Загрузка Линукса по сети на примере вычислительного кластера

М. Ю. Улейский (michael)

Линуксовка Владивостокского ЛУГ'а 30.10.2010

Содержание

1. Описание задачи
   - Как оно происходит
   - Что необходимо
   - Рабочая система
2. Настройка
   - DHCP
   - PXELINUX
   - TFTP
   - Конфигурация ядра
   - Настройка клиентов

Как оно происходит

1. Сетевой загрузчик на клиенте отправляет в сеть DHCP-запрос.
2. DHCP-сервер отвечает на запрос. В ответе содержится ip-адрес клиента, адрес TFTP-сервера и имя файла, который клиент должен загрузить и запустить.
3. Клиент загружает с TFTP-сервера и запускает загрузчик следующего уровня — PXELINUX.
4. PXELINUX загружает с TFTP-сервера соответствующий конфигурационный файл.
5. PXELINUX загружает образ ядра и передаёт ему управление.
6. Ядро монтирует корневую файловую систему по NFS и запускает /sbin/init.
7. ??????????????
8. PROFIT!

Как оно происходит

1. Сетевой загрузчик на клиенте отправляет в сеть DHCP-запрос.
2. DHCP-сервер отвечает на запрос. В ответе содержится ip-адрес клиента, адрес TFTP-сервера и имя файла, который клиент должен загрузить и запустить.
3. Клиент загружает с TFTP-сервера и запускает загрузчик следующего уровня — PXELINUX.
4. PXELINUX загружает с TFTP-сервера соответствующий конфигурационный файл.
5. PXELINUX загружает образ ядра и передаёт ему управление.
6. Ядро монтирует корневую файловую систему по NFS и запускает /sbin/init.
7. ??????????????
8. PROFIT!

Как оно происходит

1. Сетевой загрузчик на клиенте отправляет в сеть DHCP-запрос.
2. DHCP-сервер отвечает на запрос. В ответе содержится ip-адрес клиента, адрес TFTP-сервера и имя файла, который клиент должен загрузить и запустить.
3. Клиент загружает с TFTP-сервера и запускает загрузчик следующего уровня — PXELINUX.
4. PXELINUX загружает с TFTP-сервера соответствующий конфигурационный файл.
5. PXELINUX загружает образ ядра и передаёт ему управление.
6. Ядро монтирует корневую файловую систему по NFS и запускает /sbin/init.
7. ??????????????
8. PROFIT!

Как оно происходит

1. Сетевой загрузчик на клиенте отправляет в сеть DHCP-запрос.
2. DHCP-сервер отвечает на запрос. В ответе содержится ip-адрес клиента, адрес TFTP-сервера и имя файла, который клиент должен загрузить и запустить.
3. Клиент загружает с TFTP-сервера и запускает загрузчик следующего уровня — PXELINUX.
4. PXELINUX загружает с TFTP-сервера соответствующий конфигурационный файл.
5. PXELINUX загружает образ ядра и передаёт ему управление.
6. Ядро монтирует корневую файловую систему по NFS и запускает /sbin/init.
7. ??????????????
8. PROFIT!

Как оно происходит

1. Сетевой загрузчик на клиенте отправляет в сеть DHCP-запрос.
2. DHCP-сервер отвечает на запрос. В ответе содержится ip-адрес клиента, адрес TFTP-сервера и имя файла, который клиент должен загрузить и запустить.
3. Клиент загружает с TFTP-сервера и запускает загрузчик следующего уровня — PXELINUX.
4. PXELINUX загружает с TFTP-сервера соответствующий конфигурационный файл.
5. PXELINUX загружает образ ядра и передаёт ему управление.
6. Ядро монтирует корневую файловую систему по NFS и запускает /sbin/init.
7. ??????????????
8. PROFIT!

Как оно происходит

1. Сетевой загрузчик на клиенте отправляет в сеть DHCP-запрос.
2. DHCP-сервер отвечает на запрос. В ответе содержится ip-адрес клиента, адрес TFTP-сервера и имя файла, который клиент должен загрузить и запустить.
3. Клиент загружает с TFTP-сервера и запускает загрузчик следующего уровня — PXELINUX.
4. PXELINUX загружает с TFTP-сервера соответствующий конфигурационный файл.
5. PXELINUX загружает образ ядра и передаёт ему управление.
6. Ядро монтирует корневую файловую систему по NFS и запускает /sbin/init.
7. ??????????????
8. PROFIT!

