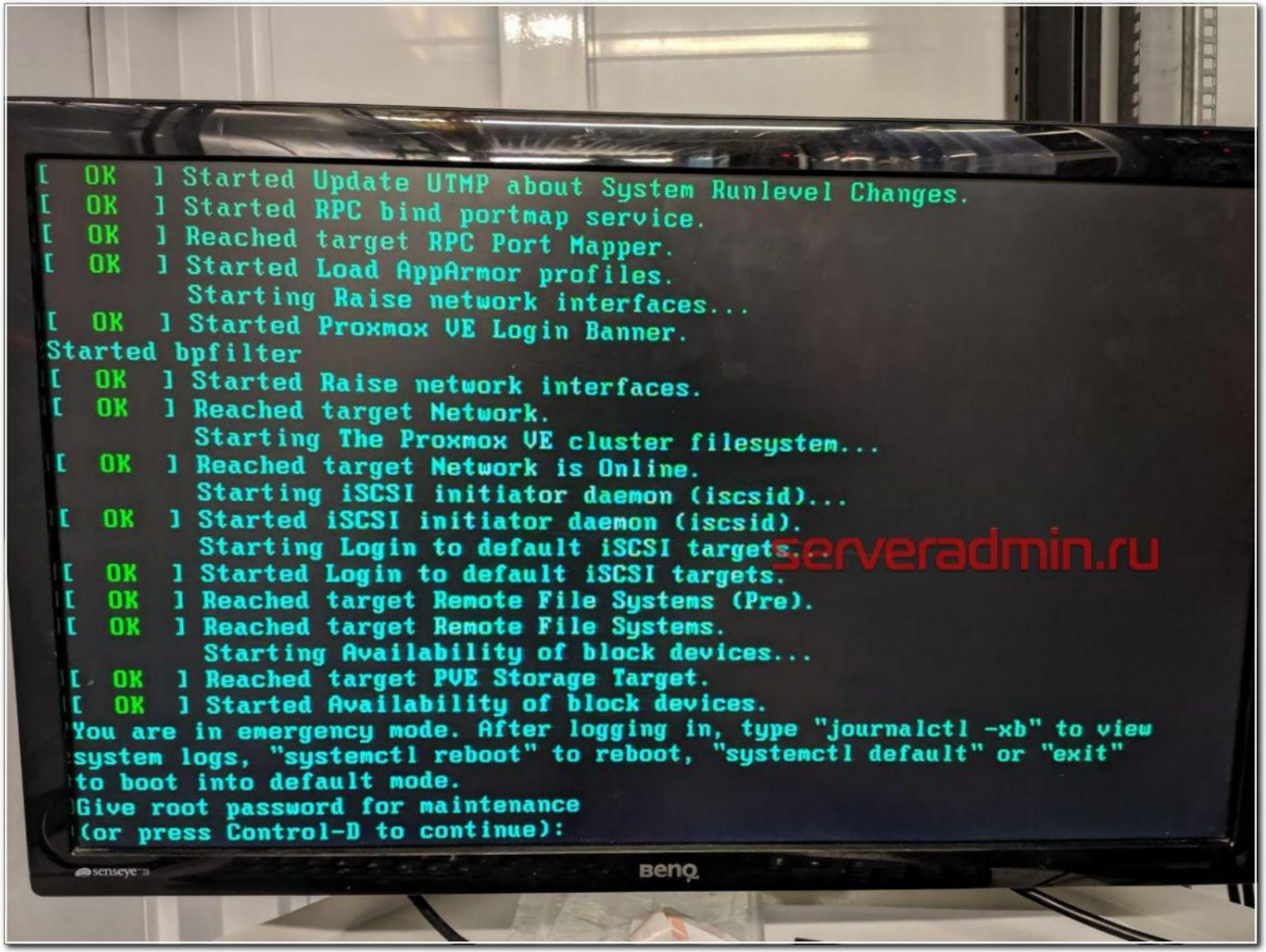

[OK] Started Update UTMP about System Runlevel Changes.
[OK] Started RPC bind portmap service.
[OK] Reached target RPC Port Mapper.
[OK] Started Load AppArmor profiles.
Starting Raise network interfaces...
[OK] Started Proxmox VE Login Banner.
Started bpfilter
[OK] Started Raise network interfaces.
[OK] Reached target Network.
Starting The Proxmox VE cluster filesystem...
[OK] Reached target Network is Online.
Starting iSCSI initiator daemon (iscsid)...
[OK] Started iSCSI initiator daemon (iscsid).
Starting Login to default iSCSI targets...
[OK] Started Login to default iSCSI targets.
[OK] Reached target Remote File Systems (Pre).
[OK] Reached target Remote File Systems.
Starting Availability of block devices...
[OK] Reached target PVE Storage Target.
[OK] Started Availability of block devices.
You are in emergency mode. After logging in, type "journalctl -xb" to view
system logs, "systemctl reboot" to reboot, "systemctl default" or "exit"
to boot into default mode.
Give root password for maintenance
(or press Control-D to continue):
serveradmin.ru
BenQ

Хочу рассказать поучительную историю, которая случилась со мной на днях. На одном из серверов в ЦОД вышел из строя диск в составе рейда mdadm. Ситуация типовая, с которой регулярно сталкиваюсь. Оставил заявку в техподдержку на замену диска с указанием диска, который надо поменять. В цоде заменили рабочий диск и оставили сбойный. Дальше история, как я решал возникшую проблему.

