Stub Domains

A Step Towards Dom0 Disaggregation

Samuel Thibault, Citrix/XenSource

The Big Domain 0

- Runs a lot of Xen components
  - Domain manager
  - Domain Builder
  - Device Models
  - PyGRUB
- These are currently running as root
  - e.g. PyGRUB to access guest's disk

- Security issues
- Scalability issues

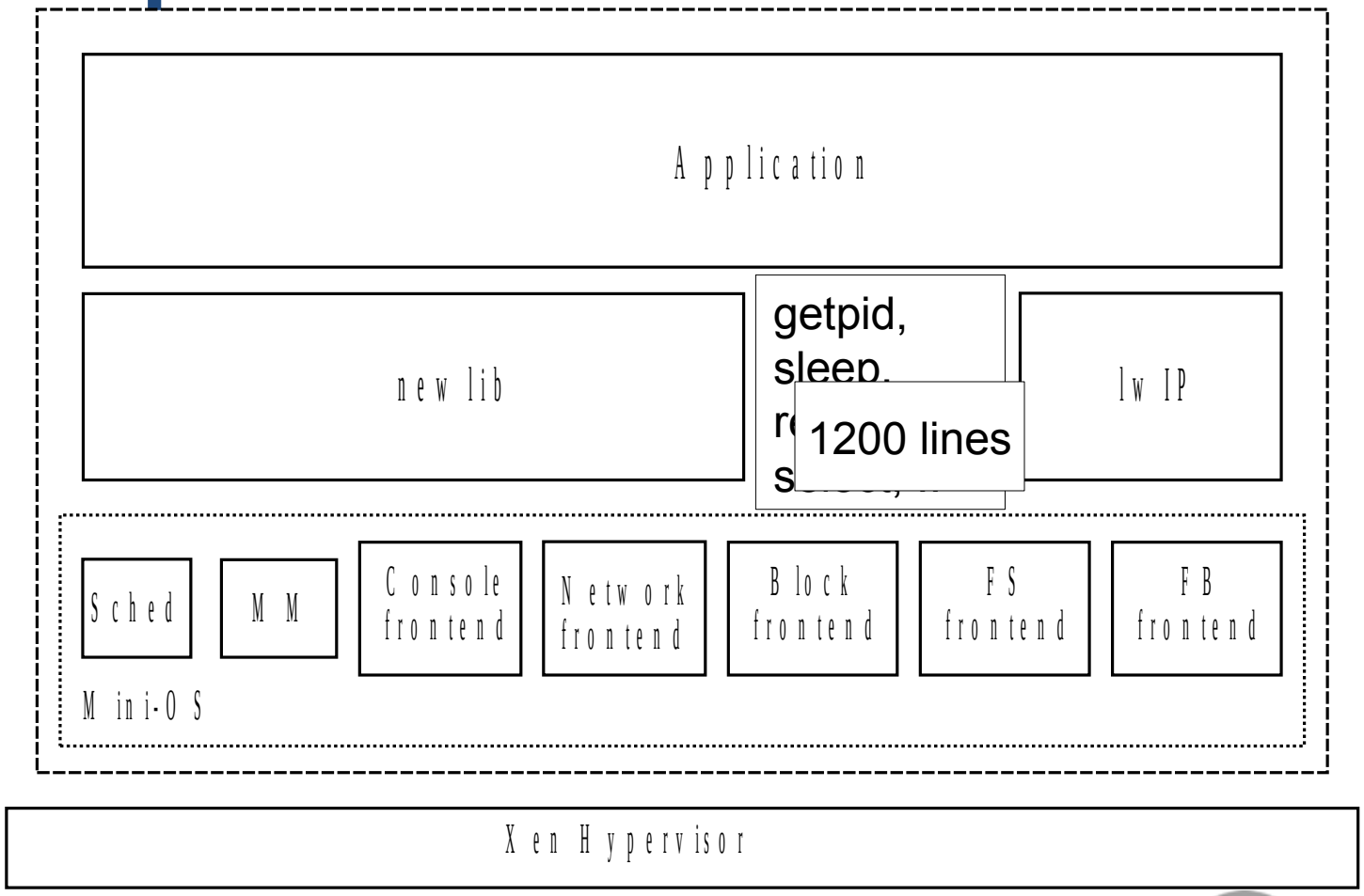
What Are Stub Domains?

- Helper domains which run Xen components
- Based on Mini-OS
- Domain Builder (Derek Murray)
- Device Model
- PV-GRUB
- ...

What Are Stub Domains?

- Helper domains which run Xen components
- Based on Mini-OS
- Domain Builder (Derek Murray)
- **Device Model**
- **PV-GRUB**
- …

POSIX Environment on Top of Mini-OS



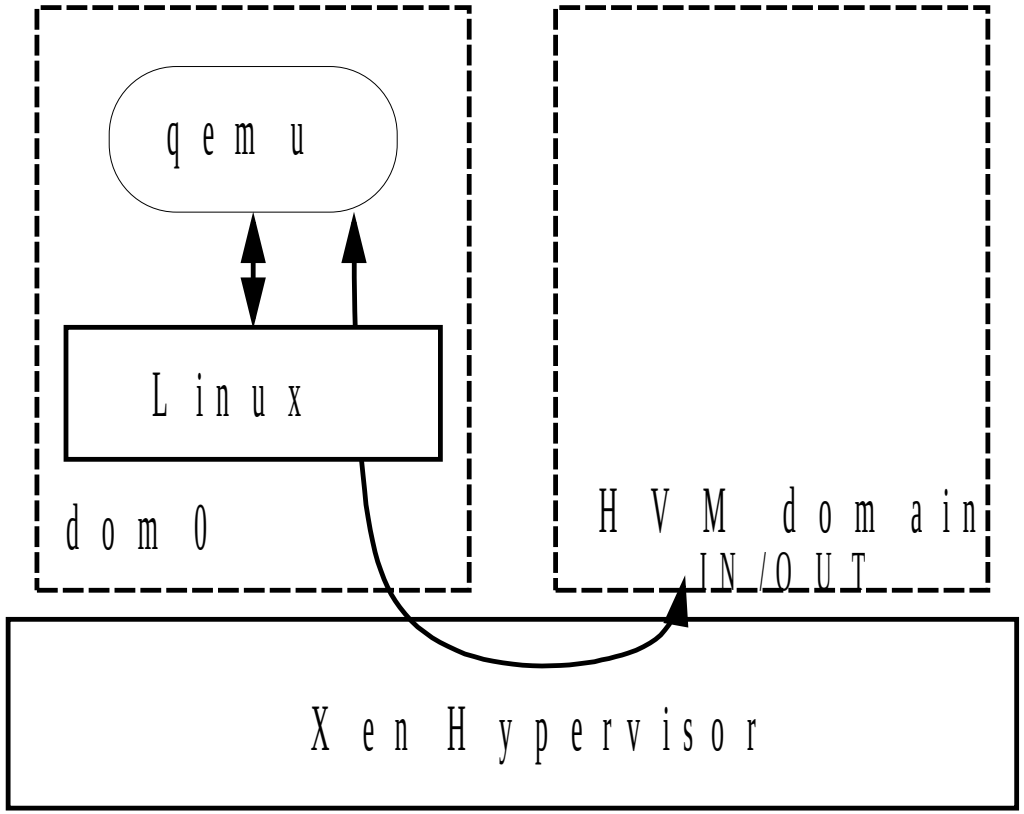
New Mini-OS Features

- Disk frontend
- FrameBuffer frontend
- FileSystem frontend
  - Imported from JavaGuest
  - Remote access to some /export (e.g. of dom0)
- More advanced MM
  - Read-Only memory
  - CoW for zeroed pages
- But still keep it simple
  - Single address space, mono-VCPU, no preemption
- Bugfixes!

stubdom/

- **Makefile**
  - Download and compile a cross-compilation environment
    - binutils, gcc, newlib, lwip
- **c/**
  - 'Hello World!' C application
- **caml/**
  - 'Hello World!' Caml application
- **README**
  - Of course :)

