. перев.: На самом деле, не все так просто. Если, в результате обновленя ядра/драйверов, в них появится ранее отсутствовавшая поддержка вашей прошивки, не исключена вероятность, что имена некоторых интерфейсов перейдут с третьей схемы на первую или вторую. Будьте готовы к такому развитию событий.

¹⁵³Прим. перев.: Там же можно изменить используемую по умолчанию схему именованя, скопировав файл `/usr/lib/systemd/network/99-default.link` и поменяв значение параметра `NamePolicy=` на нужное значение.

3. Вы можете указать параметр `net.ifnames=0` в строке параметров ядра (отключает использование предсказуемых имен для текущей загрузки).

D.5 Как именно работает новая схема?

Подробности технической реализации описаны в блоке комментариев в [исходном коде net_id built-in](#). Ознакомьтесь с ним, если у вас возникают вопросы, касающиеся расшифровки новых имен¹⁵⁴.

Предсказуемые имена сетевых интерфейсов формируются на основании:

- индексов встроенных в материнскую плату устройств (по информации от прошивки)
- индексов hotplug-слотов PCI-E (по информации от прошивки)
- физического расположения точки подключения оборудования
- MAC-адресов

<http://www.freedesktop.org/wiki/Software/systemd/PredictableNetworkInterfaceNames>

Первые два символа в имени определяют тип интерфейса:

```
en -- ethernet
sl -- SLIP (IP через последовательный порт)
wl -- WLAN
ww -- WWAN
```

Последующие символы определяются используемой схемой:

```
b<number> -- для устройств, подключенных по шине BCMA (<number> соответствует
BCMA bus core number)
c<bus_id> -- для устройств, адресуемых через Channel Command Words (CCW).
<bus_id> соответствует CCW bus-ID для простого (использующего один
субканал) или группированного (использующего несколько субканалов)
устройства, указывается без ведущих нулей [s390]
o<index>[n<phys_port_name>|d<dev_port>] -- для устройств, встроенных
в материнскую плату
s<slot>[f<function>] [n<phys_port_name>|d<dev_port>] -- для hotplug-слотов
x<MAC> -- при именовании по MAC-адресу
[P<domain>]p<bus>s<slot>[f<function>] [n<phys_port_name>|d<dev_port>]
-- на основании физического расположения PCI-устройства
[P<domain>]p<bus>s<slot>[f<function>] [u<port>] [..] [c<config>] [i<interface>]
-- идентификационная цепочка для USB-устройств
v<slot> -- номер слота для VIO-устройств (IBM PowerVM)
a<vendor><model>i<instance> -- идентификатор ACPI для устройств, подключенных
через шину платформы (ARM64)
```

Все многофункциональные (multi-function) PCI-устройства содержат в своем имени номер функции "f<function>" (отсчитываются от нуля).

При именовании по расположению устройства, индекс PCI-домена "P<domain>" указывается только в том случае, если он отличен от нулевого.

Для USB-устройства формируется полная цепочка номеров портов хабов, через которые оно подключено. Если итоговая строка будет длиннее 15 символов, она не экспортируется.

Стандартные значения `config == 1` и `interface == 0` опускаются.

Примеры:

¹⁵⁴Прим. перев.: Далее приводится перевод упомянутого блока комментариев. Последним коммитом, затронувшим данный файл, на момент перевода является 86a48 от 18 мая 2017 г.

Подключенная к PCI сетевая карта, которая идентифицируется прошивкой по индексу "1":

```
ID_NET_NAME_ONBOARD=en01
ID_NET_NAME_ONBOARD_LABEL=Ethernet Port 1
```

Сетевая карта, включенная в hotplug PCI-слот, который идентифицируется прошивкой:

```
/sys/devices/pci0000:00/0000:00:1c.3/0000:05:00.0/net/ens1
ID_NET_NAME_MAC=enx000000000466
ID_NET_NAME_PATH=enp5s0
ID_NET_NAME_SLOT=ens1
```

Multi-function PCI устройство с двумя портами:

```
/sys/devices/pci0000:00/0000:00:1c.0/0000:02:00.0/net/enp2s0f0
ID_NET_NAME_MAC=enx78e7d1ea46da
ID_NET_NAME_PATH=enp2s0f0
/sys/devices/pci0000:00/0000:00:1c.0/0000:02:00.1/net/enp2s0f1
ID_NET_NAME_MAC=enx78e7d1ea46dc
ID_NET_NAME_PATH=enp2s0f1
```