Содержание:

Цели статьи

1. Рассказать поучительную историю о том, какие могут быть проблемы при аренде серверов в ЦОД.
2. Показать на примере, как надо действовать при выходе из строя диска в рейде mdadm.
3. Простыми словами объяснить, в чем разница между программным и аппаратным рейдом.

Введение

Когда первый раз сталкиваешься с рукожопством сотрудников техподдержки дата центра, впадаешь в ступор и думаешь, ну как так то? Сейчас я спокойно отношусь к таким ситуациям и действую исходя из самых худших ожиданий. На днях я столкнулся с ситуацией, когда мне заменили не тот диск в сервере с RAID1. Вместо сбойного диска вынули рабочий и заменили чистым. К счастью все закончилось хорошо, но обо всем по порядку.

Не скажу, что у меня прям большой опыт аренды серверов, но он есть. Я регулярно обслуживаю 10-15 серверов, расположенных в разных дата центрах, как российских, так и европейских. Первый негативный опыт я получил именно в Европе и был очень сильно удивлен и озадачен. Я, как и многие, был под влиянием либеральной пропаганды на тему того, что у нас все плохо, а вот Европа образец надежности, стабильности и сервиса. Как же я ошибался. Сейчас отдам предпочтение нашим дата центрам. По моему мнению и опыту, у нас тех поддержка и сервис в целом лучше, чем там, без привязки к стоимости. В Европе дешевле схожие услуги, так как там масштабы сервисов в разы больше.

Приведу несколько примеров косяков саппорта, с которыми сталкивался.

1. При заказе приватной сети у хостера leaseweb.com развалили весь сервис на несколько часов. Был крупный проект у хостера. Рос постепенно, с нуля. Покупался сервер за сервером. Когда серверов стало много, решили, что надо объединяться в единую локалку. У хостера есть такая услуга и называется приватная сеть. Так как сервера сильно разнесены по стойкам, хостер сказал, что надо переносить все поближе друг к другу. Согласовали время для переноса серверов и все остальное. Хостер заранее выдал все сетевые настройки. После того, как хостер все сервера перенес и отчитался в тикете об успешном окончании, началась свистопляска. На части серверов указанные сетевые настройки не приводили к доступности. Часть серверов не видели друг друга. Началась длительная переписка с техподдержкой, где предлагали то включить dhcp, то отключить, и кучу всяких других бесполезных действий. В итоге оказалось, что просто напутали и с сетевыми настройками, и сервера не туда перенесли. Это была жесть. Плюс все общение на английском. С тех пор я больше никогда не заказываю подобных услуг на работающем сервисе. Если надо объединяться, то настраиваю vpn на текущих сетевых подключениях. И вообще обращаюсь к тех поддержке по минимуму. Если нужны глобальные изменения — плавных переход на дублирующую систему.
2. Как-то раз перед новым годом, 31-го декабря в 16 часов отрубился полностью крупный проект. Как оказалось, хостер выполнял какие-то работы в стойке и по ошибке вырубил питание на нашем сервере, который был балансером и точкой входа для всех запросов. В итоге весь сайт и сервис лег для посетителей. Повезло, что где-то часа за 2 они это обнаружили и отписались в тикете, что все ОК. А при первоначальном запросе сказали, что сейчас будем разбираться, но все инженеры уже празднуют, так что ничего не обещаем.
3. Ну и под конец классика. Заменили не тот диск в рейде. Вместо сбойного вынули рабочий. Каким-то чудом рейд не развалился. Все зависло, вернули обратно рабочий диск и перегрузили сервер.

Было много всяких инцидентов помельче, нет смысла описывать. Хотя нет, один все же опишу. Устанавливал свой сервер в ЦОД. Решил пойти в маш зал и проконтролировать монтаж. Если есть такая возможность, крайне рекомендую ей воспользоваться. Местный рукожоп неправильно прикрепил салазки и

сервер во время монтажа стал падать. Я его поймал, тем спас его и сервера других клиентов. В итоге помог с монтажом. Сам бы он просто не справился. Я не представляю, что было, если бы я не пошел в машзал. К чести руководства, я написал претензию, где подробно описал данный случай и попросил бесплатно месячную аренду. Мне ее предоставили. Советую всем так поступать. Зачастую, руководство может быть не в курсе того, что происходит в реальности. Надо давать обратную связь.

Уровень моего доверия к тех поддержке дата центров и хостингов вы примерно представляете :) Ну и вот случилось очередное ЧП. Подробнее остановлюсь на этой ситуации, так как она случилась вчера, свежи воспоминания.

