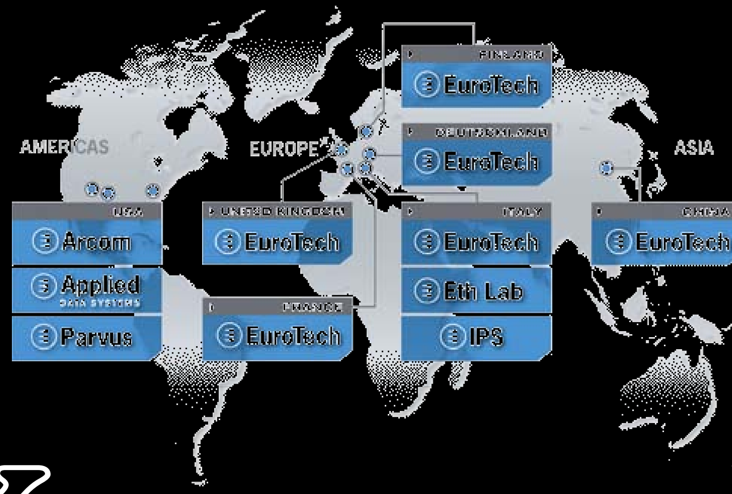
16th PDP 2008
Toulouse, France
February 13th, 2008



The VirtuaLinux

Storage Abstraction Layer for Efficient Virtual Clustering

http://sourceforge.net/projects/virtualinux/

Marco Aldinucci, Massimo Torquati, Marco Vanneschi
Computer Science Dept., University of Pisa, Italy

Pierfrancesco Zuccato
Eurotech Italy

Outline

- VirtuaLinux basics
  - motivations
  - which problems VirtuaLinux cope with
  - architecture: big picture
- VirtuaLinux features
  - high-availability
    - masterless cluster
    - diskless cluster
    - storage virtualization
  - consolidation
    - virtual cluster management tools
  - develop for cluster without a cluster
    - multi tier distribution
- Experiments & conclusions

EUROTECH GROUP

VirtuaLinux aims

- Clusters
  - a collection of homogenous but independent machines
  - ☹ are fragile
    - master node is a single point of failure
    - disks are a common source of failure
  - ☹ complex to install and maintain
    - a proper installation and configuration may require days
    - skilled administrators are required
      - root account power is a common source of misconfiguration during production
  - ☹ they are shared machines
    - a single configuration does not match user expectations
      - ... I need CentOS, I prefer Ubuntu, I believe in Windows ..
- VirtuaLinux aims to attack these problems
  - Not surprisingly, the project has been founded by an HW producer

EUROTECH GROUP

Virtualization: a brand new idea ...

Christopher Strachey published a paper titled *Time Sharing in Large Fast Computers* in the International Conference on Information Processing at UNESCO, New York, in June, 1959. Later on, in 1974, he clarified in an email to Donald Knuth that:

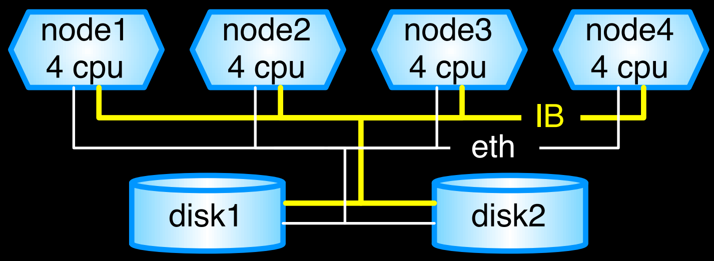
" ... [my paper] was mainly about multi-programming (to avoid waiting for peripherals) although it did envisage this going on at the same time as a programmer who was debugging his program at a console. I did not envisage the sort of console system which is now so confusingly called time sharing.". Strachey admits, however, that "time sharing" as a phrase was very much in the air in the year 1960.

