Replugging the Modern Desktop

Kay Sievers <kay.sievers@suse.de>
David Zeuthen <davidz@redhat.com>

Linux Plumbers Conference
Portland, OR, Sept 2009
History

• Back in the day
  • /sbin/hotplug, scan entire /dev, /proc/scsi/scsi, /proc/partitions
  • magicdev, supermount, subfs
  • User conf / passwords stored in /etc or hard-coded
  • Millions of LOC running as uid 0
History

● Back in the day
  ● /sbin/hotplug, scan entire /dev, /proc/scsi/scsi, /proc/partitions
  ● magicdev, supermount, subfs
  ● User conf / passwords stored in /etc or hard-coded
  ● Millions of LOC running as uid 0

● Early Desktop Integration
  ● HAL, D-Bus, PolicyKit
  ● Separate Mechanism and Policy
  ● But... Implementation too complex, not scalable, not focused, too many abstractions
• Cutting the same cake in a different way

• 1\textsuperscript{st} piece: Move device discovery/enumeration, classification, quirks, probing, event propagation to udev

• 2\textsuperscript{nd} piece: Write libudev

• 3\textsuperscript{rd} piece: Dedicated system services for major subsystems
  – DeviceKit-disks, DeviceKit-power, NetworkManager, PulseAudio, Bluez, Gypsy, ...

• 4\textsuperscript{th} piece: Port the world to subsystem services
  – Apps using simple subsystems use libudev (Cheese)
Kernel

libudev

Application

libudev

Application

libudev

Application

libdbus

Subsystem

Services

libudev

Session Tracking (ConsoleKit)

System Message Bus (D-Bus)

Authority (PolicyKit)

udevd

Other Services:

Session 1

Session 2

...
kernel devices show up in a device tree in /sys

/sys/devices
|-- pci0000:00
...  
  |-- 0000:00:1f.2
  |   |-- driver -> ../../../bus/pci/drivers/ahci
  ...  
  |   |-- host0
  |       |-- subsystem -> ../../../bus/scsi
  |       |   |-- target0:0:0
  |       |       |-- subsystem -> ../../../bus/scsi
  ...  
  |       |-- 0:0:0:0
  |       |   |-- subsystem -> ../../../bus/scsi
  ...  
  |       |-- block
  |       `-- sda
  |           |-- subsystem -> ../../../bus/scsi
  |           |   |-- alignment_offset
  |           |       |-- capability
  |           |           |-- dev
new devices and changes are announced over netlink with uevents:

recvmsg(3,
    {msg_name(12)={sa_family=AF_NETLINK, pid=0, groups=00000001},
        msg_iov(1)=[{"add@/devices/pci0000:00/0000:00:1f.2/host0/target0:0:0/...\0
            ACTION=add\0
            DEVPATH=/devices/pci0000:00/0000:00:1f.2/host0/target0:0:0/0:0:0/...\0
            SUBSYSTEM=block\0
            MAJOR=8\0
            MINOR=0\0
            DEVNAME=sda\0
            DEVTYPE=disk\0
            SEQNUM=1584\0"}, ...],
    ...}
udev rules to:
  add properties to store in database
  create meaningful symlinks
  run programs to configure/setup the device

SUBSYSTEM=="block", KERNEL=="sd*", ENV{DEVTYPE}=="disk", \
  IMPORT{program}="ata_id --export $tempnode"

/lib/udev/ata_id --export /dev/sda
ID_TYPE=disk
ID_BUS=ata
ID_MODEL=SAMSUNG_MMCQE28G8MUP-0VA
ID_MODEL_ENC=SAMSUNG\x20MMCQE28G8MUP-0VA
ID_REVISION=VAM08L1Q
ID_SERIAL=SAMSUNG_MMCQE28G8MUP-0VA_SE837A4759
ID_SERIAL_SHORT=SE837A4759