Как оно происходит

1. Сетевой загрузчик на клиенте отправляет в сеть DHCP-запрос.
2. DHCP-сервер отвечает на запрос. В ответе содержится ip-адрес клиента, адрес TFTP-сервера и имя файла, который клиент должен загрузить и запустить.
3. Клиент загружает с TFTP-сервера и запускает загрузчик следующего уровня — PXELINUX.
4. PXELINUX загружает с TFTP-сервера соответствующий конфигурационный файл.
5. PXELINUX загружает образ ядра и передаёт ему управление.
6. Ядро монтирует корневую файловую систему по NFS и запускает /sbin/init.
7. ??????????????
8. PROFIT!

Как оно происходит

1. Сетевой загрузчик на клиенте отправляет в сеть DHCP-запрос.
2. DHCP-сервер отвечает на запрос. В ответе содержится ip-адрес клиента, адрес TFTP-сервера и имя файла, который клиент должен загрузить и запустить.
3. Клиент загружает с TFTP-сервера и запускает загрузчик следующего уровня — PXELINUX.
4. PXELINUX загружает с TFTP-сервера соответствующий конфигурационный файл.
5. PXELINUX загружает образ ядра и передаёт ему управление.
6. Ядро монтирует корневую файловую систему по NFS и запускает /sbin/init.
7. ??????????????
8. PROFIT!

Что необходимо

1. Загрузчик в биосе или на сетевой карте клиента.
2. DHCP- или BOOTP-сервер (dhcp).
3. TFTP-сервер (atftp).
4. PXELINUX (syslinux).
5. Соответствующим образом сконфигурированное ядро.
6. NFS-сервер.

Что необходимо

1. Загрузчик в биосе или на сетевой карте клиента.
2. DHCP- или BOOTP-сервер (dhcp).
3. TFTP-сервер (atftp).
4. PXELINUX (syslinux).
5. Соответствующим образом сконфигурированное ядро.
6. NFS-сервер.

Что необходимо

1. Загрузчик в биосе или на сетевой карте клиента.
2. DHCP- или BOOTP-сервер (dhcp).
3. TFTP-сервер (atftp).
4. PXELINUX (syslinux).
5. Соответствующим образом сконфигурированное ядро.
6. NFS-сервер.

Что необходимо

1. Загрузчик в биосе или на сетевой карте клиента.
2. DHCP- или BOOTP-сервер (dhcp).
3. TFTP-сервер (atftp).
4. PXELINUX (syslinux).
5. Соответствующим образом сконфигурированное ядро.
6. NFS-сервер.

Что необходимо

1. Загрузчик в биосе или на сетевой карте клиента.
2. DHCP- или BOOTP-сервер (dhcp).
3. TFTP-сервер (atftp).
4. PXELINUX (syslinux).
5. Соответствующим образом сконфигурированное ядро.
6. NFS-сервер.

Что необходимо

1. Загрузчик в биосе или на сетевой карте клиента.
2. DHCP- или BOOTP-сервер (dhcp).
3. TFTP-сервер (atftp).
4. PXELINUX (syslinux).
5. Соответствующим образом сконфигурированное ядро.
6. NFS-сервер.

Рабочая система

- Сервер 172.16.1.1 с установленными dhcp-3.0.5, atftp-0.7 и pxelinux-3.11.
- 12 клиентов 172.16.1.10–21 — двухядерные Athlon 64.
- 20 клиентов 172.16.1.30–49 — четырёхядерные Intel Core2 Quad.