Robert P. Goldberg describes the then state of things in his 1974 paper titled *Survey of Virtual Machines Research*. He says: *"Virtual machine systems were originally developed to correct some of the shortcomings of the typical third generation architectures and multi-programming operating systems - e.g., OS/360."*

- ❖ Anyway, it works (quite often)
  - ✦ at the bottom line it is a well known tool: abstraction
    - ✴ high-level (e.g. JVM); medium-level (e.g. FreeBDS jails); low-level (Simulazione [e.g. Cell], Binary translation [e.g. WMware, Qemu, ...], paravirtualization [Xen, KVM, ...])
- ❖ makes it possible
  - ☺ consolidate different OSes in a single HW
  - ☺ share HW and SW resources
  - ☺ insulate classes of users and resources

EUROTECH GROUP

Cluster: a quite classic configuration



Physical Cluster + external SAN
InfiniBand + Ethernet
4 Nodes x 4 CPUs
Cluster InfiniBand 192.0.0.0/24
Cluster Ethernet 192.0.1.0/24
Internet Gateway 131.1.7.6

- Diskless blades + external storage (SAN/NAS)
  - Fiber/Infiniband SAN-RAID are fast and robust
    - they are enforced at HW level, irrespectively of the OS
  - sometime enforced by law (e.g. USA's Sarbenes-Oxley)
- Any existing Linux distribution for this configuration?
  - A plethora of them, but ...
  - they are not standard distributions
    - typically services and their paths require substantial re-configuration
    - complex, require specialized initrd
    - SO update not easy (cannot rely on standard update tools)

VirtuaLinux approach

- A meta-distribution, conceptually (➽ standardization)
  - choose a Linux distribution and then configure it for clusters
    - Ubuntu, Debian, CentOS, ...
  - the guest OS is not modified, just properly configured
- Master-less (➽ robustness)
  - no master node (all nodes cooperatively behave as master node)
- Disk-less (➽ robustness, flexibility)
  - each physical node access to a private and a cluster shared volume
    - volumes on the iSCSI-attached SAN are virtualized by way of VirtuaLinux storage virtualization layer
- Transparently supports Virtual Clusters (VC) (➽flexibility)
  - tools for VC deployment, mapping, lifecycle control, etc.
  - currently based on Xen paravirtualization
  - VCs are insulated one each other, just share physical resources

EUROTECH GROUP

IN SUPREMÆ DIGNITATIS 1343

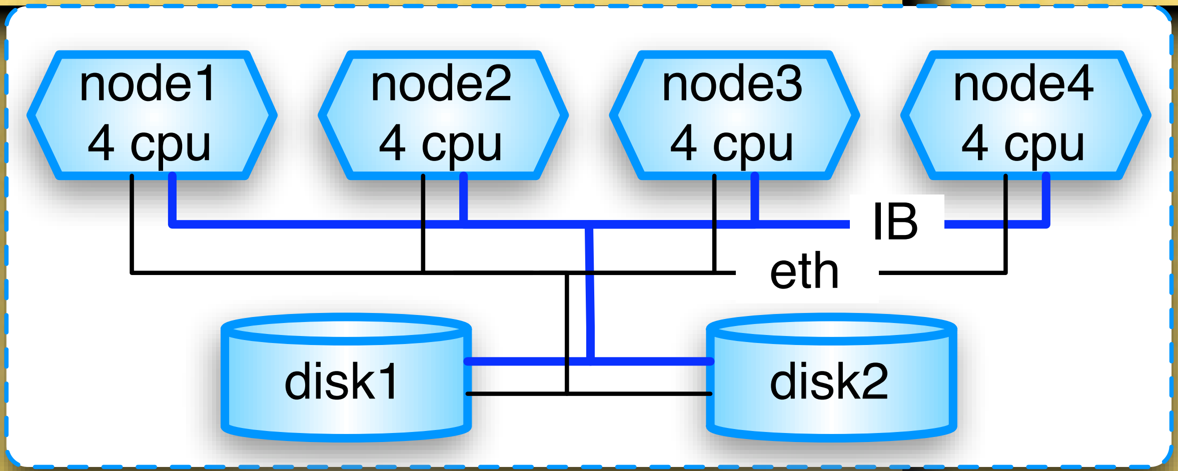
Virtual Clusters (VC)

- Natural evolution of VM idea to cluster level
- a collection of coordinated of virtual nodes
  - each one being a VM with its own
    - Vcpu
    - Vstorage: private and VC-wide shared
    - virtual networking: VC-private and inter VCs
    - VM technology neutral
      - several options are possible: VMware, Xen, QEMU, ...
      - independent from VM tecnology
      - quality of VC improves while quality of VM improves

