Linux kernel 3.0 release

IO Data Flow Hook On Xen

Chunjie_zhu@trendmicro.com.cn
Jerry_zhang@trendmicro.com.cn

What?

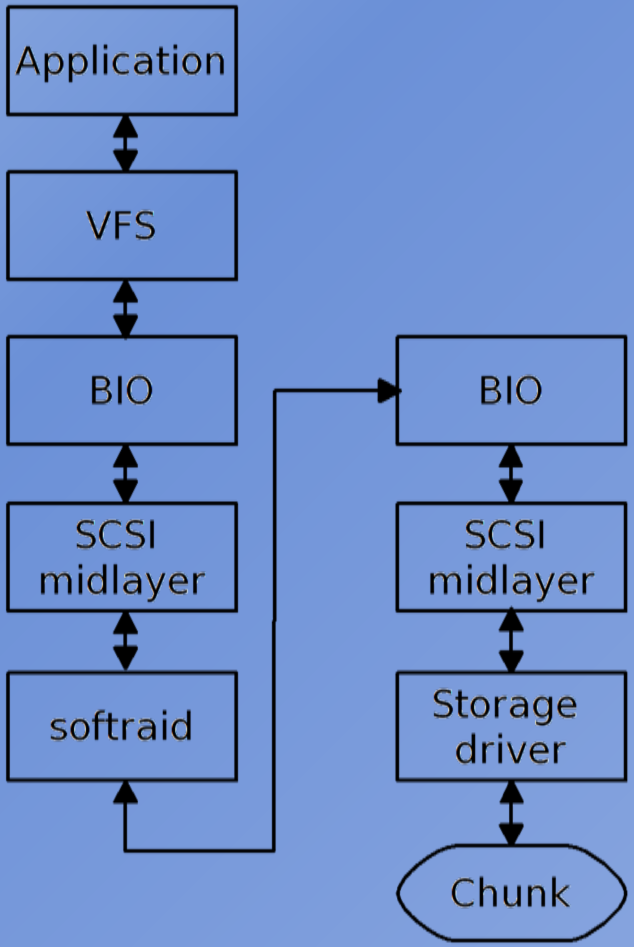
Why?

How?

Agenda

- IO hook general idea
- IO hook on virtualization platform
- IO hook achievement on Xen

IO Hook Philosophy



Application

VFS

BIO

SCSI midlayer

softraid

BIO

SCSI midlayer

Storage driver

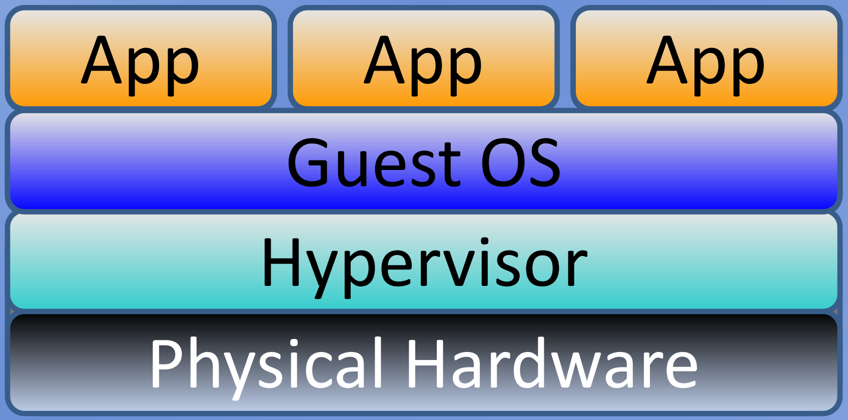
Chunk

<table>
<tr><th>OSI</th><th>TCP/IP</th></tr>
<tr><td>Application</td><td rowspan="2">Application</td></tr>
<tr><td>Presentation</td></tr>
<tr><td>Session</td><td rowspan="2">Transport</td></tr>
<tr><td>Transport</td></tr>
<tr><td>Network</td><td>Network</td></tr>
<tr><td>Data link</td><td rowspan="2">Physical</td></tr>
<tr><td>Physical</td></tr>
</table>

