

XEN - Virtualisation and Grid Computing

Marcus Hardt

Forschungszentrum Karlsruhe GmbH



- **Xen**
 - What is Xen
 - Paravirtualisation
 - Configuration / Starting / Migrating
- **Why Virtualization – Use cases**
 - Virtual Cluster
 - How installed
- **Xen Performance**
 - What to measure
 - How to measure
 - Results
- **Xen Life Demo**
 - Starting, Stopping
 - Migrating

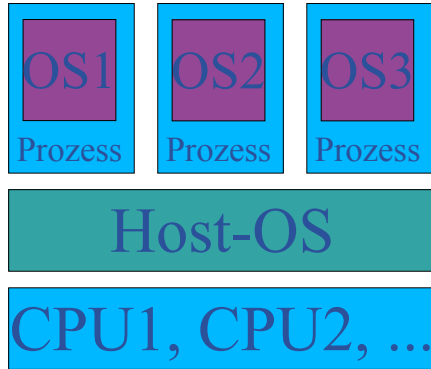
- Approx. 2 years old
- Started by the *Systems Research Group* of the University of Cambridge, UK
- Originally part of the Xenoserver project
 - Idea: A distributed network of OS environments tailored to the user's needs
- **Xen is thus closely related to the ideas of Grid Computing !**
- Now available in Version 2.07
- 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, ...)



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 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 !**

