Xen – using virtualisation techniques in a Grid environment

Dr.Rüdiger Berlich,
Marcus Hardt, Dr. Marcel Kunze
Forschungszentrum Karlsruhe GmbH

www.eu-egee.org

What is Virtualisation ?

- Inspired by mainframes
- Allows you to partition your hardware
- Let's you run more than one OS concurrently

Virtualisation: Some Use Cases

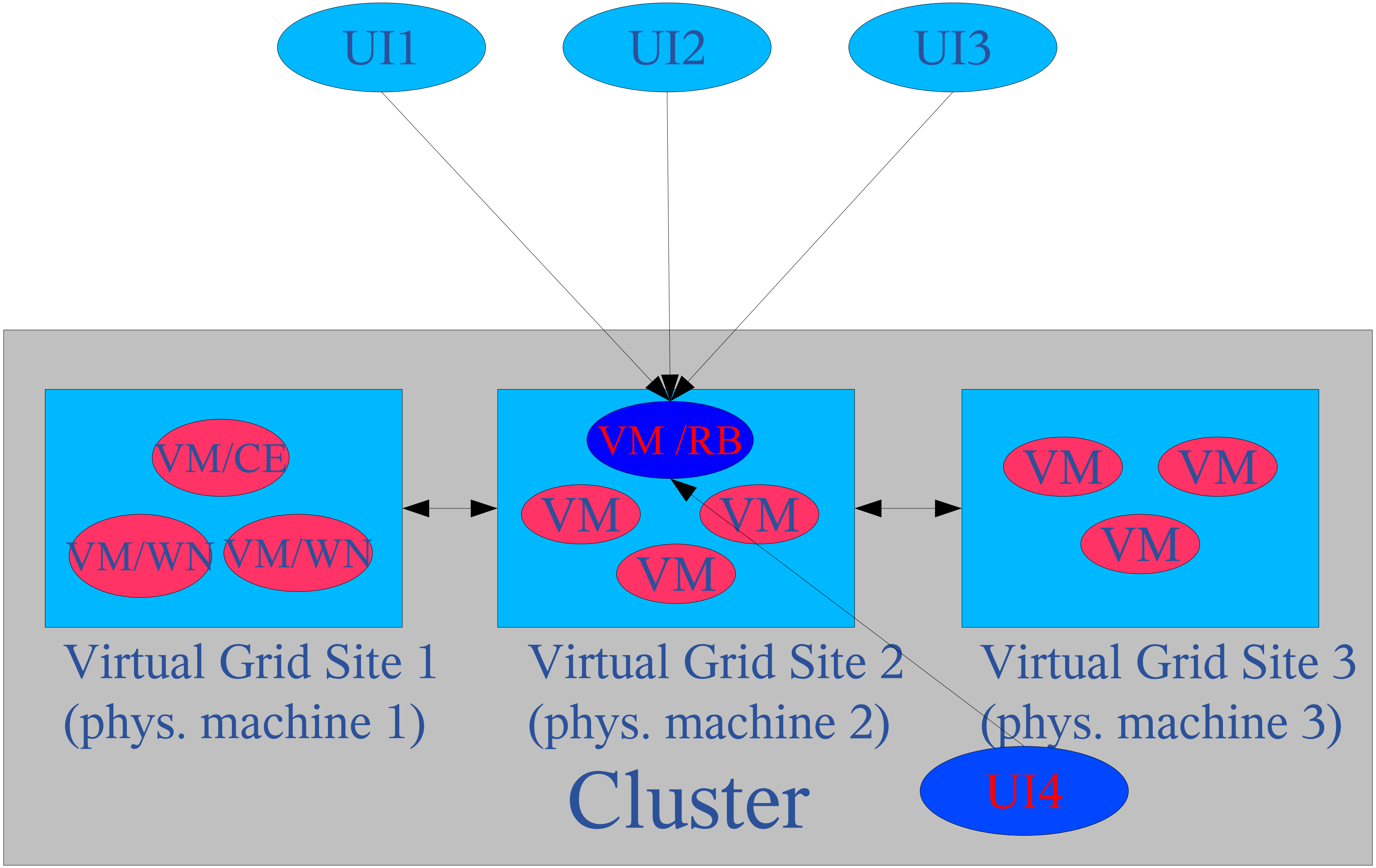
- Consolidation of different services on one machine
  - Compute Center, ISP, Webhoster, …
  - Sandbox – secure environments
- More efficient usage of resources
  - Cluster, farms
- Additional layer of abstraction