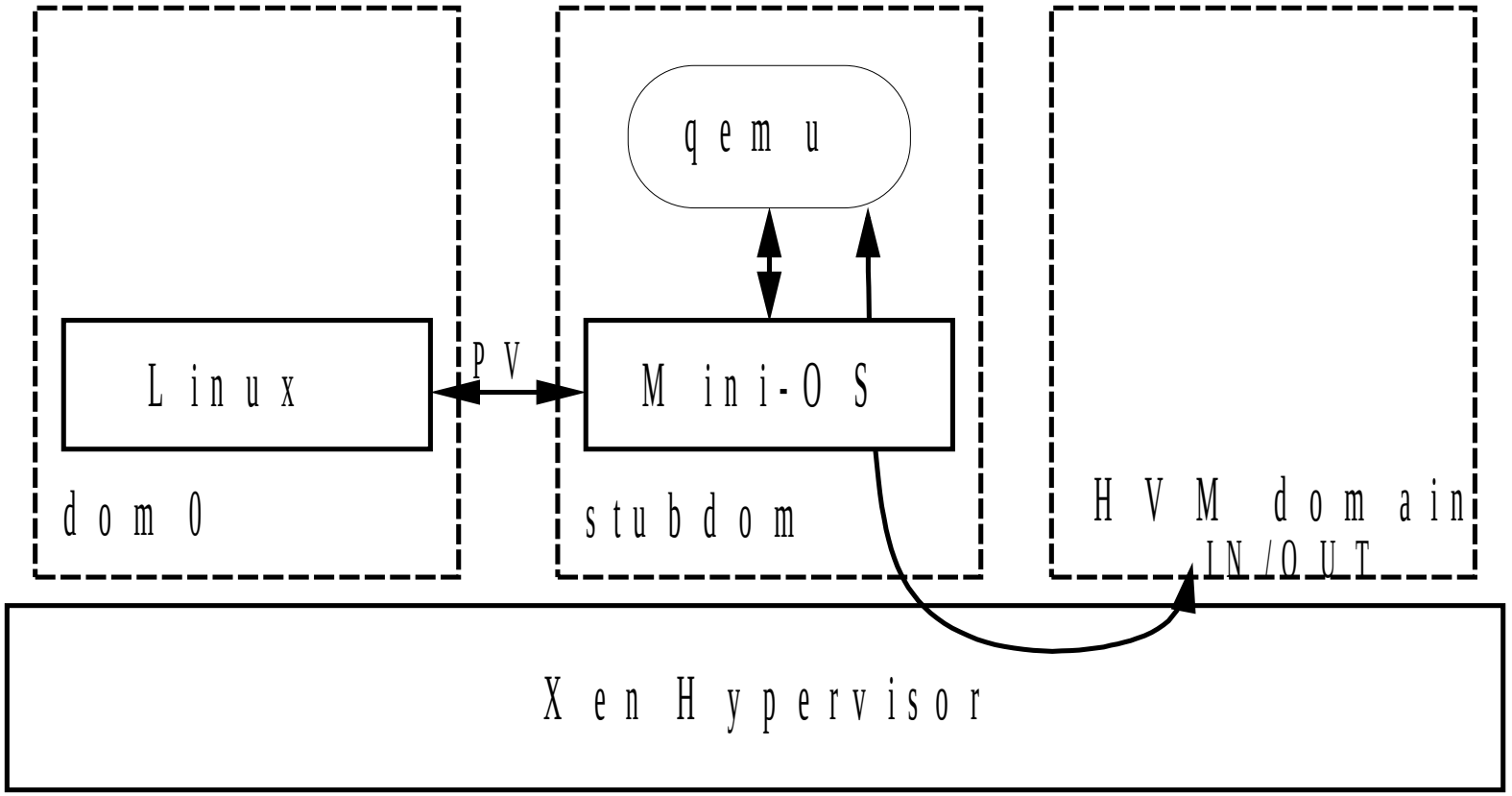
Current HVM device model



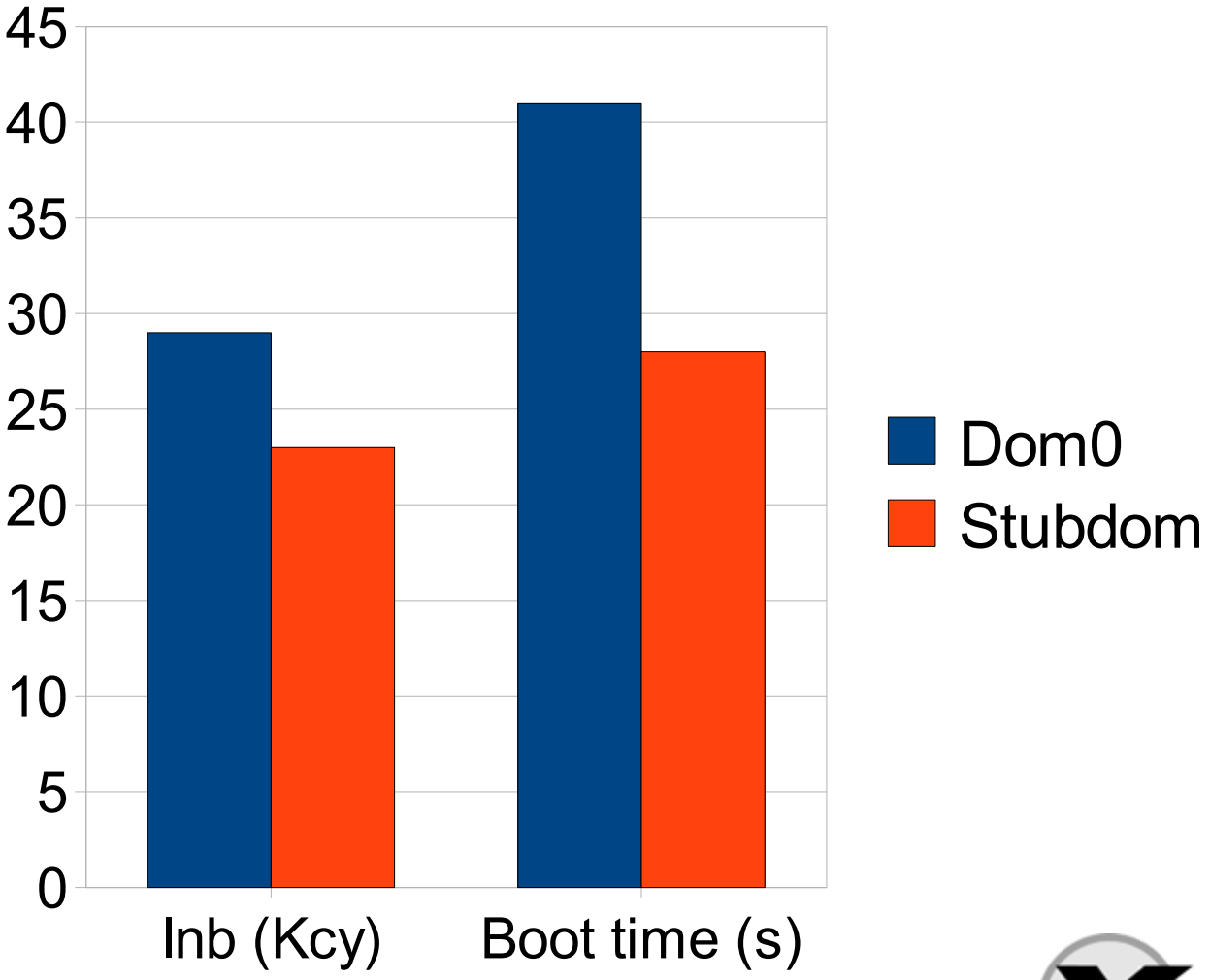
Current HVM dm

- Not always responsive
  - Have to wait for dom0 Linux to schedule qemu
- Eats dom0 CPU time
- Uses dom0 resources from userland
  - Disk, tap network
  - Hence runs as root