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

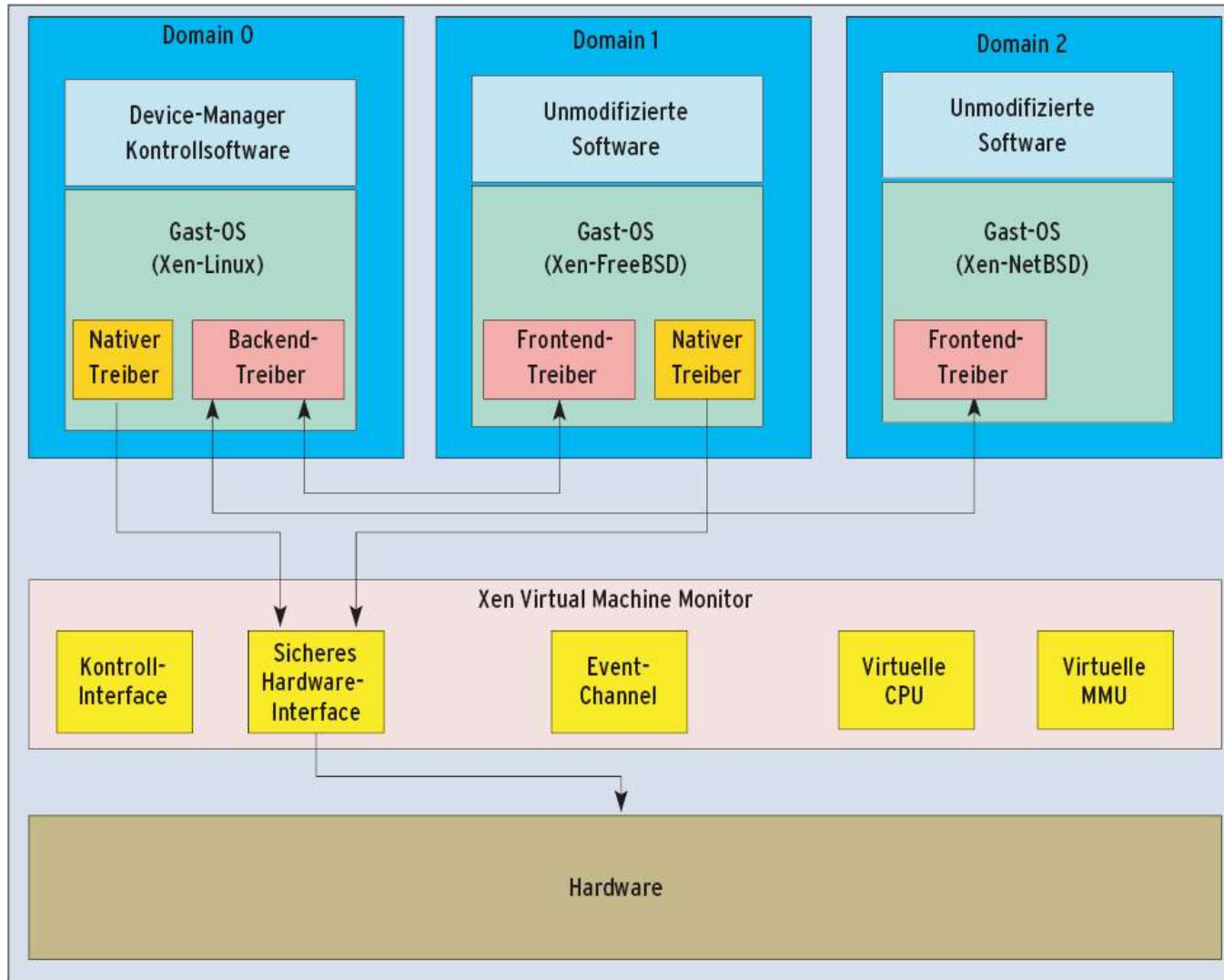


VMWare Workst.
GSX Server
Usermode Linux
Win4Lin
Bochs
Virtual PC

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

IBM zSeries
XEN
ESX Server





- Privileged 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 $O(3000 \text{ LOC})$

- 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

```

emacs@euridike.fzk.de
File Edit Options Buffers Tools IM-Python Python Help
# 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 host
# IP config doesn't, e.g. in ifcfg-eth0 or via DHCP
# You can use 'extra' to set the runlevel and custom
# variables used by custom rc scripts (e.g. VMID=,
# 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 rd"
--:% xmexample1 (Python)--L46--37%-----
(No changes need to be saved)

```

```

ruediger@orpheus:~ - Befehlsfenster - Konsole
Sitzung Bearbeiten Ansicht Lesezeichen Einstellungen Hilfe
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
Enabling syn flood protection done
Disabling IP forwarding done
Setting up hostname 'linux' done
Setting up loopback interface lo done
lo IP address: 127.0.0.1/8
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:

```




TightVNC: root's x11 desktop (xendemo-8:0)

Welcome to xendemo-8

Login:

Password:



TightVNC: root's x11 desktop (xendemo-7:0)

```

Bash
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:~#
  
```

Debian - The Universal Operating System - Mozilla

http://www.debian.org/

debian

Select a server near you:
United States

About News Getting Debian Support Development Site map Search

What is Debian?

Debian is a [free](#) operating system for your computer. An operating system is a collection of programs and utilities that make it possible to run other programs. Debian uses the [Linux](#) kernel (operating system), but most of the software comes from the [GNU project](#); hence the name GNU/Linux.

ruediger@orpheus:~ - Befehlsfenster - Konsole <5>

```

Sitzung Bearbeiten Ansicht Lesezeichen Einstellungen Hilfe
xendemo-freebsd# uname -a
FreeBSD xendemo-freebsd 5.3-RELEASE FreeBSD 5.3-RELEASE #37: Mon Jan 24 16:11:53 PST 2005 kmacy@bldf1.eng.netapp.com:/t/niners/users/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) (vifname eth0) (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) (vifname eth0) (index 0))
orpheus:~ # xm vif-list NetBSD-8
(vif (idx 0) (vif 0) (mac aa:00:00:16:68:0e) (vifname eth0) (index 0))
orpheus:~ #
  
```

ruediger@orpheus:~ - Befehlsfenster

```

Sitzung Bearbeiten Ansicht Lesezeichen Einstellungen Hilfe
xendemo-freebsd# uname -a
FreeBSD xendemo-freebsd 5.3-RELEASE FreeBSD 5.3-RELEASE #37: Mon Jan 24 16:11:53 PST 2005 kmacy@bldf1.eng.netapp.com:/t/niners/users/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#
  
```

- **Installation Course on cluster/grid computing:**
 - Summerschool on Gridcomputing at FZK
 - ~40 Students vs. 16 available PCs
 - PCs required for max 3 days
 - => My boss won't buy missing PCs for that time (est. 75)
 - Virtualisation provides:
 - No need to buy additional 60 PCs (obvious)
 - No need to install 60 additional PCs
 - Students can check output of booted Xen domains via ssh
 - last year we moved and installed 40 PCs (1.5 Racks) in the offices