Замена диска в рейде mdadm

Речь пойдет о дешевых дедиках от selectel. Я их много где использую и в целом готов рекомендовать. Это обычные десктопные системники за скромные деньги. Свое мнение об этих серверах, а так же сравнение с полноценными серверами сделаю в конце, в отдельном разделе.

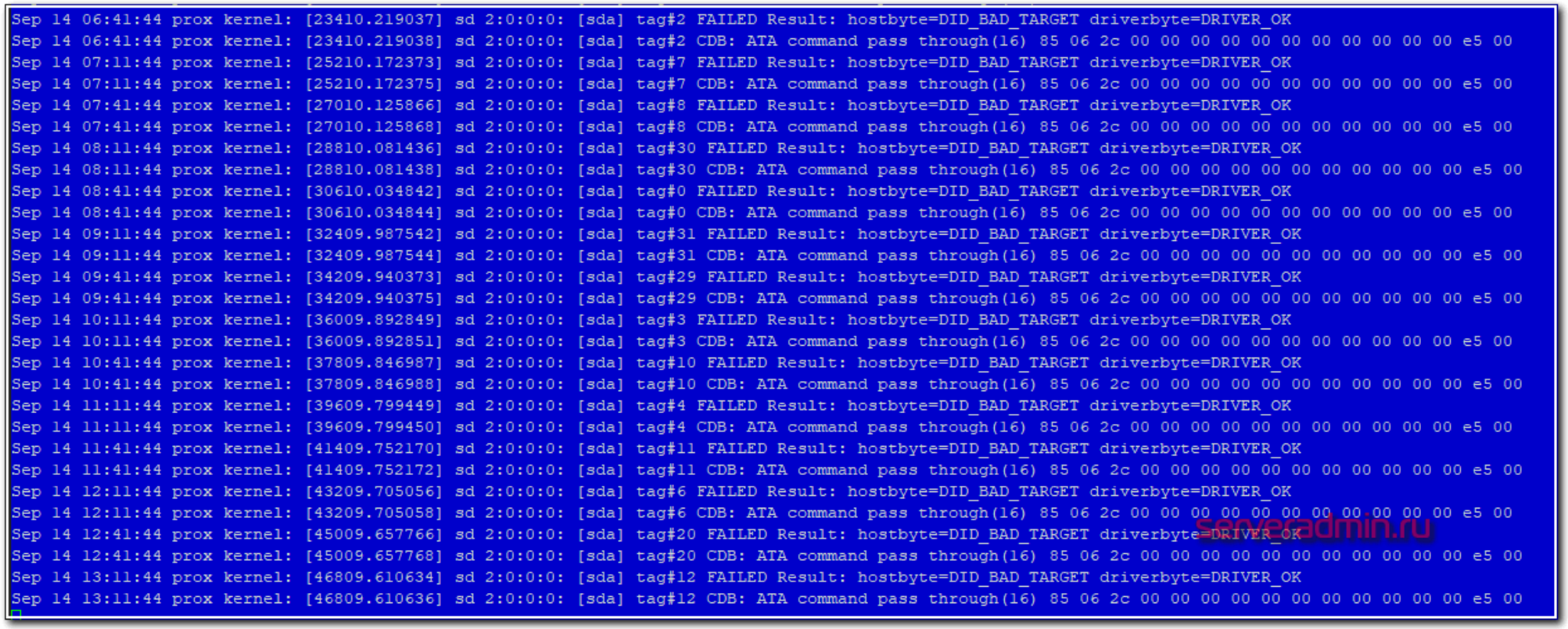
На сервере была установлена система Debian из стандартного шаблона Selectel. Вот особенности дисковой подсистемы этих серверов и шаблона.

- 2 ssd диска, объединенные в mdadm
- /boot раздел на /dev/md0 размером 1G
- корень / на /dev/md1 и поверх lvm на весь массив

В целом, хорошая и надежная разбивка, чему будет подтверждение дальше. На сервере был установлен proxmox, настроен мониторинг mdadm. Мониторинг дисков не сделал. В какой-то момент получил уведомление в zabbix, что mdadm развалился. Сервер при этом продолжал работать. Ситуация штатная. Пошел в консоль сервера, чтобы все проверить. Посмотрел состояние рейда.

```
# cat /proc/mdstat
```

Убедился, что один диск выпал из массива. В системном логе увидел следующее.



Попробовал посмотреть информацию о выпавшем диске.

```
# smartctl -i /dev/sda
```

Информации не было, утилита показывала ошибку обращения к диску. Получилось посмотреть модель и серийный номер только работающего диска.