  -> taylored OS environment
- “Grid in a box”

Grid in a Box: The Vision

- First outlined in paper “A single-computer Grid gateway using virtual machines” by Univ. Dublin
- Basically: Server-Konsolidation using virtual machines (Install Server, CE, SE, UI)
- When thinking the thought further: Build an entire Grid in a cluster, running multiple virtual machines, provide easy access from private machines.
- Biggest advantage: In environments where performance is not the biggest concern, one can multiply the available ressources
- See also http://public.eu-egee.org/files/xen-grid-in-a-box-fzk.pdf

Grid in a Box: Components



Grid in a Box: Characteristics

Advantages:

- Allows the creation of virtual Grids that are for the user indistinguishable (except for performance) from a “real” Grid (at least in theory ...)
- Do this with a fraction of the typical ressources
- Take down or break single ressources
- Experimentation in a safe environment

Ship a whole Grid as disk images to a customer

Disadvantages:

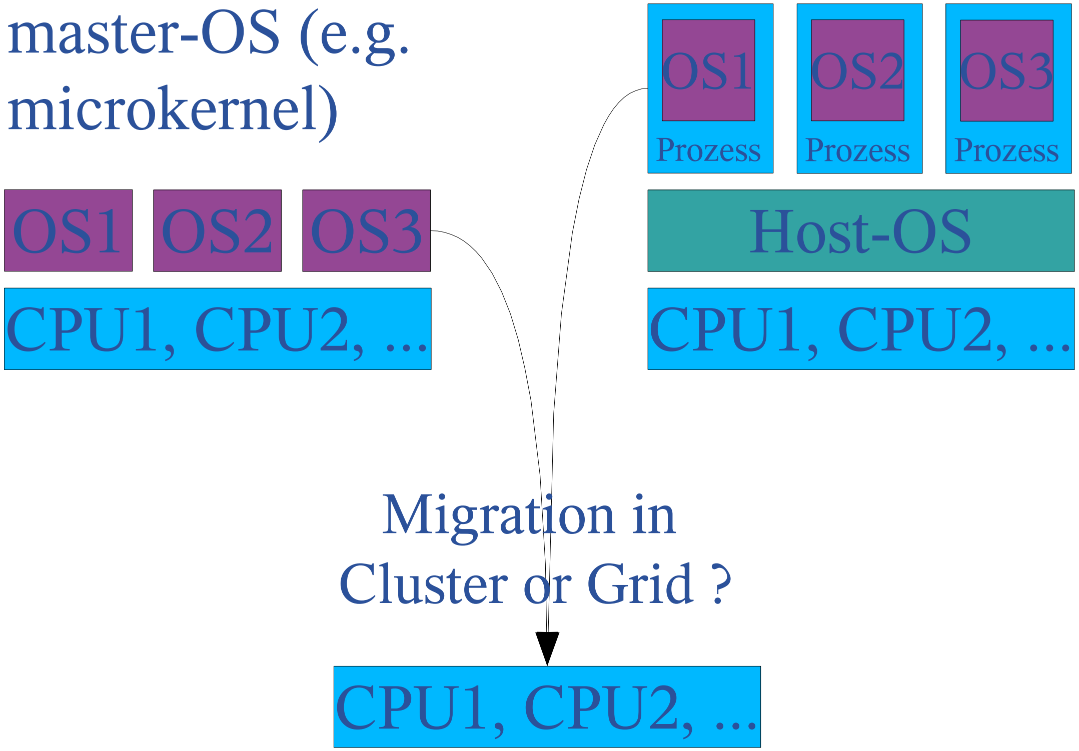
- Stability; Maintenance
- Single point of failure
- Not the “real thing”
- Easier said than done ...