- **Simple installation of a virtual cluster:**
 - Linux installation:
 - `mount -o loop image mnt`
 - `ssh <installed machine> tar csp / | (cd mnt;tar xsp)`
 - Additional modifications:
 - `/etc/fstab`
 - `/etc/passwd`
 - `/lib/tls`
 - Image duplication
 - `for i in `seq 1 100`;do cp image image-$i; done`
 - Booting
 - `for i in `seq 1 100`;do xm create <conf> id=$i; done`
 - Network via dhcp

- **Submitting a job to “the grid”**
 - The grid =
Scattered heterogenous resources with different admins
 - Developers can hardly cope with that
 - Virtualisation allows:
 - Developer is given an OS image
 - Image is transported to resource
 - ... booted ... processing ... executed ... results returned

- **IT Consolidation**

- Start/stop servers on demand
 - maybe based on monitoring information (load, response time, ...)
- Self healing hardware
 - One standby server per IT department
vs. one per (important) service
- Easy provisioning of machines
 - `cp Debian-stable.img webserver-ukuug.org.img`
 - `xm create ...`
- Concentration of rarely used machines to one
 - est. 100-200 EUR per machine per year. (1 EUR/W/a)
- Migration of domains helpful for administration

- **Load Balancing in Cluster Systems**
 - Oversubscription of the cluster
 - Some jobs do I/O, while others compute
 - Individual Operating Systems provided
 - Easier administration, especially of SMP machines
 - Migration helps administration
 - Python based configuration increases flexibility

- **Using Windows Desktops** ~~(for reasonable work)~~ (to sell CPU cycles)
 - We have est 4000 Windows Desktops at FZK
 - Idle 76% of their time (used 8h on 5days a week)
 - Doesn't require air-condition
 - Once Xen supports Windows in domU....
 - Run two domains on every Desktop:
 - *Windows Desktop*
 - *Cluster Node*
 - Image supplied by customer
 - When Desktop is used, cluster can be suspended or migrated away

- **My rootserver**
 - 70 EUR/month:
 - Xeon 2,8GHz, 1GB RAM, 80GB HDD, 0.5TB traffic
 - Big enough to be shared with 4+ people
 - Xen provides enough security to do this
 - Issues:
 - Provider only allows one MAC Adress on the port of the switch :-(
 - Provider runs all subnets with only one IP
 - *After some work it works :-)*
=> I save 60EUR/month

