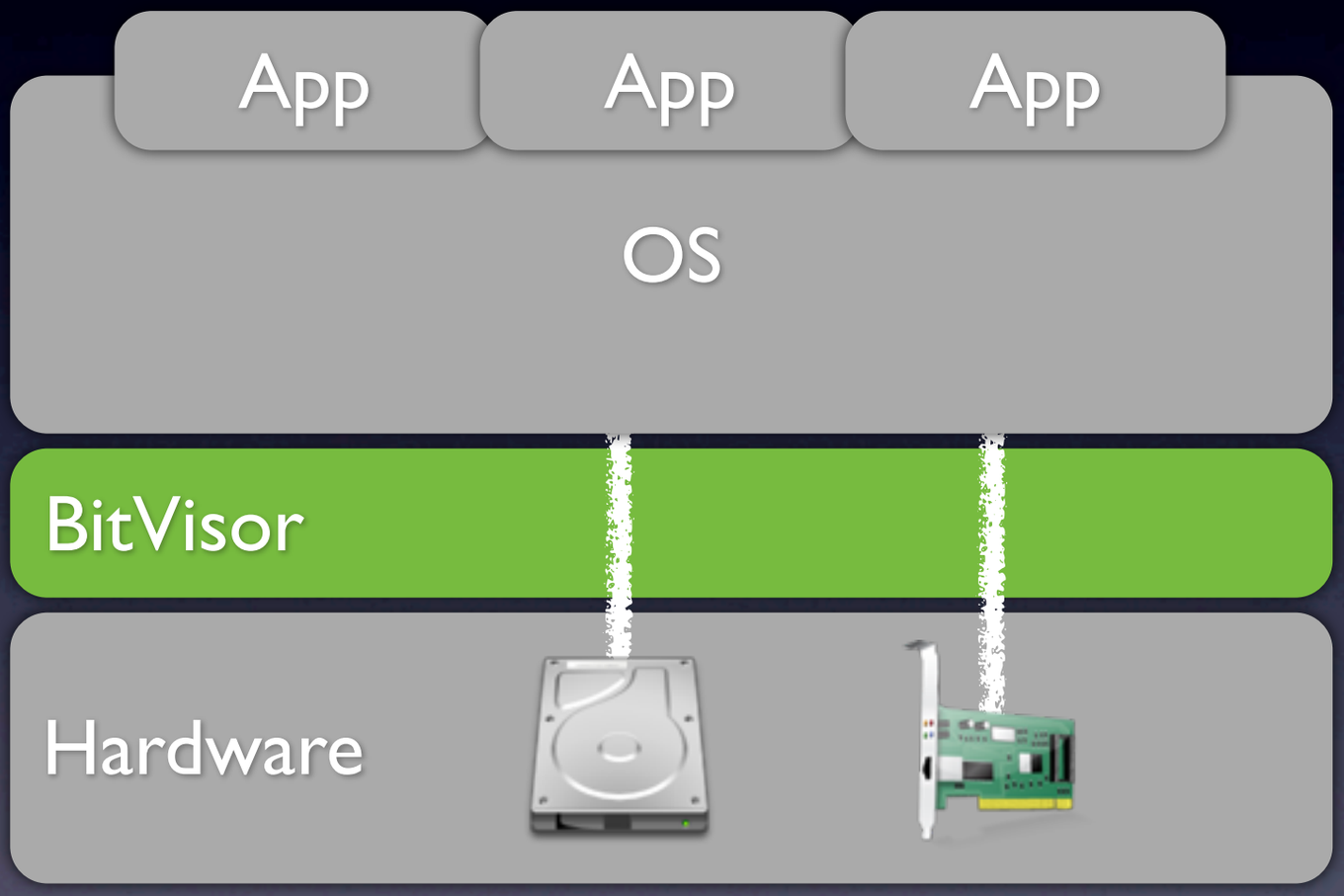
OSb: OSv on BitVisor

(2014/11/21 BitVisor Summit 3)

Yushi Omote
University of Tsukuba

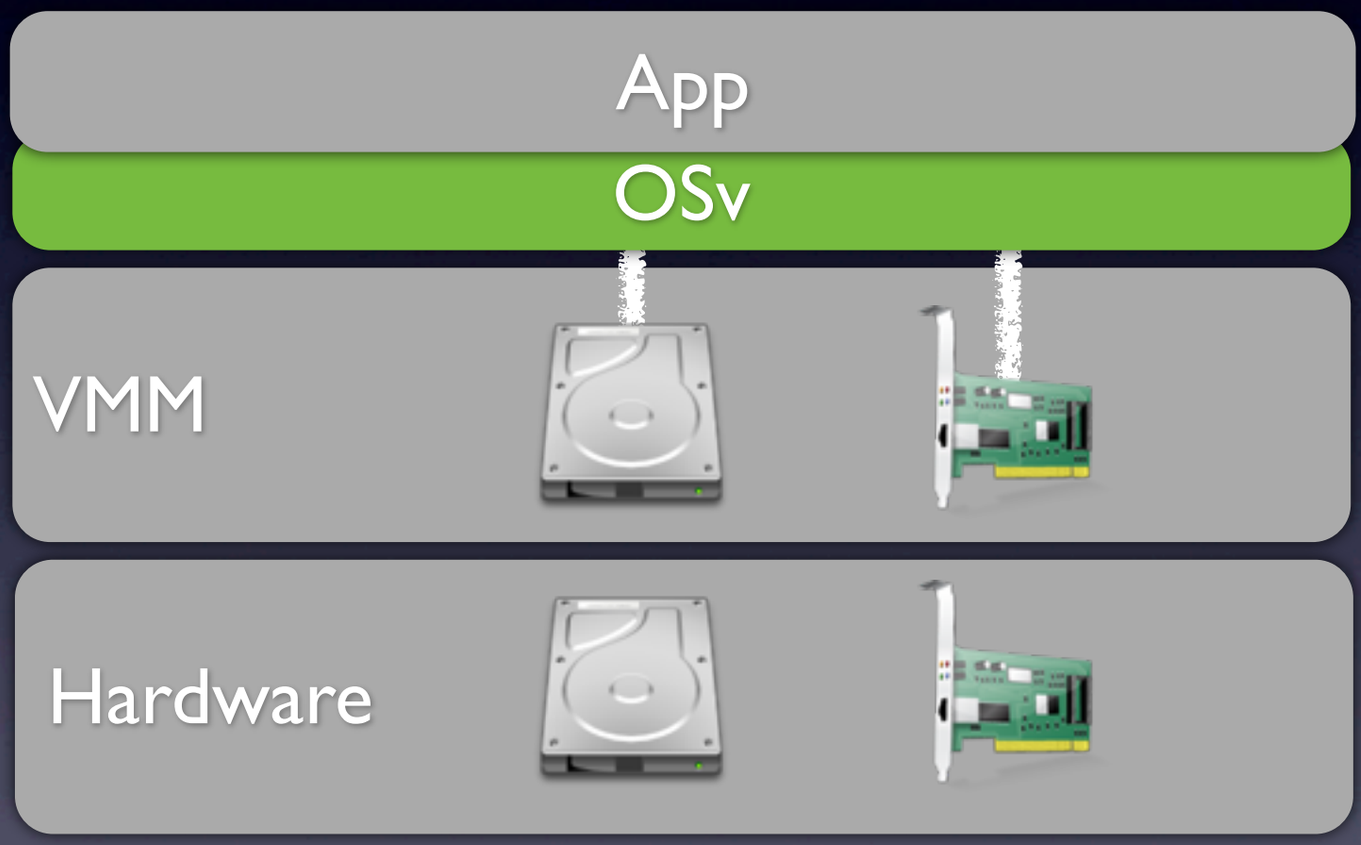
Takahiro Shinagawa
The University of Tokyo