Utility

- transparent encryption (e.g. Linux dm-crypt)
- virtual block device driver (e.g. Linux softraid)
- file hidden
- virtual filesystem (e.g. FUSE)
- firewall (e.g. Netfilter)
- virtual network device driver (e.g. bond & vlan)

IT World Is Changing …



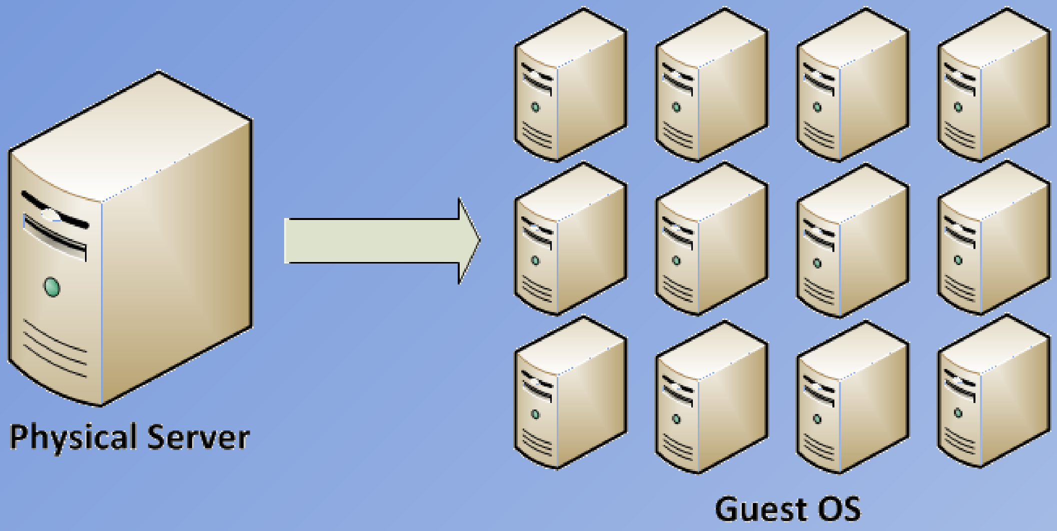
App App App

Guest OS

Hypervisor

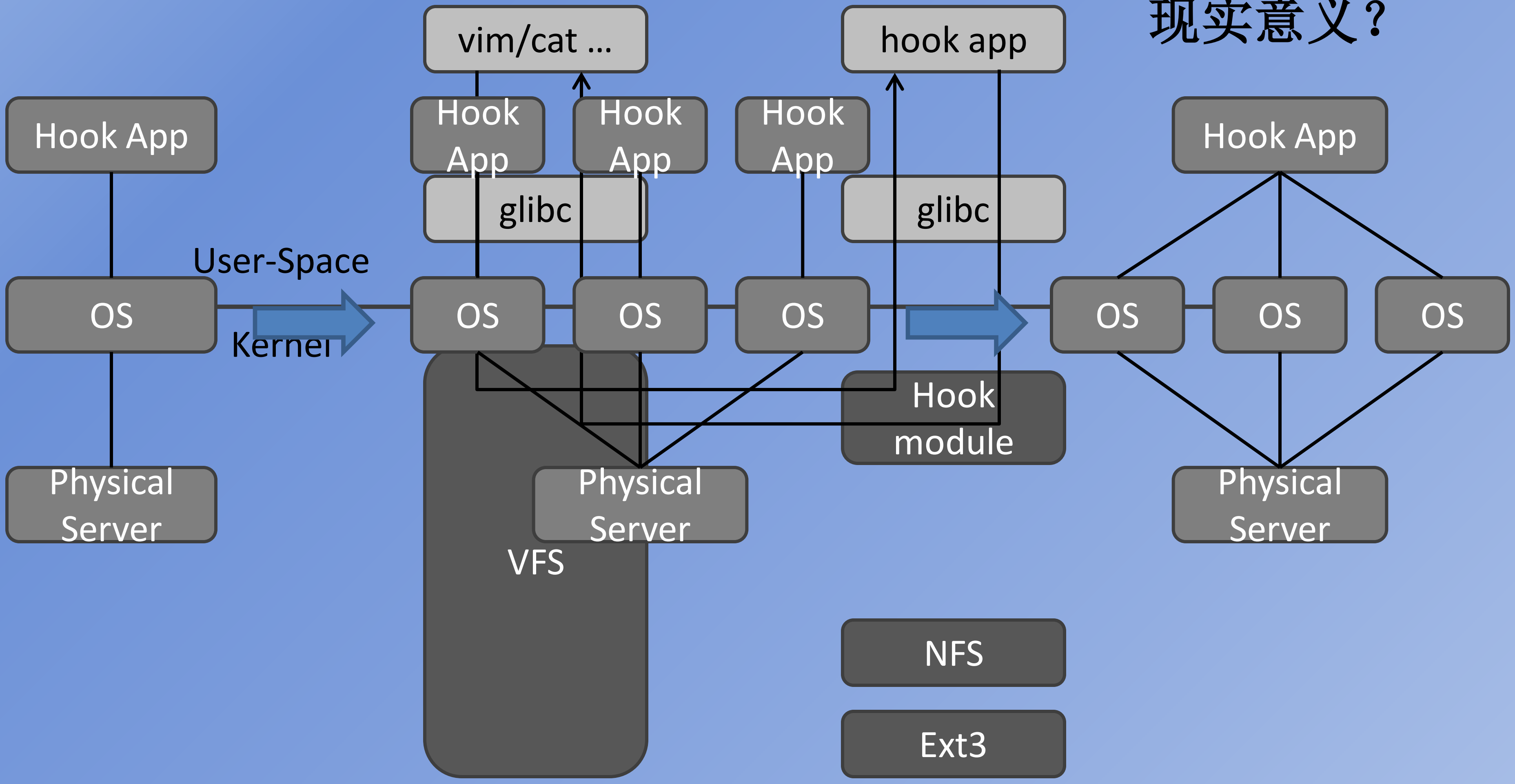
Physical Hardware

从对于变化的研究
来研究技术的变化



Change Our Cognition

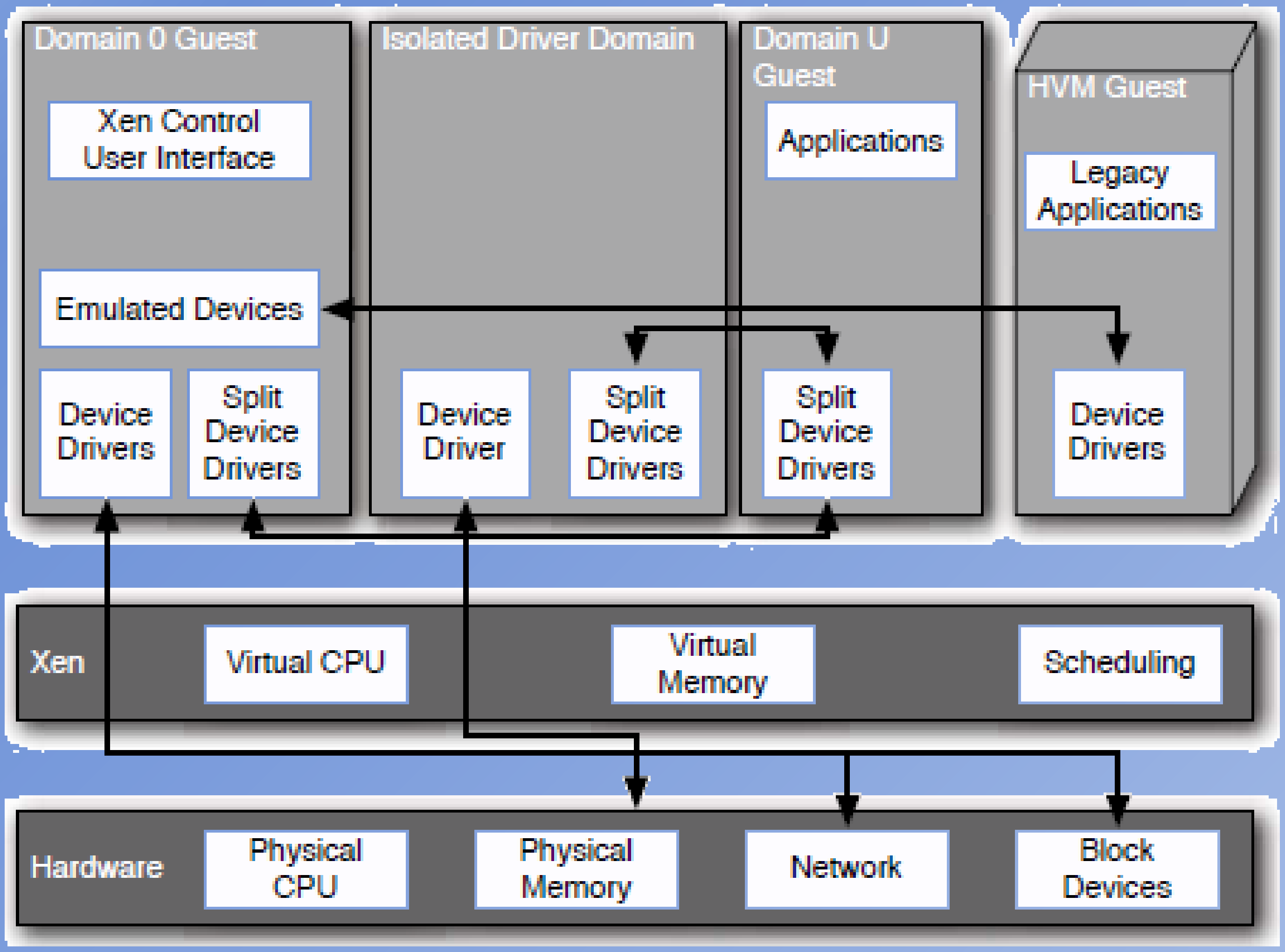
现实意义？