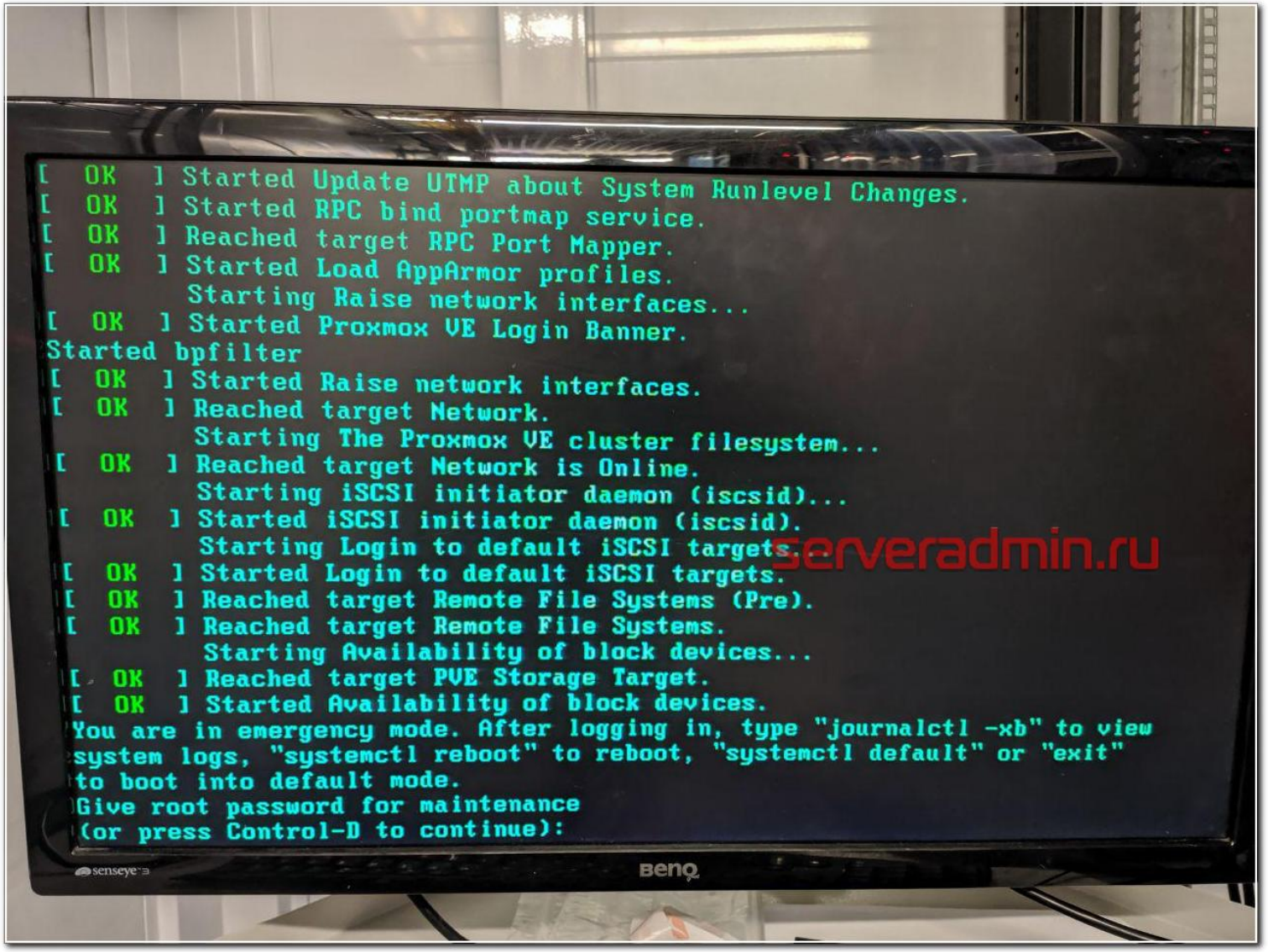
Я не стал разбираться, что там к чему с диском. Если вижу проблемы, сразу меняю. Предупредил заказчика, что с диском проблемы, нужно планировать

замену. Так как железо десктопное, «сервер» надо выключать. Согласовали время после 22 часов. Я в это время уже сплю, поэтому написал тикет в тех поддержку, где указал время и серийный номер диска, который нужно было **оставить**. Я сделал на этом акцент, объяснил, что сбойный диск не отвечает, поэтому его серийник посмотреть не могу. Расписал все очень подробно, чтобы не оставить почвы для недопонимания или двойного толкования. Я в этом уже спец, но все равно не помогло.

Я спокойно согласился на эту операцию, потому что часто делаются бэкапы и они гарантированно рабочие. Настроен мониторинг бэкапов и делается регулярное полуручное восстановление из них. Договоренность была такая, что хостер после замены дожидается появления окна логина, а заказчик проверяет, что сайт работает. Все так и получилось — сервер загрузился, виртуалки поднялись, сайт заработал. На том завершили работы.

Утром я встал и увидел, что весь системный лог в ошибках диска, рабочего диска в системе нет, а есть один глючный и один новый. Сразу же запустил на всякий случай ребилд массива и он вроде как даже прошел без ошибок. Перезагрузка временно оживила сбойный диск. В принципе, на этом можно было бы остановиться, заменить таки сбойный диск и успокоиться. Но смысл в том, что этот сбойный диск почти сутки не был в работе и данные на нем старые. Это не устраивало. Потом пришлось бы как-то склеивать эти данные с данными из бэкапов. В случае с базой данных это не тривиальная процедура. Созвонился с заказчиком и решили откатываться на рабочий диск, который вытащили накануне ночью.