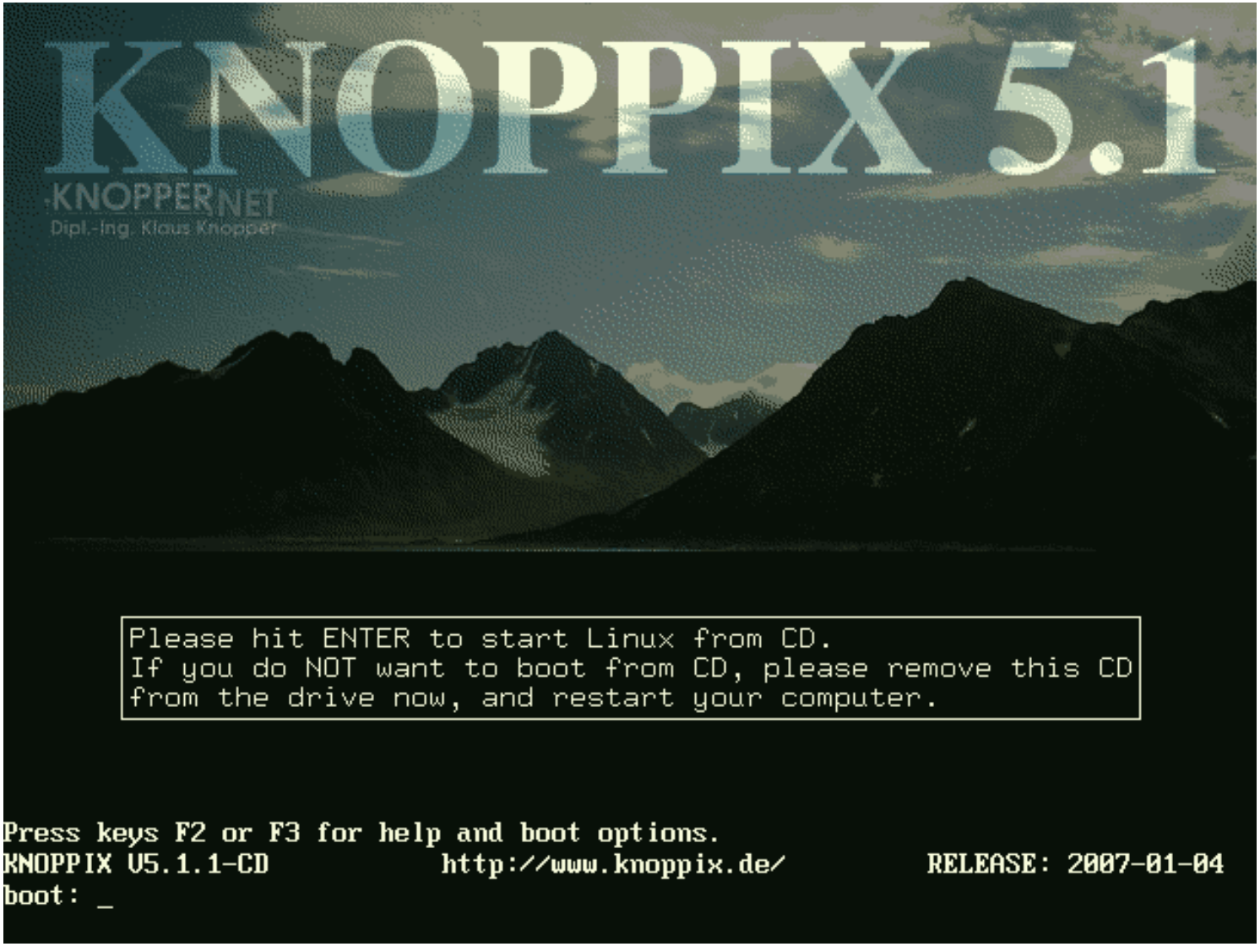
Как это выглядит

DHCP

Фрагмент /etc/dhcp/dhcpd.conf:

```
ddns-update-style ad-hoc;
subnet 172.16.1.0 netmask 255.255.255.0 {
}
# First cluster, first rack
host cartman {
 hardware ethernet 00:18:f3:70:91:b9;
 fixed-address 172.16.1.10;
 filename "/tftpboot/pxelinux.0";
 next-server 172.16.1.1;
}
host stan {
 hardware ethernet 00:18:f3:70:91:c1;
 fixed-address 172.16.1.11;
 filename "/tftpboot/pxelinux.0";
 next-server 172.16.1.1;
}
```