Techniques and Products (OS)

Virtualisierung with hardware or specialised master-OS (e.g. microkernel)

Guest-OS is a process: higher overhead, but easier to implement

IBM zSeries
XEN
ESX Server



Migration in Cluster or Grid ?

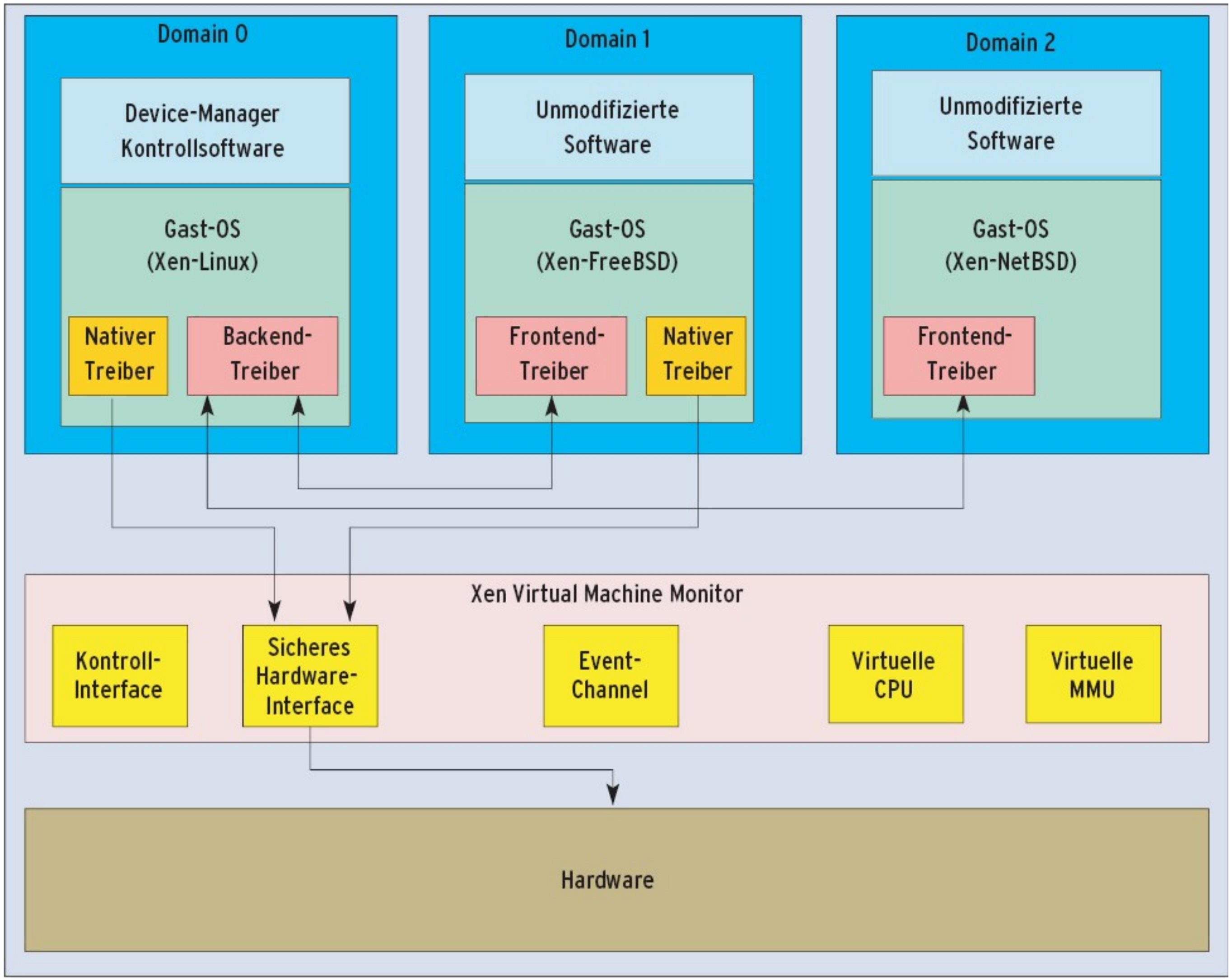
VMWare Workst.
GSX Server
Usermode Linux
Win4Lin
Bochs
qemu (many different processors)
Virtual PC