SUBSYSTEM=="block", KERNEL=="sd*", ENV{DEVTYPE}=="disk", \
  ENV{ID_SERIAL}=="?*", SYMLINK+="disk/by-id/$env{ID_BUS}-$env{ID_SERIAL}"
send event back to multiple listeners:

recvmsg(3,
    {msg_name(12)={sa_family=AF_NETLINK, pid=-4226, groups=00000002},
    msg_iov(1)={["udev-147\0\0\0\0\0\0\0\0\312\376\35\352 \0m\3\20\306\320B\1\214\272\31
        UDEV_LOG=3\0
        ACTION=add\0
        DEVPATH=/devices/pci0000:00/0000:00:1f.2/host0/target0:0:0/0:0:0:0/block/sda\0
        SUBSYSTEM=block\0
        DEVNAME=/dev/sda\0
        DEVTYPE=disk\0
        SEQNUM=1584\0
        MAJOR=8\0
        MINOR=0\0
        DEVLINKS=/dev/block/8:0 \n        /dev/disk/by-id/ata-SAMSUNG_MMCQE28G8MUP-0VA_SE837A4759 \n        /dev/disk/by-path/pci-0000:00:1f.2-scsi-0:0:0:0:0\0
        ID_TYPE=disk\0
        ID_BUS=ata\0
        ID_MODEL=SAMSUNG_MMCQE28G8MUP-0VA\0
        ID_MODEL_ENC=SAMSUNG\x20MMCQE28G8MUP-0VA\0
        ID_REVISION=VAM08L1Q\0
        ID_SERIAL=SAMSUNG_MMCQE28G8MUP-0VA_SE837A4759\0
        ID_SERIAL_SHORT=SE837A4759\0
        ID_PATH=pci-0000:00:1f.2-scsi-0:0:0:0\0
        DKD_ATA_SMART_IS_AVAILABLE=1\0}, ...],},
receive event with libudev:

```c
struct udev_monitor *monitor;
struct udev_device *device;

monitor = udev_monitor_new_from_netlink(udev, "udev");
udev_monitor_enable_receiving(monitor);
udev_monitor_filter_add_match_subsystem_devtype(monitor, "block", "disk");
device = udev_monitor_receive_device(monitor);
```

message multiplexing in the kernel
messages filtered inside the kernel with berkeley packet filter
libgudev javascript example
Storage Subsystem Daemon

• DeviceKit-disks

• Consumes udev information

• Started on demand

• High-level API w/ progress reporting
  • Mount, Unmount, Eject, Poll, Fsck
  • Partitioning, Formatting, FS Label
  • ATA SMART monitoring
  • MD-RAID (Create, Start/Stop, Check, ...)
  • Drive spindown
Palimpsest Demo

(d-feet, fs labels, mkfs, ATA SMART from USB, new-ui)
Every 2.0s: tree /dev/disk/

```
/dev/disk/
|-- ata-INTEL_SSDSA2H080G1GC_CVEMB42101HD080DG1 -> /dev/sda
|   |-- ata-INTEL_SSDSA2H080G1GC_CVEMB42101HD080DG1-part1 -> /dev/sda1
|   |-- ata-INTEL_SSDSA2H080G1GC_CVEMB42101HD080DG1-part2 -> /dev/sda2
|   |   |-- scsi-ATA_INTEL_SSDSA2H0C6VEMB42101HD080DG1 -> /dev/sda
|   |   |   |-- scsi-ATA_INTEL_SSDSA2H0C6VEMB42101HD080DG1-part1 -> /dev/sda1
|   |   |   |   |-- scsi-ATA_INTEL_SSDSA2H0C6VEMB42101HD080DG1-part2 -> /dev/sda2
|   |   |   |   |   |-- usb-Generic STORAGE DEVICE_0000000000400-0:0:part1 -> /dev/sdb1
|   |   |-- by-label
|   |   |   |   |   |   |-- FIRST\x20\x20NAME -> /dev/sdb1
|   |   |   |   |   |   |   |-- Fedora\x20\x28Rawhide\x29 -> /dev/sdal
|   |   |-- by-path
|   |   |   |   |   |   |   |   |-- FIRST\x20\x20NAME -> /dev/sdb1
|   |   |   |   |   |   |   |   |-- Fedora\x20\x282\x20x28Rawhide\x29\x20\x20x29 -> /dev/sdal
|   |   |   |   |   |   |   |   |   |-- FIRST\x20\x20NAME -> /dev/sdb1
|   |   |   |   |   |   |   |   |   |   |-- Fedora\x20\x282\x20x28Rawhide\x29\x20\x20x29 -> /dev/sdal
```

---

```
root@x61:~ # udevadm monitor --udev
monitor will print the received events for:
UDEV - the event which udev sends out after rule processing
```