BitVisor



http://www.justis.as-1.co.jp

OSv



http://medical-care.feed.jp

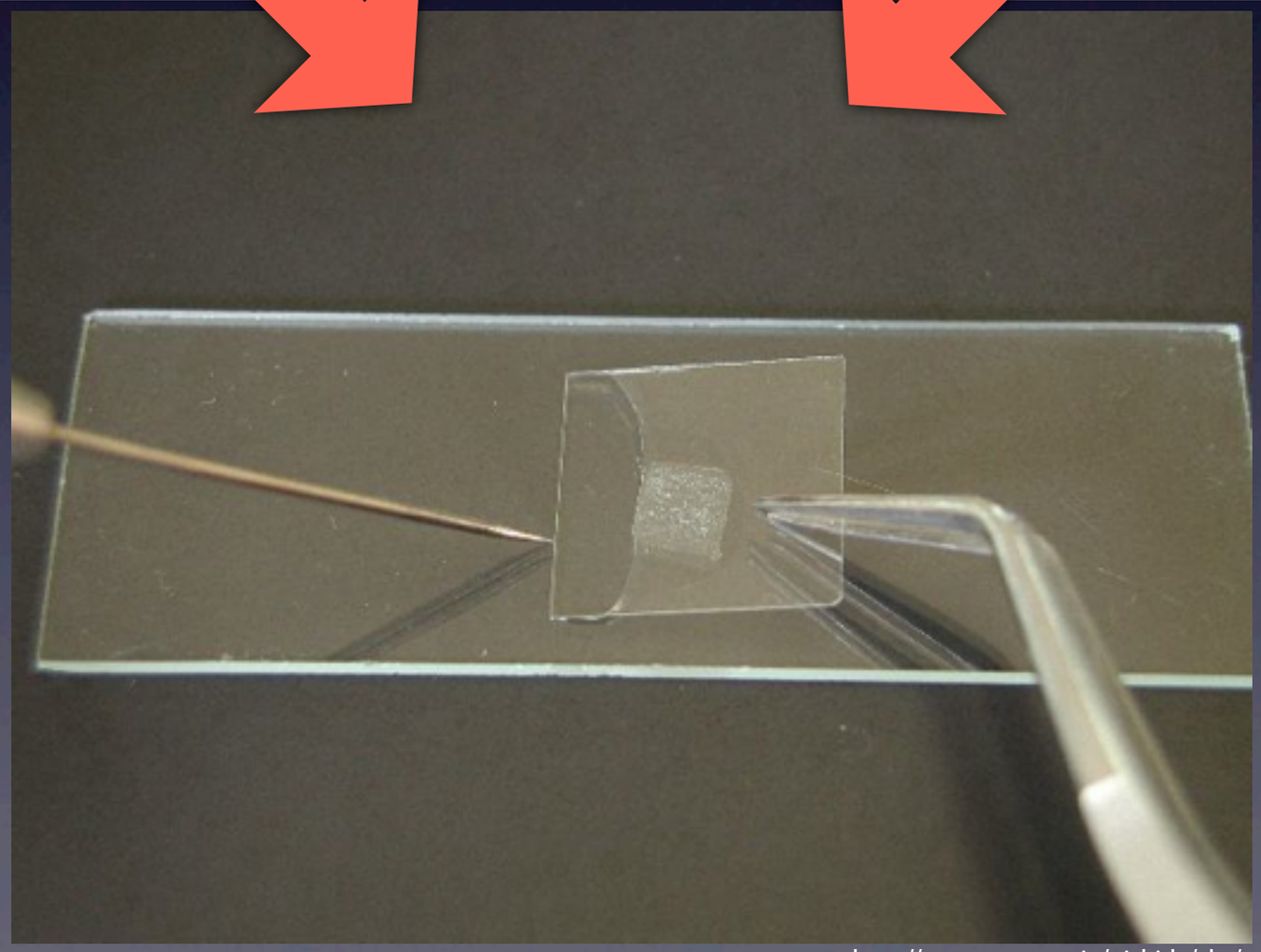
BitVisor

BitVisor

OSv

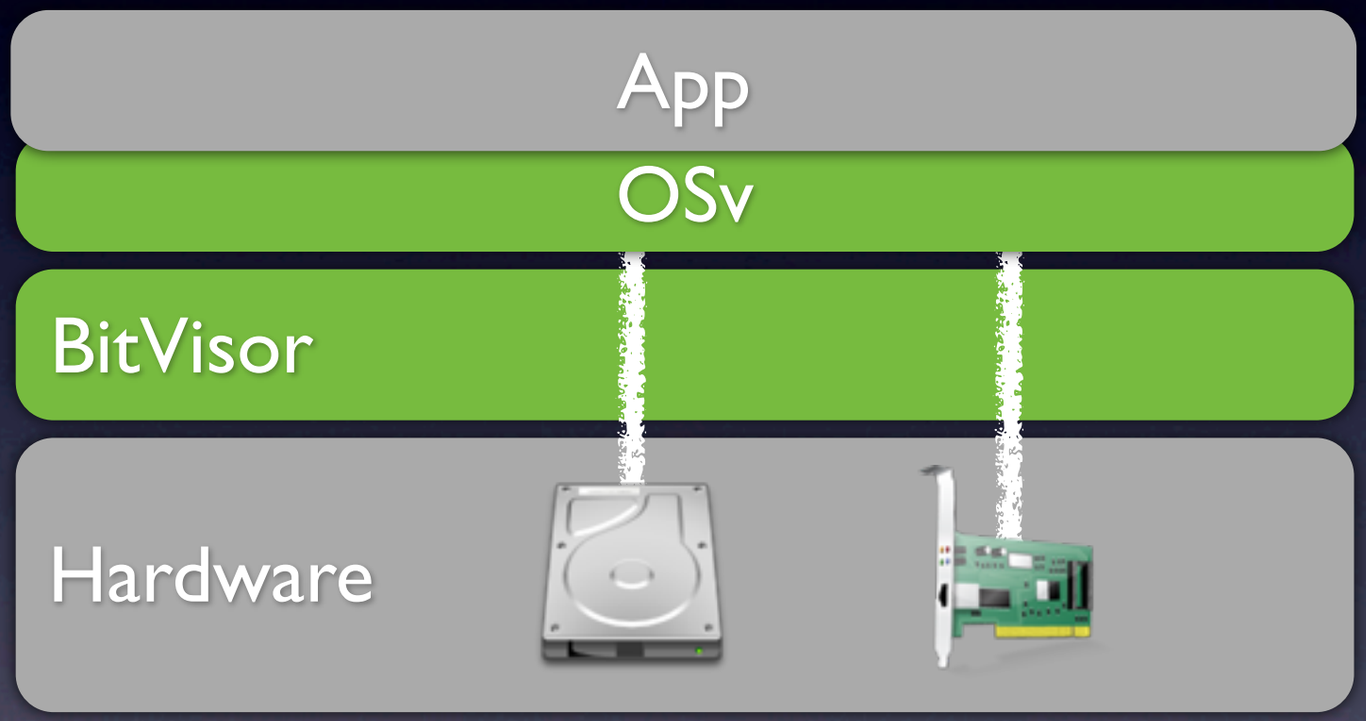
BitVisor

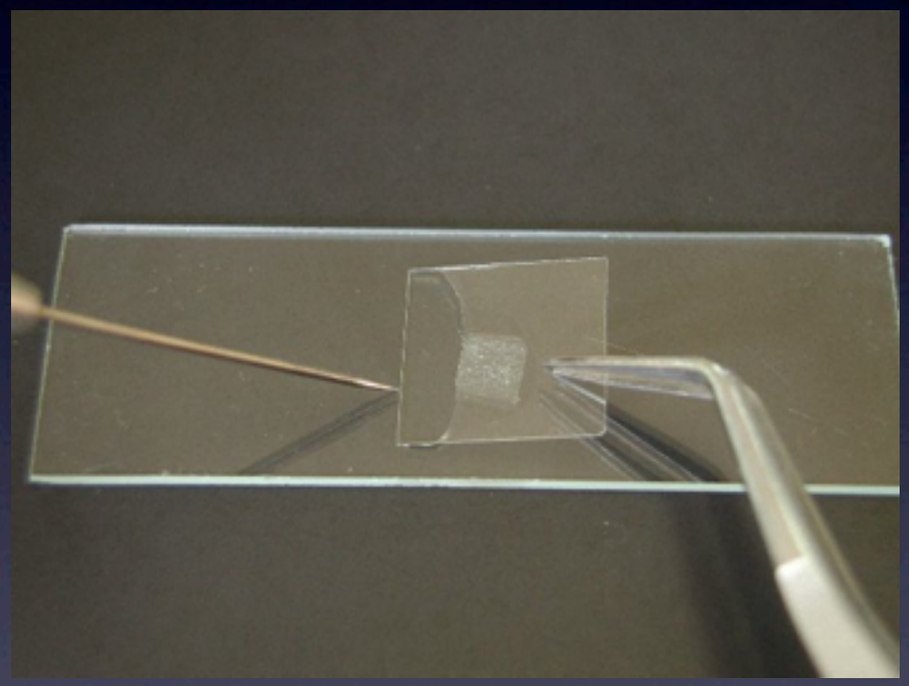
OSv



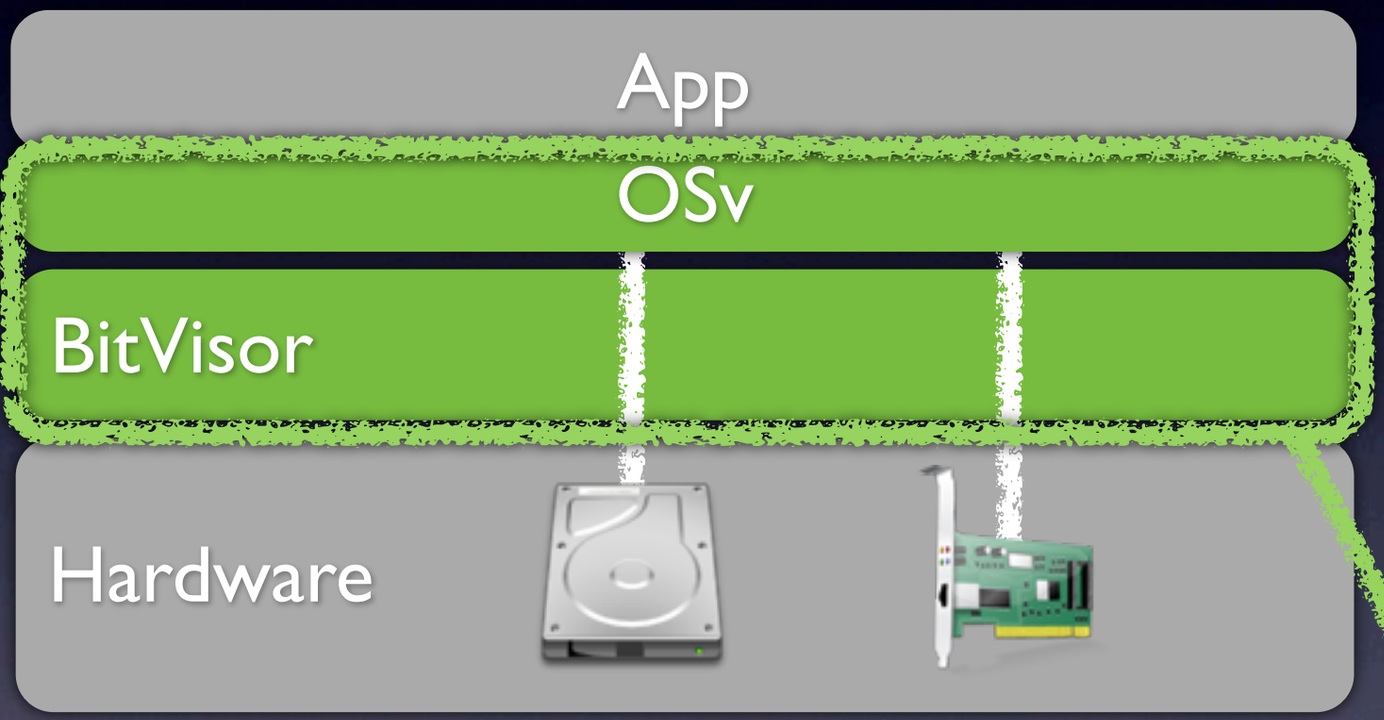
http://www.root.ne.jp/nishide/shs/

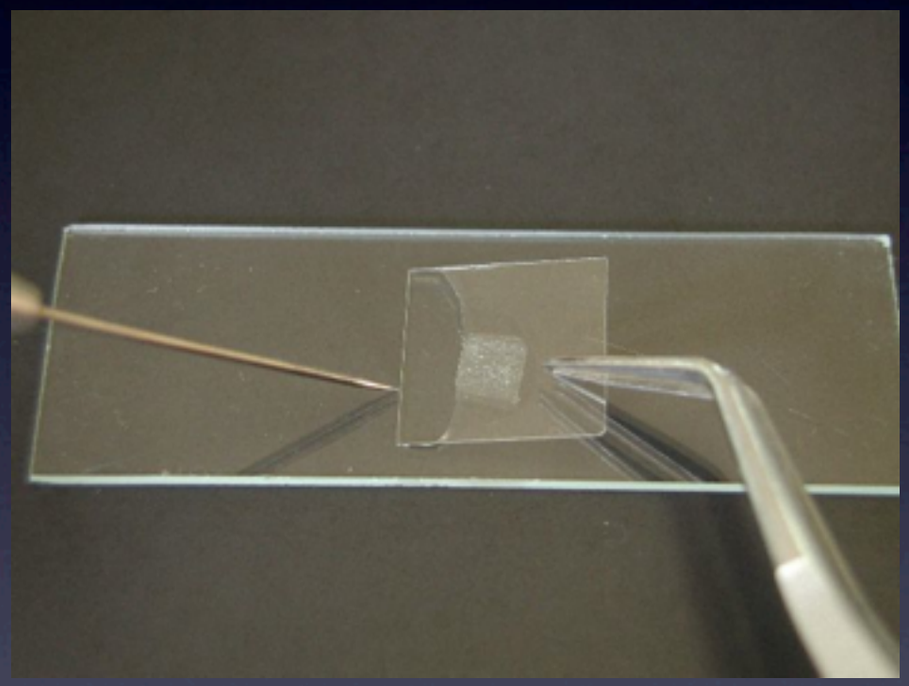
OSv on BitVisor





OSv on BitVisor





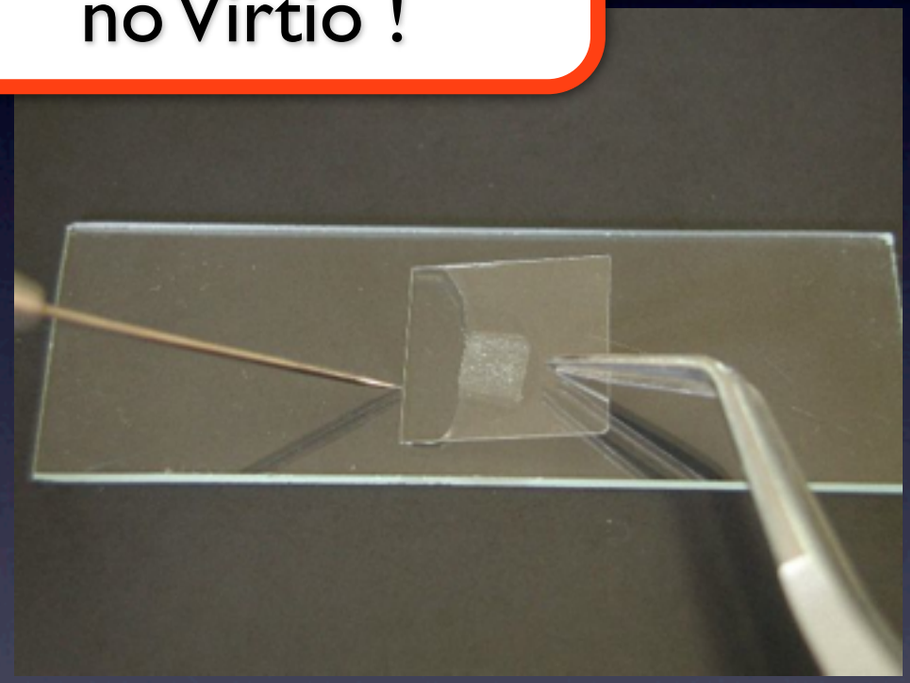
OSb

OSv on BitVisor

Physical interface sucks !