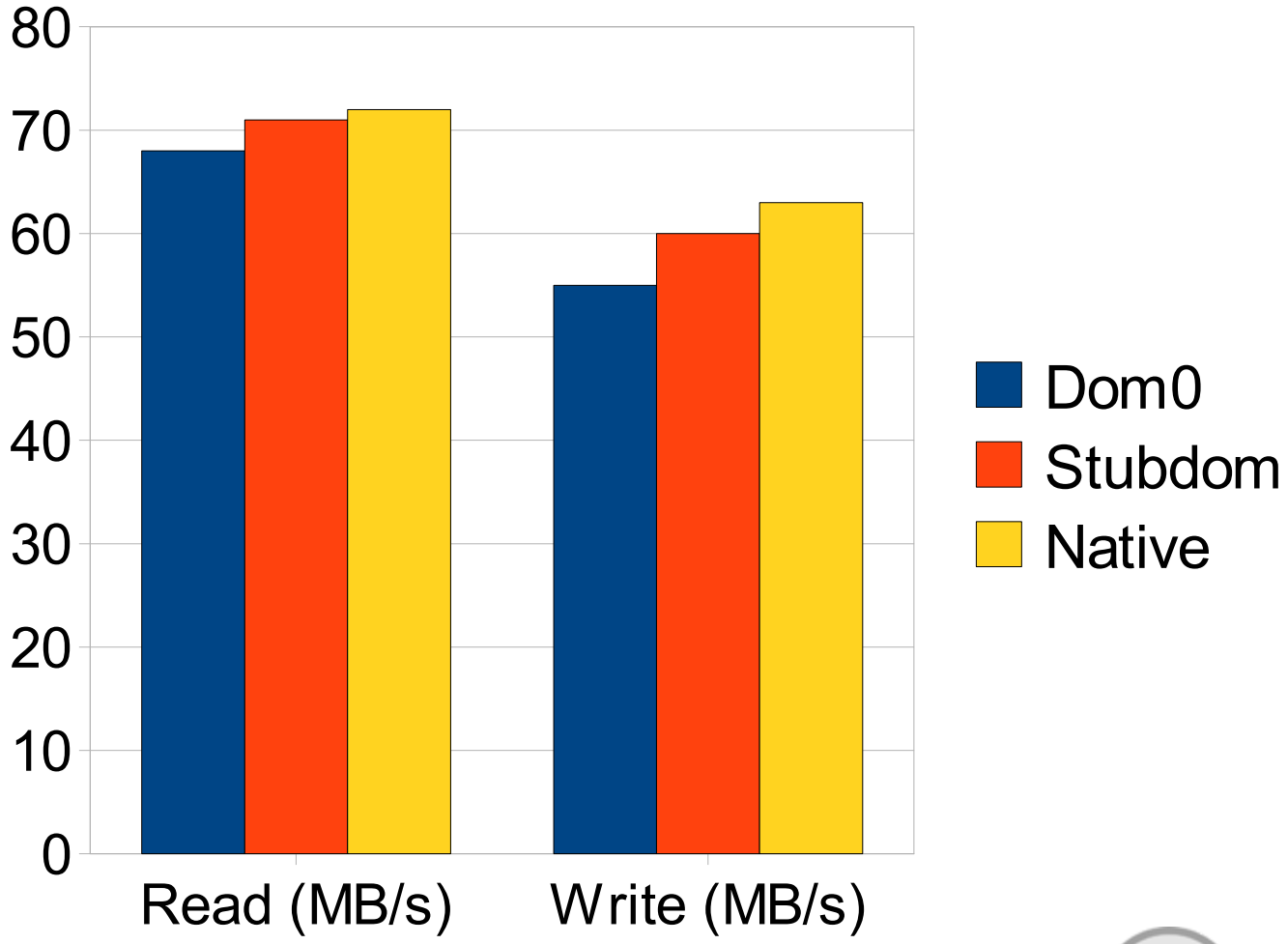
HVM dm domain



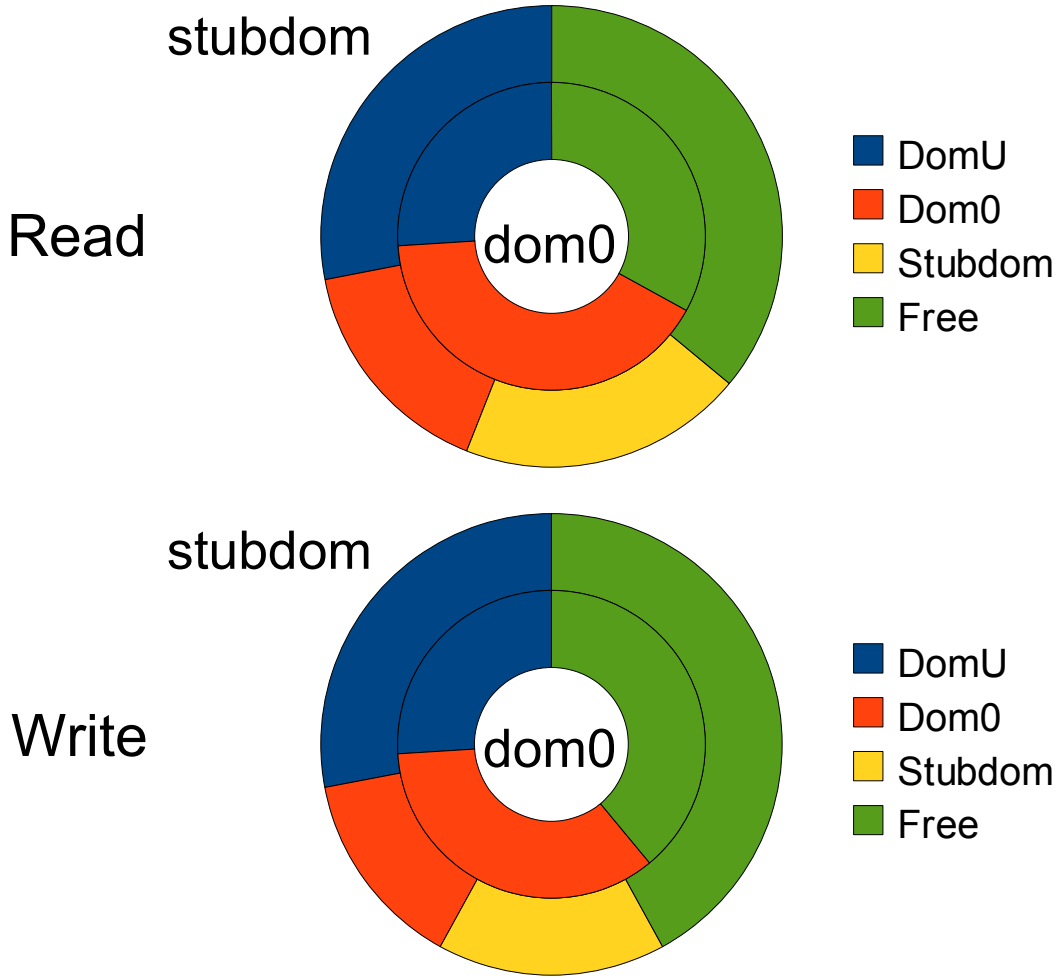
HVM dm domain



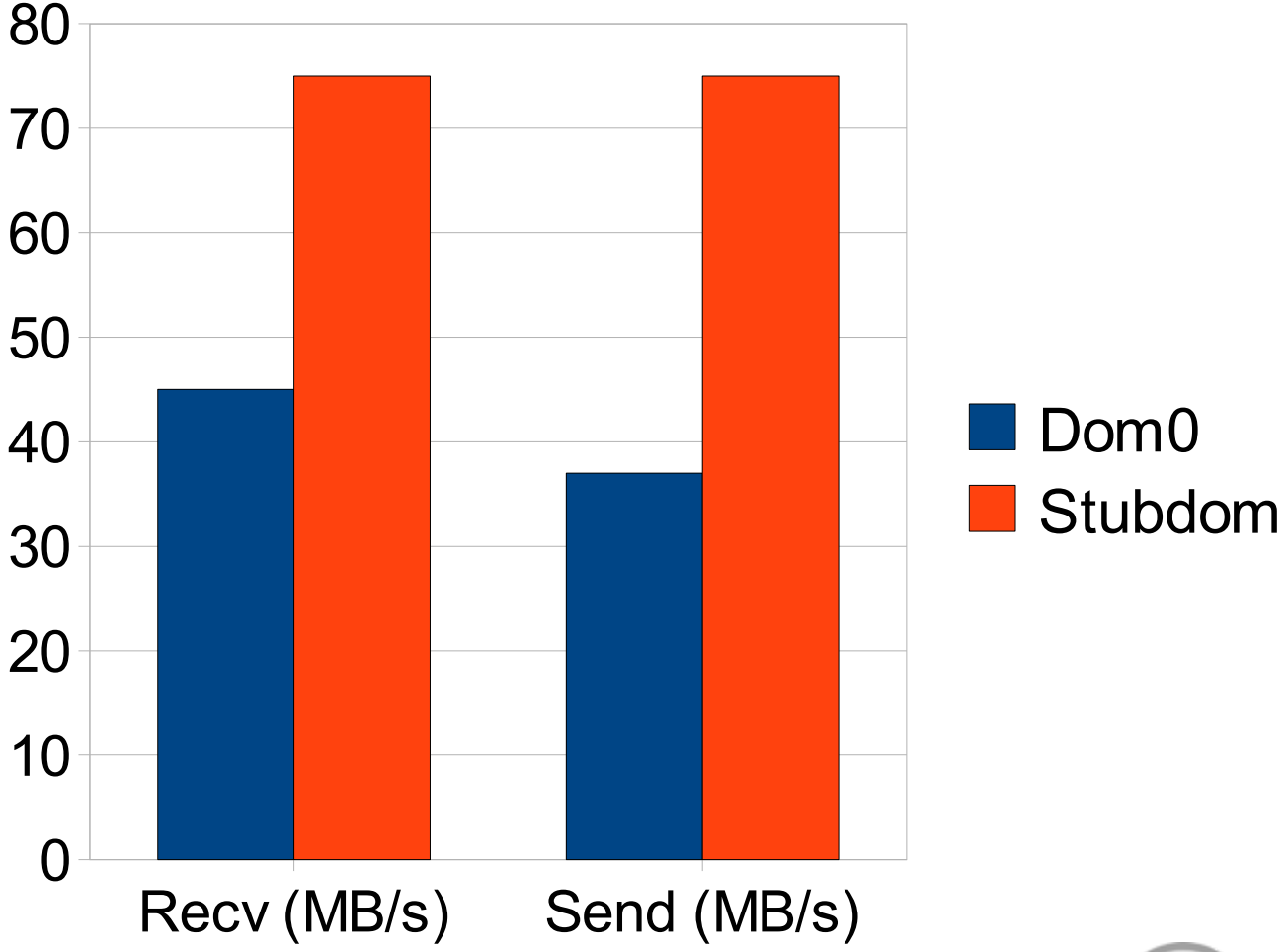
HVM dm domain Disk Perfs



HVM dm domain Disk CPU%

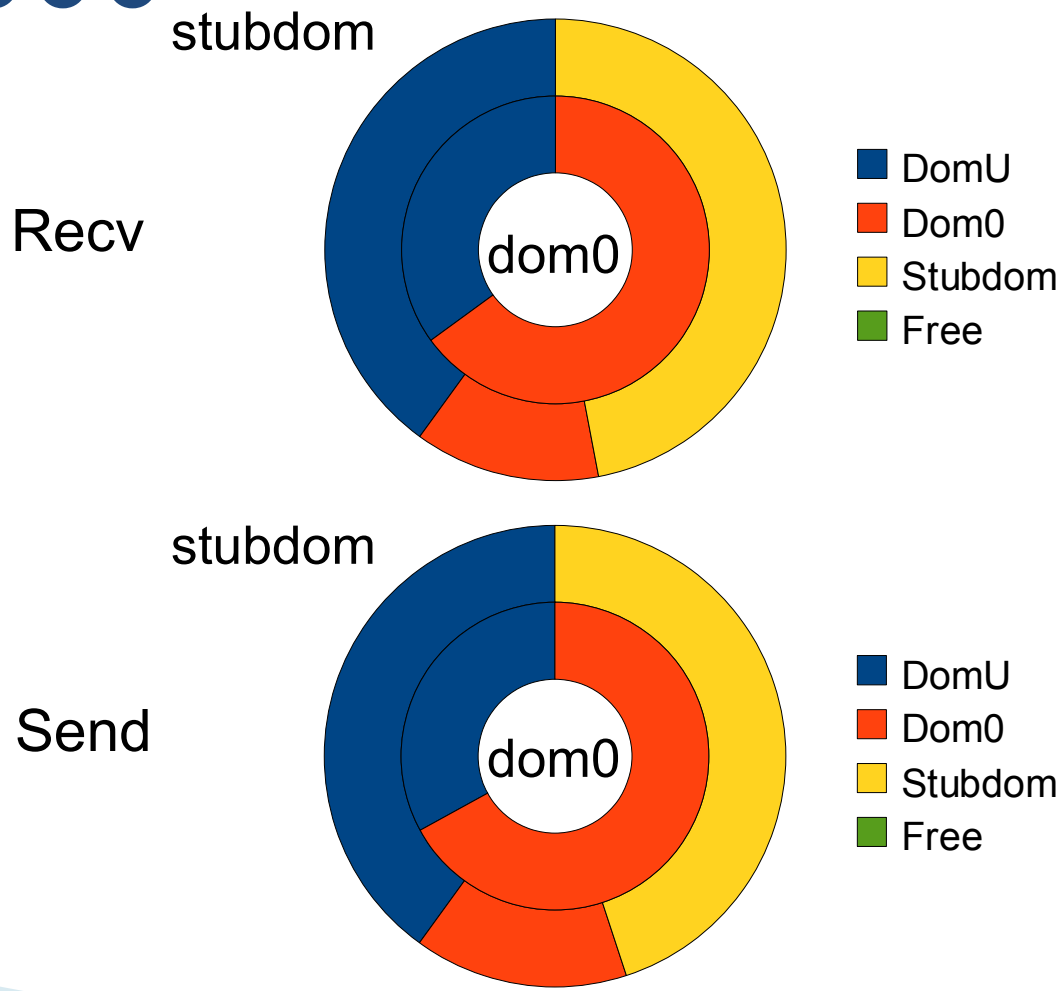


HVM dm domain Net Perfs e1000



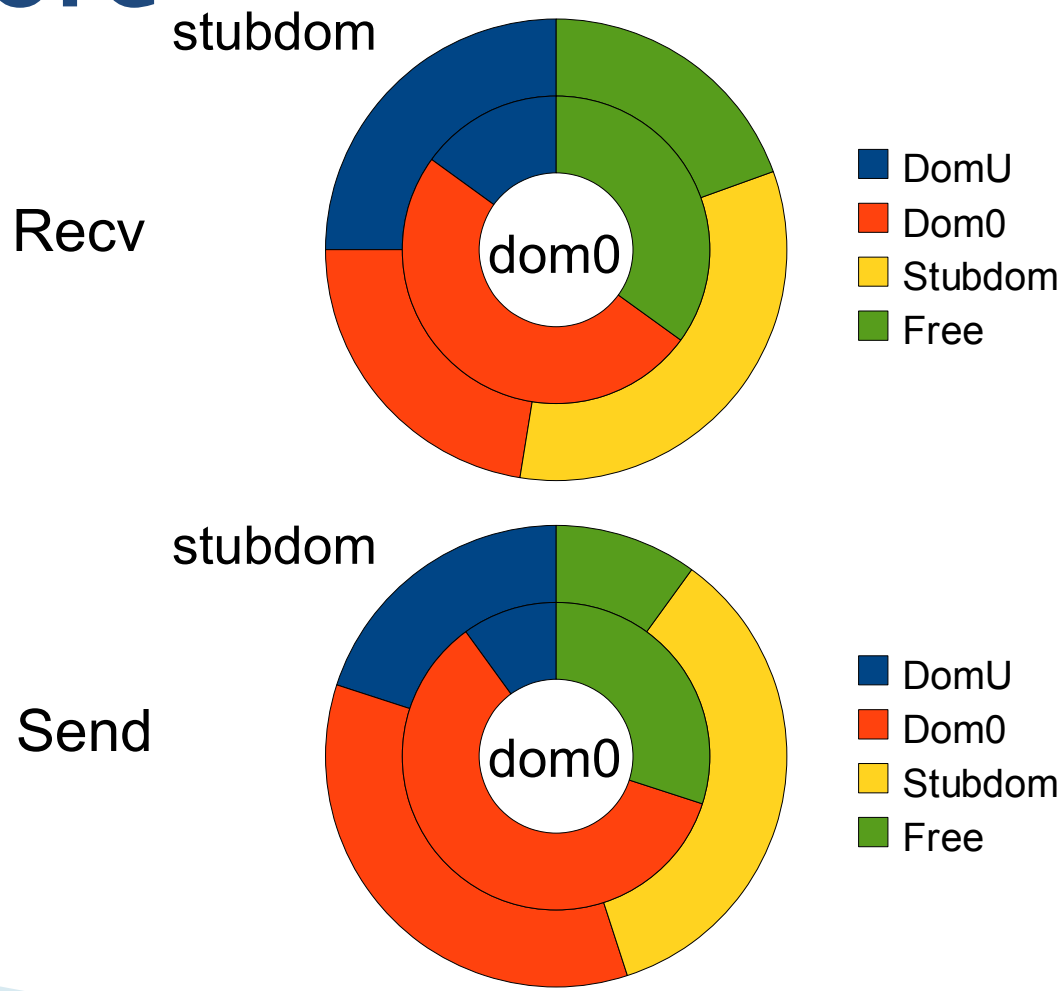
HVM dm domain Net CPU%

e1000



HVM dm domain Net Perfs bicore

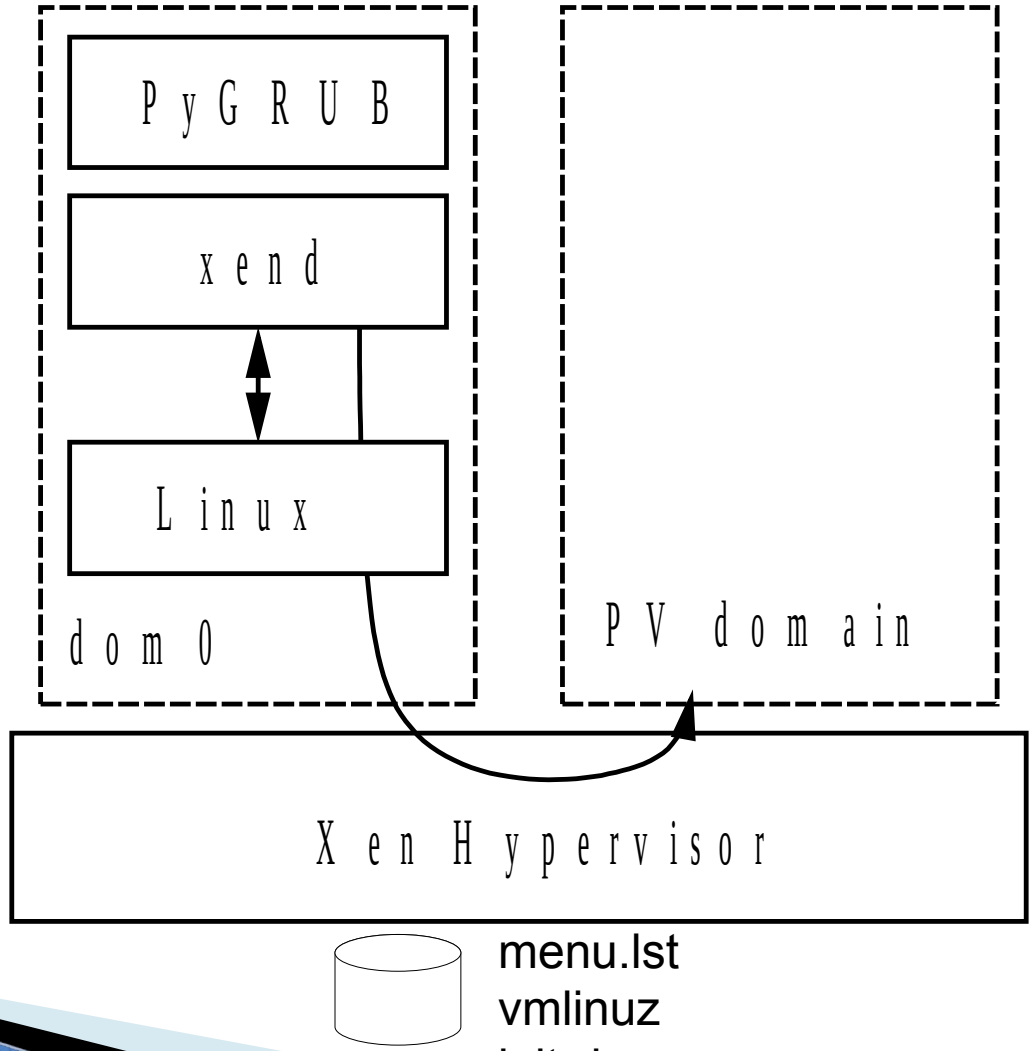
HVM dm domain Net CPU% bicore



HVM dm domain

- Almost unmodified qemu
  - Disable e.g. sound support, plug Mini-OS PV drivers
- Relieves dom0
- Provides better CPU usage accounting
  - Can charge HVM domain with dm domain time
- A lot safer
  - Only privilege is having the HVM dom as **target**
  - Uses same resource access as PV guests
- More efficient
  - Let the hypervisor schedule it directly
  - More lightweight OS