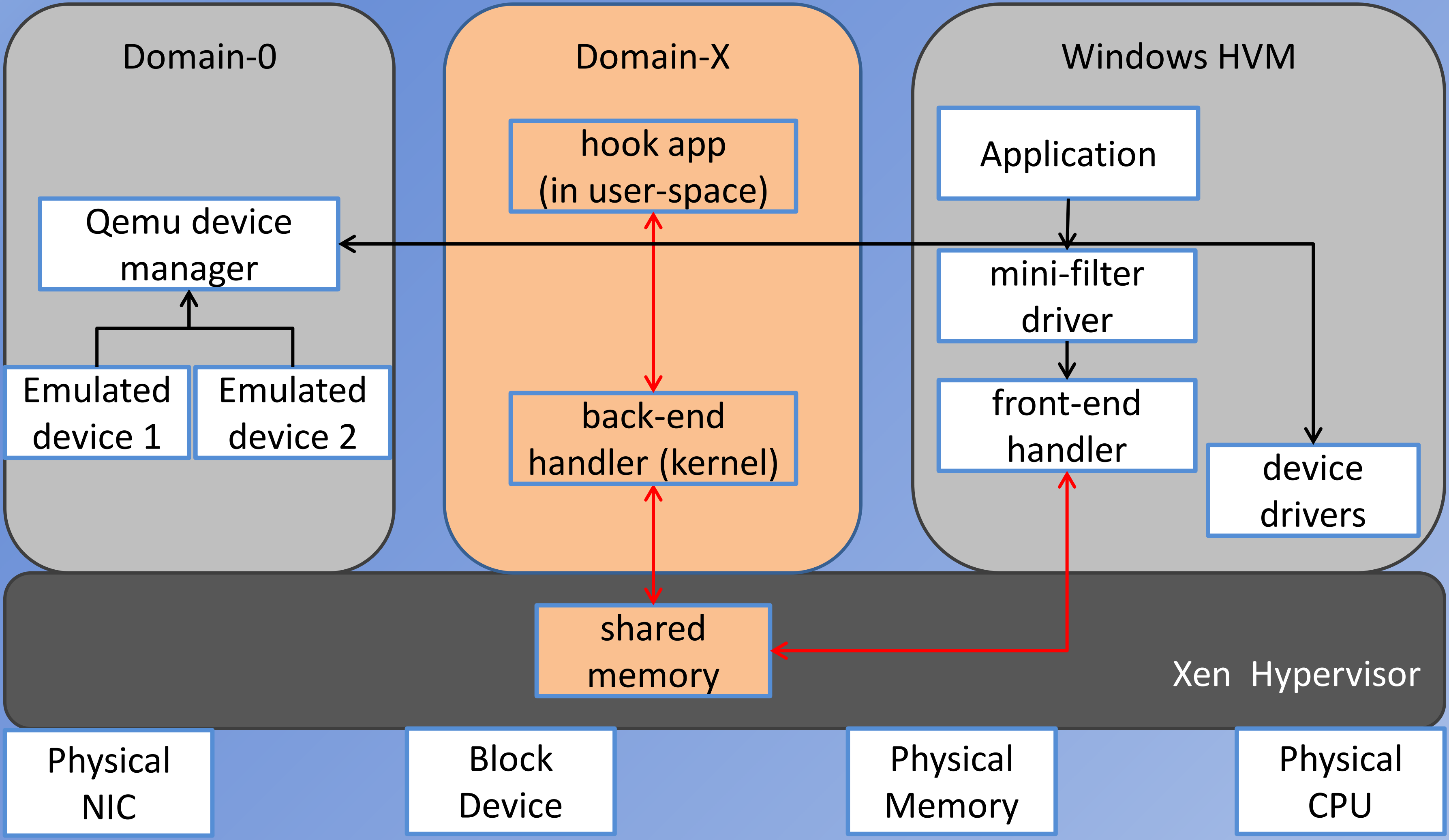
Advantage

- reduce management cost
  - ✓ uniform configuration interface
  - ✓ frequent patch/hot fix
  - ✓ migration
  - ✓ virtual appliance shipping
- management task
  - ✓ heterogeneous -> uniform