What the hell is PRO/1000 !

What the fuck no Virtio !

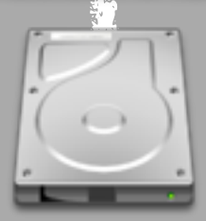


Ap

OSv

BitVisor

Hardware



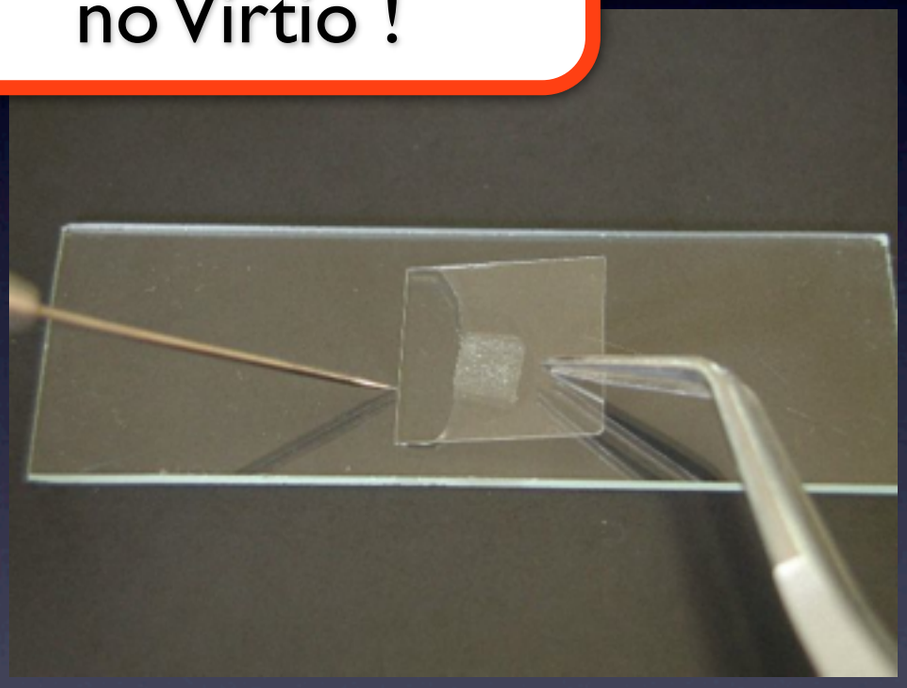
OSb

OSv on BitVisor

Physical interface sucks !

What the hell is PRO/1000 !

What the fuck no Virtio !



Ap

OSv

BitVisor

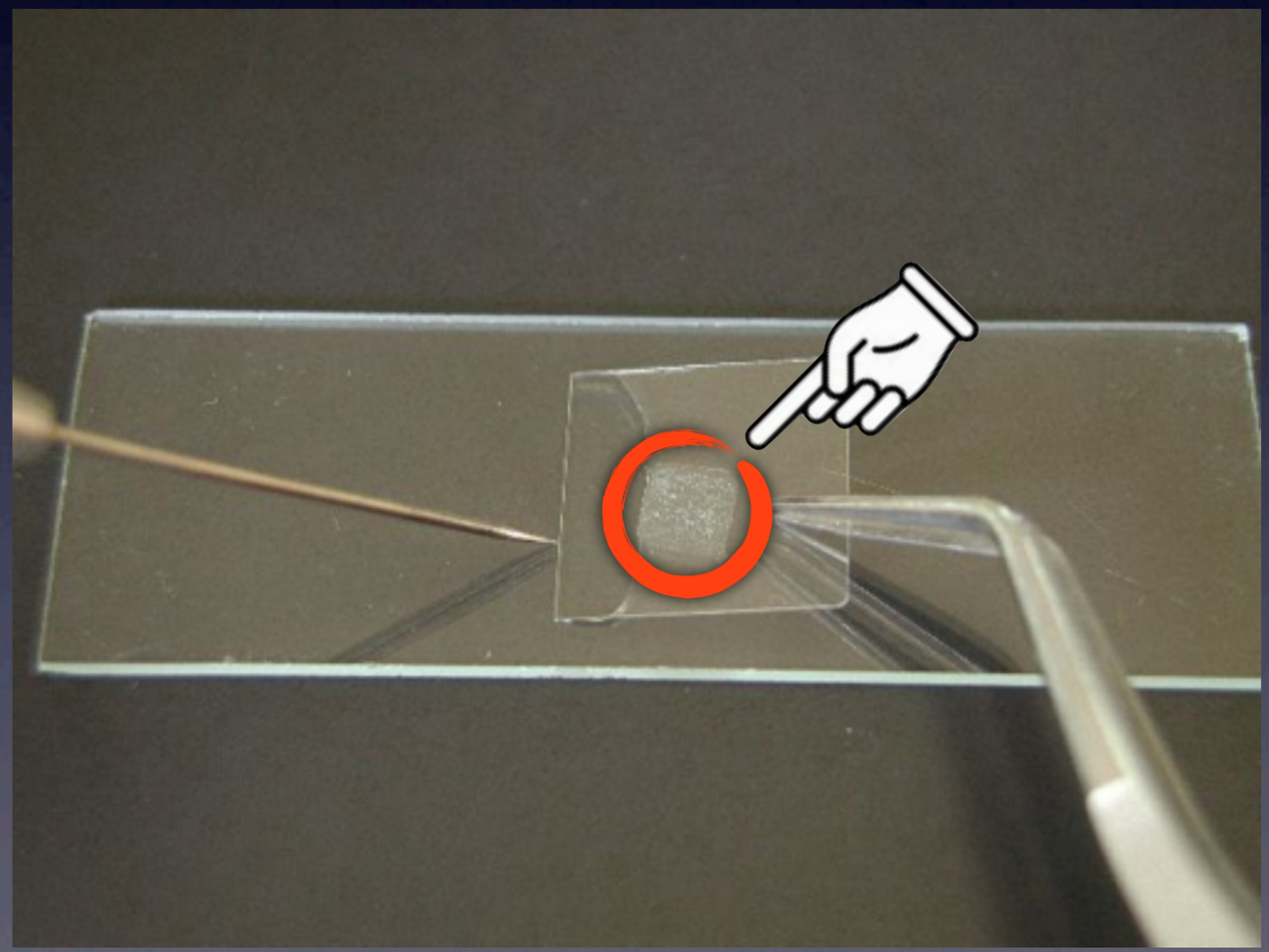
Virtio

Hardware

You have no choice.

OSb

~ The Road to OSb ~



OSv Code Reading

- Boot Process
- Some Drivers

- MBR
- Read local disk with INT13/42.
- Load command line, boot loader, OSv kernel.
- Get memory map with INT15/E820.
- Setup segment descriptors/page tables.
- Switch to 64-bit mode.
- premain()
- main()

...

OSv may run on physical machine. (=BitVisor)

- Serial / VGA output.
- SATA.
- Virtio NIC/BLK/RNG/SCSI.
- ACPI.
- APIC.

...

Only one thing to fix

Skip on BitVisor

Triple fault ★

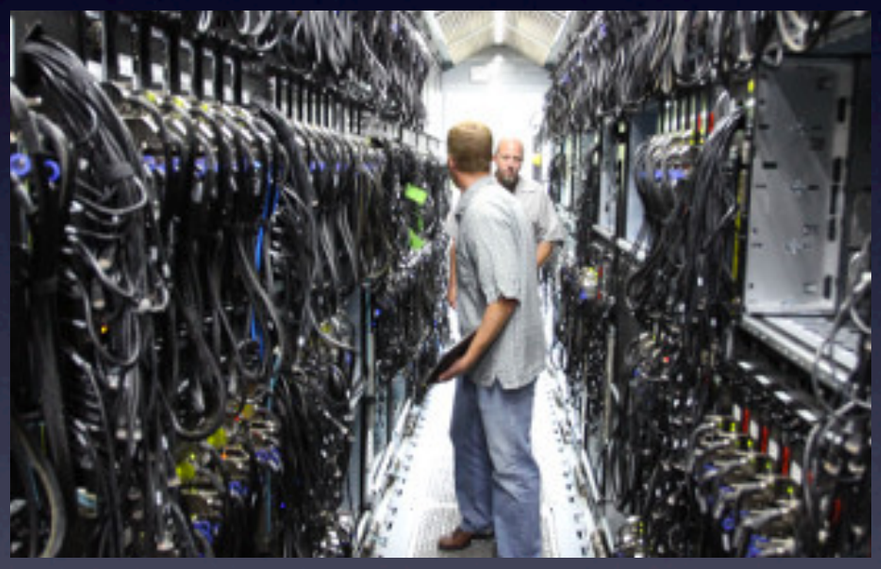
drivers/acpi.c:

```
    // Copy the root table list to dynamic memory
    if (!is_bitvisor()) {
        status = AcpiReallocateRootTable();
        if (ACPI_FAILURE(status)) {
            acpi_e("AcpiReallocateRootTable failed: %s\n",
AcpiFormatException(status));
            return;
        }
    }
```


RiDxdrRiDxdrRiDxdrRiDxdrRiDxdrRiDxdrRiDxdrRiDxdrRiDxdrRiDxdr
drRiDxdrRiDxdrRiDxdrRiDxdrRiDxdrRiDxdrRiDxdrRiDxdrRiDx
DxdrRiDxdrRiDxdrRiDxdrRiDxdrRiDxdrRiDxdrRiDxdrRiDxdrRi
RiDxdrRiDxdrRiDxdrRiDxdrRiDxdrRiDxdrRiDxdrRiDxdrRiDxdr
drRiDxdrRiDxdrRiDxdrRiDxdrRiDxdrRiDxdrRiDxdrRiDxdrRiDx
I: error
12 CPUs detected
Firmware vendor: FUJITSU // Phoenix Technologies Ltd.
bsd: initializing - done
VFS: mounting ramfs at /
VFS: mounting devfs at /dev
RAM disk at 0x0ffff800113a92040 (4096K bytes)
net: initializing - done
vga: Add VGA device instance
AHCI: Disabled
AHCI: Enabled
AHCI 0:1 initia
AHCI 0:2 initia
AHCI 0:3 initia
AHCI 0:4 initia
AHCI 0:5 initi ized
Warning: No h dware source of entropy available to your platform,
CSPR G will rely on software source of entropy to provide
random: <Software, Yarrow> initialized
Hello world
main() returned.
random: device unblocked.
Hello world !