Я создал тикет и попросил вернуть рабочий диск на место. К счастью, он сохранился. К нему добавить еще один полностью чистый. Хостер оперативно все сделал и извинился. В завершении прислал скриншот экрана сервера.


[OK] Started Update UTMP about System Runlevel Changes.
[OK] Started RPC bind portmap service.
[OK] Reached target RPC Port Mapper.
[OK] Started Load AppArmor profiles.
Starting Raise network interfaces...
[OK] Started Proxmox VE Login Banner.
Started bpfilter
[OK] Started Raise network interfaces.
[OK] Reached target Network.
Starting The Proxmox VE cluster filesystem...
[OK] Reached target Network is Online.
Starting iSCSI initiator daemon (iscsid)...
[OK] Started iSCSI initiator daemon (iscsid).
Starting Login to default iSCSI targets...
[OK] Started Login to default iSCSI targets.
[OK] Reached target Remote File Systems (Pre).
[OK] Reached target Remote File Systems.
Starting Availability of block devices...
[OK] Reached target PVE Storage Target.
[OK] Started Availability of block devices.
You are in emergency mode. After logging in, type "journalctl -xb" to view
system logs, "systemctl reboot" to reboot, "systemctl default" or "exit"
to boot into default mode.
Give root password for maintenance
(or press Control-D to continue):
serveradmin.ru
BenQ

И самоустранился. Дальше решать проблему загрузки он предложил загрузившись в **режиме rescue**. Этот режим доступен через панель управления сервером в админке, даже если сервер не имеет ipmi консоли. Как я понял, по сети загружается какой-то live cd для восстановления. Я в нем загрузился, убедился, что данные на месте, но понять причину ошибки не смог. Может быть и смог бы, если бы дольше покопался, но это очень неудобно делать, не видя реальной консоли сервера. Я попросил подключить к серверу kvm over ip, чтобы я мог подключиться к консоли. Тех поддержка без лишних вопросов оперативно это сделала.

К слову, мне известны случаи, когда техподдержка selectel потом сама чинила загрузку и возвращала mdadm в рабочее состояние. Видел такие переписки в тикетах у своих клиентов до того, как они обращались ко мне. Но я не стал настаивать на таком решении проблемы, так как боялся, что будет хуже. К тому же это было утро воскресенья и специалистов, способных это сделать, могло просто не быть. Плюс, я не думаю, что они обладали бы большими компетенциями, чем я. Я бы за их зарплату не пошел работать в ЦОД.

После того, как я подключился к консоли сервера, восстановление загрузки было делом техники.

You are in emergency mode

У меня много примеров того, как я восстанавливал загрузку сломавшихся linux дистрибутивов.

- Kernel panic not syncing: VFS: Unable to mount root fs
- Booting from Hard Disk error, Entering rescue mode
- Так же сломавшуюся загрузку я чинил при переносе виртуальных машин с одного гипервизора на другой — восстановление загрузки linux сервера
- В статье про бэкап и перенос linux сервера я тоже касаюсь темы починки загрузчика

В данной ситуации с mdadm я был уверен, что все получится, так как сам массив с системой жив, данные доступны. Надо только разобраться, почему система не загружается. Напомню, что ошибка загрузки была следующая.

```
You are in emergency mode. After logging in, type "journalctl -xb" to view system logs, "systemctl reboot" to reboot,
"systemctl default" to try again to boot into default mode.
Give root password for maintenance
(or type Control-D to continue):
```

Дальше нужно ввести пароль root и вы окажетесь в системной консоли. Первым делом я проверил состояние массива mdadm.

```
# cat /proc/mdstat
Personalities : [raid1] [linear] [multipath] [raid0] [raid6] [raid5] [raid4] [raid10]
md1 : active raid1 sda3[1]
      467716096 blocks super 1.2 [2/2] [U_]
      bitmap: 2/4 pages [8KB], 65536KB chunk

md0 : inactive raid1 sda2[1](S)
      999424 blocks super 1.2 [2/2] [U_]
```

Состояние массива md0, на котором располагается раздел /boot — **inactive**. Вот, собственно, и причина того, почему сервер не загружается. Судя по всему, когда был подключен сбойный диск, mdadm отключил массив, чтобы предотвратить повреждение данных. Не понятно, почему именно на разделе /boot, но по факту было именно это. Из-за того, что массив остановлен, загрузиться с него не получалось. Я остановил массив и запустил снова.

```
# mdadm -S /dev/md0
# mdadm -R /dev/md0
```

После этого массив вышел из режима inactive и стал доступен для дальнейшей работы с ним. Я перезагрузил сервер и убедился, что он нормально загружается. Сервер фактически был в рабочем состоянии, просто с развалившимся массивом mdadm, без одного диска.