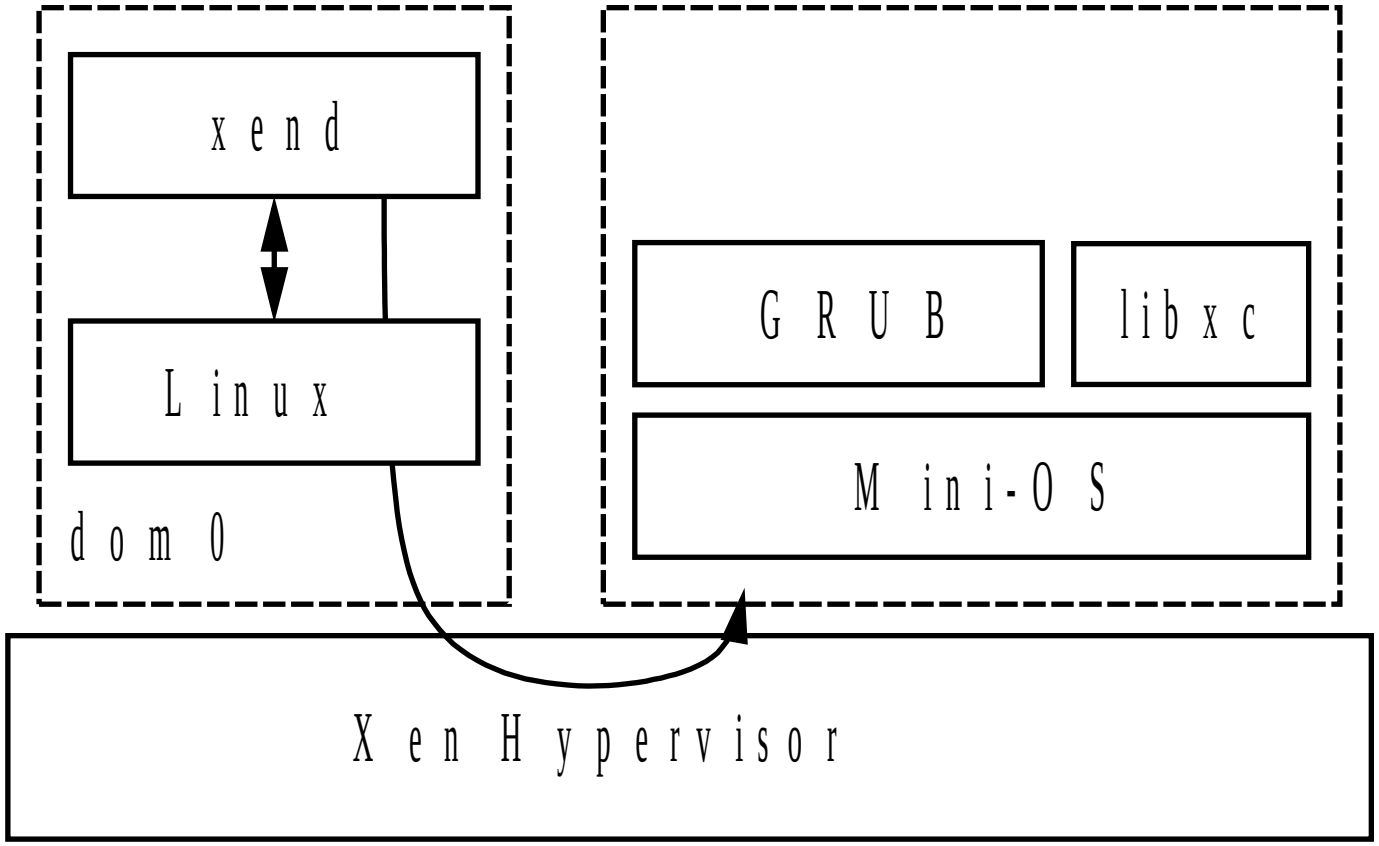
PyGRUB


PyGRUB
xend
Linux
dom 0
PV domain
Xen Hypervisor
menu.lst
vmlinuz
initrd

PyGRUB

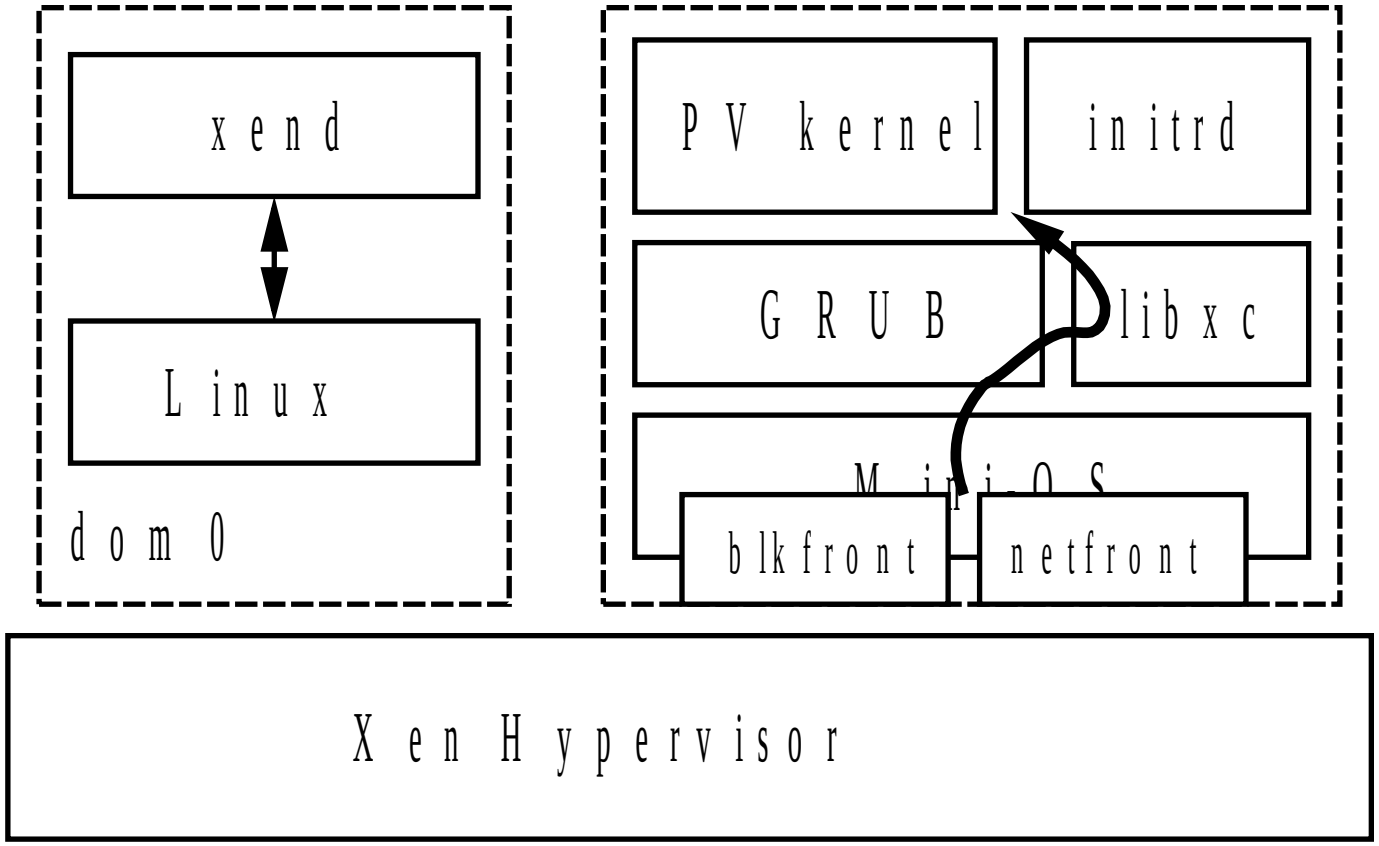
- Needs to be root to access guest disk
  - Security issues
- Does not currently provide network boot
- Reimplements GRUB