EUROTECH GROUP

IN SUPREMÆ DIGNITATIS 1343

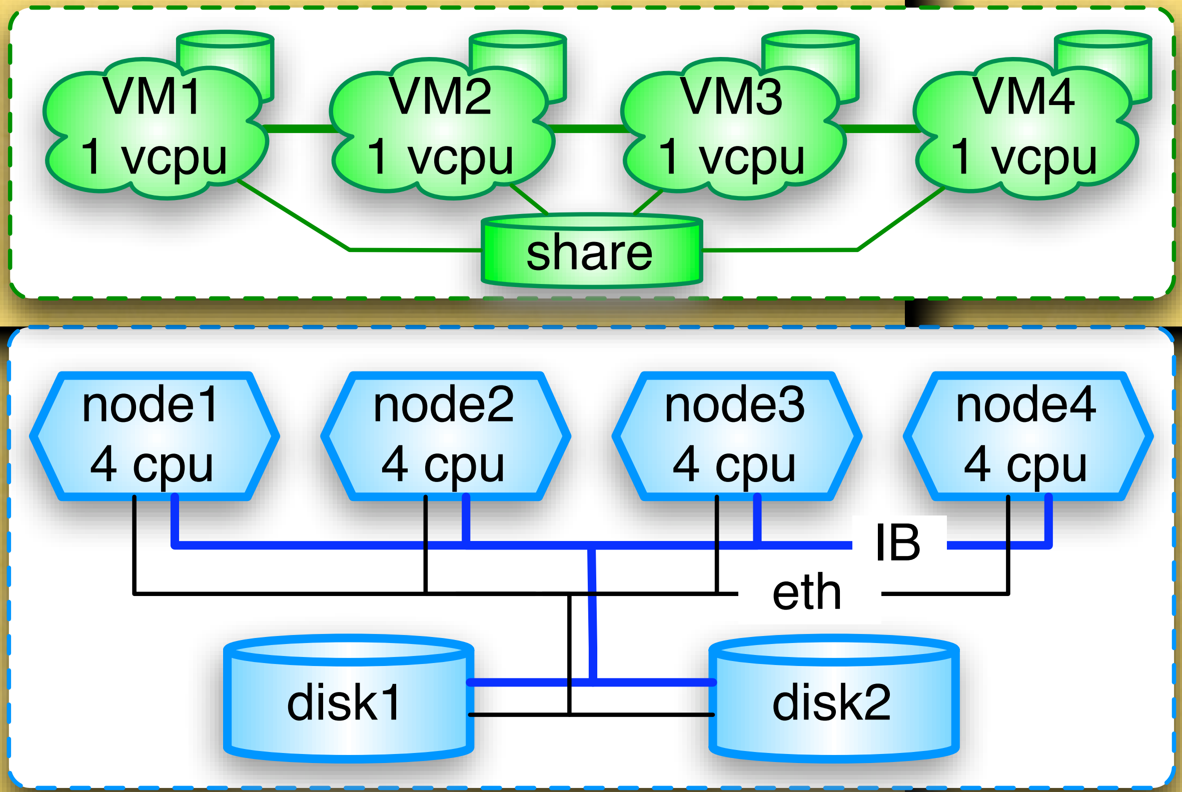
Virtual Clusters (VC)



Physical Cluster + external SAN
InfiniBand + Ethernet
4 Nodes x 4 CPUs
Cluster InfiniBand 192.0.0.0/24
Cluster Ethernet 192.0.1.0/24
Internet Gateway 131.1.7.6

Virtual Clusters (VC)