Если вам это не поможет, предлагаю еще несколько советов, что можно предпринять, чтобы починить загрузку. Первым делом проверьте файл */etc/fstab* и посмотрите, какие разделы и как там монтируются. Вот мой пример этого файла.

```
/dev/mapper/vg0-root /               ext4    errors=remount-ro 0       1
UUID=789184ea-50e4-4788-98f4-b500928d35c8 /boot           ext3    defaults        0       2
/dev/mapper/vg0-swap_1 none            swap    sw              0       0
```

Вам нужно убедиться, что указанные lvm разделы */dev/mapper/vg0-root* и */dev/mapper/vg0-swap_1* действительно существуют. Для этого используйте команду:

```
# lvs
  LV     VG  Attr       LSize   Pool Origin Data%  Meta%  Move Log Cpy%Sync Convert
  root   vg0 -wi-ao-- – 441.28g
  swap_1 vg0 -wi-ao-- –  <4.77g
```

Подробно об этой команде, о работе с lvm и вообще с дисками я рассказываю в отдельной статье — настройка диска в debian. Если с lvm разделами все нормально, проверьте /boot. У меня он монтируется по uuid. Посмотреть список uuid всех разделов можно командой.

```
# blkid
/dev/sda1: PARTUUID="5668dd38-a5e2-495e-856f-af0547a9d907"
/dev/sda2: UUID="3f8c654b-5c1d-cb5c-3b13-6bd5925c995f" UUID_SUB="8bf2478f-ec17-a055-1f70-d20dd13a19b3"
LABEL="Jellicent:0" TYPE="linux_raid_member" PARTUUID="7e3210cc-f267-4372-85e2-1dae7731a6bb"
/dev/sda3: UUID="c123309f-fc71-7b99-2fee-9cd567bd6f9d" UUID_SUB="e0697294-88dc-d6f5-b61c-bdf6091bfceb"
LABEL="Jellicent:1" TYPE="linux_raid_member" PARTUUID="df1e1100-a01a-46da-8fd3-81ed4c010c11"
/dev/sdb1: PARTUUID="5668dd38-a5e2-495e-856f-af0547a9d907"
/dev/sdb2: UUID="3f8c654b-5c1d-cb5c-3b13-6bd5925c995f" UUID_SUB="a8431f0f-6d98-3ca5-1dc7-da62082a4a8c"
LABEL="Jellicent:0" TYPE="linux_raid_member" PARTUUID="7e3210cc-f267-4372-85e2-1dae7731a6bb"
/dev/sdb3: UUID="c123309f-fc71-7b99-2fee-9cd567bd6f9d" UUID_SUB="ea65601c-8a17-c654-735a-1e5c892e6584"
LABEL="Jellicent:1" TYPE="linux_raid_member" PARTUUID="df1e1100-a01a-46da-8fd3-81ed4c010c11"
/dev/md0: UUID="789184ea-50e4-4788-98f4-b500928d35c8" TYPE="ext3"
/dev/md1: UUID="O9Qq20-35Uk-n1Lx-d993-xnkr-9jCi-l0dNVy" TYPE="LVM2_member"
/dev/mapper/vg0-root: UUID="8eccb650-dd7e-4643-8898-9dd5befea121" TYPE="ext4"
/dev/mapper/vg0-swap_1: UUID="186dfb1a-7c7e-4750-804c-cc507a38f514" TYPE="swap"
```

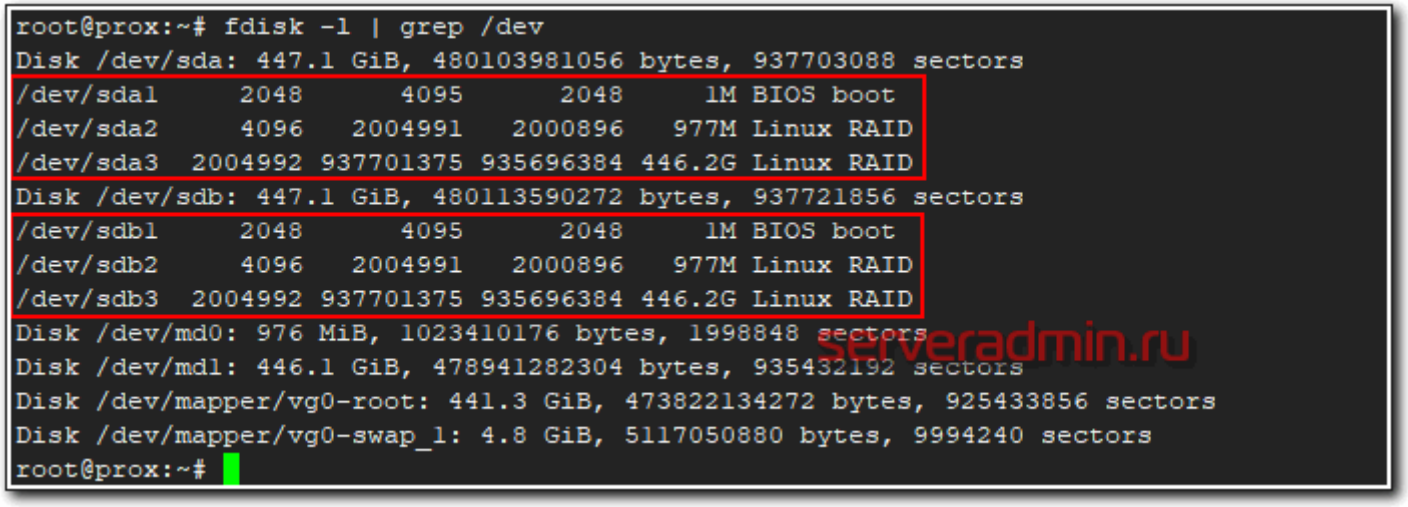
Как вы видите, у меня uuid раздела для загрузки полностью совпадает с тем, что указано в fstab. Если по какой-то причине uuid изменился (разобрали и собрали новый массив), отредактируйте fstab.