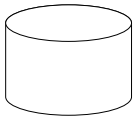
PV-GRUB start



menu.lst
vmlinuz
initrd

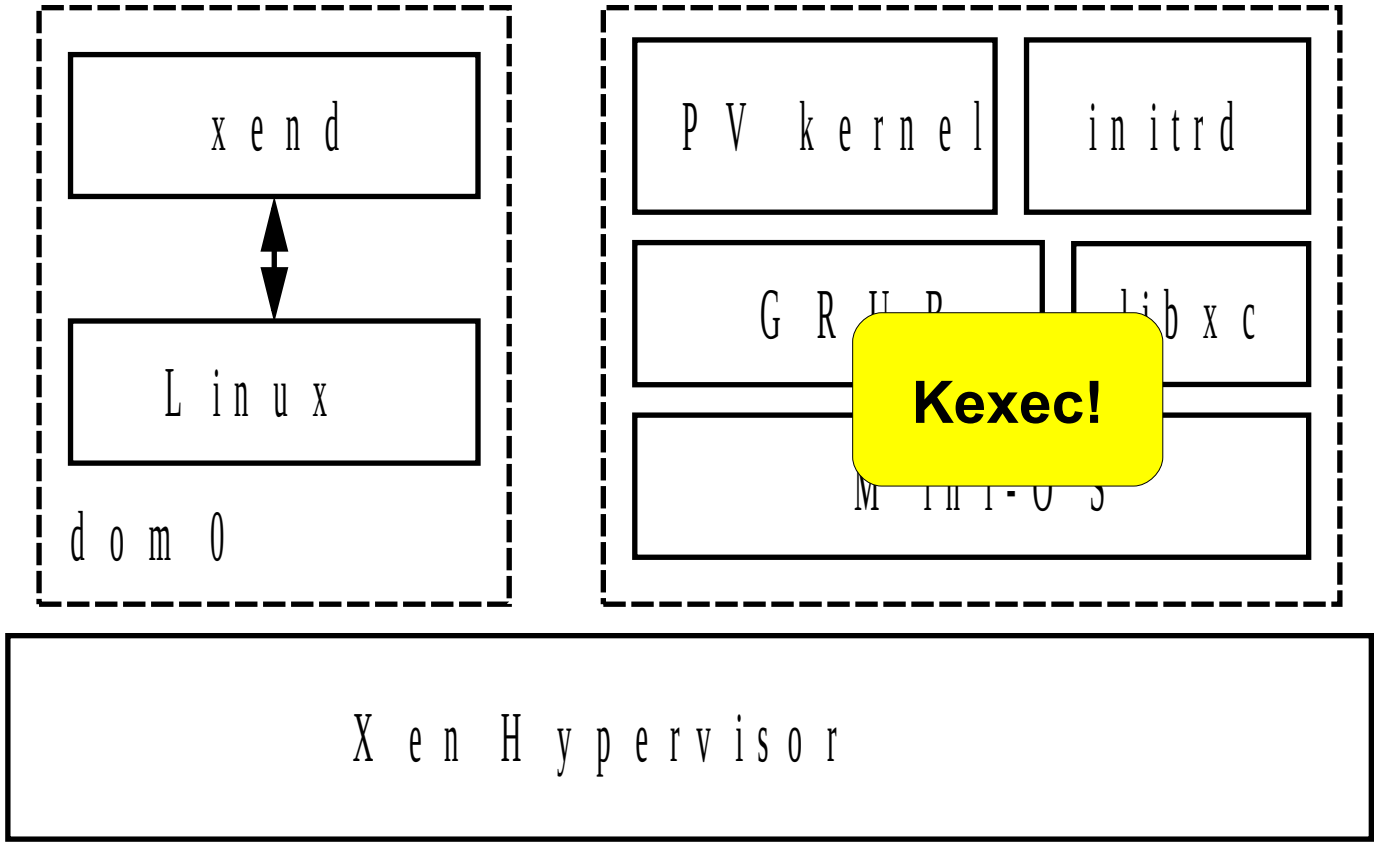
PV-GRUB loading



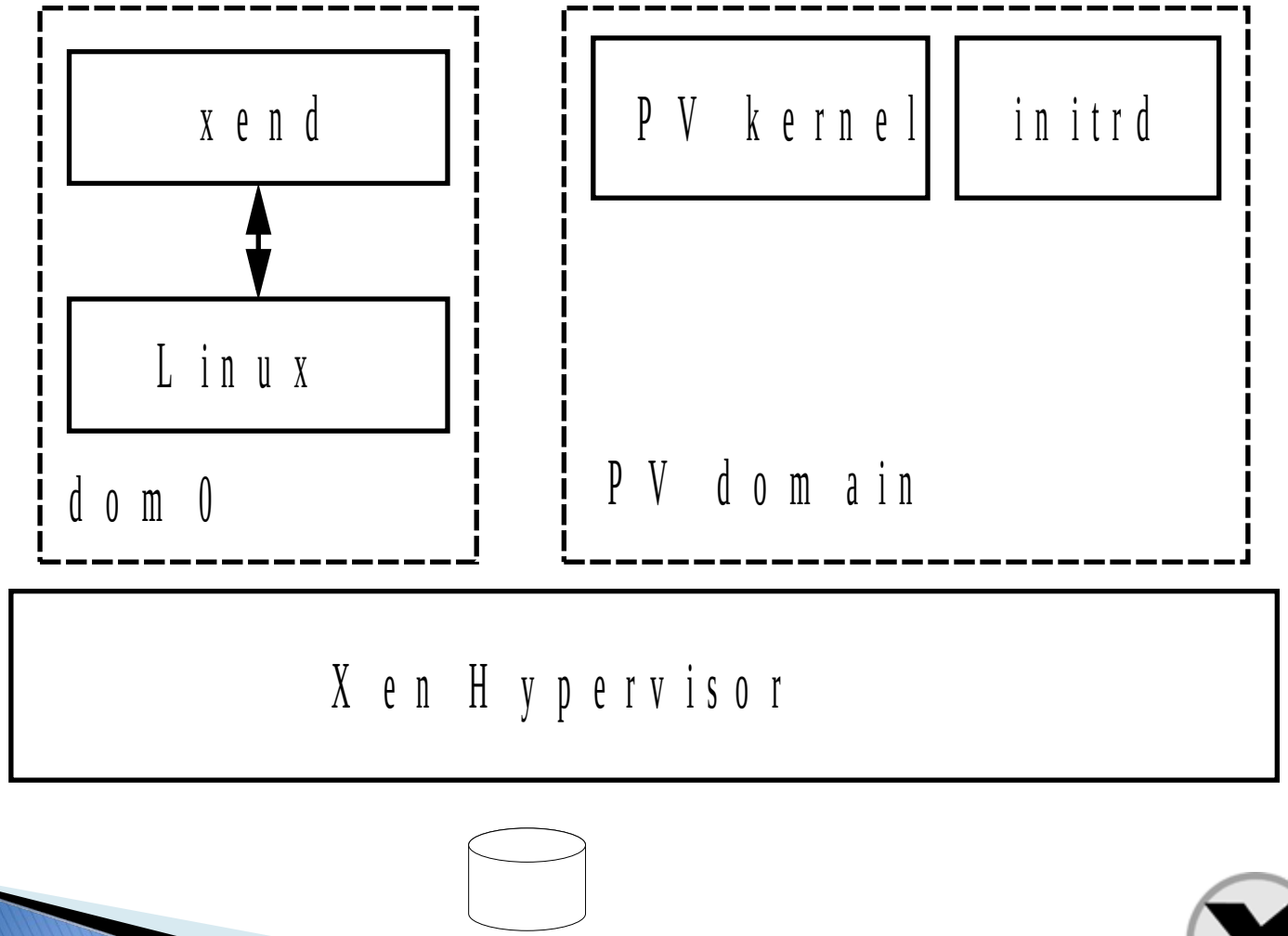


menu.lst
vmlinuz
initrd

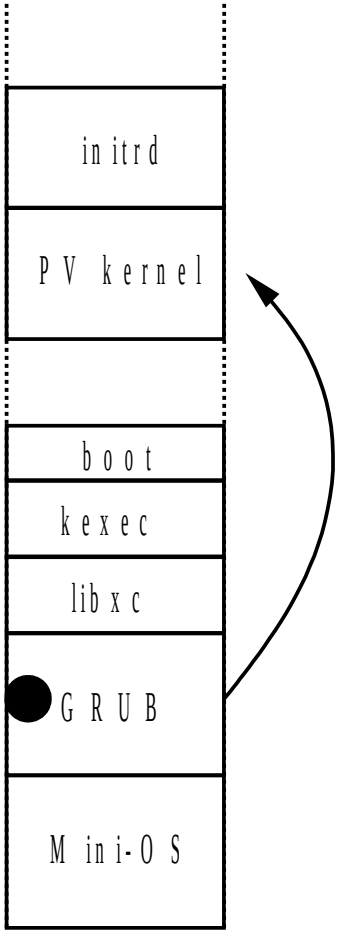
PV-GRUB loaded



PV-GRUB