PXELINUX — возможности



PXELINUX — конфигурация (начало)

Пусть клиент имеет MAC-адрес 00:18:F3:70:91:B9 и ip-адрес 172.16.1.10. Тогда PXELINUX будет искать на TFTP-сервере свой конфигурационный файл в следующем порядке:

1. /tftpboot/pxelinux.cfg/01-00-18-f3-70-91-b9 (буквы в нижнем регистре!)
2. /tftpboot/pxelinux.cfg/AC10010A (ip-адрес в hex, буквы в верхнем регистре!)
3. /tftpboot/pxelinux.cfg/AC10010
4. . . .
5. /tftpboot/pxelinux.cfg/A
6. /tftpboot/pxelinux.cfg/default

PXELINUX — конфигурация (конец)

В нашем случае возможности PXELINUX задействованы по минимуму, поэтому каждый клиент имеет простейший конфиг. Файл /tftpboot/pxelinux.cfg/01-00-18-f3-70-91-b9:

```
TIMEOUT 1
TOTALTIMEOUT 1
ONTIMEOUT linux-cls1 root=/dev/nfsroot\
 nfsroot=172.16.1.1:/home/common/cluster\
 ip=172.16.1.10:172.16.1.1:172.16.1.255:255.255.255.0:::'off'

LABEL Default
 KERNEL linux-cls1
 APPEND root=/dev/nfsroot\
  nfsroot=172.16.1.1:/home/common/cluster\
  ip=172.16.1.10:172.16.1.1:172.16.1.255:255.255.255.0:::'off'
```

TFTP

Читаем мануал:

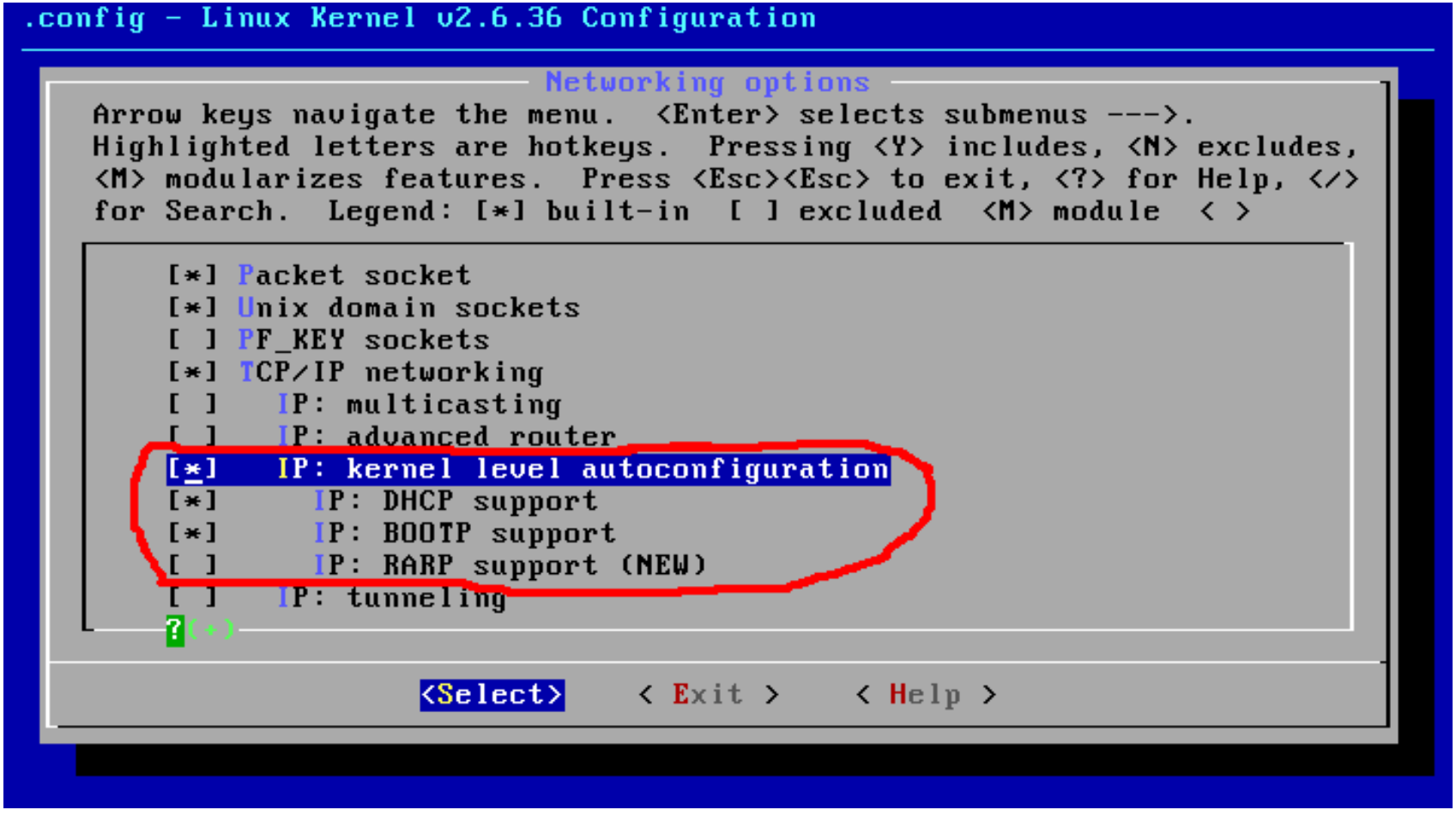
```
ATFTPD(8)                                                    ATFTPD(8)

NAME
       atftpd - Trivial File Transfer Protocol Server.

SYNOPSIS
       atftpd [options] directory

DESCRIPTION

       atftpd is a  TFTP  (RFC1350) server.  By default it is started by
       inetd on most sytems,  but may run as a stand alone daemon.  This
       server  is  multi-threaded and supports all options described  in
       RFC2347 (option extension), RFC2348 (blksize), RFC2349 (tsize and
       timeout) and RFC2090  (multicast option).  It also supports mtftp
       as defined in the PXE specification.
```