Start VC green

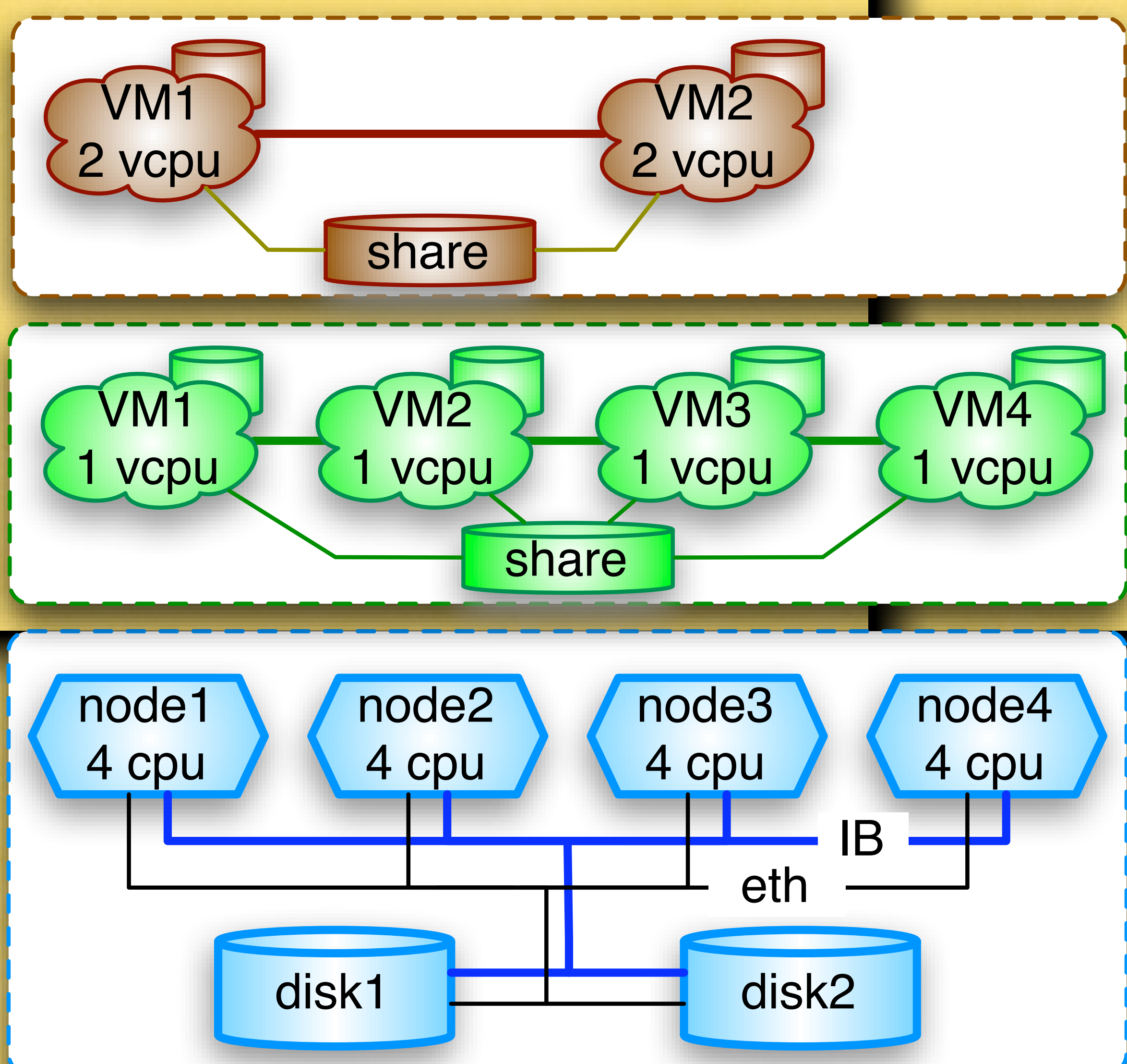


Virtual Cluster "green"
4VMs x 1VCPUs
10.0.3.0/24

Physical Cluster + external SAN
InfiniBand + Ethernet
4 Nodes x 4 CPUs
Cluster InfiniBand 192.0.0.0/24
Cluster Ethernet 192.0.1.0/24
Internet Gateway 131.1.7.6

Virtual Clusters (VC)