Xen: History

- Approx. 2 years old
- Started by the *Systems Research Group* of the University of Cambridge, UK
- Originally part of the Xenoserver project, which aims to build a public infrastructure for wide-area distributed computing.
- Idea: Provide a distributed network of OS environments tailored to the user's needs
- Xen is thus closely related to the ideas of Grid Computing !
- Now available in Version 2.07 (3.0 will be released soon !)
- Outlook: Native execution of arbitrary Intel-based OS feasible using hardware virtualisation features (Intel Vanderpool)
- Ports to 64 bit platforms underway (with the help of AMD, Intel, ...)

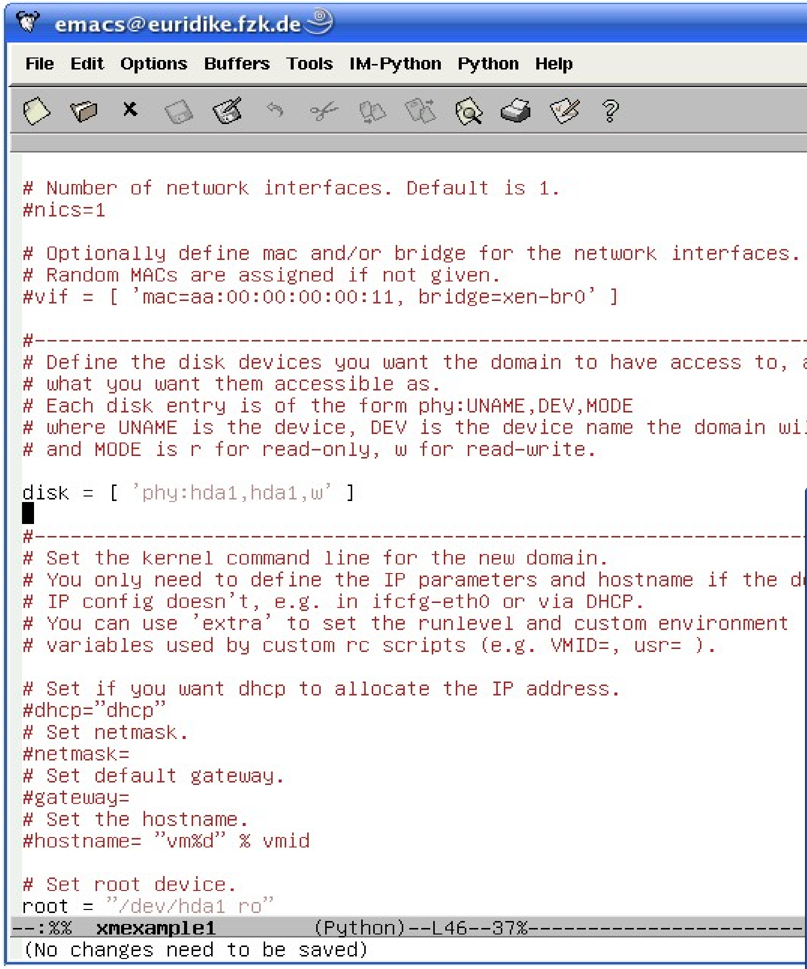
ξένος

Xen: Paravirtualisation



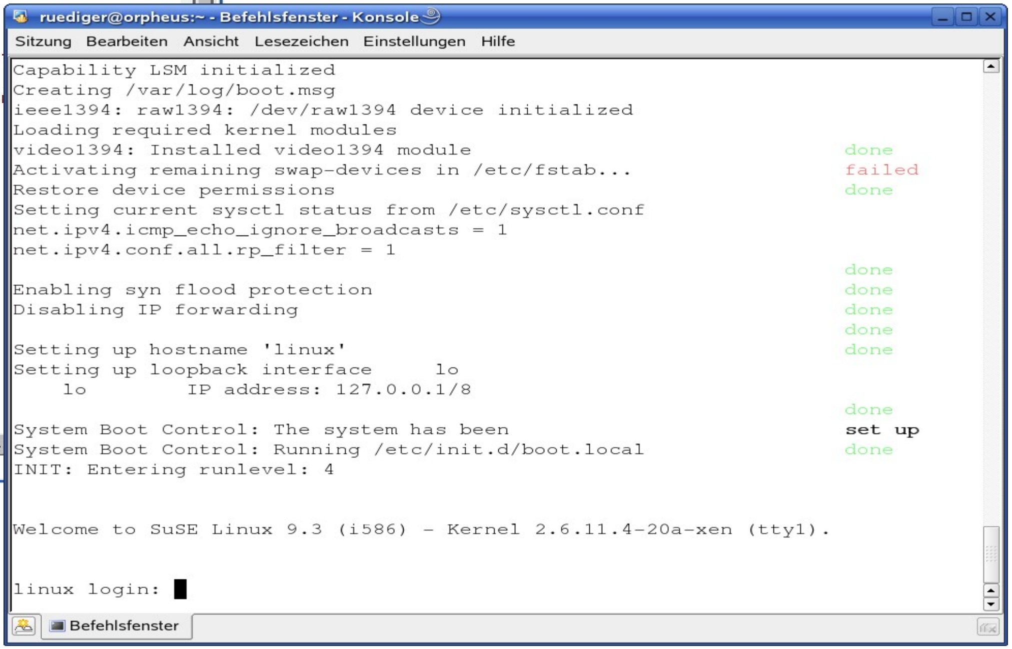
- Priviledged calls are done through dedicated interface in domain 0
- Advantage: Very high performance (low overhead, very little emulation necessary)
- Disadvantage: Guest-OS must be ported to Xen (but not the applications !)
- But: very minor adaptations, in the range of *O*(3000 LOC)