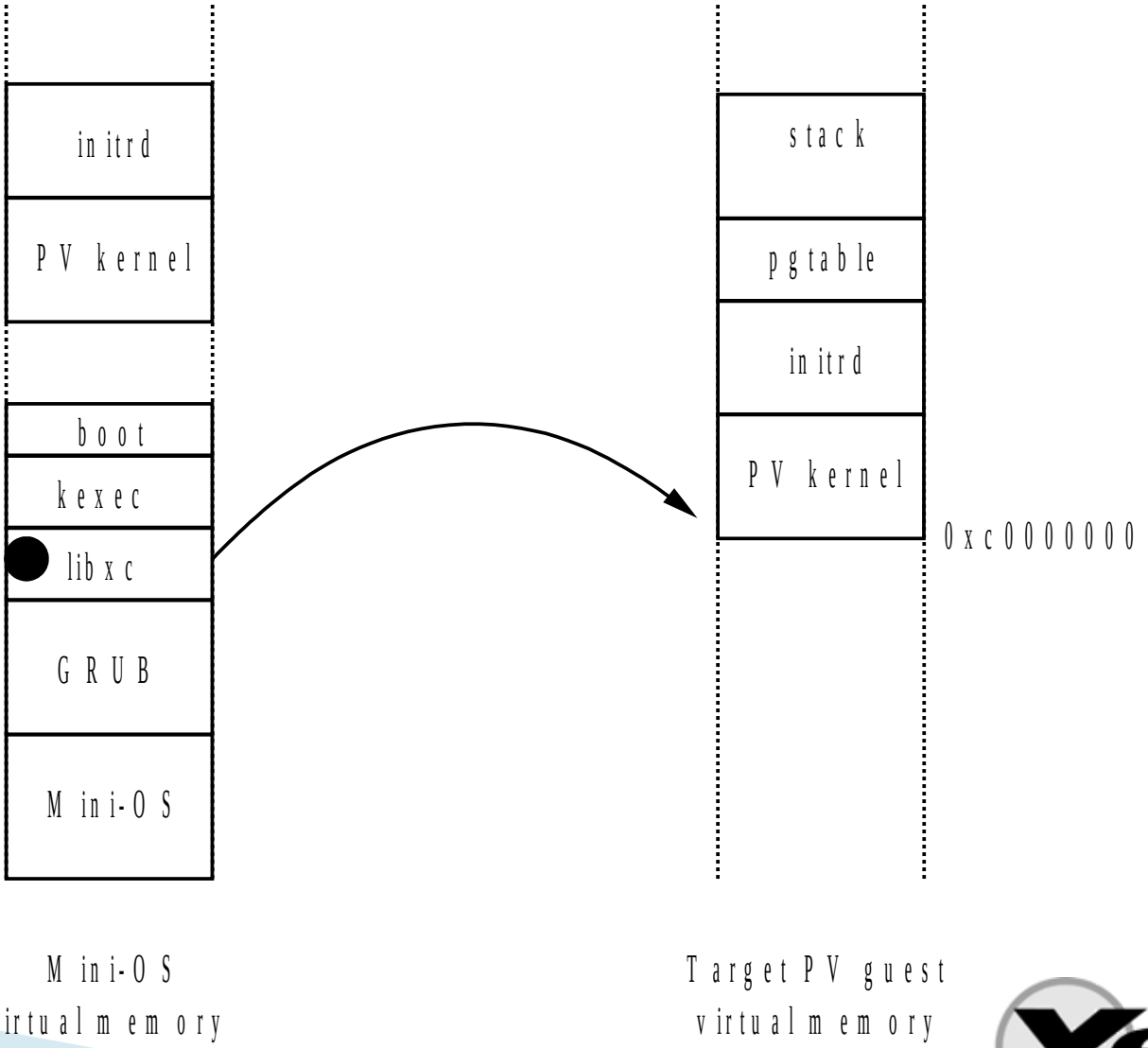


PV-kexec

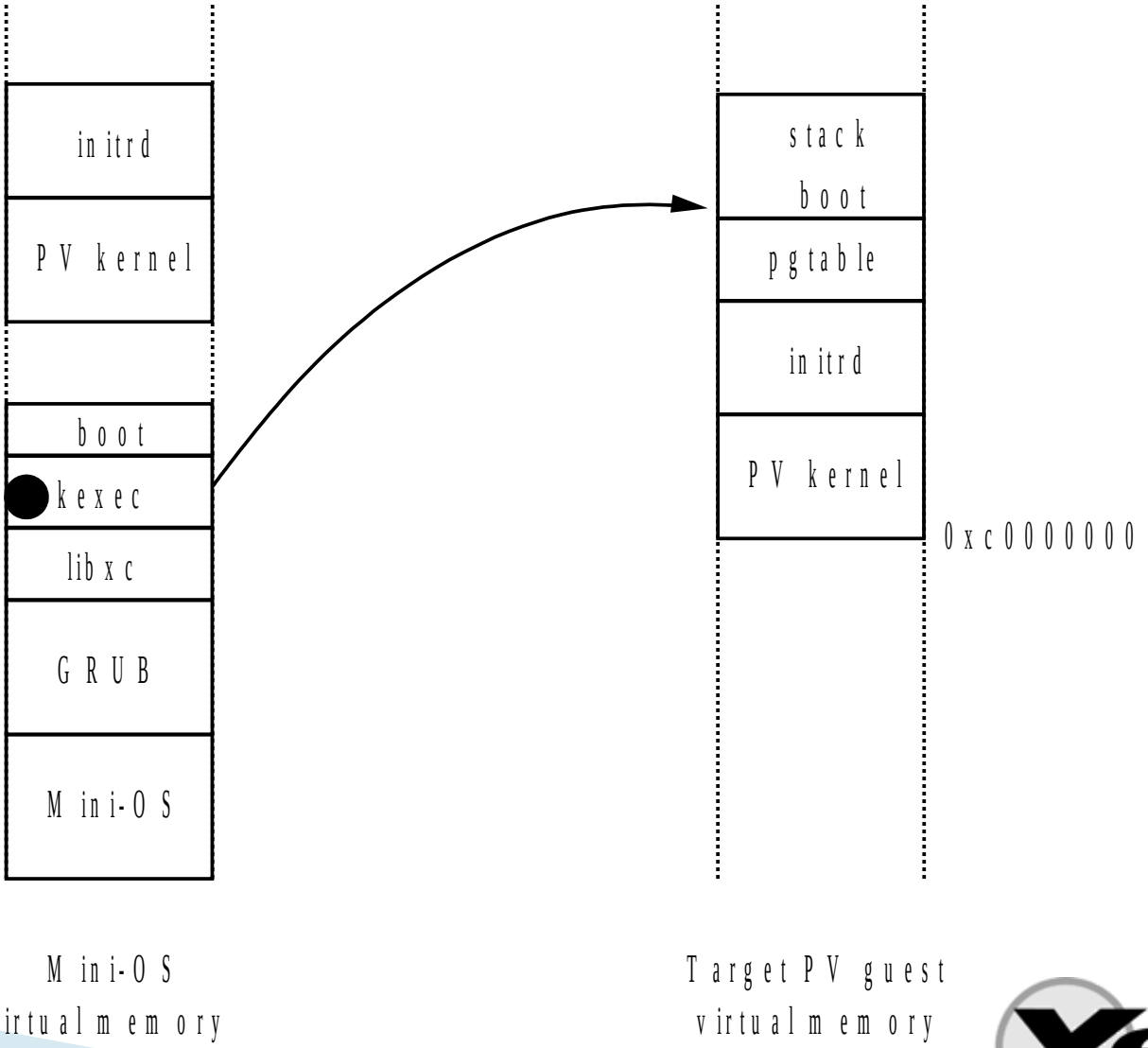


Mini-OS
virtual memory

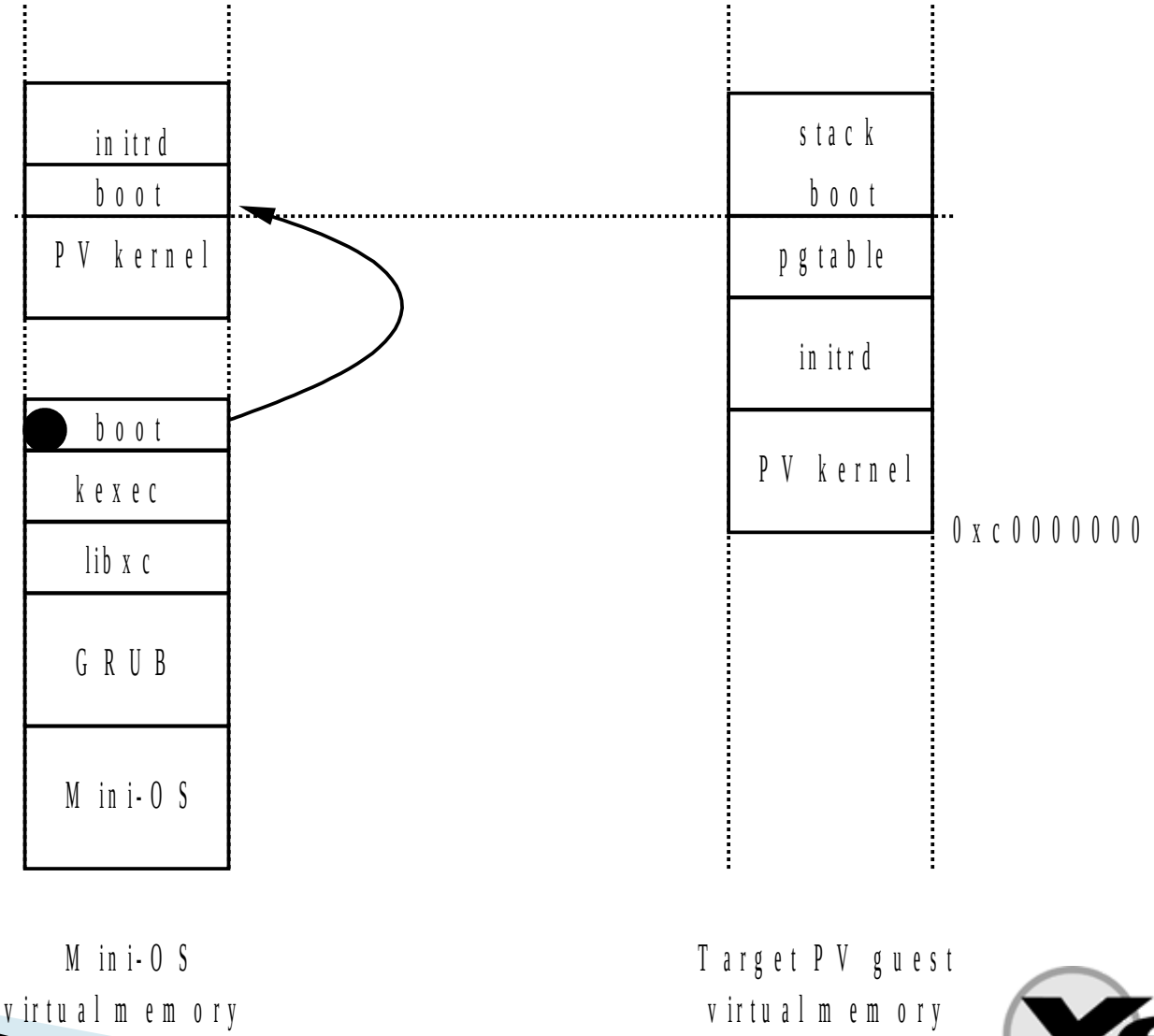
PV-kexec


initrd
PV kernel
boot
kexec
libxc
GRUB
Mini-OS
stack
pgtable
initrd
PV kernel
0xc0000000
Mini-OS
virtual memory
Target PV guest
virtual memory