Подключенная к PCI WLAN-карта:

```
/sys/devices/pci0000:00/0000:00:1c.1/0000:03:00.0/net/wlp3s0
ID_NET_NAME_MAC=wlx0024d7e31130
ID_NET_NAME_PATH=wlp3s0
```

Встроенный USB 3G модем:

```
/sys/devices/pci0000:00/0000:00:1d.0/usb2/2-1/2-1.4/2-1.4:1.6/net/wwp0s29u1u4i6
ID_NET_NAME_MAC=wwx028037ec0200
ID_NET_NAME_PATH=wwp0s29u1u4i6
```

Подключенный по USB Android-смартфон:

```
/sys/devices/pci0000:00/0000:00:1d.0/usb2/2-1/2-1.2/2-1.2:1.0/net/enp0s29u1u2
ID_NET_NAME_MAC=enxd626b3450fb5
ID_NET_NAME_PATH=enp0s29u1u2
```

Группированный CCW-интерфейс на s390:

```
/sys/devices/css0/0.0.0007/0.0.f5f0/group_device/net/encf5f0
ID_NET_NAME_MAC=enx026d3c00000a
ID_NET_NAME_PATH=encf5f0
```

Е Специальные файловые системы¹⁵⁵

Итак, вы запустили программу `mount(8)`, и увидели в ее выводе множество странных файловых систем, не указанных в `/etc/fstab`. У вас могут возникнуть вопросы: «Как от них избавиться?» или «Как задать для них параметры монтирования?».

В Linux предусмотрено несколько способов взаимодействия программ из пространства пользователя с ядром. Наиболее популярными механизмами являются системные вызовы (`syscall`), интерфейсы Netlink, а также виртуальные файловые системы (ФС), такие, как `/proc` и `/sys`. Подобные ФС являются лишь программными интерфейсами, и не обеспечивают долговременного хранения данных (в частности, между перезагрузками системы). Они просто используют классические механизмы работы с файлами для предоставления доступа к различным данным и настройкам. Кроме того, существуют специальные ФС, используемые программами для собственных нужд, в частности, для

¹⁵⁵Прим. перев.: Перевод статьи «API File Systems» с официального сайта проекта, по состоянию на 2015-02-20 15:48:02 (коммит e2db9).

хранения сегментов разделяемой памяти (shared memory), временных файлов и сокетов. В данной статье мы обсудим все эти разновидности *специальных файловых систем*. Ниже представлен список таких ФС, поддерживаемых в типовых Linux-системах:

- `/sys` предоставляет доступ к драйверам и устройствам, а также некоторым другим параметрам ядра;
- `/proc` дает доступ к информации о выполняемых процессах, настройкам ядра, а также другим параметрам;
- `/dev` отображает файлы устройств (device nodes);
- `/run` содержит файлы и сокеты, используемые программами;
- `/tmp` хранит временные и часто изменяемые объекты (X);
- `/sys/fs/cgroup` (и другие файловые системы, смонтированные в подкаталогах этого каталога) позволяют работать с иерархией контрольных групп;
- `/sys/kernel/security`, `/sys/kernel/debug` (X), `/sys/kernel/config` (X) предоставляют доступ к специализированным механизмам ядра;
- `/sys/fs/selinux` используется для взаимодействия с SELinux;
- `/dev/shm` содержит объекты разделяемой памяти;
- `/dev/pts` обеспечивает доступ к псевдо-TTY устройствам;
- `/proc/sys/fs/binfmt_misc` используется для регистрации в ядре дополнительных бинарных форматов (X);
- `/dev/mqueue` содержит объекты IPC-механизма mqueue (X);
- `/dev/hugepages` позволяет программам запрашивать выделение «гигантских» страниц памяти (X);
- `/sys/fs/fuse/connections` обеспечивает доступ к FUSE-соединениям (X);
- `/sys/firmware/efi/efivars` предоставляет доступ к переменным EFI;

systemd монтирует все эти файловые системы на ранних стадиях загрузки, даже если они не указаны явно в `/etc/fstab`. В зависимости от настроек вашего ядра, некоторые из перечисленных выше ФС могут быть недоступны, и наоборот, могут присутствовать другие специальные ФС, не приведенные в этом списке. Все эти механизмы играют важную роль во взаимодействии ядра с программами и программ между собой. Именно поэтому они подключаются автоматически, безо всякого участия пользователя. Неожиданное вмешательство в их работу (отключение или изменение параметров) может нарушить нормальную работу ваших программ, так как они утратят доступ к необходимым для них интерфейсам.