Start VC
tan



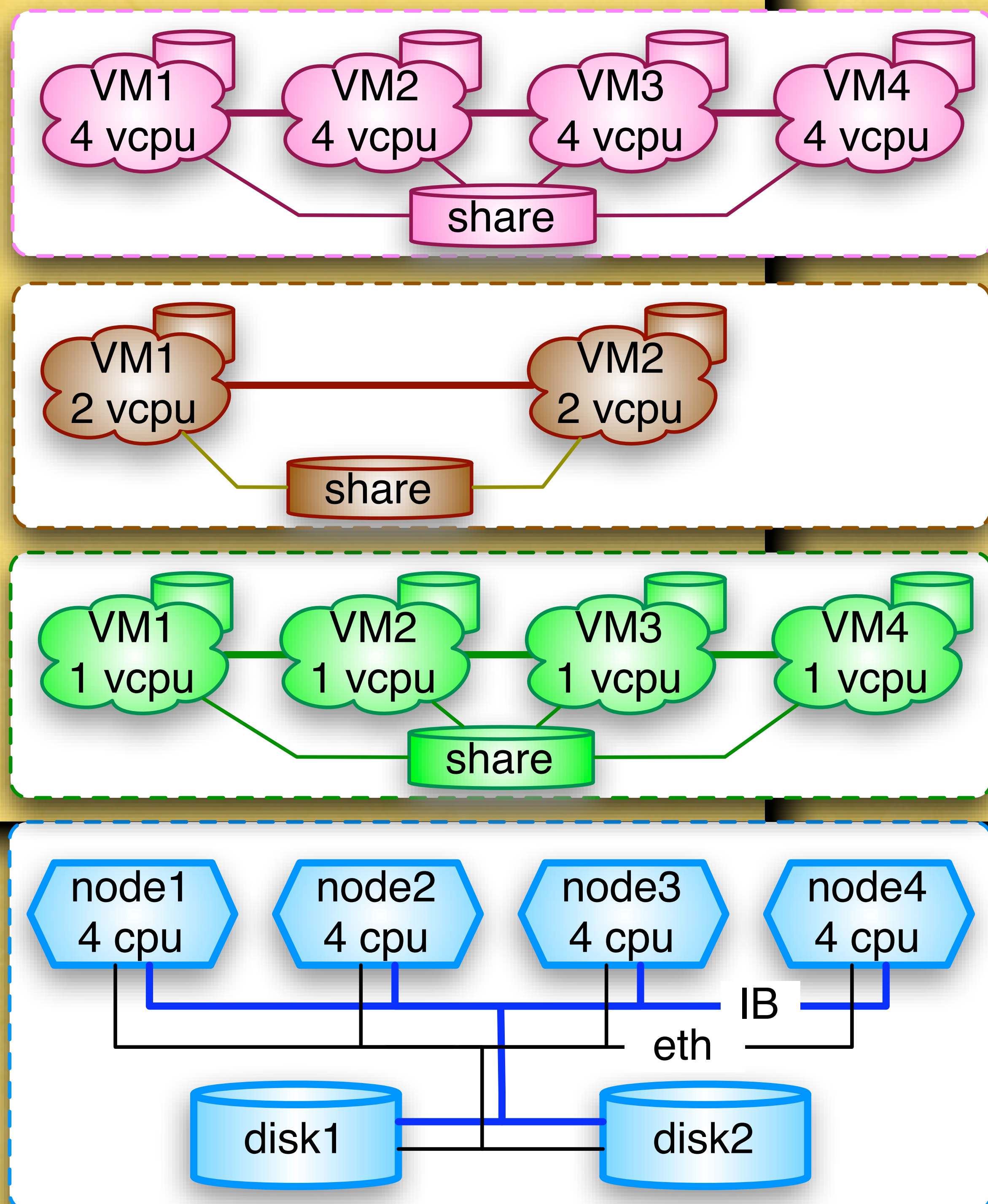
Virtual Cluster "tan"
2VMs x 2VCPUs
10.0.1.0/24

Virtual Cluster "green"
4VMs x 1VCPUs
10.0.3.0/24

Physical Cluster + external SAN
InfiniBand + Ethernet
4 Nodes x 4 CPUs
Cluster InfiniBand 192.0.0.0/24
Cluster Ethernet 192.0.1.0/24
Internet Gateway 131.1.7.6

Virtual Clusters (VC)