Xen: Configuration, Starting

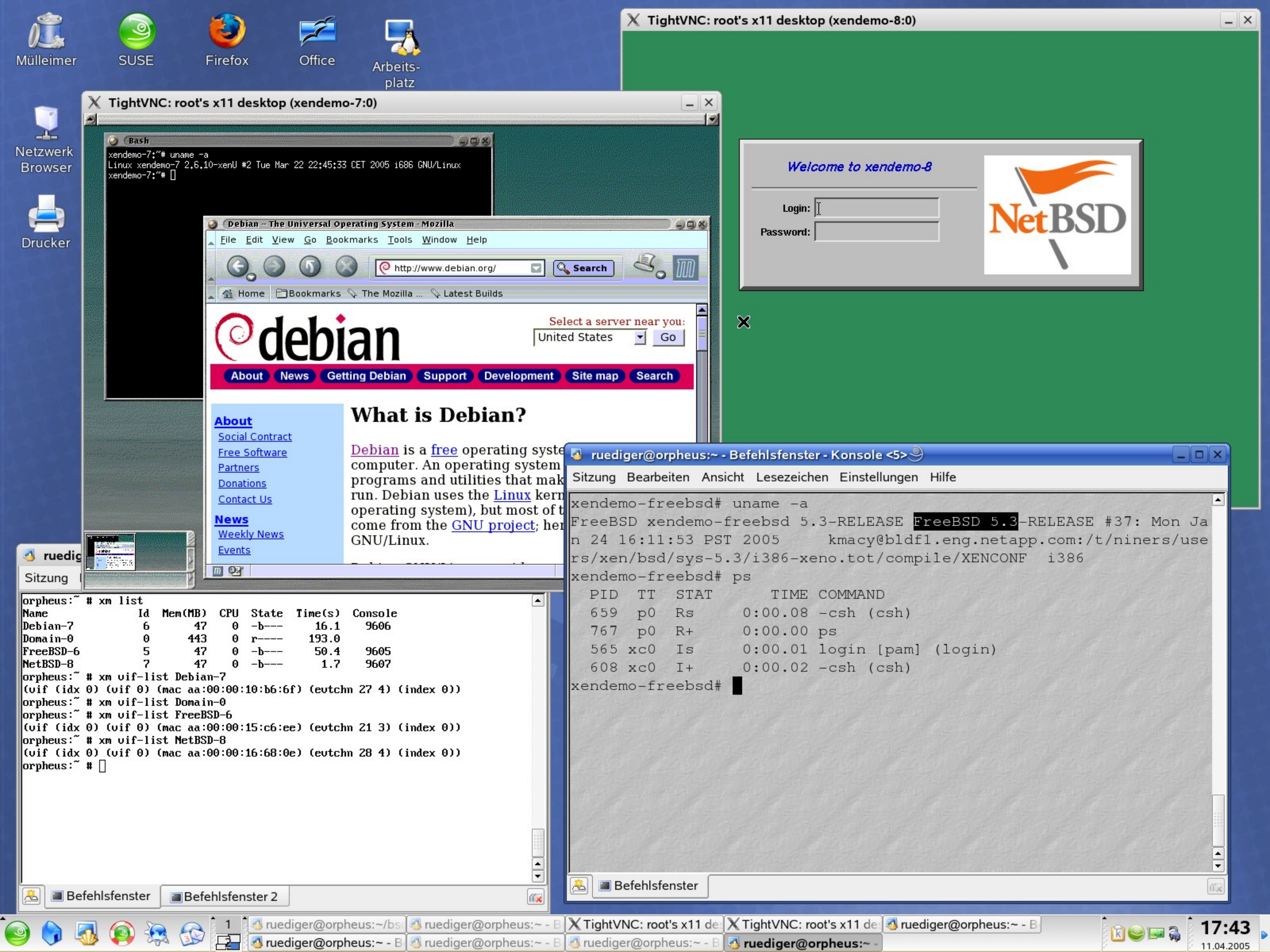


```
# Number of network interfaces. Default is 1.
#nics=1

# Optionally define mac and/or bridge for the network interfaces.
# Random MACs are assigned if not given.
#vif = [ 'mac=aa:00:00:00:00:11, bridge=xen-br0' ]

#----------------------------------------------------------------------------
# Define the disk devices you want the domain to have access to, and
# what you want them accessible as.
# Each disk entry is of the form phy:UNAME,DEV,MODE
# where UNAME is the device, DEV is the device name the domain will see,
# and MODE is r for read-only, w for read-write.

disk = [ 'phy:hda1,hda1,w' ]

#------------------------------------------------------------------
# Set the kernel command line for the new domain.
# You only need to define the IP parameters and hostname if the d
# IP config doesn't, e.g. in ifcfg-eth0 or via DHCP.
# You can use 'extra' to set the runlevel and custom environment
# variables used by custom rc scripts (e.g. VMID=, usr= ).

# Set if you want dhcp to allocate the IP address.
#dhcp="dhcp"
# Set netmask.
#netmask=
# Set default gateway.
#gateway=
# Set the hostname.
#hostname= "vm%d" % vmid

# Set root device.
root = "/dev/hda1 ro"
--:%%  xmexample1         (Python)--L46--37%----------------------
(No changes need to be saved)
```

- Configuration with Python Script
- Starting with the command "xm create -c myconfig"
- Possibility to attach X output, e.g. with VNC
- External IP assigned e.g. via DHCP
- From the outside, domains cannot be distinguished from physical hosts