В абсолютном большинстве случаев достаточно настроек, используемых по умолчанию. Однако, могут возникнуть ситуации, когда необходимо изменить параметры монтирования специальных ФС, либо полностью отключить монтирование некоторых из них.

Несмотря на то, что эти ФС по умолчанию отсутствуют в `/etc/fstab`, ничто не мешает вам их туда добавить. Параметры монтирования, которые вы при этом укажете, будут автоматически применяться к соответствующим ФС. Проще говоря: если вам нужно изменить параметры монтирования для специальных ФС, просто добавьте их в `/etc/fstab` с указанием соответствующих опций. Кроме параметров монтирования, так можно изменить и сам тип файловой системы (в частности, перенести `/tmp` из `tmpfs` на раздел физического диска).

Также вы можете полностью отключить монтирование некоторых (но не всех) специальных систем, если это действительно необходимо. Отключаемые ФС в списке выше отмечены (X). Для их отключения, достаточно заблокировать (замаскировать) соответствующий юнит:

```
systemctl mask dev-hugepages.mount
```

В результате выполнения такой операции, соответствующая ФС больше не будет монтироваться по умолчанию, начиная со следующей загрузки системы. О том, что такое блокировка юнита, вы можете прочитать в главе 5.

На всякий случай отметим, что применение к специальным ФС параметров монтирования, указанных в `/etc/fstab`, обеспечивается службой `systemd-remount-fs.service`.

Зачем вы мне все это рассказываете? Я просто хочу убрать tmpfs из /tmp!

У вас есть три варианта:

1. Вы можете отключить любое монтирование в `/tmp`, в результате чего содержимое данного каталога будет храниться на том же разделе, что и корень. Для этого достаточно выполнить команду

```
systemctl mask tmp.mount
```

2. Вы можете смонтировать в `/tmp` обычную файловую систему, размещенную на диске. Для этого достаточно создать соответствующую запись в `/etc/fstab`.
3. Если вы хотите оставить в `/tmp` tmpfs, но при этом задать для нее другой размер — это тоже делается через внесение в `/etc/fstab` соответствующей записи, предписывающей монтирование tmpfs в `/tmp` с нужным вам значением параметра `size=`.

F Запуск служб после появления сети¹⁵⁶

Итак, ваша служба настроена на запуск только после достижения цели `network.target`, однако, несмотря на это, она все равно запускается до появления сети. У вас возникают вопросы: «Почему так происходит?» и «Как это исправить?».

В некоторых LSB-совместимых скриптах инициализации, при описания зависимостей используется сущность (facility) `$network`. Определение этой сущности в стандарте довольно расплывчато и оставляет широкий простор для различных трактовок. Вот примеры некоторых из них:

- Запущен демон управления сетью (например, NetworkManager или systemd-networkd).
- Все заданные в настройках сетевые интерфейсы переведены в состояние UP и получили IP-адреса.
- Все имеющиеся физические сетевые интерфейсы, на которых зарегистрирована несущая (link beat), получили IP-адреса. Наличие или отсутствие явных настроек роли не играет.
- Доступен DNS-сервер.
- Доступна некоторая выбранная сетевая служба.
- Доступен некий абстрактный «интернет».
- Все заданные в настройках ethernet-интерфейсы переведены в состояние UP, однако процесс настройки PPP-интерфейсов еще не начался.
- Активирован некий профиль сетевых настроек, задающий одно из условий выше. В разных профилях могут использоваться разные условия.

¹⁵⁶Прим. перев.: Перевод статьи «Running Services After the Network is up» с официального сайта проекта, по состоянию на 2016-08-21 13:14:53 (коммит 4d5d5).

- Выполняется некоторый набор условий (выбор условий определяется физическим расположением системы).
- Настроен как минимум один глобальный адрес IPv4.
- Настроен как минимум один глобальный адрес IPv6.
- Настроен как минимум один глобальный адрес IPv4 или IPv6.

Этот список можно продолжать и дальше. Каждый предложенный в нем вариант можно использовать в качестве критерия появления сети, однако ни один из них не подходит в качестве варианта по умолчанию, пригодного для абсолютного большинства задач.