There is no such thing as a free lunch

or

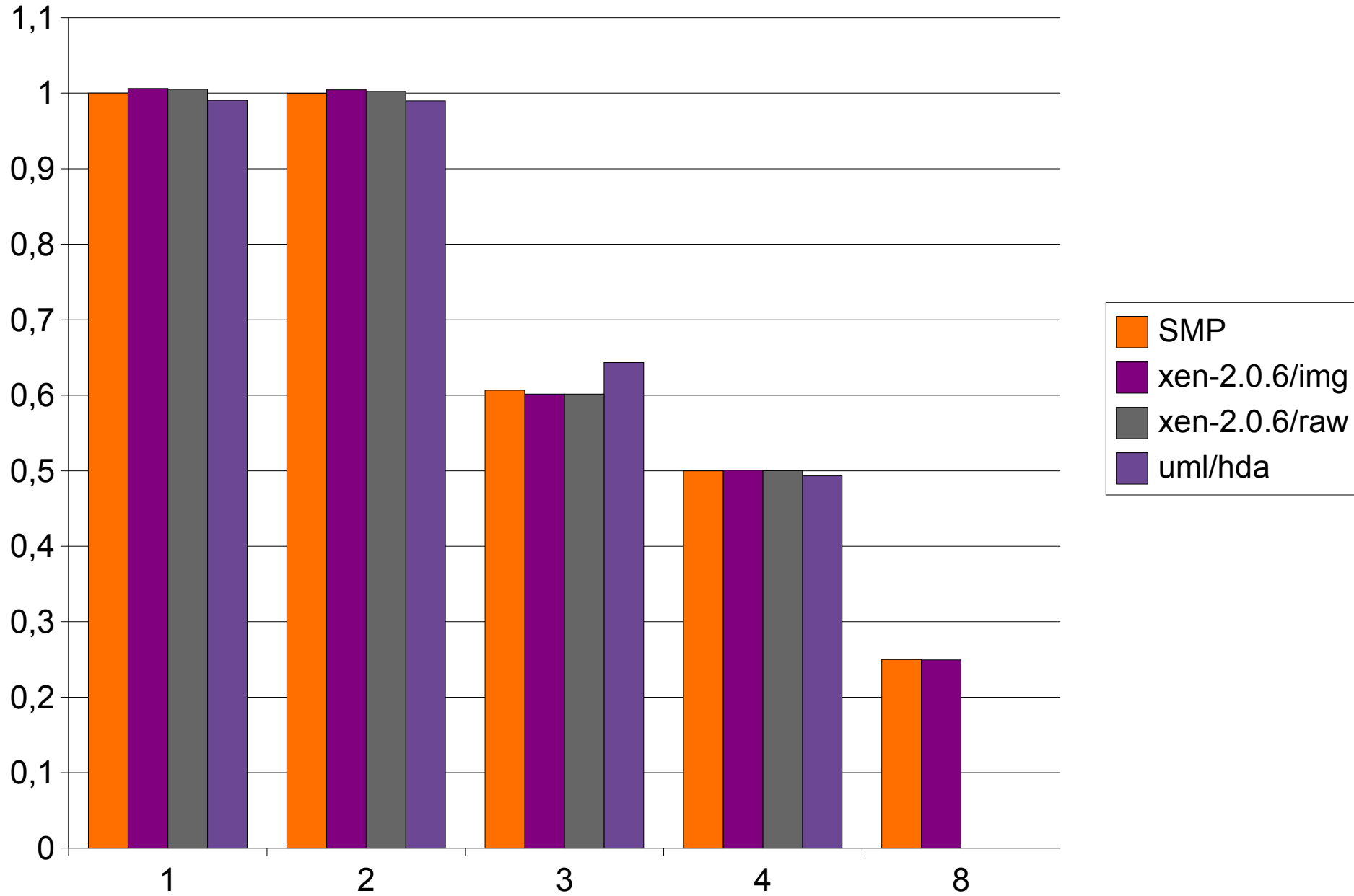
How much performance does virtualisation cost?

- **How measured**
 - Hardware
 - Dual-PIII-700MHz / 1GB RAM / 40GB Disk / 100Mbit/s
 - Reference Measurement 1-4,8 parallel runs on plain smp
 - Benchmark installation booted and run on 1-4,8 xen domains
 - I/O: Partition and Image backed instances
 - Comparison Measurement on 1-4 UML instances

 - Commercial Products yet to be compared
 - MS Virtual PC
 - VMWare, cannot be published, license restrictions