On BitVisor,

Whenever updating OSv images,
you update the entire Physical Disk.



http://a.fsdn.com/sd/articles/14/11/12/1946208-1.jpg

http://www.dreamstime.com

Troublesome...
(especially for development)

Network-boot of OSb

./scripts/run.py

Network-boot of OSb

./scripts/run.py

Hardware

Network-boot of OSb

./scripts/run.py

Image

Update

Hardware

Network-boot of OSb

./scripts/run.py

Image

Update

IPMI

Hardware

Network-boot of OSb

./scripts/run.py

Image Update

PXE Boot

IPMI

Hardware

Network-boot of OSb

./scripts/run.py

Image

Update

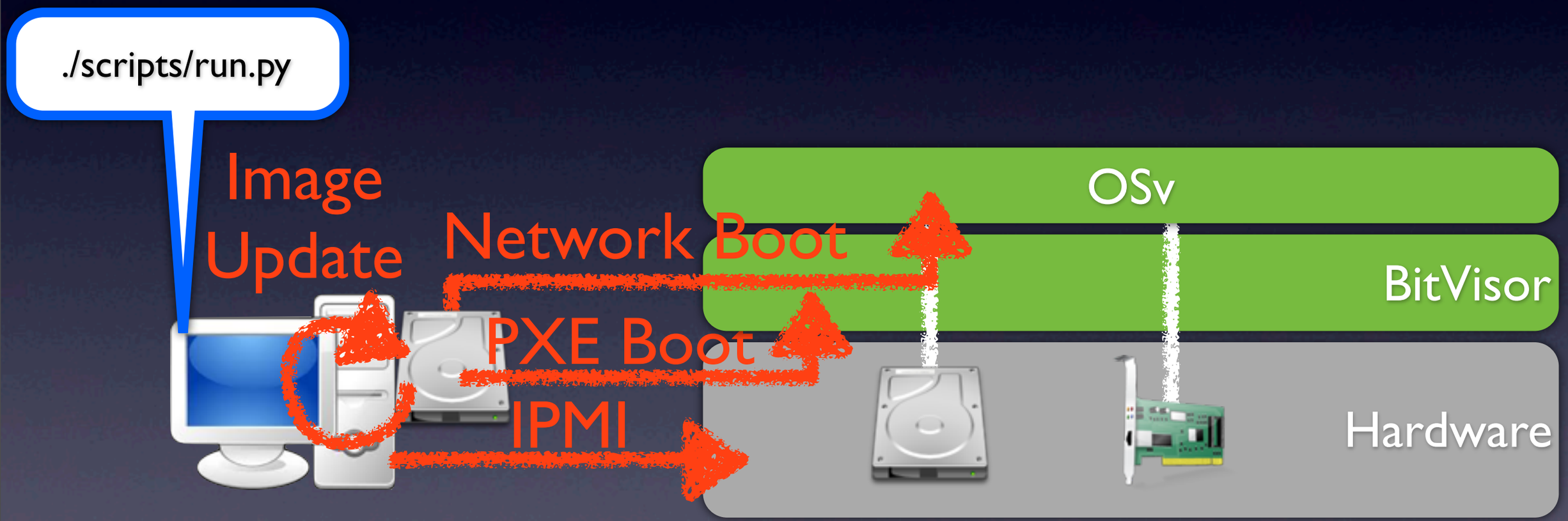
PXE Boot

IPMI

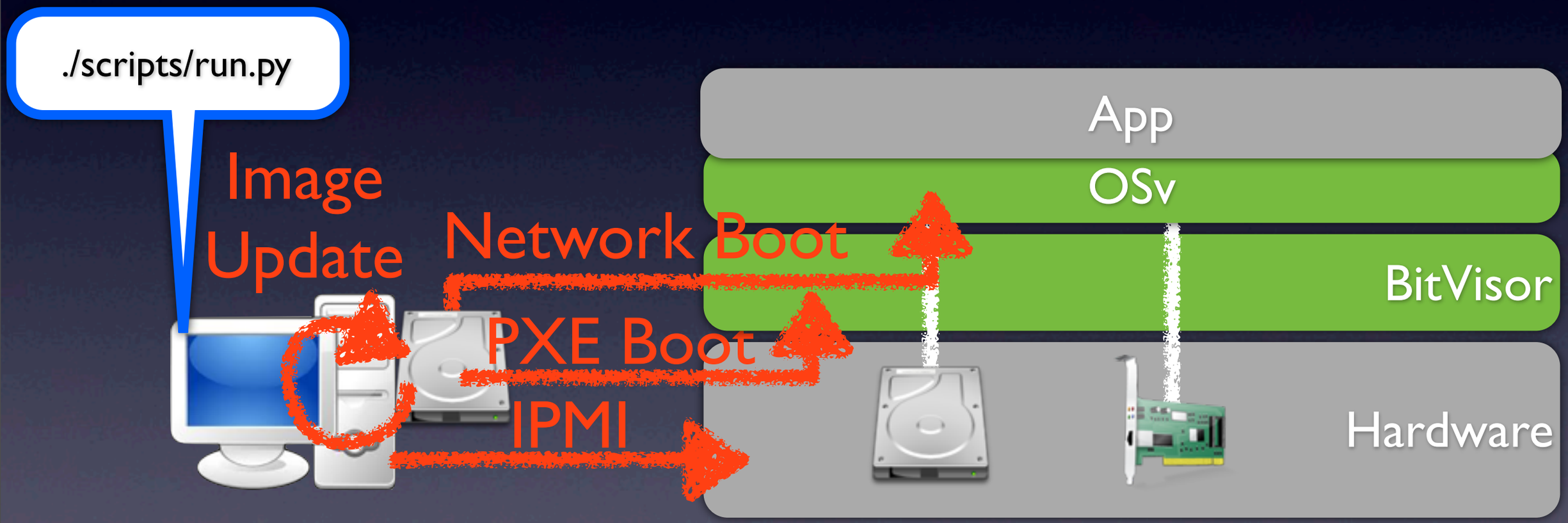
BitVisor

Hardware

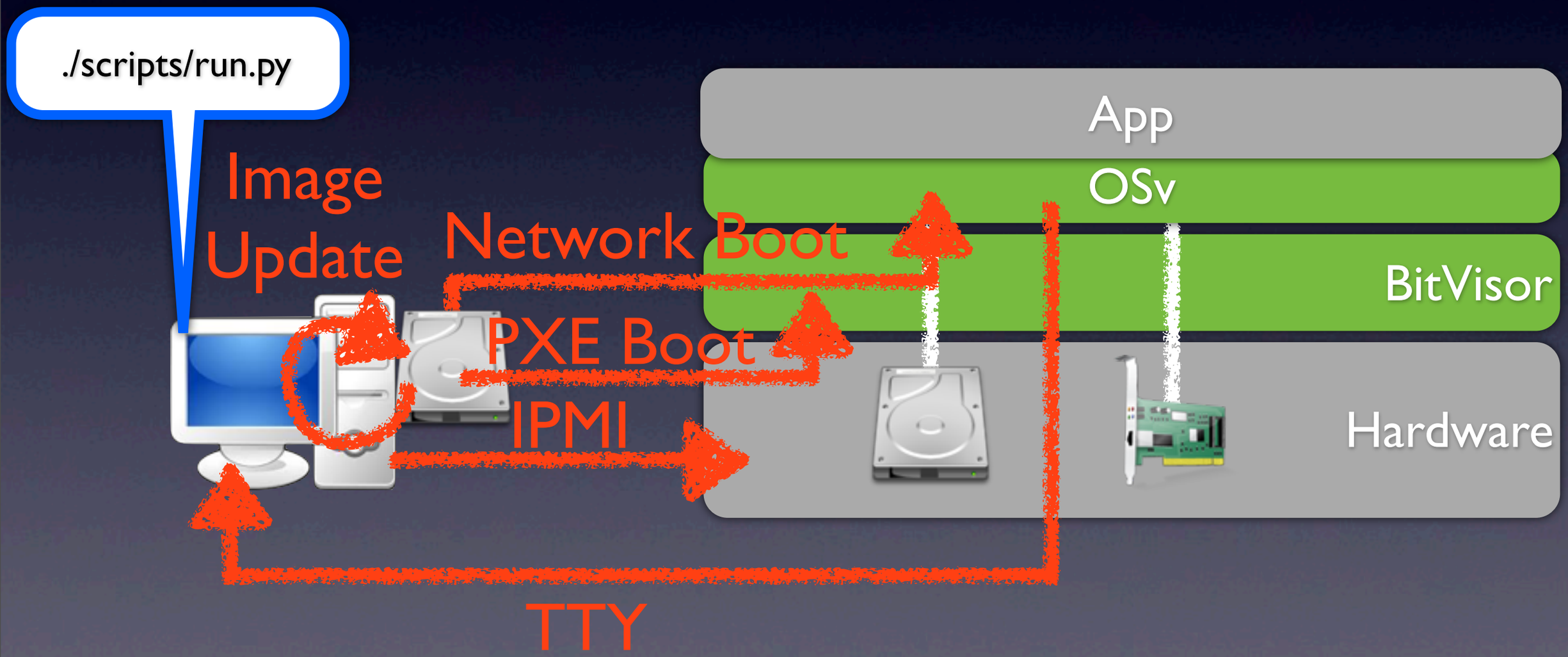
Network-boot of OSb



Network-boot of OSb

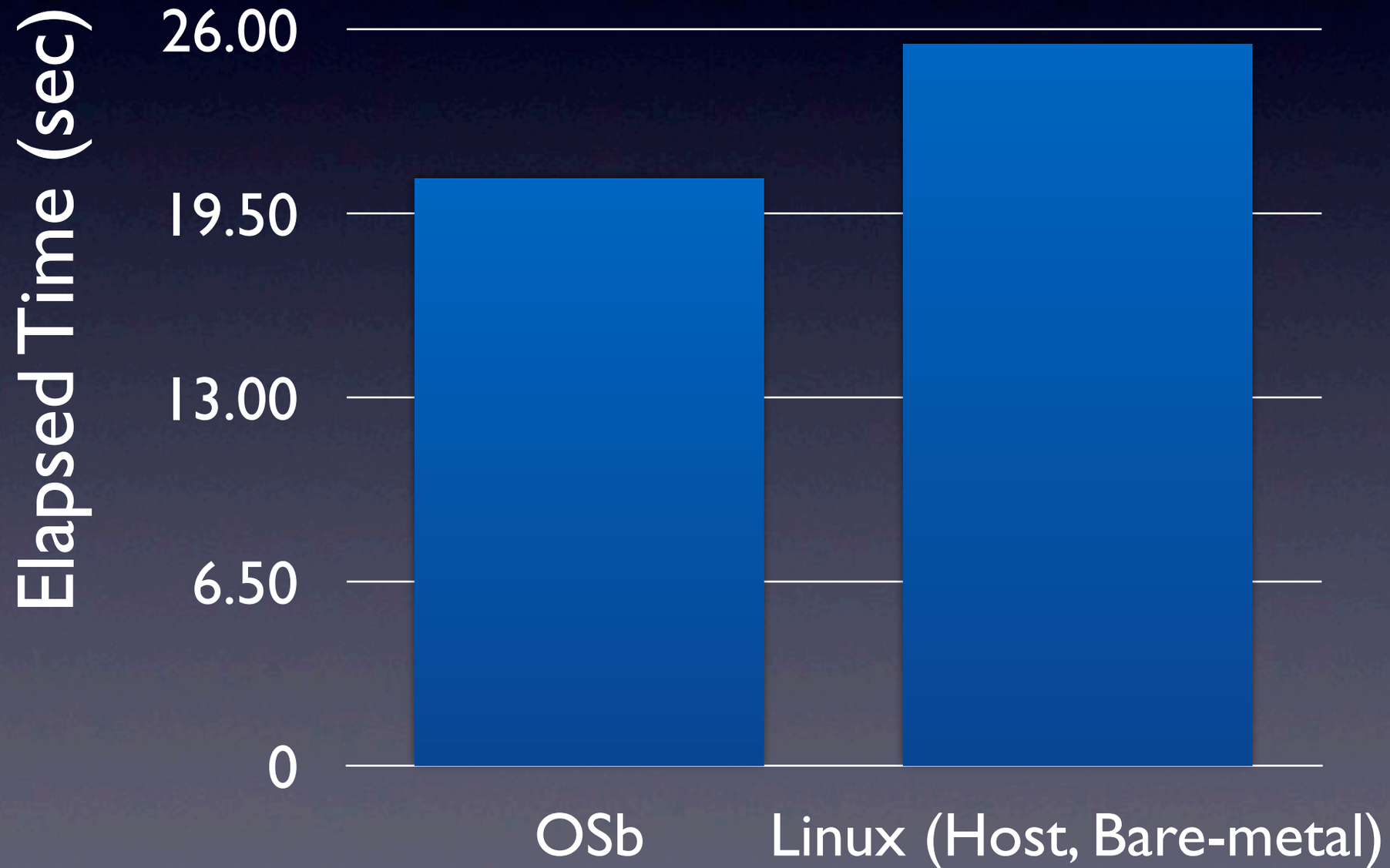


Network-boot of OSb



Performance?