Современные компьютерные сети живут в очень динамичном ритме: компьютеры переменяются между сетями, сетевые настройки меняются, устройства подключаются и отключаются, виртуальные сети настраиваются, перенастраиваются и выключаются. Наличие сетевого соединения не является чем-то постоянным и безусловным. В разные моменты времени компьютер может быть подключен к разным сетям. И это справедливо не только для мобильных устройств (смартфонов, планшетов и ноутбуков), но и, в определенной мере, для встраиваемых и серверных систем. Программы, которые считают, что сеть является чем-то неизменным и непрерывно доступным, не могут нормально функционировать в таких условиях. Хорошо написанная программа должна корректно реагировать на изменение состояния сети, и не унывать в беде. Если нужный ей сервер не доступен в настоящий момент, она должна не зависать по тайм-ауту, а пытаться достучаться к нему снова и снова. Если сетевое соединение утрачено, она должна отреагировать на это. Реализовать такое реагирование в коде демона, на самом деле, не так уж и сложно. Существует множество хорошо известных сетевых служб, которые уже давно поддерживают такую возможность. Подобные службы можно запускать в любой момент, они устойчивы к сбоям, и работают корректно во всех возможных ситуациях.

`$network` предназначена для поддержки программ, созданных в предположении, что сеть доступна постоянно (т.е. написанных не очень аккуратно). Конкретные требования таких программ могут сильно отличаться. Например, IMAP-серверу будет достаточно наличия IP-адреса, на котором можно слушать. В то время как для клиентов сетевых файловых систем требуется доступность и работоспособность сервера, а также, опционально, наличие работоспособного DNS. Таким образом, конкретные требования к `$network` сильно зависят от решаемой задачи.

По соображениям надежности, загрузка системы не должна зависеть от второстепенных служб. В частности, отсутствие ответа от DHCP-сервера не должно завешивать процесс загрузки системы (за исключением ситуаций, когда сетевое соединение действительно необходимо для работы системы, например, при загрузке по сети).

F.1 Как это реализовано в `systemd`

В `systemd` существует сразу три юнита (целевых состояния, `target units`), в совокупности берущих на себя роль LSB-сущности `$network`¹⁵⁷:

- `network.target` не играет существенной роли в процессе загрузки системы. Активация данного целевого состояния лишь показывает, что программа управления сетью была запущена. А вот будет ли к этому моменту настроена сеть — не регламентируется. Основное назначение данного юнита — синхронизация операций при *остановке* системы. Так как порядок остановки юнитов обратен порядку их запуска, все юниты, в описании которых указано `After=network.target`, при выключении системы будут остановлены до того, как отключится сеть. Это позволит программам корректно закрывать все сетевые соединения, а не оставлять их «повисшими». Обратите внимание, что `network.target` является *пассивным* юнитом: вы не можете запустить его ни вручную (через `systemctl start`),

¹⁵⁷Прим. перев.: Приведенные здесь сведения применимы только к `systemd` версий 213 и выше, в которых появился юнит `network-online.target`. О том, как сущность `$network` работала в предыдущих версиях `systemd`, желающие могут прочитать в примечании 161.

ни через зависимости своих служб (`Wants` и `Requires`). Активацией и деактивацией данного юнита занимается программа, управляющая сетью в вашей системе. В юнит-файлах служб, использующих сеть, целесообразно указывать `After=network.target`, но ни в коем случае не `Wants=network.target`, и уж тем более не `Requires=network.target`¹⁵⁸.