ruediger@orpheus:~ - Befehlsfenster - Konsole
Sitzung Bearbeiten Ansicht Lesezeichen Einstellungen Hilfe
Befehlsfenster

```
Capability LSM initialized
Creating /var/log/boot.msg
ieee1394: raw1394: /dev/raw1394 device initialized
Loading required kernel modules
video1394: Installed video1394 module                               done
Activating remaining swap-devices in /etc/fstab...                  failed
Restore device permissions                                          done
Setting current sysctl status from /etc/sysctl.conf
net.ipv4.icmp_echo_ignore_broadcasts = 1
net.ipv4.conf.all.rp_filter = 1
                                                                    done
Enabling syn flood protection                                       done
Disabling IP forwarding                                             done
                                                                    done
Setting up hostname 'linux'                                         done
Setting up loopback interface     lo
    lo        IP address: 127.0.0.1/8
                                                                    done
System Boot Control: The system has been                            set up
System Boot Control: Running /etc/init.d/boot.local                 done
INIT: Entering runlevel: 4


Welcome to SuSE Linux 9.3 (i586) - Kernel 2.6.11.4-20a-xen (tty1).


linux login:
```



```
xendemo-7:~# uname -a
Linux xendemo-7 2.6.10-xenU #2 Tue Mar 22 22:45:33 CET 2005 i686 GNU/Linux
xendemo-7:~# 
```

What is Debian?

Debian is a free operating syste
computer. An operating system
programs and utilities that mak
run. Debian uses the Linux kern
operating system), but most of t
come from the GNU project; her
GNU/Linux.

```
xendemo-freebsd# uname -a
FreeBSD xendemo-freebsd 5.3-RELEASE FreeBSD 5.3-RELEASE #37: Mon Ja
n 24 16:11:53 PST 2005     kmacy@bldf1.eng.netapp.com:/t/niners/use
rs/xen/bsd/sys-5.3/i386-xeno.tot/compile/XENCONF  i386
xendemo-freebsd# ps
  PID  TT  STAT      TIME COMMAND
  659  p0  Rs     0:00.08 -csh (csh)
  767  p0  R+     0:00.00 ps
  565 xc0  Is     0:00.01 login [pam] (login)
  608 xc0  I+     0:00.02 -csh (csh)
xendemo-freebsd# 
```

```
orpheus:~ # xm list
Name              Id  Mem(MB)  CPU  State  Time(s)  Console
Debian-7           6       47    0  -b---     16.1    9606
Domain-0           0      443    0  r----    193.0
FreeBSD-6          5       47    0  -b---     50.4    9605
NetBSD-8           7       47    0  -b---      1.7    9607
orpheus:~ # xm vif-list Debian-7
(vif (idx 0) (vif 0) (mac aa:00:00:10:b6:6f) (evtchn 27 4) (index 0))
orpheus:~ # xm vif-list Domain-0
orpheus:~ # xm vif-list FreeBSD-6
(vif (idx 0) (vif 0) (mac aa:00:00:15:c6:ee) (evtchn 21 3) (index 0))
orpheus:~ # xm vif-list NetBSD-8
(vif (idx 0) (vif 0) (mac aa:00:00:16:68:0e) (evtchn 28 4) (index 0))
orpheus:~ # 
```

Xen: Networking

- Domain 0 provides bridged networking to DomU's (i.e. guest-OSs)
- DomU's get access to external networking environment
- Can be assigned IP by external DHCP server
- DomU's can be reached from external hosts, appear like standard physical hosts
- A physical network card can be assigned to a DomU (if available)