Start VC pink



Virtual Cluster "pink"
4VMs x 4VCPUs
10.0.0.0/24

Virtual Cluster "tan"
2VMs x 2VCPUs
10.0.1.0/24

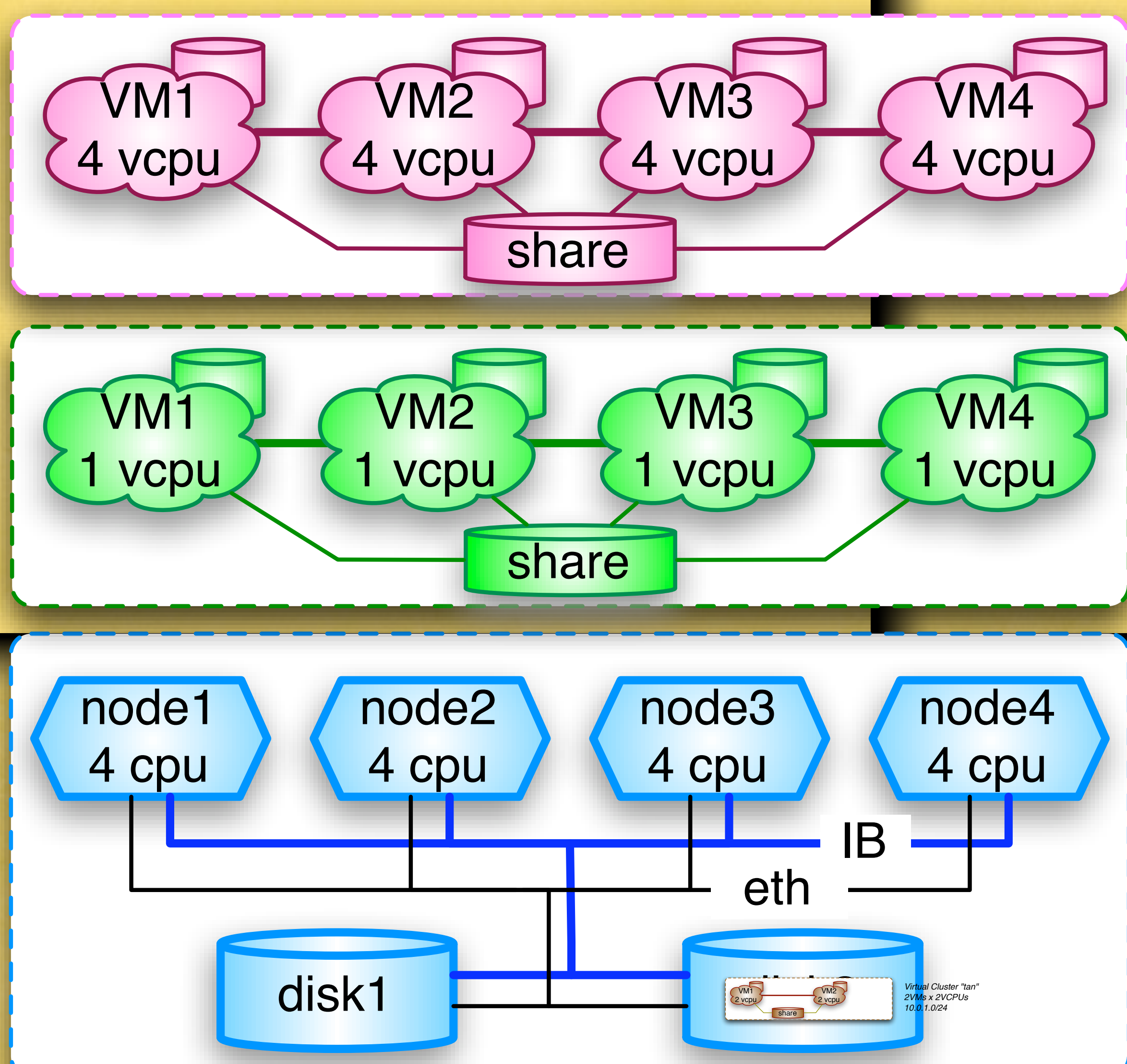
Virtual Cluster "green"
4VMs x 1VCPUs
10.0.3.0/24

Physical Cluster + external SAN
InfiniBand + Ethernet
4 Nodes x 4 CPUs
Cluster InfiniBand 192.0.0.0/24
Cluster Ethernet 192.0.1.0/24
Internet Gateway 131.1.7.6

EUROTECH GROUP

Virtual Clusters (VC)