- `network-online.target` активируется только после появления сети. Трактовка понятия «появилась сеть» остается на совести разработчиков программы, управляющей сетью¹⁵⁹. Обычно оно подразумевает, что одному или нескольким сетевым интерфейсам присвоены маршрутизируемые IP-адреса. Основная задача обсуждаемого юнита — обеспечить задержку запуска отдельных служб до того момента, как появится сеть. Данный юнит является *активным*, т.е. его можно указывать в зависимостях служб, которым необходима запущенная сеть (и, в то же время, нельзя указывать в зависимостях у самой службы управления сетью). По умолчанию, этот юнит автоматически прописывается в зависимости к точкам монтирования всех удаленных файловых систем (например, NFS, SMBFS/CIFS), приведенным в файле `/etc/fstab`. Таким образом, монтирование этих файловых систем начнется только после того, как появится сеть. Однако, если таких точек монтирования не указано, а также отсутствуют службы, явно требующие по зависимостям данный юнит, он вообще не активируется в процессе загрузки системы, что позволяет избежать нежелательных задержек в случае проблем с сетью. Настоятельно рекомендуется не злоупотреблять зависимостями от этого юнита. В частности, для серверных приложений, как правило, такая зависимость избыточна (обычно, сервер может работать даже в отсутствие внешней сети, обслуживая локальные соединения). Основная задача обсуждаемого юнита — своевременный запуск клиентских программ, которые не могут работать без сети.
- `network-pre.target`¹⁶⁰ активируется до того, как начнется настройка сетевых интерфейсов. Основная функция этого юнита — своевременный запуск служб, выполняющих настройку брандмауэра. Таким образом, к моменту появления сети брандмауэр будет уже готов к отражению возможных атак. Указанный юнит является *пассивным* — вы не можете запустить его вручную, и его нельзя указывать в качестве `Requires/Wants`-зависимости в юнит-файлах службы управления сетью. Напротив, такие зависимости должны прописываться у тех служб, которые должны быть запущены до появления сети. В юнит-файле службы управления сетью целесообразно указать `After=network-pre.target`, но не `Wants=network-pre.target/Requires=network-pre.target`. В то же время, в юнит-файлах служб, которые должны быть запущены до появления сети (например, уже обсуждавшиеся выше службы брандмауэра), наоборот, рекомендуется указывать `Before=network-pre.target` и `Wants=network-pre.target`. Таким образом, данное целевое состояние будет активироваться в нужный момент только

¹⁵⁸Прим. перев.: Концепция активных и пассивных юнитов более подробно пояснена на странице руководства `systemd.special(7)`. Она описывает взаимосвязь двух классов служб: потребители (`consumers`) и поставщики (`providers`). Например, служба управления сетью является поставщиком (сети), а демон торрент-клиента — потребителем. Синхронизация между ними может осуществляться посредством как пассивных, так и активных целевых состояний. Разница между активными и пассивными целевыми юнитами состоит в том, что для активных целей `Requires/Wants` зависимости прописываются в юнит-файле службы-потребителя, а для пассивных — в юнит-файле службы-поставщика. Никакой магии здесь нет — это просто соглашение между авторами юнит-файлов. В рамках этого же принципа, в конфигурационные файлы пассивных юнитов добавляется директива `RefuseManualStart=yes`, запрещающая их активацию «вручную».

¹⁵⁹Прим. перев.: Определением момента готовности сети обычно занимается отдельная утилита. В `NetworkManager` это `nm-online`, в `systemd-networkd` — `systemd-networkd-wait-online(8)`. В случае `NM`, сеть считается готовой после появления на любом из управляемых им устройств корректного IP-адреса (глобального или локального), в случае `networkd` — после того, как для всех интерфейсов, прописанных в его конфигурации, завершится процесс настройки (успешно или с ошибкой), причем как минимум для одного интерфейса настройка должна завершиться успешно.

¹⁶⁰Прим. перев.: Поддержка данного юнита добавлена в `systemd-networkd` начиная с `systemd 214`. Тогда же в официальной документации появились первые упоминания о нем. Однако, ничто не мешает использовать этот юнит и в более ранних версиях `systemd` — достаточно скопировать соответствующий [конфигурационный файл](#) в каталог `/etc/systemd/system`.

в том случае, если у какой-либо из ваших служб действительно имеется такая зависимость. Если же подобных служб нет, `network-pre.target` не будет активироваться вообще.

Когда `systemd` встречает в LSB-заголовках `init`-скриптов зависимость `$network`, он преобразовывает ее в зависимости `Wants=network-online.target` и `After=network-online.target`, что позволяет обеспечить поведение, более или менее соответствующее требованиям LSB¹⁶¹.

За дальнейшими подробностями вы можете обратиться к странице руководства `systemd.special(7)`.

F.2 Короче, как заставить `network.target` работать?

Это зависит от конфигурации вашей системы и конкретных требований ваших служб (см. выше). Многие службы управления сетью предоставляют возможность принудительной активации `network-online.target`, в результате чего `network.target` по своему эффекту становится практически эквивалентной `network-online.target`.