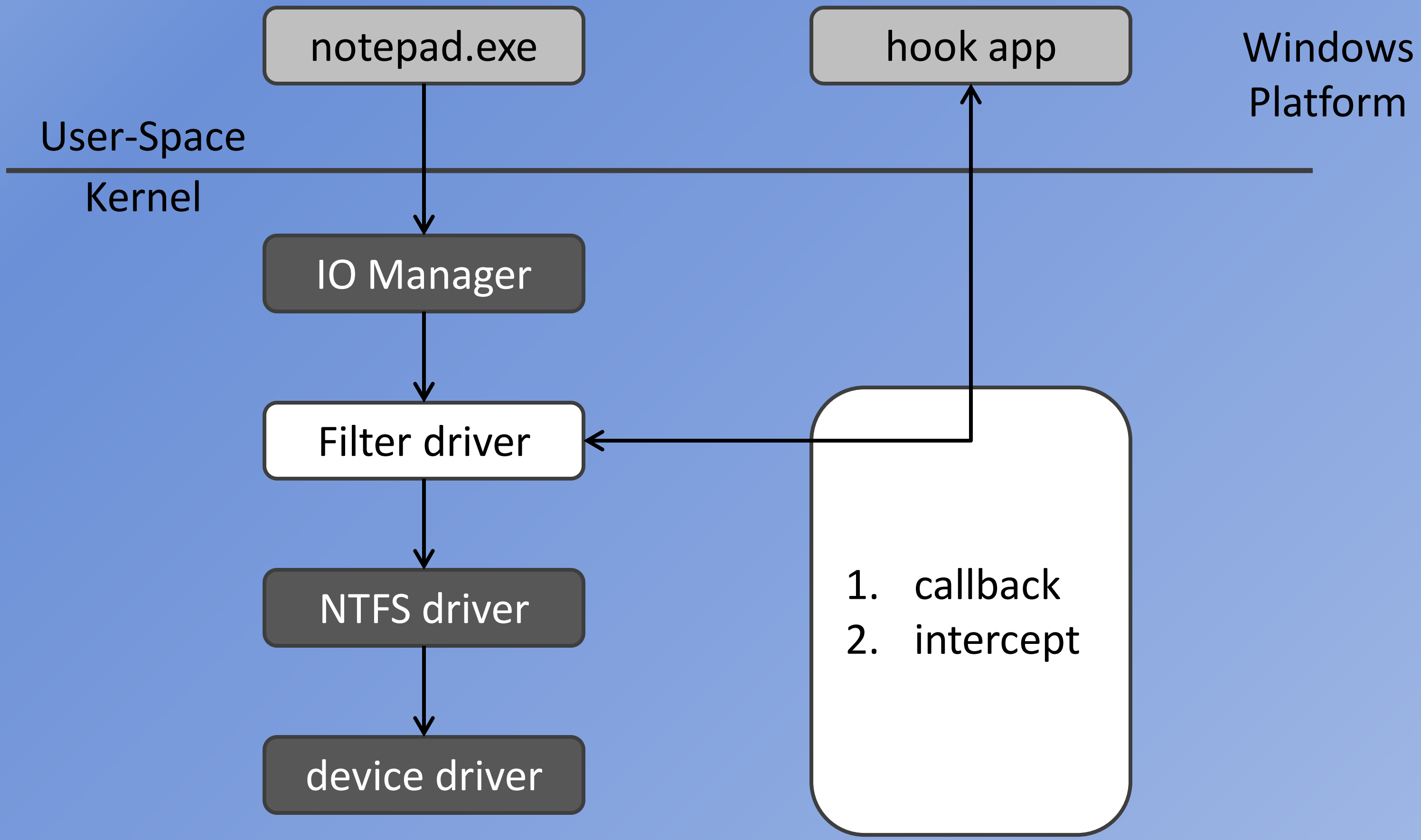
Xen IO Overview



Filesystem Hook Overview



Xen Filesystem IO Hook (1)