---

```
[root@x61 ~]# devkit-disks --monitor
Monitoring activity from the disks daemon. Press Ctrl+C to cancel.
```

---

```
[davidz@x61 Hacking]$ devkit-disks --monitor
Monitoring activity from the disks daemon. Press Ctrl+C to cancel.
```

---

```
80 GB Solid-State Disk
ATA INTEL SSDSA2H080G1GC
MBR Partition Table
Fedora (Rawhide)
21 GB Linux E43 (version 1.0)
2.0 GB Unrecognized
Unknown or Unused
46 GB Free
Unallocated Space

Generic STORAGE DEVICE
Generic STORAGE DEVICE
4.1 GB Media, MBR Partition Table
FIRST NAME
4.1 GB Raw (32-bit version)

Partition
The attributes of the partition can be edited or can also be deleted to make room for others.
Partition Label: 
Type: W95 FAT32 (0x0b) 

- Bootable
Delete
Resize
Apply

Mountable Filesystem
The volume contains a mountable filesystem.
Label: FIRST NAME
```
Every 2.0s: tree /dev/disk/

```bash
/dev/disk/
```

```bash
|-- ata-Intel_SSDSA2M080GIGC_CVEMB42101H0D0BGW -> ../../sda
 |  |-- ata-Intel_SSDSA2M080GIGC_CVEMB42101H0D0BGW-part1 -> ../../sda1
 |  |-- ata-Intel_SSDSA2M080GIGC_CVEMB42101H0D0BGW-part2 -> ../../sda2
 |  |-- scsi-SATA-Intel_SSDSA2M080GIGC_CVEMB42101H0D0BGW -> ../../sda
 |  |-- scsi-SATA-Intel_SSDSA2M080GIGC_CVEMB42101H0D0BGW-part1 -> ../../sda1
 |  |-- scsi-SATA-Intel_SSDSA2M080GIGC_CVEMB42101H0D0BGW-part2 -> ../../sda2
 |  |-- usb-Generic_STORAGE_DEVICE_000000009407-0:0:part1 -> ../../sda1
 |     |-- by-id
 |     |  |-- Fedora\x20\x28x20Rawhide\x29 -> ../../sda1
 |     |  |-- M\x20\x28x20Super\x20\x28x20Pe -> ../../sda1
 |     |     |-- by-path
 |     |     |  |-- pci-0000:00:1.d.7-usb-0.1.0:1.6-scsi-0:0:0:0 -> ../../sdb
 |     |     |  |-- pci-0000:00:1.d.7-usb-0.1.0:1.6-scsi-0:0:0:0:part1 -> ../../sda1
 |     |     |  |-- pci-0000:00:1.d.7-usb-0.1.0:1.6-scsi-0:0:0:0:part2 -> ../../sda2
 |     |     |     |-- by-uuid
 |     |     |     |  |-- 786263c1-5e28-4dc6-97b8-1ab6e2213444 -> ../../sda1
 |     |     |     |  |-- F588-8E30 -> ../../sdb1
 |     |     |     |  |-- by-label
 |     |     |     |     |-- My\x20\x20\x28x20Super\x20\x28x20Pe -> ../../sda1
 |     |     |     |     |  |-- by-path
 |     |     |     |     |     |-- pci-0000:00:1.d.7-usb-0.1.0:1.6-scsi-0:0:0:0 -> ../../sdb
 |     |     |     |     |     |  |-- pci-0000:00:1.d.7-usb-0.1.0:1.6-scsi-0:0:0:0:part1 -> ../../sda1
 |     |     |     |     |     |  |-- pci-0000:00:1.d.7-usb-0.1.0:1.6-scsi-0:0:0:0:part2 -> ../../sda2
 |     |     |     |     |     |  |-- by-uuid
 |     |     |     |     |     |     |-- 786263c1-5e28-4dc6-97b8-1ab6e2213444 -> ../../sda1
 |     |     |     |     |     |     |  |-- F588-8E30 -> ../../sdb1
```

4 directories, 17 files
SMART Data

Drive: 320 GB Hard Disk - ST932042 1ASG
Disk is healthy

Status
Updated: Less than a minute ago - Update Now
Self-tests: Last self-test completed OK - Run self-test
Model: ST9320421ASG
Firmware Version: SD13
Serial Number: 5TJ08ZSC
Powered On: 81.2 days
Temperature: 32° C / 90° F
Bad Sectors: None
Self Assessment: Passed
Overall Assessment: Disk is healthy