Запускаем:

```
atftpd --daemon /tftpboot
```

Конфигурация ядра — kernel level autoconfiguration

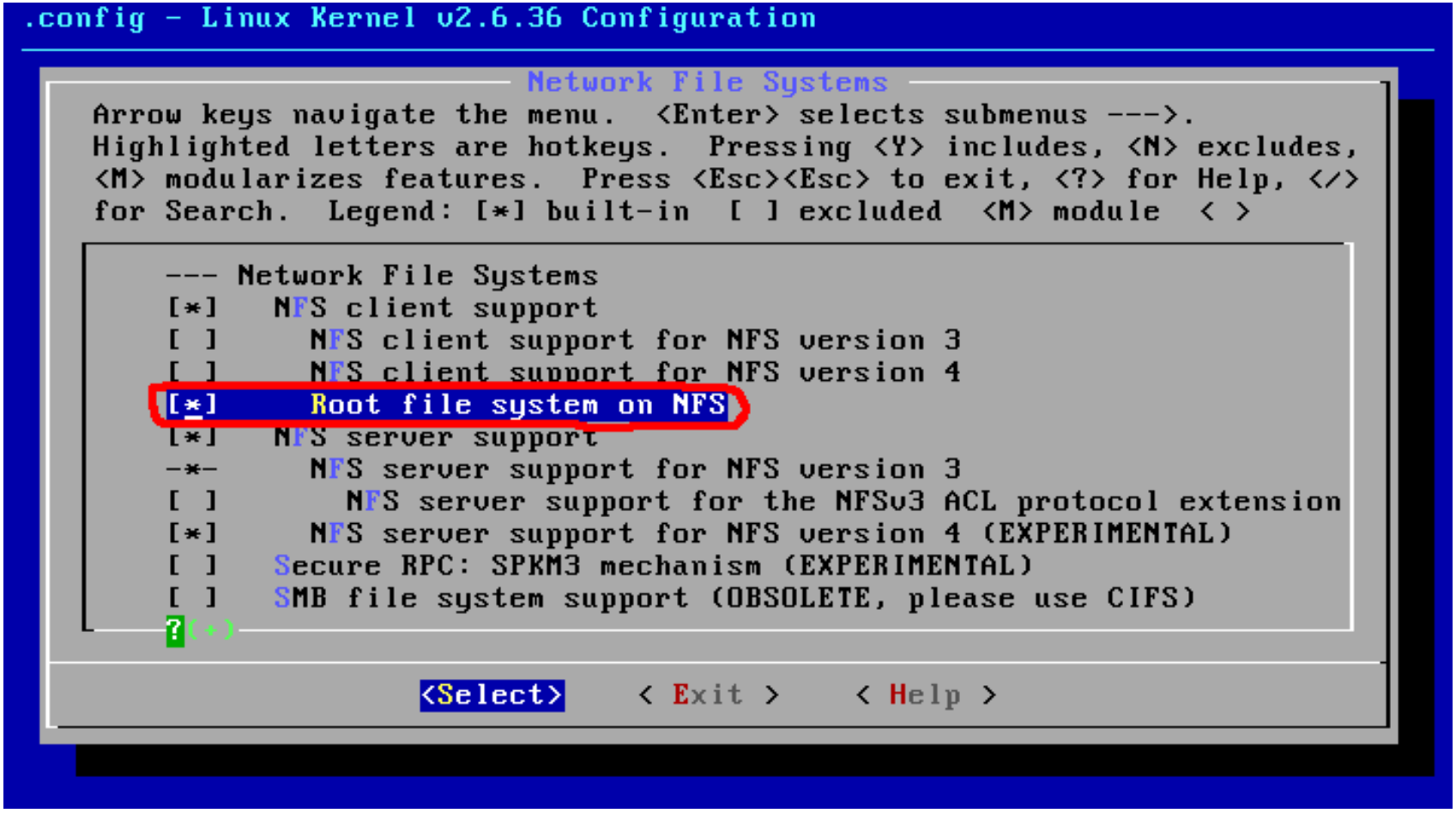
Networking support ⇒ Networking options




```
IP_PNP=y
IP_PNP_DHCP=y
IP_PNP_BOOTP=y
```