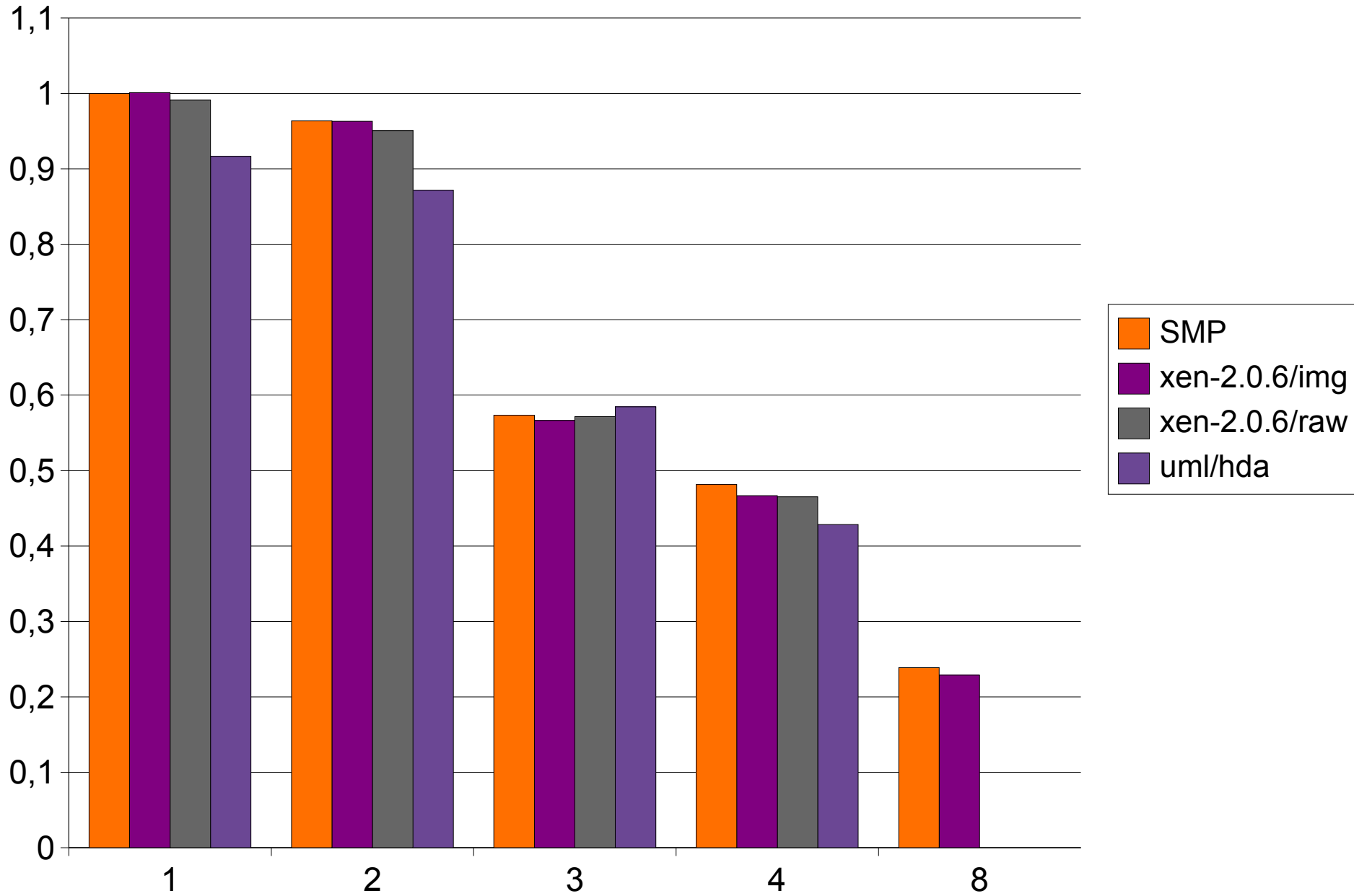
- CPU Benchmark
 - “Four in a row” of <http://www.freebench.org>
 - Play four in a row against itself
 - Integer bound
 - Why?
 - It's free, and doesn't need much RAM (SPEC does)
 - It doesn't run too long