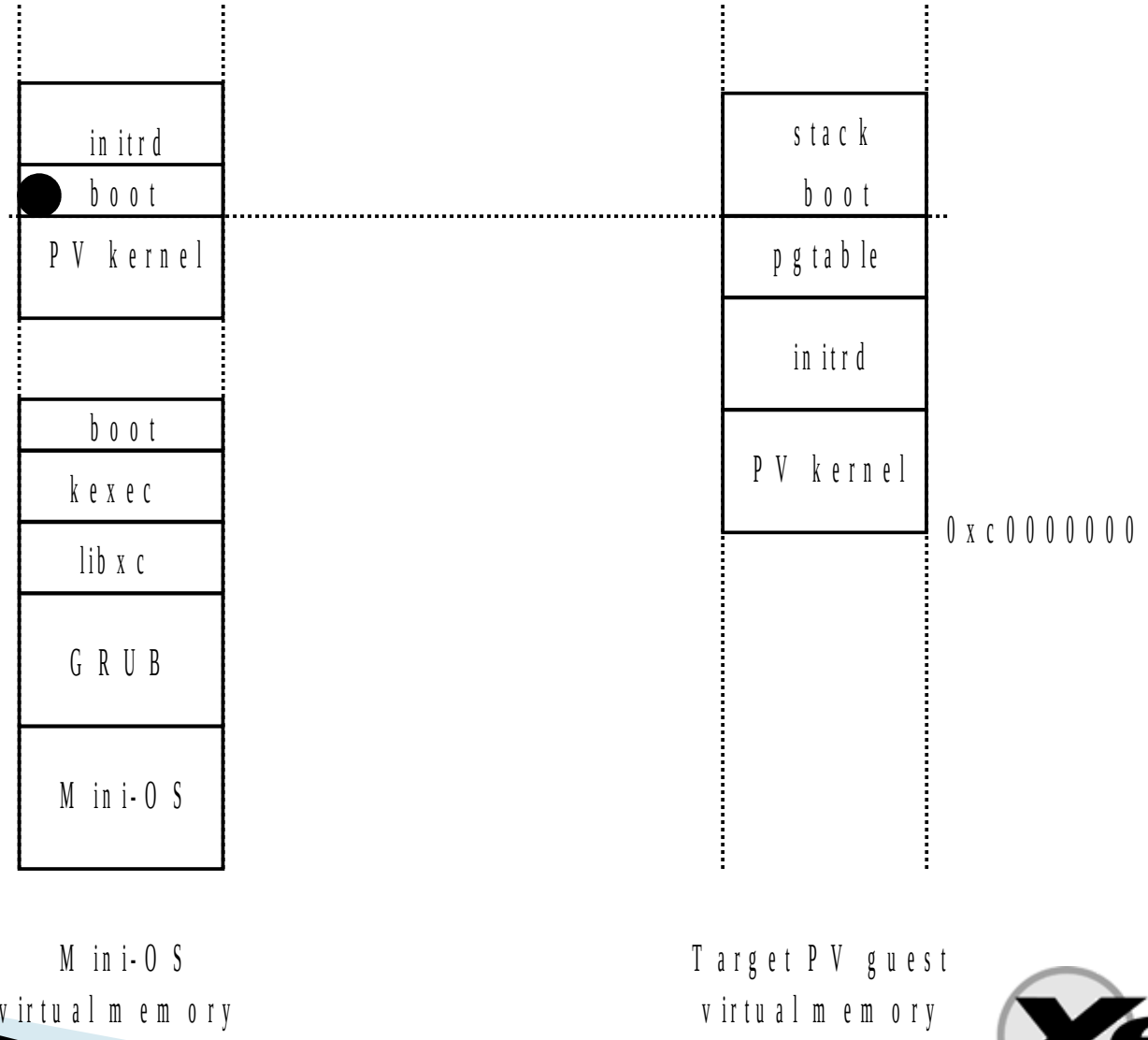
PV-kexec



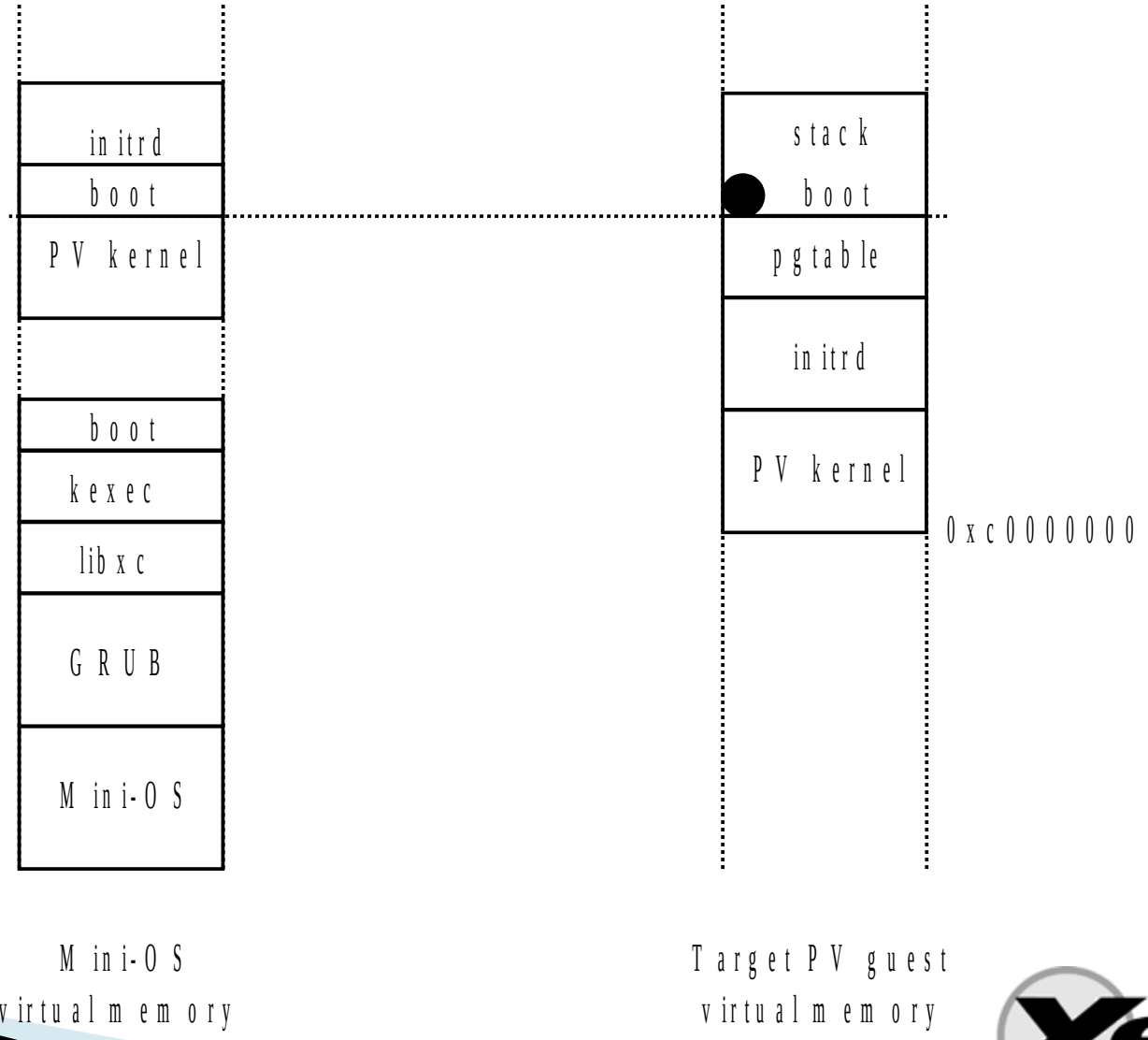
PV-kexec



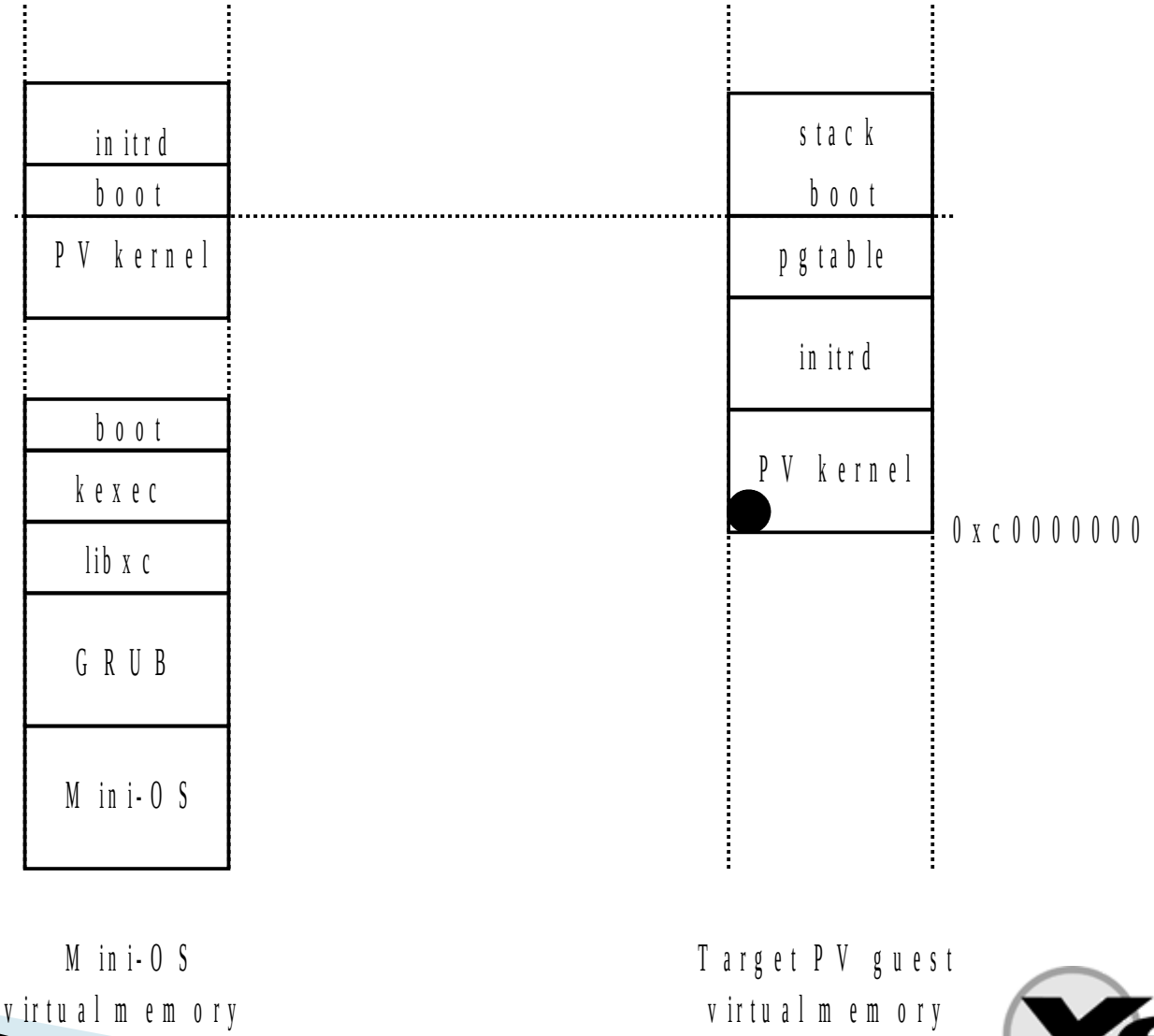
PV-kexec


initrd
boot
PV kernel
boot
kexec
libxc
GRUB
Mini-OS
stack
boot
pgtable
initrd
PV kernel
0xc0000000
Mini-OS
virtual memory
Target PV guest
virtual memory

PV-kexec


initrd
boot
PV kernel
boot
kexec
libxc
GRUB
Mini-OS
stack
boot
pgtable
initrd
PV kernel
0xc0000000
Mini-OS
virtual memory
Target PV guest
virtual memory

PV-kexec


initrd
boot
PV kernel
boot
kexec
libxc
GRUB
Mini-OS
stack
boot
pgtable
initrd
PV kernel
0xc0000000
Mini-OS
virtual memory
Target PV guest
virtual memory

PV-GRUB

- Executes upstream GRUB
  - Replace native drivers with Mini-OS drivers
  - Add PV kexec implementation
- Just uses the target PV guest resources
- Supports network
- Supports graphical menu

Conclusion

- Dm domain
  - Improves security
  - Improves accounting
  - Improves scalability
  - Improves performances
- PV-GRUB
  - Improves security
  - Provides network boot
- Mini-OS also being tested at Cisco for IOS
- Available in the unstable tree

Future Work

- Dm domain
  - *Live* migration, PCI PT
  - IA-64 support
  - Group scheduling with HVM domain
- PV-GRUB
  - Kexec 64bit guest from 32bit PV-GRUB
  - PVFB shutdown/restart
- OCaml support
  - 'Hello World!' works
  - Needs runtime rebuild to properly hook into POSIX layer


Xen
Summit™