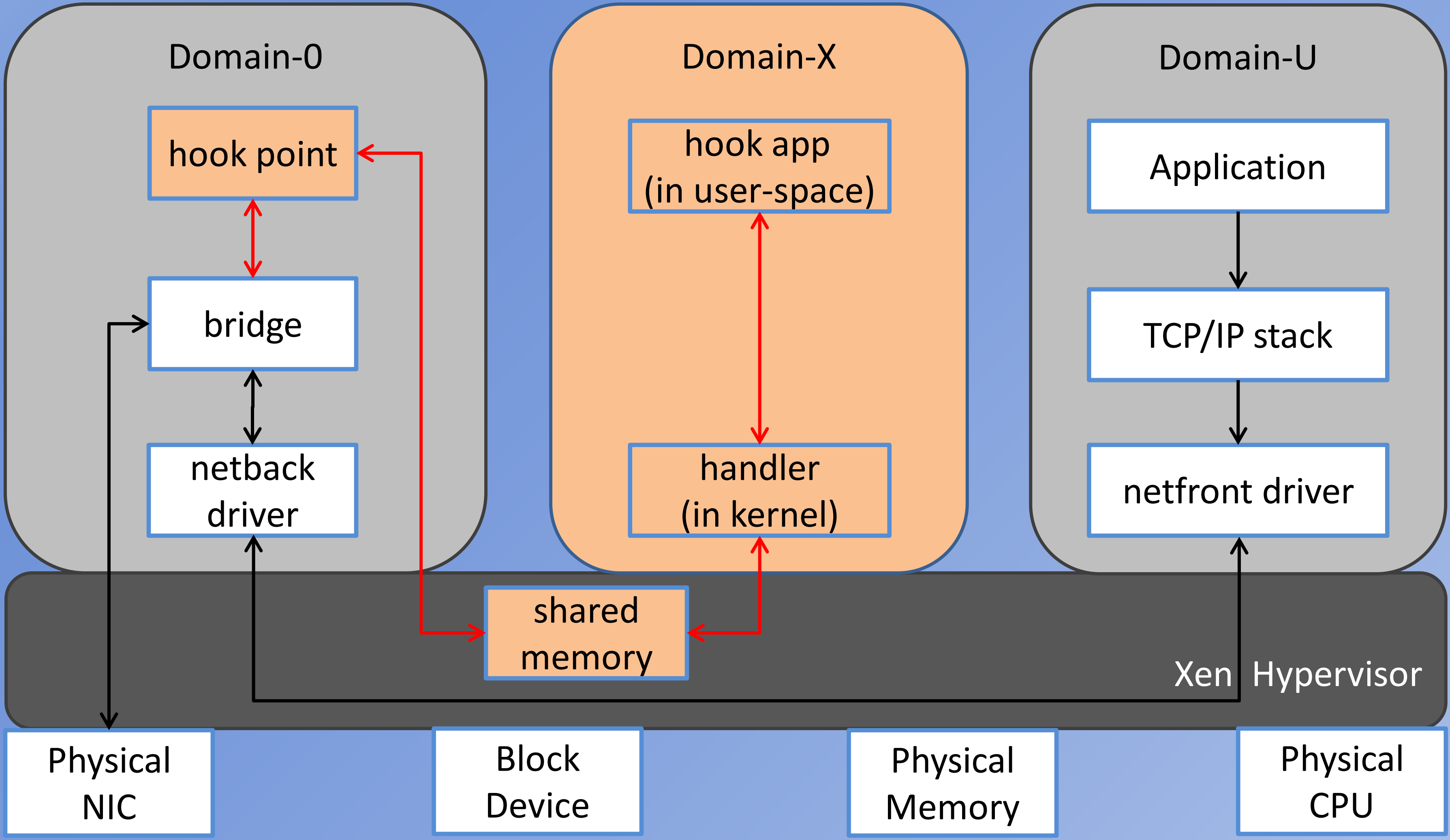


1. callback
2. intercept