for (12-billion times) val++;



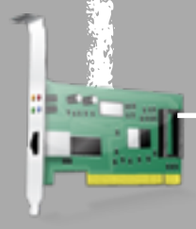
Virtio NIC

App

OSv

BitVisor

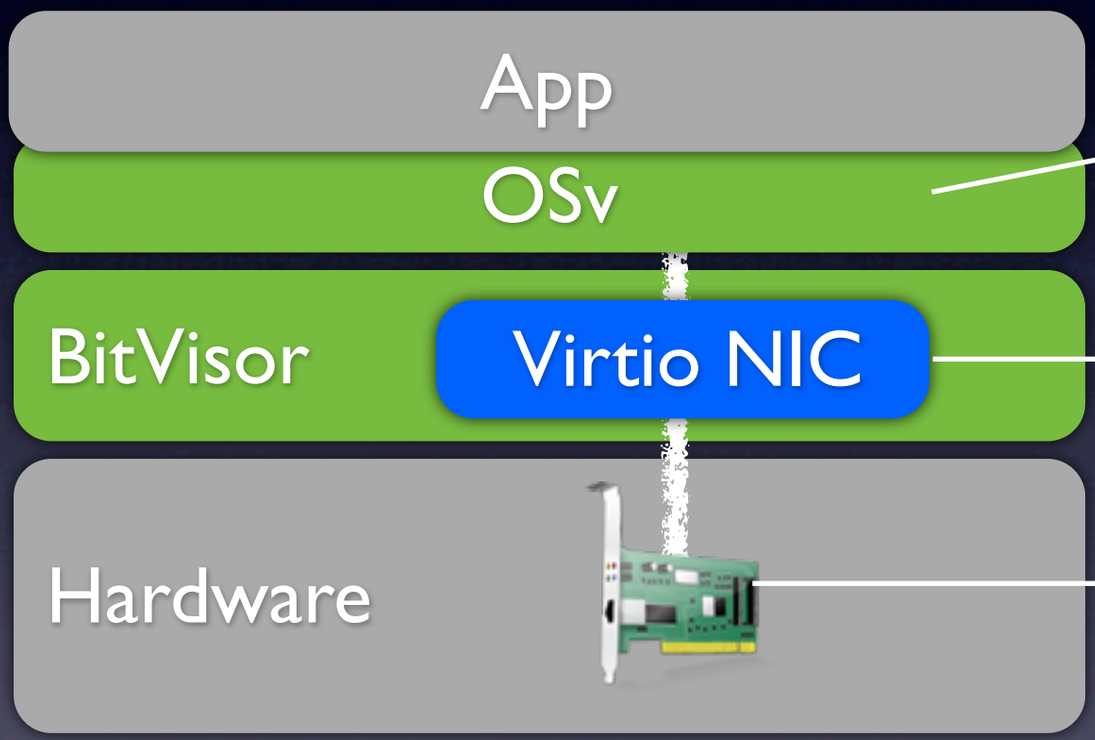
Hardware



No PRO/1000...
Virtio NIC please...

PRO/1000...

Virtio NIC

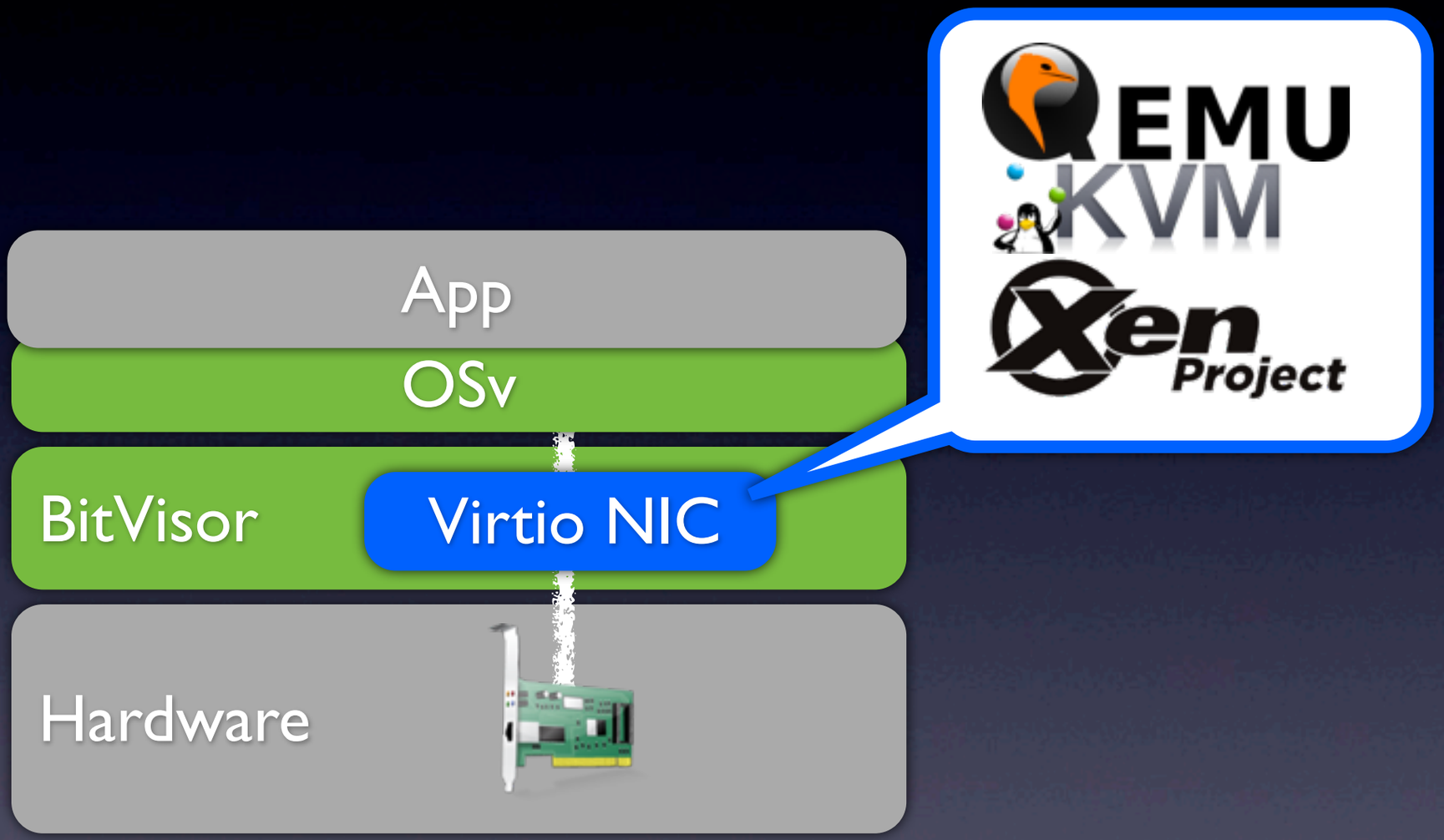


No PRO/1000...
Virtio NIC please...

Implementing Virtio

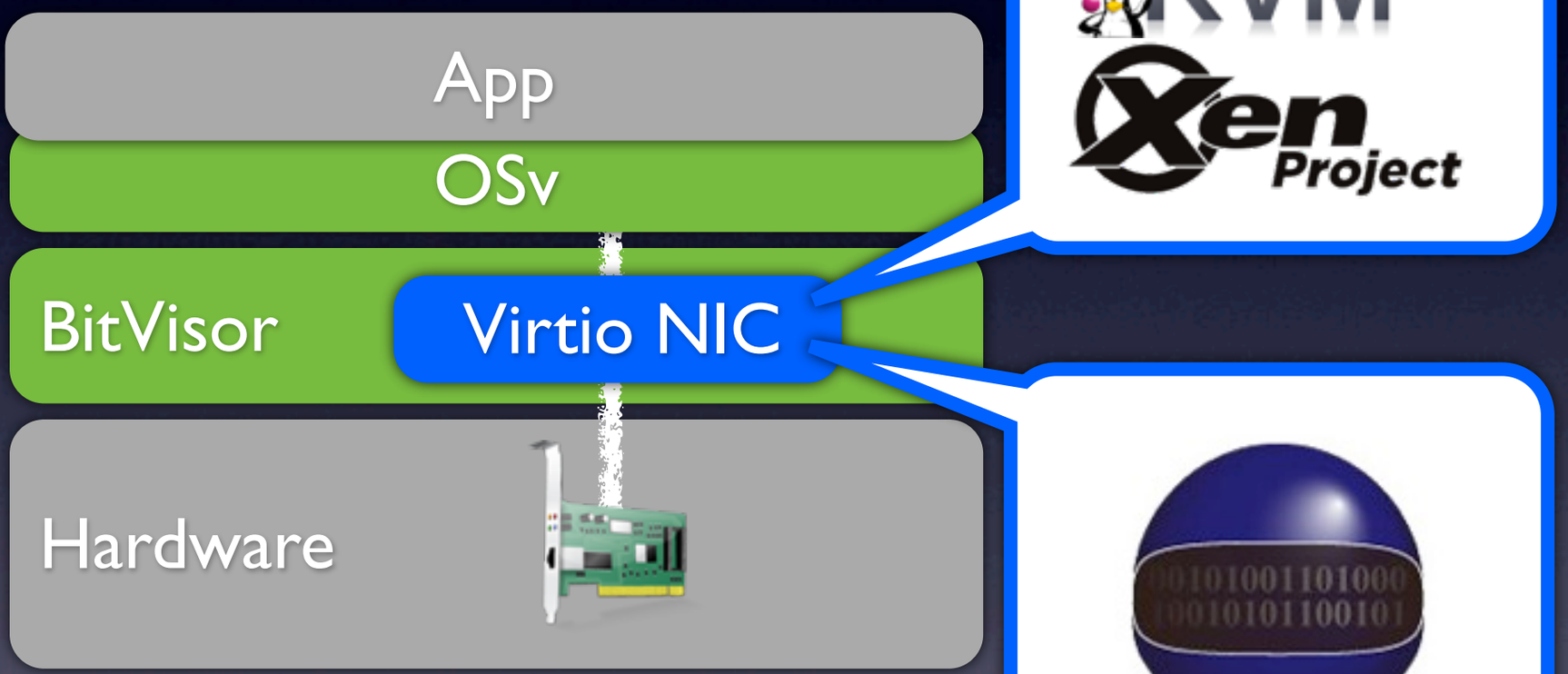
PRO/1000...

How to implement?



Finally, Our Virtualization?

How to implement?