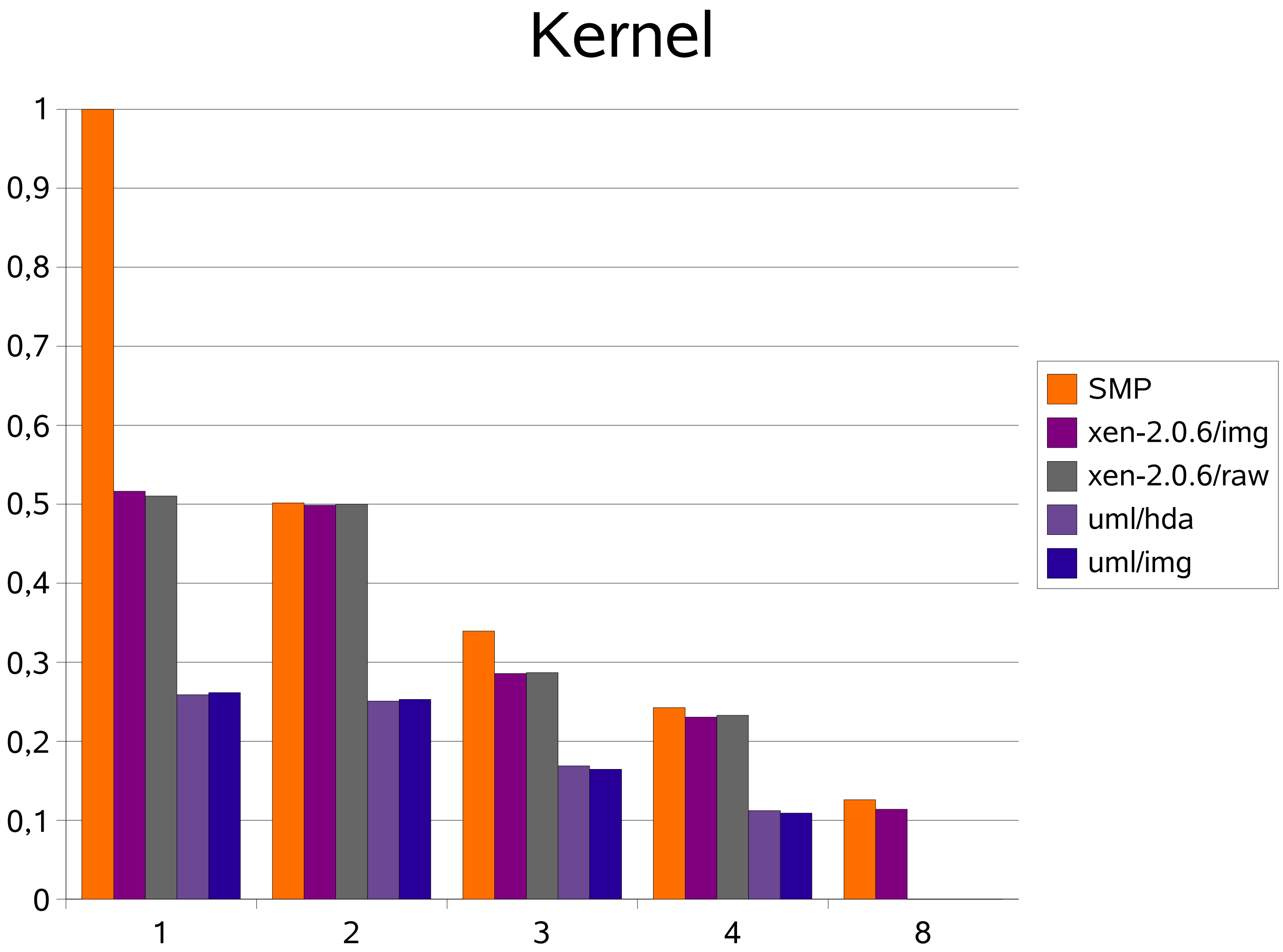
Xen: Migration

Xen can migrate domains between different physical hosts while keeping the network connection alive !

- Create a copy of the memory allocated to a given domain, while the it is still running
- During migration, only an incremental backup of the domain's memory needs to be copied
- Network connections are kept alive, including IP
- No check-pointing needed !!!
- Downtime in the range of milliseconds
- Disadvantage: disk image must be on shared storage !

Results – Kernel compilation

Kernel benchmark: make -j 4



Conclusion

- Stable, high-performance environment
- Very active user community
- Commercial support available
- Supported by large processor manufacturers
- Unique live-migration capability ("stay tuned ...")
- Proven ability to serve as the basis of a "Grid in a box"
- Can inspire a new kind of Grid Computing !
- Will use a gLite-based "Grid in a box" using Xen for GridKa School 2005 (see http://gks05.fzk.de)
- Try it out ! It is easy to use !

Thank you !

We'd like to thank the German Federal Ministry of Education and Research, BMB+F, the EGEE project and its representatives as well as Forschungszentrum Karlsruhe / Germany for their continuous interest and support !

bmb+f - Förderschwerpunkt
Hadronen -
und Kernphysik
Großgeräte der physikalischen
Grundlagenforschung