Все дальнейшие действия я делал уже по ssh. Скопировал таблицу разделов с рабочего диска sda на чистый sdb.

```
# sfdisk -d /dev/sda | sfdisk /dev/sdb
```

Проверил таблицы разделов и убедился, что они идентичные.

```
# fdisk -l | grep /dev
```



Скопировал раздел **BIOS boot partition** с рабочего диска на новый.

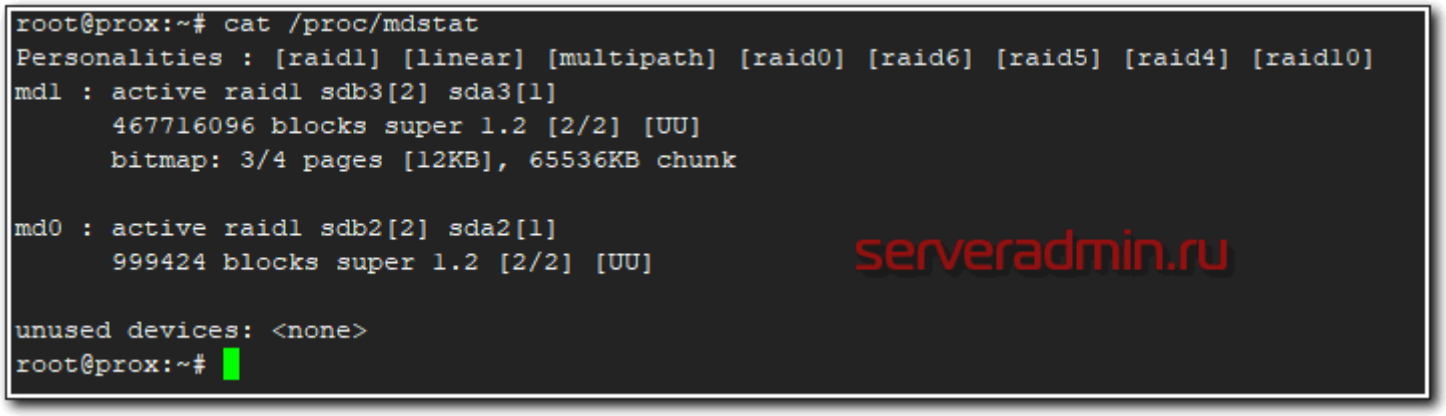
```
# dd if=/dev/sda1 of=/dev/sdb1 bs=512
```

Потом добавил разделы диска sdb2 и sdb3 в рейд массив.

```
# mdadm – add /dev/md0 /dev/sdb2
# mdadm – add /dev/md1 /dev/sdb3
```

Дождался окончания ребилда и убедился, что он прошел. Проверил состояние массива.

```
# cat /proc/mdstat
```



В завершении устанавливаем загрузчик на оба диска.

```
# dpkg-reconfigure grub-pc
```

После этого я перезагрузился и убедился, что все работает нормально. По хорошему, теперь надо было бы поменять загрузочный диск с первого на второй и убедиться, что со второго тоже нормально грузится. Я не стал этого делать, и так простой и так был велик. Главное, чтобы массив был на месте, а починить загрузку, если что, дело техники.

Вот и все по замене диска в массиве mdadm. После доступа к консоли сервера, мне потребовалось минут 10, чтобы вернуть сервер в рабочее состояние.

В чем отличия программного и аппаратного рейда

Сейчас расскажу, чем принципиально отличается программный рейд контроллер (mdadm) от аппаратного, для тех, кто этого до конца не понимает. Если бы у меня вышел из строя диск на аппаратном рейд контроллере, установленном в полноценный сервер, проблема по замене сбойного диска в RAID решалась бы в следующей последовательности:

1. Рейд контроллер оповещает о том, что с диском проблемы и выводит его из работы. В случае с софтовым рейдом система может зависнуть в случае

проблем с диском, прежде чем пометит его как проблемный и перестанет к нему обращаться.
2. Я оставляю тикет в тех поддержку, где прошу заменить сбойный диск. Информацию о нем я посмотрю в панели управления рейд контроллером.
3. Сотрудник тех поддержки видит сбойный диск, так как индикация на нем, скорее всего, будет мигать красной лампочкой. Это не гарантия того, что рукожоп все сделает правильно, но тем не менее, шансов, что он ошибется, меньше. Я сталкивался с ситуацией, когда и в этом случае диск меняли не тот.
4. При появлении нового диска raid контроллер автоматически начинает ребил массива.