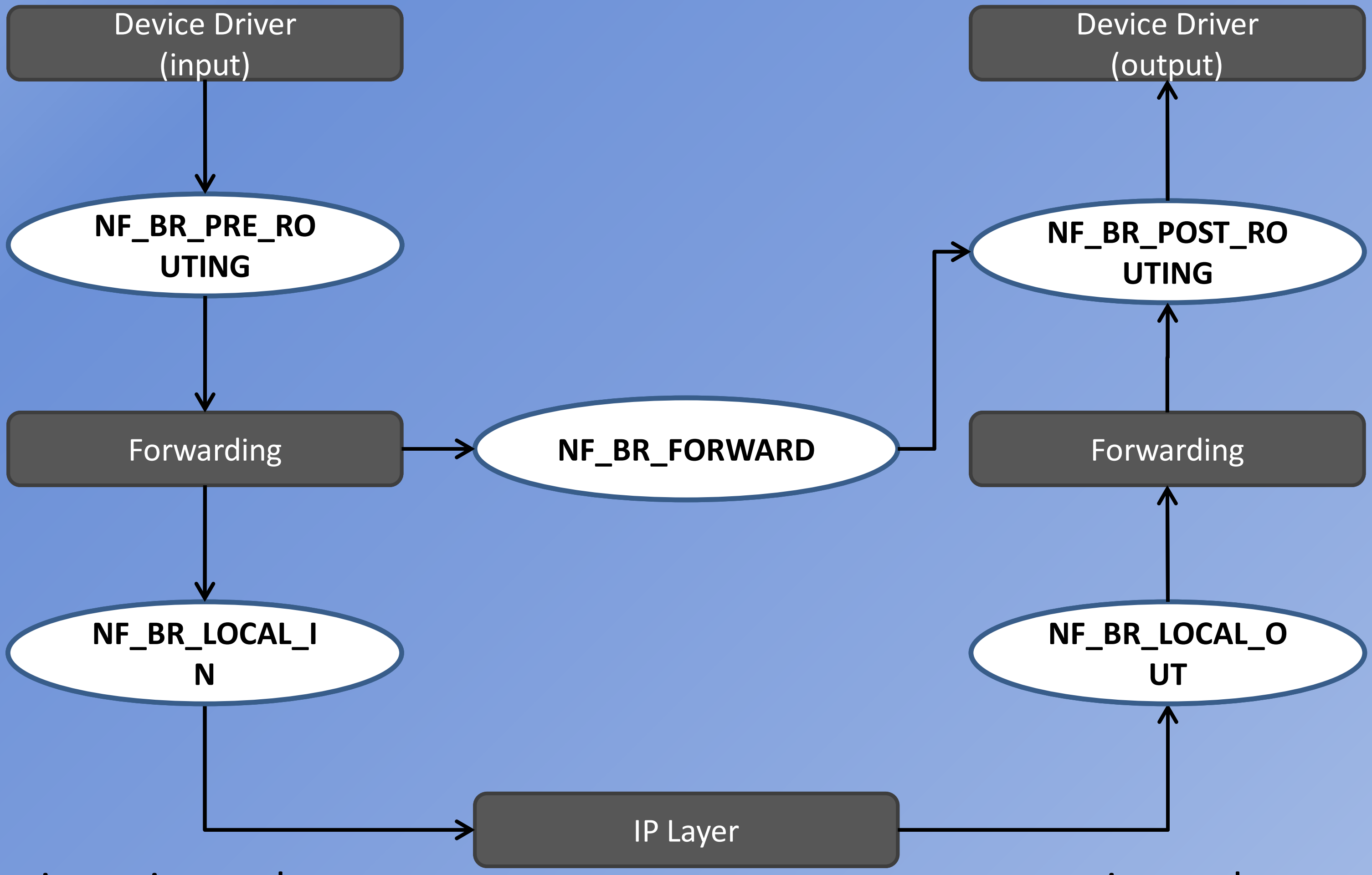
Xen Filesystem IO Hook (2)