CPU



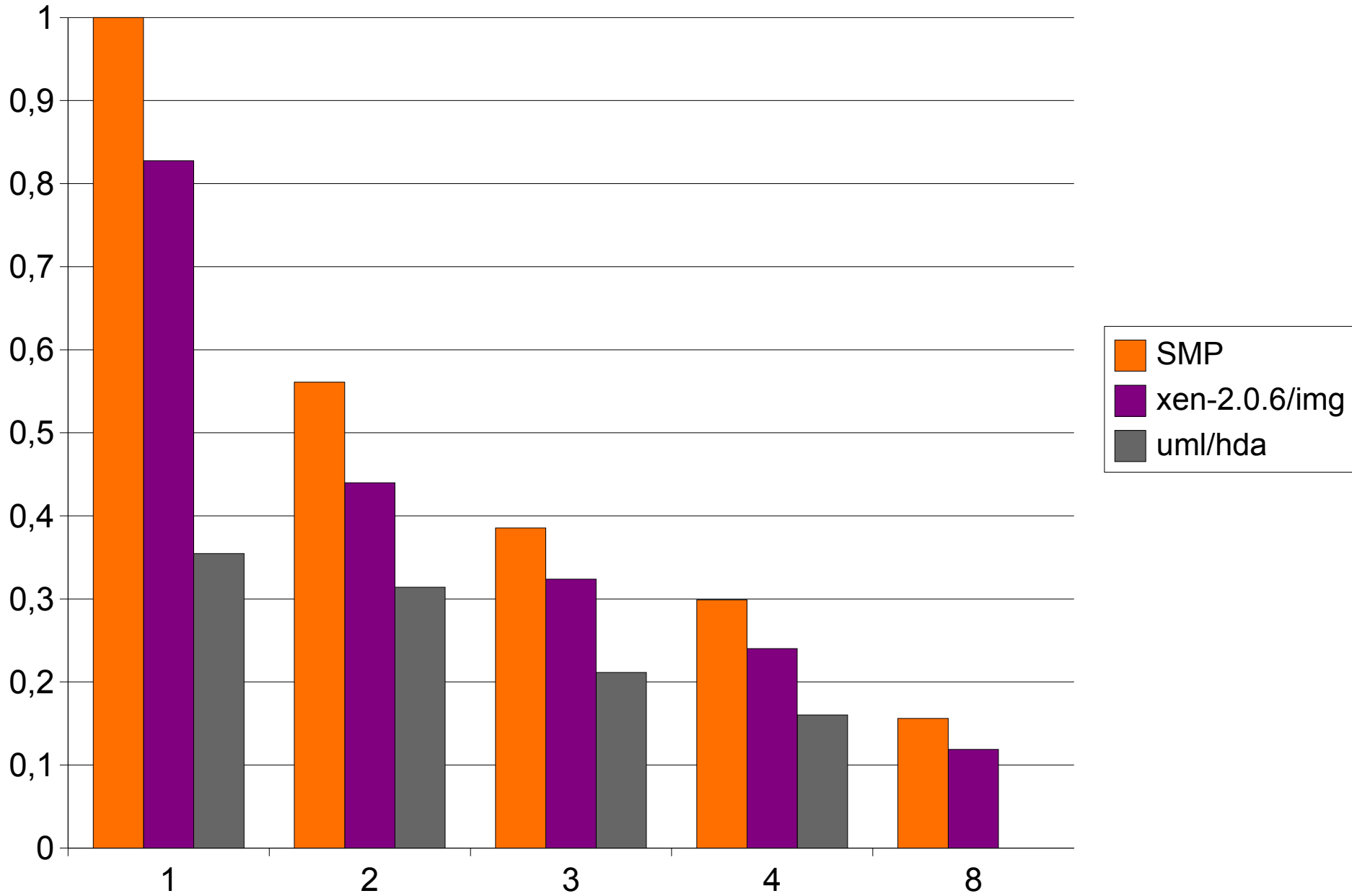
- **Memory Benchmark**
 - pCompress2, <http://www.freebench.org>
 - Integer bound, but also memory intensive
 - Intensive use of memcmp, memcpy and qsort
 - Why?
 - First one I came across
 - considering streamer