Если же у вас в сервере уже установлен запасной диск на случай выхода из строя диска в составе raid массива, то все еще проще:

1. При выходе из строя диска, контроллер помечает его как сбойный, вводит в работу запасной диск и начинает ребилд.
2. Вы получаете оповещение о том, что вышел из строя диск и оставляете тикет в тех поддержку на замену запасного диска.

И это все. В обоих случаях у вас вообще нет простоя. Вот принципиальная разница между mdadm и железным raid контроллером. Стоимость полноценного сервера с контроллером и постоянным ipmi доступом к консоли в среднем в 3 раза выше, чем у сервера на десткопном железе с софтовым рейдом при схожей производительности. Это все при условии, что вам достаточно одного процессора и 64G памяти. Это потолок для десктопных конфигураций. Дальше считайте сами, что вам выгоднее. Если возможен простой в несколько часов на замену диска или других комплектующих, то смело можно использовать десктопное железо. Mdadm обеспечивает сопоставимую гарантию сохранности данных в сравнении с железным контроллером. Вопрос лишь в простое и производительности. Ну и своевременные бэкапы добавляют уверенности в том, что вы переживете неполадки с железом.

При использовании железного рейда на hdd дисках, есть возможно получить очень значительный прирост скорости за счет кэша контроллера. Для ssd дисков я особо не замечал разницы. Но это все на глазок, никаких замеров и сравнений я не делал. Нужно еще понимать, что десктопное железо в целом менее надежное. К примеру, в том же селектеле на дешевых серверах я ловил перегрев или очень высокую температуру дисков. Прыгала в районе 55-65 градусов. Все, что ниже 60-ти, тех поддержка футболила, говоря, что это допустимая температура, судя по документации к дискам. Это так и есть, но мы же понимаем, что диск, постоянно работающий на 59 градусах с бОльшей долей вероятности выйдет из строя.

Вот еще пример разницы в железе. Если у вас в нормальном сервере выйдет из строя планка памяти, сервер просто пометит ее как сбойную и выведет из работы. Информацию об этом вы увидите в консоли управления — ilo, idrac и т.д. В десктопном железе у вас просто будет постоянно виснуть сервер и вам придется долго выяснять, в чем же проблема, так как доступа к железу у вас нет, чтобы проще было запланировать тестирование сервера. А если вы закажете это у тех поддержки, то есть ненулевая вероятность, что станет хуже — сервер уронят, перепутают провода подключения дисков и т.д. В общем, это всегда риск. Проще сразу съезжать с такой железки на другую.

Заключение

Надеюсь, моя статья была интересной. Для тех, кто никогда не работал с ЦОДами будет полезно узнать, чего можно от них ожидать. Я скучаю по временам, когда все сервера, которые я администрировал, были в серверной, куда никому не было доступа и куда я мог в любой момент попасть и проверить их. Сейчас все стало не так. И твои сервера уже не твои. Их может сломать, уронить, что-то перепутать сотрудник тех поддержки дата центра.

Сейчас большой тренд на переход в облака. Я смотрю на эти облака и не понимаю, как с ними можно нормально взаимодействовать. Заявленная производительность не гарантированная, нагрузка плавает в течении суток. Упасть может в любой момент и ты не будешь понимать вообще в чем проблема. Твои виртуалки могут быть по ошибке удалены и кроме извинений и компенсации в 3 копейки ты ничего не получить. Каждое обращение в ТП как лотерея. Думаешь, что сломают в этот раз. Если сервера железные, то когда пишу тикет на доступ к железу, я морально и технически всегда готов к тому, что этот сервер сейчас отключится и я больше не смогу к нему подключиться.

В целом, опыт работы с облаками у меня негативный. Несколько раз пробовал для сайтов и все время съезжал. Нет гарантированного времени отклика. А это сейчас фактор ранжирования. Для очень быстрого сайта остается только один вариант — свое железо, а дальше уже кому какое по карману. Зависит от надежности и допустимого времени простоя.

Я про облака заговорил, потому что тенденции к тому, что от железных серверов надо отказываться и все переносить в облака. С одной стороны удобно должно быть. Как минимум, не будет указанных выше в статье проблем. А с другой стороны добавляется куча других проблем. Я пока сижу на железяках разного качества и стоимости. А у вас как?

- Практические навыки внедрения виртуализации KVM, oVirt, Xen.
- Кластеризация сервисов на базе pacemaker,k8s, nomad и построение дисковых кластеров на базе ceph, glaster, linstore.

Проверьте себя на вступительном тесте и смотрите подробнее программу по .