Конфигурация ядра — root file system on NFS

File systems ⇒ Network File Systems




```
NFS_FS=y
ROOT_NFS=y
```

Да, драйвер сетевой карты не должен быть модулем.

Настройка клиентов — начало

/etc/fstab:

```
proc    /proc   proc    defaults,noauto                         0 0
none    /tmp    tmpfs   defaults,noauto,size=10M,nodev,nosuid   0 0
none    /sys    sysfs   defaults,noauto                         0 0
none    /dev    tmpfs   defaults,noauto,size=1M,nosuid          0 0
172.16.1.1:/home/common/culc /culc nfs\
 hard,intr,nodev,rw,defaults,proto=tcp 0 0
```

/etc/inittab:

```
id:1:initdefault:
boot::bootwait:/bin/Rcinit
sshd:1:respawn:/bin/sshd -D
```

Настройка клиентов — конец

/bin/Rcinit:

```
#!/bin/bash
mount /proc
mount /tmp
mount /dev
mount /sys
echo ""> /proc/sys/kernel/hotplug
udevd -d
udevtrigger && udevsettle
mkdir /tmp/run && chmod 755 /tmp/run && mkdir /tmp/lock &&\
 chmod 755 /tmp/lock && mkdir /var/empty
ln -s /proc/self/fd /dev/fd
ln -s /proc/self/fd/0 /dev/stdin
ln -s /proc/self/fd/1 /dev/stdout
ln -s /proc/self/fd/2 /dev/stderr
touch /var/run/utmp && chmod 644 /var/run/utmp
portmap
gmond
mount /culc
date -s "`ssh -o ConnectTimeout=5 -i /etc/date.key 172.16.1.1`"
```

RTFM

- Описание PXELINUX.
  http://syslinux.zytor.com/wiki/index.php/PXELINUX
- Diskless Nodes with Gentoo.
  http://www.gentoo.org/doc/en/diskless-howto.xml
- Mounting the root filesystem via NFS (nfsroot).
  http://tinyurl.com/36l8ehw
  или Documentation/filesystems/nfs/nfsroot.txt в исходниках ядра.
- Mans, google, etc.

Такие дела!