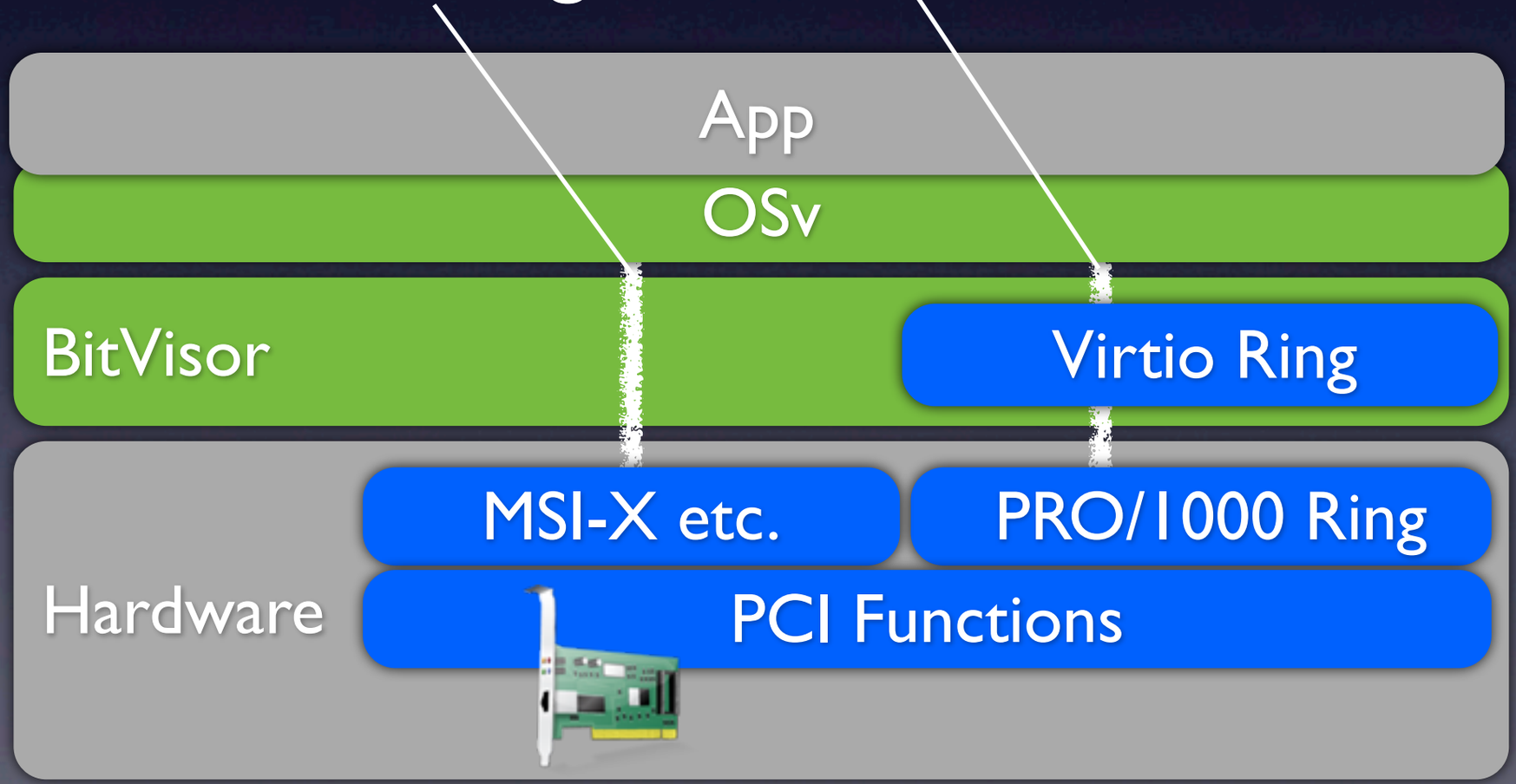
Finally, Our Virtualization?

No, BitVisor's Way!

Para Pass-through Virtio

Intercept only interesting I/Os!

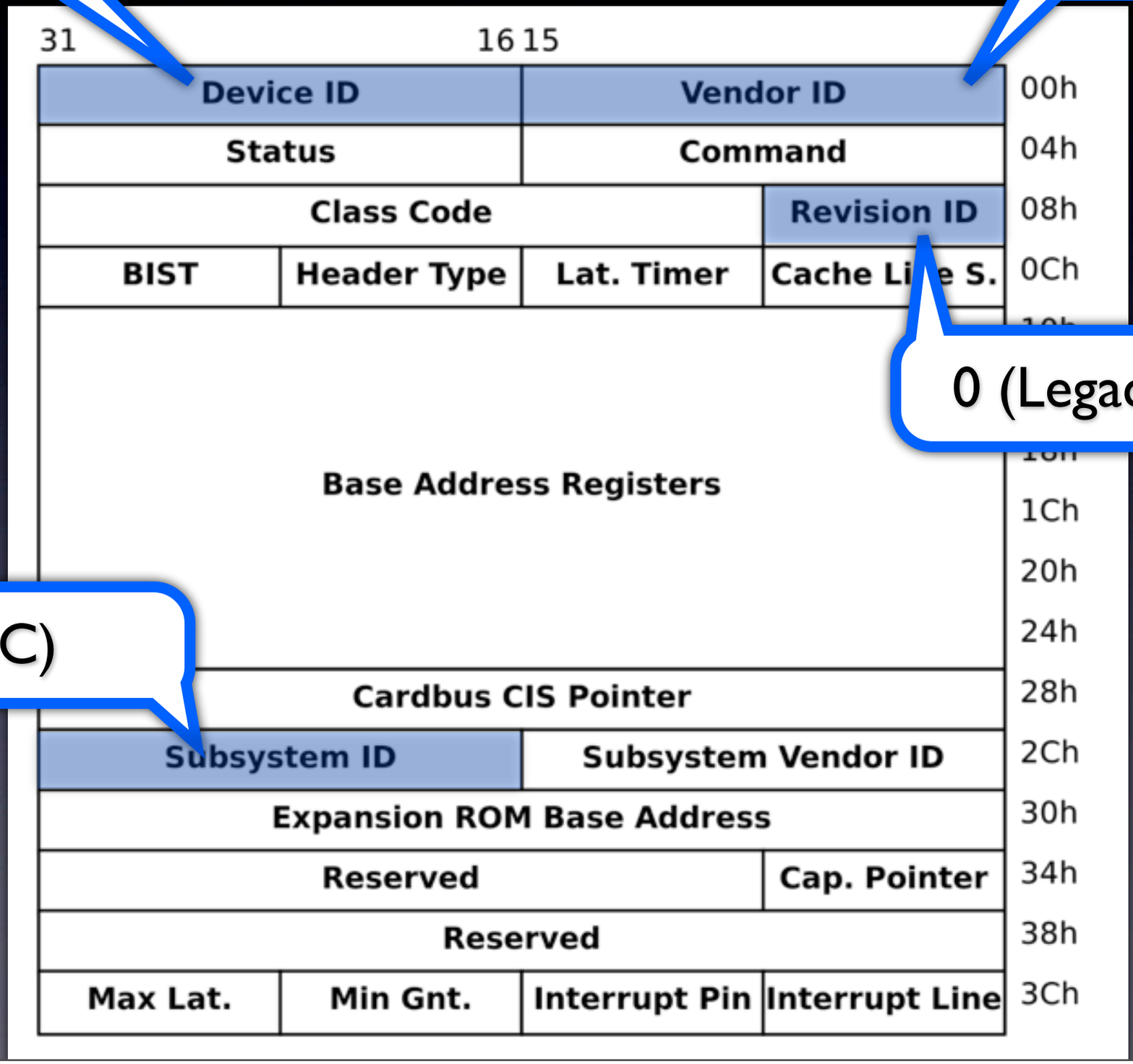
Pass-through!



Faking PCI IDs

0x1000
(Virtio NIC)

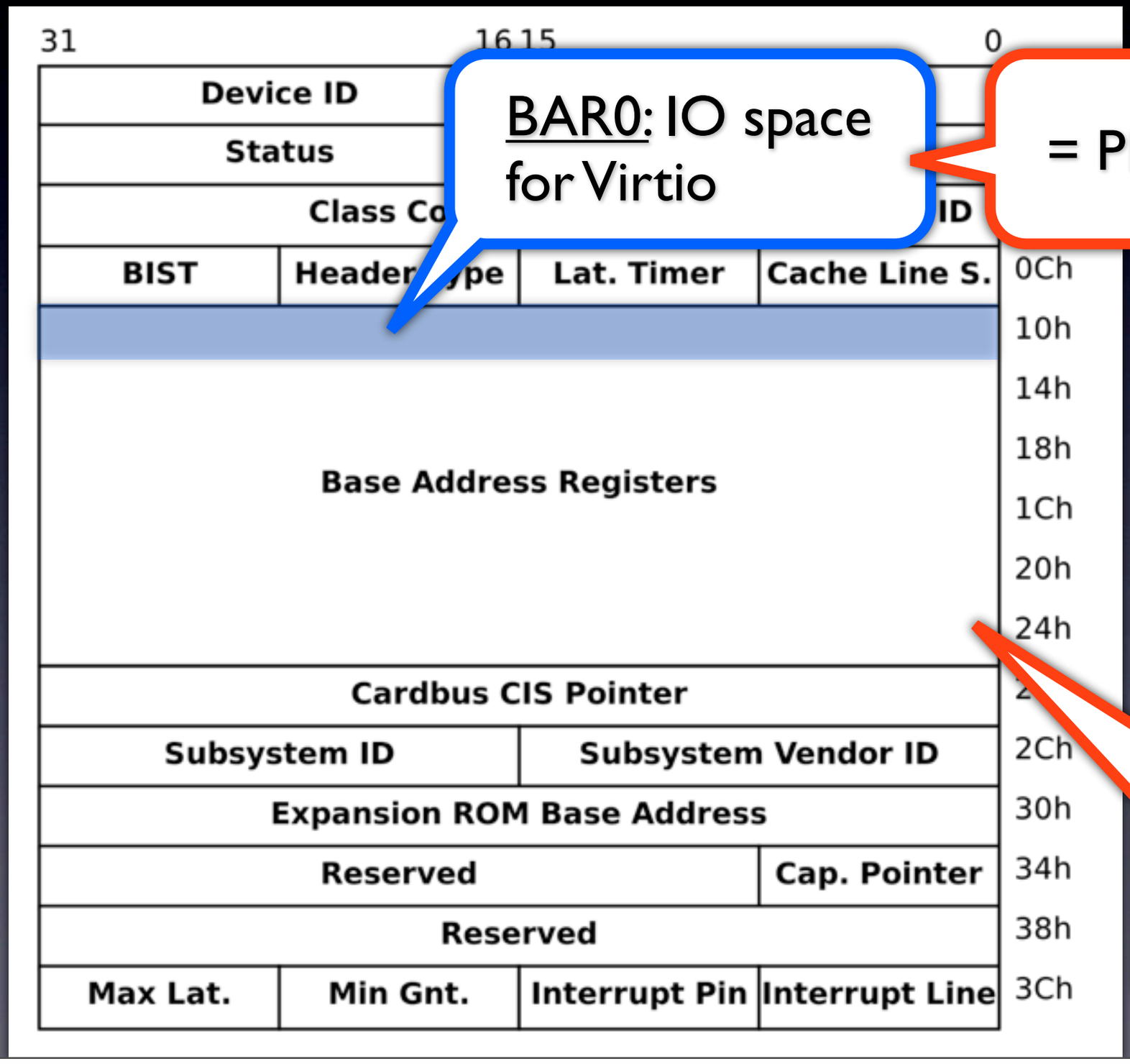
0x1AF4
(Virtio Device)



0 (Legacy Virtio)

1 (Virtio NIC)

Faking PCI BARs



BAR0: IO space for Virtio

= PRO/1000 IO space

Everything else Pass-through (e.g. MSI-X)

Virtio NIC Operations

- Virtio Ring
- Packet Transmission
- Packet Reception

Virtio Ring

BAR0 (IO)