- One agent on each windows Guest OS
  - ✓ mini-filter driver
  - ✓ filter/delete/quarantine
  - ✓ do not need binary update
- Deployment challenge
  - ✓ convince user “no harm”
  - ✓ VM template? Good idea?

Network Hook Overview



Xen Network IO Hook (1)



incoming packets

outgoing packets

Xen Network IO Hook (2)

- Where to hook?
  - ✓ Layer-2 (bridge) [√]
  - ✓ Layer-3 (ip) [√]
  - ✓ Layer-4 (tcp) [X]
- Xen uses bridge-network by default
  - ✓ /etc/xen/xend-config.sxp

    (network-script network-bridge)

Xen Network IO Hook (3)

- Layer-2 hook vs. Layer-3 hook
  - ✓ mac address permanent while ip address maybe dynamic (DHCP)
  - ✓ ARP packet to Dom0 cannot be hooked in IP Layer
    - proxy ARP & ARP spoof
  - ✓ easy to cooperate with Open vSwitch

Data Handling (1)

- ➢ Where to handle these hooked data?
  - ✓ Dom0
  - ✓ one dedicated PV domain, “DomX” [√]
- ➢ Data transfer between domains
  - ✓ TCP/IP socket transmit?
  - ✓ memory sharing? [√]
    - event notification?
    - synchronization?

Data Handling (2)

- ➢ Difference in filesystem hook & network hook
  - ✓ Filesystem hook
    - Domain U <-> share memory <-> Domain X
  - ✓ Network hook
    - Domain 0 <-> share memory <-> Domain X

Data Handling (3)

- ➢ Memory sharing between 2 domains
  - ✓ grant table provided by Xen
    - allocate page & grant reference id on initiator side
    - map grant reference id on other side
    - who should be initiator?
  - ✓ alternative channel organization
    - place metadata & data in the channel
    - place metadata in the channel while put data out-band

Data Handling (4)

- Event notification between 2 domains
  - event channel provided by Xen
    - similar as POSIX signal
    - local port <-> remote port
    - bind local port with one virtual irq handler
    - initialization
      1. where to keep remote domid & port? xenstore
    - when to trigger virtual irq handler?

domain switch to -> ret_from_intr -> test_all_events ->
event_do_upcall -> virtual irq handler (Xen-3.4.0)

Data Handling (5)

- Memory access sync between 2 domains
  - ✓ shared memory organized as ring-buffer
  - ✓ xen/include/public/io/ring.h (xen-4.0.1)
    - one reader & one writer
    - memory barrier
  - ✓ filesystem hook
    - one reader & multiple writer

Xen Programming Interface

- Xen hypercall
  - similar as Linux system call
    - event channel
    - grant table
    - domain control
    - …
  - Linux wrapper interfaces
  - trap Guest OS kernel to Xen hypervisor
    - normal kernel routines may trap to Xen hypervisor

schedule -> update_rq_clock -> native_read_tsc -> “rdtsc” -> invalid op exception -> trap into Xen (linux-2.6.24-29-xen)

Potential Issue

- PV driver in HVM
- PCI through
- VMDq
- …

Q & A