Attributes

<table>
<thead>
<tr>
<th>ID</th>
<th>Attribute</th>
<th>Assessment</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Start/Stop Count</td>
<td>Good</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Number of spindle start/stop cycles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Reallocated Sector Count</td>
<td>Good</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Count of remapped sectors. When the hard drive finds a read/write/verification error, it mark the sector as &quot;reallocated&quot; and transfers data to a special reserved area (spare area)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Seek Error Rate</td>
<td>Good</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Frequency of errors while positioning</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Don't warn me if the disk is failing
A hard disk may be failing

One or more hard disks report health problems. Click the icon to get more information.
Add /sys/class/block/sda1

DeviceAdded(sda1)

signal

Mount(sda1, ...)

BeginAuthentication()

AuthenticationAgentResponse()

CheckAuthorization()

result

Mount()

result
Desktop Integration

- GVfs volume monitor
- GNOME Power Manager
Formatting

Create new filesystem on the selected device

Warning: All data on the volume will be irrevocably lost.

Type: Encrypted, compatible with Linux (FAT)
Name: My Little Pony
MD RAID
### Power Management Preferences

**On AC Power**

**Actions**
- Put computer to **sleep** when inactive for: Never
- When laptop lid is closed: Suspend
- **Spin down hard disks when possible**

**Display**
- Put **display** to sleep when inactive for: 30 minutes
- Set display **brightness** to: 100%
- **Dim display when idle**

**Buttons:**
- Help
- Make Default
- Close
Questions?
Docs / References

- http://hal.freedesktop.org/docs/DeviceKit-disks/
- http://hal.freedesktop.org/docs/DeviceKit-power/
- http://hal.freedesktop.org/docs/polkit/
Replugging the Modern Desktop

Kay Sievers <kay.sievers@suse.de>
David Zeuthen <davidz@redhat.com>

Linux Plumbers Conference
Portland, OR, Sept 2009
History

• Back in the day
  • /sbin/hotplug, scan entire /dev, /proc/scsi/scsi, /proc/partitions
  • magicdev, supermount, subfs
  • User conf / passwords stored in /etc or hard-coded
  • Millions of LOC running as uid 0
History

• Back in the day
  • /sbin/hotplug, scan entire /dev, /proc/scsi/scsi, /proc/partitions
  • magicdev, supermount, subfs
  • User conf / passwords stored in /etc or hard-coded
  • Millions of LOC running as uid 0

• Early Desktop Integration
  • HAL, D-Bus, PolicyKit
  • Separate Mechanism and Policy
  • But... Implementation too complex, not scalable, not focused, too many abstractions
• Cutting the same cake in a different way

• 1\textsuperscript{st} piece: Move device discovery/enumeration, classification, quirks, probing, event propagation to udev

• 2\textsuperscript{nd} piece: Write libudev

• 3\textsuperscript{rd} piece: Dedicated system services for major subsystems
  – DeviceKit-disks, DeviceKit-power, NetworkManager, PulseAudio, Bluez, Gypsy, ...

• 4\textsuperscript{th} piece: Port the world to subsystem services
  – Apps using simple subsystems use libudev (Cheese)
Kernel

Application
libudev
Session 1

Application
libudev
Session 2

Application
libdbus

Subsystem
Services
libudev

System Space

Kernel Space

udev

7
kernel devices show up in a device tree in /sys

/sys/devices
   |-- pci0000:00
   ...   
   |    |-- 0000:00:1f.2
   |       |   |-- driver -> ../../bus/pci/drivers/ahci
   ...   
   |    |-- host0
   |    |    |-- subsystem -> ../../bus/scsi
   |    |    |-- target0:0:0
   |    |    |    |-- subsystem -> ../../bus/scsi
   ...   
   |    |    |-- 0:0:0:0
   |    |    |    |-- subsystem -> ../../bus/scsi
   ...   
   |    |-- block
   |    |    |-- sda
   |    |    |    |-- subsystem -> ../../class/block
   |    |    |    |-- alignment_offset
   |    |    |    |-- capability
   |    |    |    |-- dev
new devices and changes are announced over netlink with uevents:

recvmsg(3,
    {msg_name(12)={sa_family=AF_NETLINK, pid=0, groups=00000001},
    msg_iov(1)={["add@/devices/pci0000:00/0000:00:1f.2/host0/target0:0:0:0:0/...
    ACTION=add
    DEVPATH=/devices/pci0000:00/0000:00:1f.2/host0/target0:0:0/0:0:0:0/...
    SUBSYSTEM=block
    MAJOR=8
    MINOR=0
    DEVNAME=sda
    DEVTYPE=disk
    SEQNUM=1584"}, ...],
    ...
udev rules to:
add properties to store in database
create meaningful symlinks
run programs to configure/setup the device