Если вы используете `NetworkManager`, вы можете задействовать специальную службу `NetworkManager-wait-online.service`:

```
systemctl enable NetworkManager-wait-online.service
```

Если же вы используете `systemd-networkd`, соответствующая служба будет называться `systemd-networkd-wait-online.service`:

```
systemctl enable systemd-networkd-wait-online.service
```

Включение такой службы позволит гарантировать, что загрузка продолжится только после того, как все заданные в настройках сетевые интерфейсы будут переведены в состояние UP и получают IP-адреса. Максимальное время ожидания — 90 секунд¹⁶². Обратите внимание, что включение подобных служб может сильно замедлить загрузку. Обе эти службы по умолчанию отключены¹⁶³.

Как альтернативный вариант, вы можете поменять непосредственно юнит-файл той службы, которая нуждается в сети, добавив туда опции `After=network-online.target` и `Wants=network-online.target`¹⁶⁴.

¹⁶¹Прим. перев.: Трансляция LSB-сущности `$network` в `network-online.target` введена в `systemd` начиная с `systemd 213`. В `systemd` версий до 212 включительно, `$network` транслировалась в `network.target`. Что приводило к довольно неожиданным эффектам — как уже упоминалось выше, активация `network.target` вовсе не означает, что сеть уже настроена. В связи с этим, официальная документация рекомендовала привязывать `network.target` к моменту запуска сети, например, через `NetworkManager-wait-online.service`. Соответствующие команды приведены чуть ниже по тексту. Они действуют даже в новых версиях `systemd`, но особого смысла уже не несут — задача синхронизации решается юнитом `network-online.target`, который в случае необходимости активируется автоматически (реализовано в юнит-файлах `systemd-networkd 213` и выше, `NetworkManager 0.9.9.95` и выше).

¹⁶²Прим. перев.: У `NetworkManager` в апстримной конфигурации по умолчанию сейчас используется значение 30 секунд. См. параметр `--timeout` программы `nm-online` в файле настроек службы `/usr/lib/systemd/system/NetworkManager-wait-online.service`. У `systemd-networkd` время ожидания ограничивается тайм-аутом `systemd` для запуска служб, по умолчанию 90 секунд.

¹⁶³Прим. перев.: У внимательного читателя может возникнуть вопрос: как же определяется момент активации `network-online.target`, если юниты, отвечающие за ожидание сети, отключены? На самом деле все довольно просто: при установке `NetworkManager/systemd-networkd` соответствующие `*-wait-online`-юниты прописываются в зависимости к `network-online.target` при помощи симлинков в `/etc/systemd/system/network-online.target.wants`. В результате, когда какая-нибудь служба укажет зависимость от `network-online.target`, будет автоматически активирована и соответствующая служба `*-wait-online`. Если таких служб нет, то и активации не произойдет. Однако, если вы принудительно включите службу `*-wait-online` при помощи приведенных выше команд, она будет прописана в зависимости уже к `multi-user.target`, а значит, будет активироваться при каждой загрузке.

¹⁶⁴Прим. перев.: Собственно, этот путь и является единственно разумным в большинстве случаев. Привязка `network.target` к `network-online.target`, описанная выше, фактически, является пережитком проблем старых версий `systemd` (212 и ниже), когда, по воле его разработчиков, LSB-сущность `$network` транслировалась в `network.target`.

F.3 А что делать нам, разработчикам?

Если вы — не администратор, а разработчик сетевой службы, то вам стоит задуматься не о том, что делать с `network.target`, а о том, как можно исправить вашу службу, чтобы она нормально реагировала на изменение состояния сети. Это позволит порадовать ваших пользователей (когда программа работает без дополнительной возни — это не может не радовать), а также уменьшит количество сообщений об ошибках (так как ваша программа уже не будет паниковать по пустякам). Кроме того, системы ваших пользователей будут грузиться быстрее, потому что их уже не будет задерживать необходимость ожидать появления сети (в случае с медленным DHCP-сервером, прогресс может быть весьма ощутимым).

Можно предложить несколько подходов к решению этой задачи:

- Отслеживайте изменения конфигурации сети при помощи [rtnetlink](#) и реагируйте соответствующим образом. Это наиболее правильный, но далеко не всегда самый простой способ.
- Если вы разрабатываете серверное приложение: слушайте только адреса `:::`, `:::1`, `0.0.0.0` и `127.0.0.1`. Все эти псевдо-адреса должны быть доступны постоянно. Если ваша программа будет слушать только их, ей будут глубоко безразличны изменения конфигурации сети.
- Если вы разрабатываете серверное приложение и вам нужно слушать некий заданный адрес: воспользуйтесь опцией `IP_FREEBIND`, доступной на Linux-системах. Она позволит вашей программе слушать даже те адреса, которые не присвоены локальным сетевым интерфейсам (в данный момент или вообще), что также сделает ваш код устойчивым к изменению сетевой конфигурации.