BitVisor emulates these registers

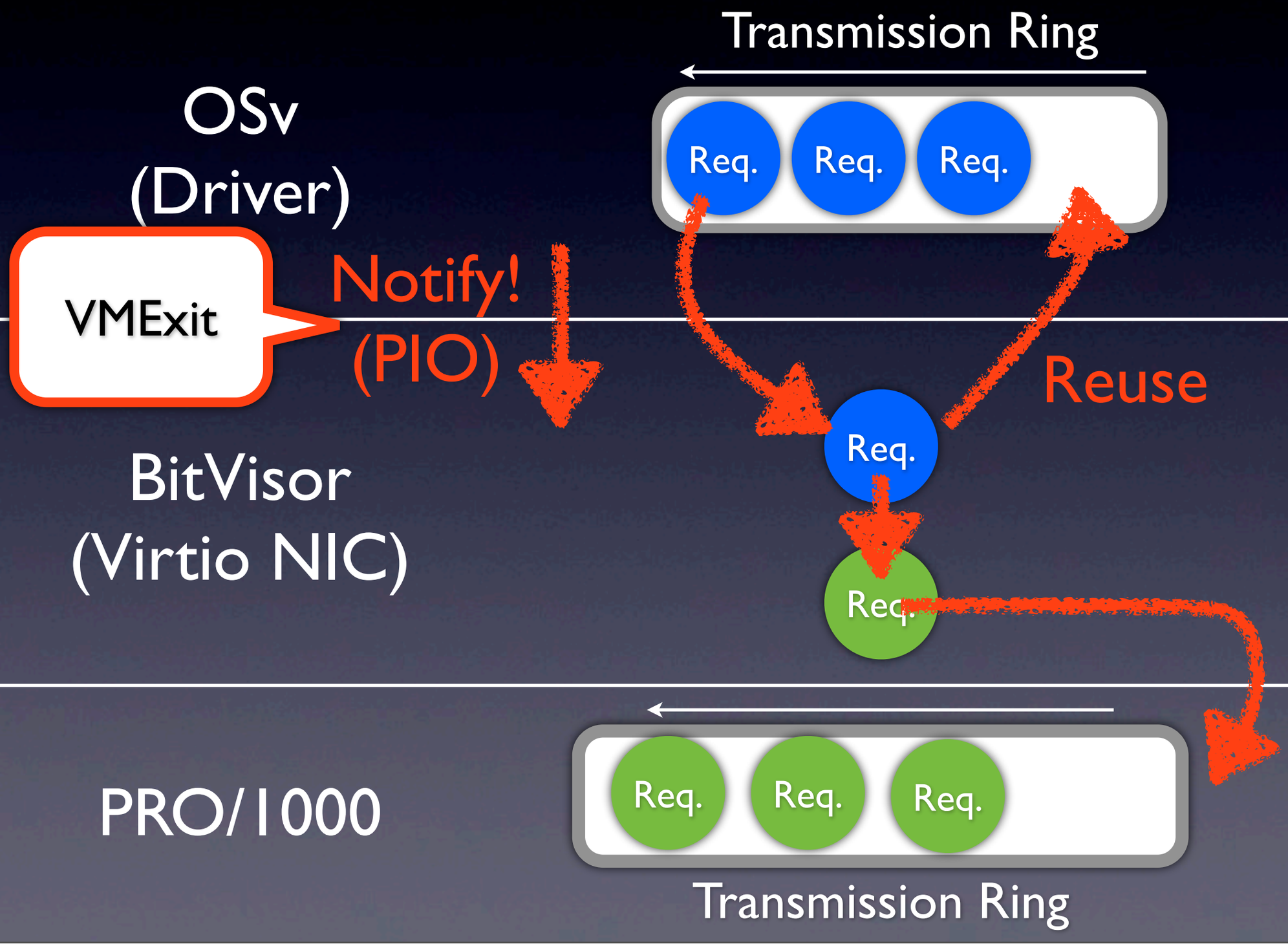
Guest selects rings:
* Transmission Ring
* Reception Ring

Bits	32	32	32	16		16	8	8
Read / Write	R	R+W	R+W	R	R+W	R+W	R+W	R
Purpose	Device Features bits 0:31	Driver Features bits 0:31	Queue Size	Queue Select	Queue Notify	Queue Address	Device Status	ISR Status

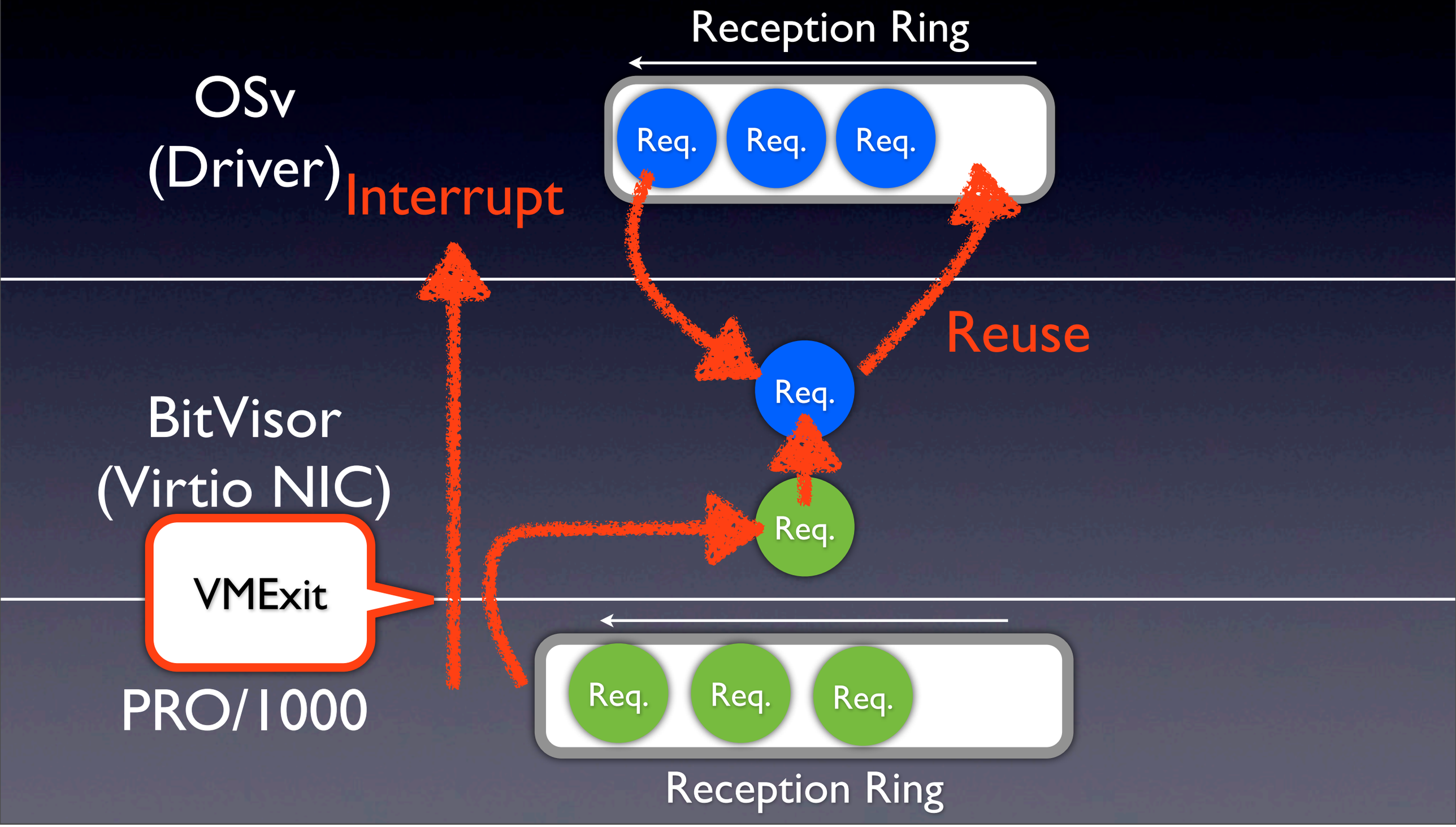
Memory space

Virtio Ring (Available + Used)

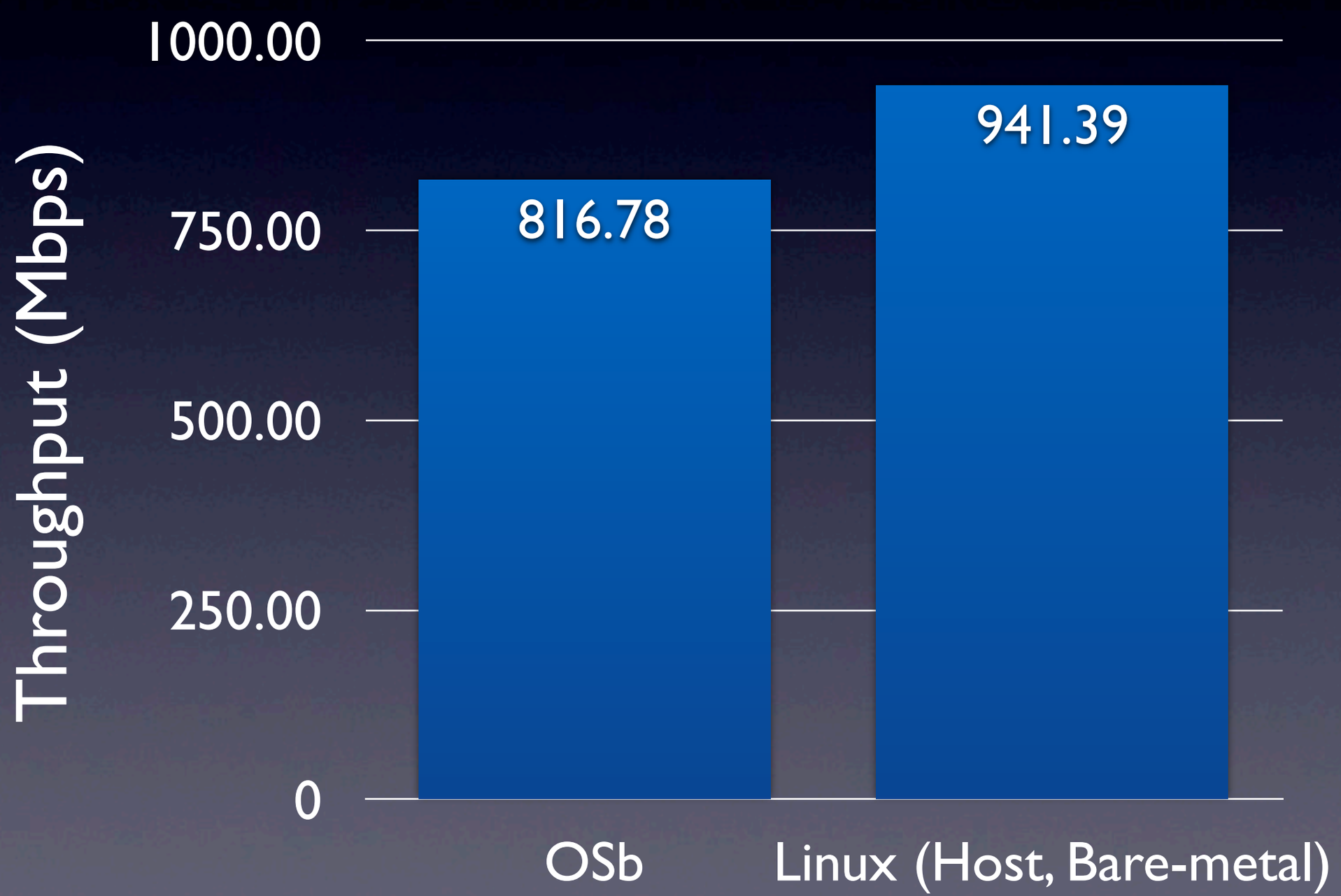
Packet Transmission



Packet Reception



Netperf TCP_STREAM



VMExits are Costly...