SUBSYSTEM=="block", KERNEL=="sd***", ENV{DEVTYPE}=="disk", \nIMPORT{program}="ata_id --export $tempnode"

/lib/udev/ata_id --export /dev/sda
ID_TYPE=disk
ID_BUS=ata
ID_MODEL=SAMSUNG_MMCQE28G8MUP-0VA
ID_MODEL_ENC=SAMSUNG\x20MMCQE28G8MUP-0VA
ID_REVISION=VAM08L1Q
ID_SERIAL=SAMSUNG_MMCQE28G8MUP-0VA_SE837A4759
ID_SERIAL_SHORT=SE837A4759

SUBSYSTEM=="block", KERNEL=="sd***", ENV{DEVTYPE}=="disk", \nENV{ID_SERIAL}=="?***", SYMLINK+="disk/by-id/$env{ID_BUS}-$env{ID_SERIAL}"
send event back to multiple listeners:

recvmsg(3,
   {msg_name(12)={sa_family=AF_NETLINK, pid=-4226, groups=00000002},
   msg_iov(1)=[{"udev-147\0\0\0\0\0\0\0\0\312\376\35\352\0m\3\20\306\320B\1\214\272\31
UDEV_LOG=3\0
ACTION=add\0
DEVPATH=/devices/pci0000:00/0000:00:1f.2/host0/target0:0:0:0:0:0:0:/block/sda\0
SUBSYSTEM=block\0
DEVNAME=/dev/sda\0
DEVTYPE=disk\0
SEQNUM=1584\0
MAJOR=8\0
MINOR=0\0
DEVLINKS=/dev/block/8:0 \n/dev/disk/by-id/ata-SAMSUNG_MMCQE28G8MUP-0VA_SE837A4759 \n/dev/disk/by-path/pci-0000:00:1f:2-scsi-0:0:0:0:0:0:0
ID_TYPE=disk\0
ID_BUS=ata\0
ID_MODEL=SAMSUNG_MMCQE28G8MUP-0VA\0
ID_MODEL_ENC=SAMSUNG\x20MMCQE28G8MUP-0VA\0
ID_REVISION=VAM08L1Q\0
ID_SERIAL=SAMSUNG_MMCQE28G8MUP-0VA_SE837A4759\0
ID_SERIAL_SHORT=SE837A4759\0
ID_PATH=pci-0000:00:1f:2-scsi-0:0:0:0:0:0
DKD_ATA_SMART_IS_AVAILABLE=1\0}, ...],
...}
receive event with libudev:

```c
struct udev_monitor *monitor;
struct udev_device *device;

monitor = udev_monitor_new_from_netlink(udev, "udev");
udev_monitor_enable_receiving(monitor);
udev_monitor_filter_add_match_subsystem_devtype(monitor, "block", "disk");
device = udev_monitor_receive_device(monitor);
```

message multiplexing in the kernel
messages filtered inside the kernel with berkeley packet filter
libgudev javascript example
Storage Subsystem Daemon

- DeviceKit-disks

- Consumes udev information
- Started on demand

- High-level API w/ progress reporting
  - Mount, Unmount, Eject, Poll, Fsck
  - Partitioning, Formatting, FS Label
  - ATA SMART monitoring
  - MD-RAID (Create, Start/Stop, Check, …)
  - Drive spindown
Palimpsest Demo

(d-feet, fs labels, mkfs, ATA SMART from USB, new-ui)
A hard disk may be failing

One or more hard disks report health problems. Click the icon to get more information.
Desktop Integration

- GVfs volume monitor
- GNOME Power Manager
Formatting
MD RAID
Questions?
Docs / References

- http://hal.freedesktop.org/docs/DeviceKit-disks/
- http://hal.freedesktop.org/docs/DeviceKit-power/
- http://hal.freedesktop.org/docs/polkit/