G Моя служба не может получить приоритет realtime¹⁶⁵

Итак, у вас есть служба, которая требует приоритет реального времени (realtime). Однако, когда вы пытаетесь запустить ее на системе, работающей под управлением systemd, ваша служба не может получить этот приоритет, хотя обладает всеми необходимыми для этого привилегиями. Вы хотите понять, почему так происходит, и как это можно исправить.

В чем же дело?

По умолчанию, `systemd` помещает каждую системную службу в свою контрольную группу в иерархии контроллера `cru`. Такое поведение дает очень полезный эффект: службы с множеством рабочих потоков (например, Apache с сотнями CGI-процессов), получают примерно такую же долю процессорного времени, как и службы с небольшим количеством рабочих потоков (например, MySQL). Проще говоря, процессорное время распределяется уже не *между процессами*, а *между службами*.

К сожалению, в настоящее время реализация `cru`-контроллера в Linux имеет один существенный недостаток: она требует явного задания `realtime`-бюджета времени (RT budget) для своих контрольных групп. Если же бюджет группы не задан, то ее процессы не смогут получить приоритет реального времени (соответствующая операция будет завершаться с ошибкой `EPERM` — «недостаточно полномочий»). `systemd` не может присваивать такой бюджет *каждой* группе, просто потому, что не знает, как его правильно делить между ними. Бюджет выдается в абсолютных единицах времени, и их общее количество ограничено. Мы не можем предложить механизм распределения бюджета, который бы подходил по умолчанию для большинства случаев. Поэтому, в конфигурации по умолчанию, приоритет реального времени недоступен для системных служб.

¹⁶⁵Прим. перев.: Перевод статьи «[My Service Can't Get Realtime!](#)» с официального сайта проекта, по состоянию на 2013-05-18 08:20:36 (коммит 2f77b).

Как это исправить?

Обойти это ограничение несложно. Существует несколько различных путей:

- Можно просто отключить использование спу-контроллера по умолчанию для системных служб. Для этого, задайте в файле `/etc/systemd/system.conf` параметр `DefaultControllers=` равным пустой строке, после чего перезагрузите систему. (Либо вы можете отключить контроллер `cpu` на этапе сборки ядра. `systemd` не пытается использовать контроллеры, которые не поддерживаются ядром.)
- Также вы можете отключить группировку по `cpu` только для конкретных служб. Для этого отредактируйте конфигурацию службы, добавив параметр

```
ControlGroup=cpu:/
```

в секцию `[Service]`. В результате, процессы данной службы будут помещены не в собственную контрольную группу (как это делалось по умолчанию), а в корневую контрольную группу иерархии `cpu`. Процессы из корневой группы располагают полным `realtime`-бюджетом.

- И наконец, вы можете явно выделить своей службе некоторый бюджет. Для этого добавьте в секцию `[Service]` строку наподобие

```
ControlGroupAttribute=cpu.rt_runtime_us 500000
```

Подробнее о правильном распределении бюджета читайте в [документации к ядру](#).

Последние две опции неприменимы к `SysV`-службам. Тем не менее, вы можете подготовить для таких служб соответствующие `service`-файлы, которые, помимо упомянутых выше параметров, содержат вызов соответствующего `init`-скрипта с аргументом `start` в `ExecStart=`, и с аргументом `stop` — в `ExecStop=`. (Также имеет смысл задать для них параметры `RemainAfterExit=1` и `Type=forking`.)

Отметим, что все вышесказанное касается только системных служб, и не затрагивает пользовательские сеансы. По умолчанию, программы пользователя размещаются в корневой группе контроллера `cpu`, и поэтому вышеописанные ограничения их не касаются.

Мы надеемся, что в будущем ядро будет исправлено таким образом, чтобы не требовать явного задания `realtime`-бюджета для получения приоритета реального времени (а при получении такого приоритета, бюджет процесса должен автоматически выделяться из бюджета ближайшей родительской группы). В идеале, мы хотели бы распространить практику выделения `cpu`-групп не только на системные службы, но и на пользовательские сеансы, чтобы уравнивать пользователей в правах на процессорное время, вне зависимости от того, сколько процессов запускает каждый конкретный пользователь.