MEM



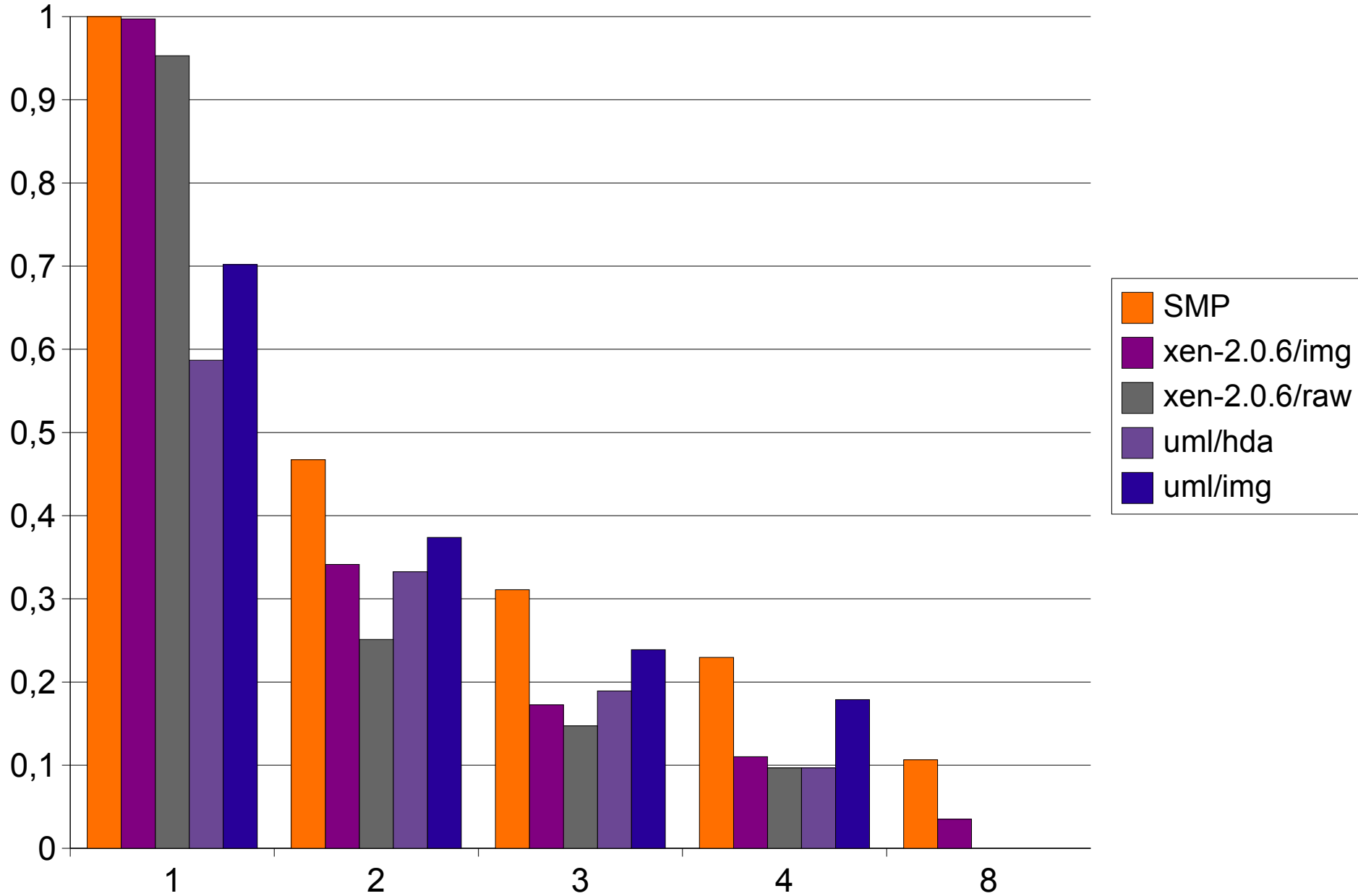
- **Network Benchmark:**
 - tbench <http://samba.org/ftp/tridge/dbench>
 - Network bound
 - dbench and tbench simulate the load of the netbench benchmark
 - tbench produces the network load (dbench=disk)
 - Why?
 - Free
 - Easy to use
 - Fast