OSv
(Driver)

VMExit

Notify!
(PIO)

Guest
mode

Switching
Overhead

Interrupt

Host
mode

BitVisor
(Virtio NIC)

VMExit

PRO/1000

Exitless Virtio with Dedicated Core

ELVIS [Har'El et al. ATC'13]

Always Guest Mode

Always Host Mode

Transmission Ring

Reception Ring

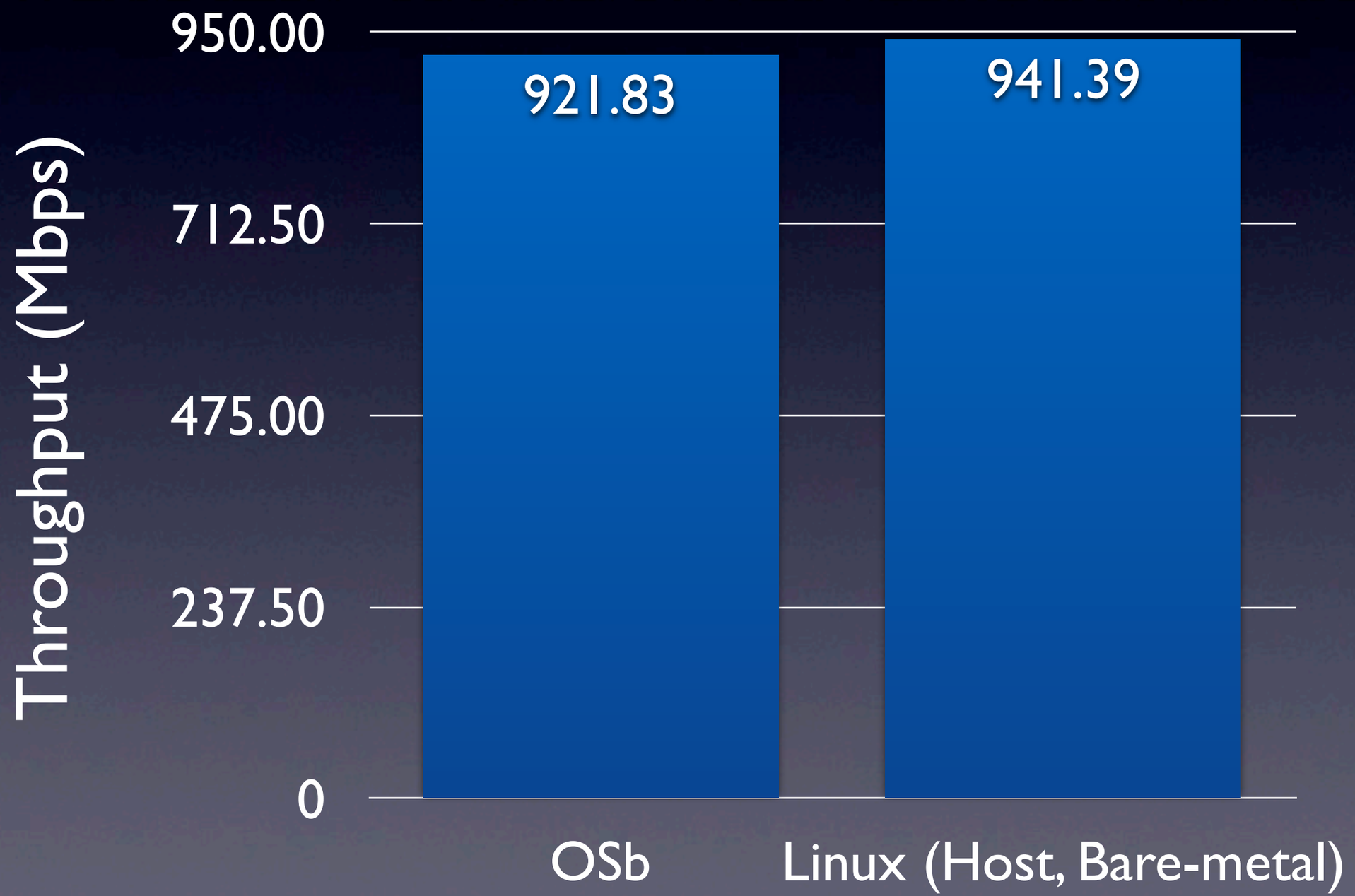
Polling

Interrupt (IPI)

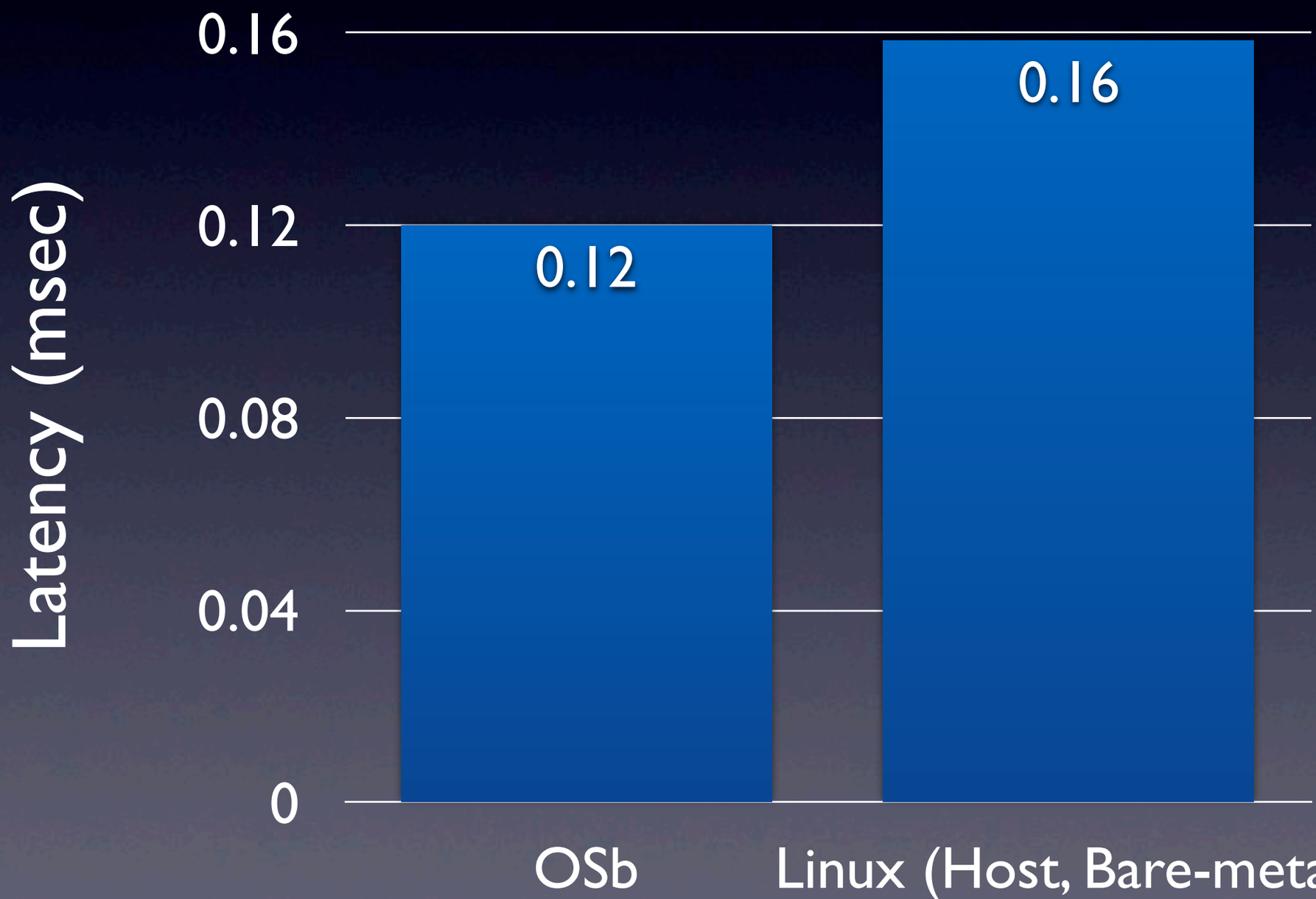
Implementation Summary

- Core Concealing
  - Modifying ACPI MADT Table
- Notification PIO Pass-through
  - Pass-through to ineffective PIO
- Interrupt
  - Get vector # from MSI-X table and Send IPI
  - (Or ask PRO/1000 to trigger)

Netperf TCP_STREAM (Exitless Virtio)



Ping (Exitless Virtio)



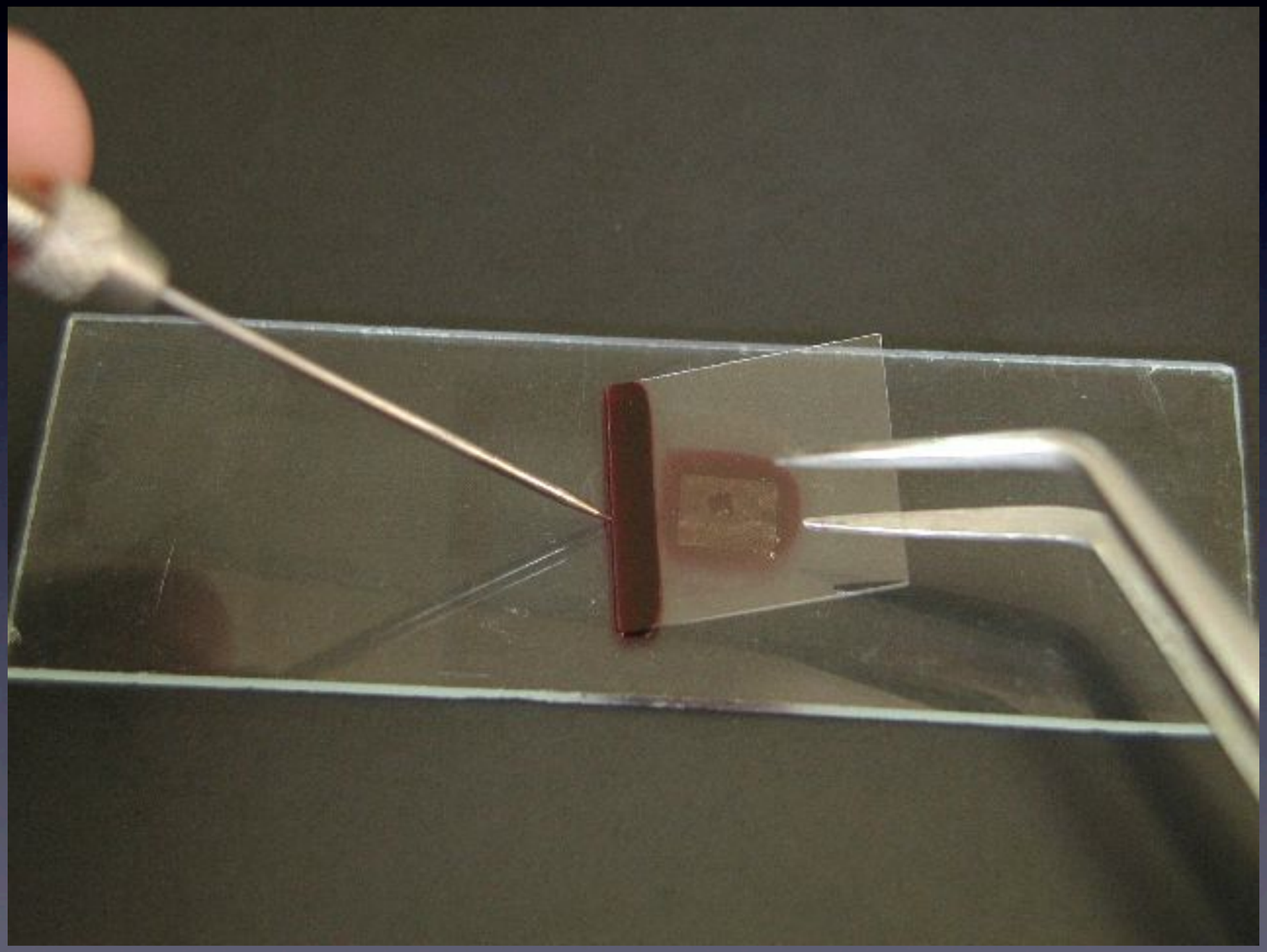
Still Under Optimization

- TODOs
- Interrupt Moderation
- Zero Copy/VMDq
- Advanced PRO/1000 Descriptor
- TCP Offloading

Summary & Future Work

- Summary
  - Running OSv on BitVisor
    - Virtio NIC + Optimization
- Future work
  - Further Optimization & Evaluation
  - Other Virtio Devices (BLK, RNG...)

Thank you !



http://www.root.ne.jp/nishide/shs/