NET



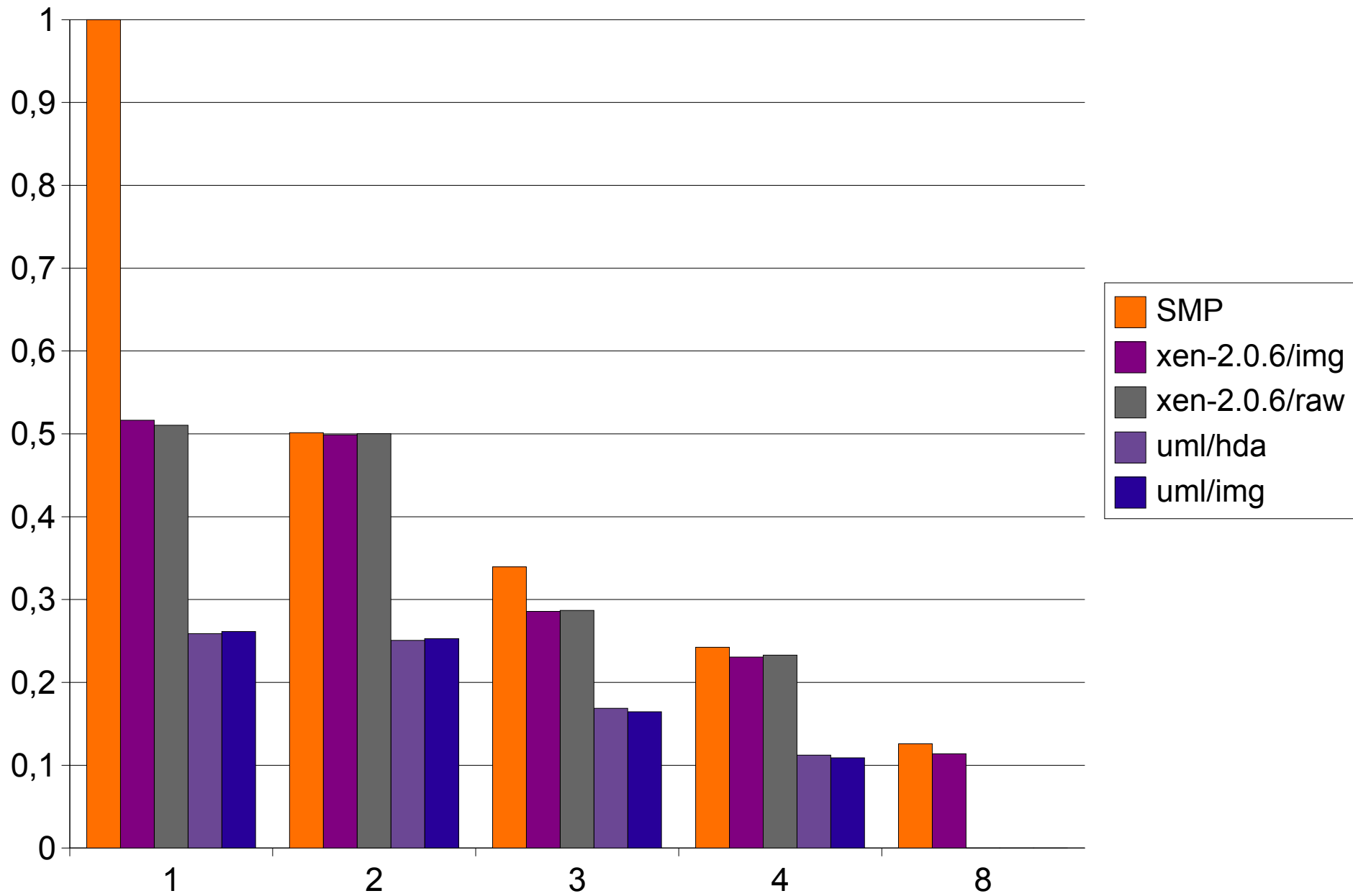
- **Disk I/O Benchmark**
 - DD:
 - `dd if=/dev/hdaX of=/dev/null bs=32k count=32k`
 - Image backed vs. Partition backed
 - Why?
 - Simple
 - Flexible
 - Free ;-)

DD-2.0.6



- Kernel Benchmark:
 - `make -j4`
 - Why?
 - Standard application benchmark

Kernel



- **Linux cannot keep images on NFS**
 - Use SAN, GNBD or iSCSI instead
- **/lib/tls problem**
 - `mv /lib/tls /lib/tls.disabled.for.xen`

- **Stable Virtualisation Environment**
 - Proven to serve as “grid in a box”
- **Good Performance**
 - Less than 10% virtualisation cost (except net i/o: 20%)
 - Better than userspace tools (UML, VMWare Workstation)
- **Easy to use and install**
- **Very active user community**
 - Fast and good answers via mailinglist
 - Up to date with recent changes in Debian
- **Commercial Support available additionally**
- **Supported by hardware manufacturers**
- **Unique live migration capability**
- **Try it out**