Suspend tan



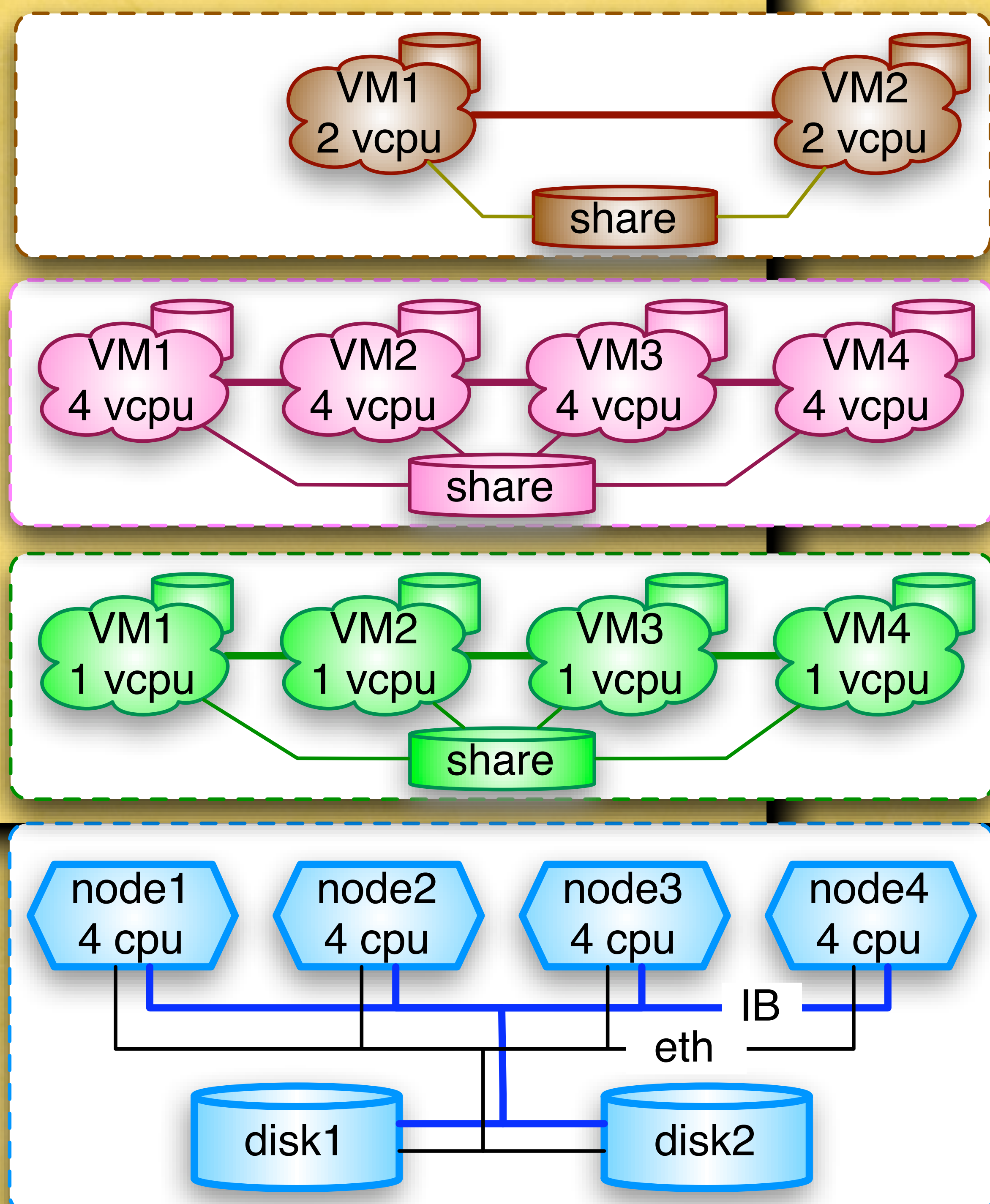
Virtual Cluster "pink"
4VMs x 4VCPUs
10.0.0.0/24

Virtual Cluster "green"
4VMs x 1VCPUs
10.0.3.0/24

Physical Cluster + external SAN
InfiniBand + Ethernet
4 Nodes x 4 CPUs
Cluster InfiniBand 192.0.0.0/24
Cluster Ethernet 192.0.1.0/24
Internet Gateway 131.1.7.6

Virtual Clusters (VC)

Restart &
Remap tan



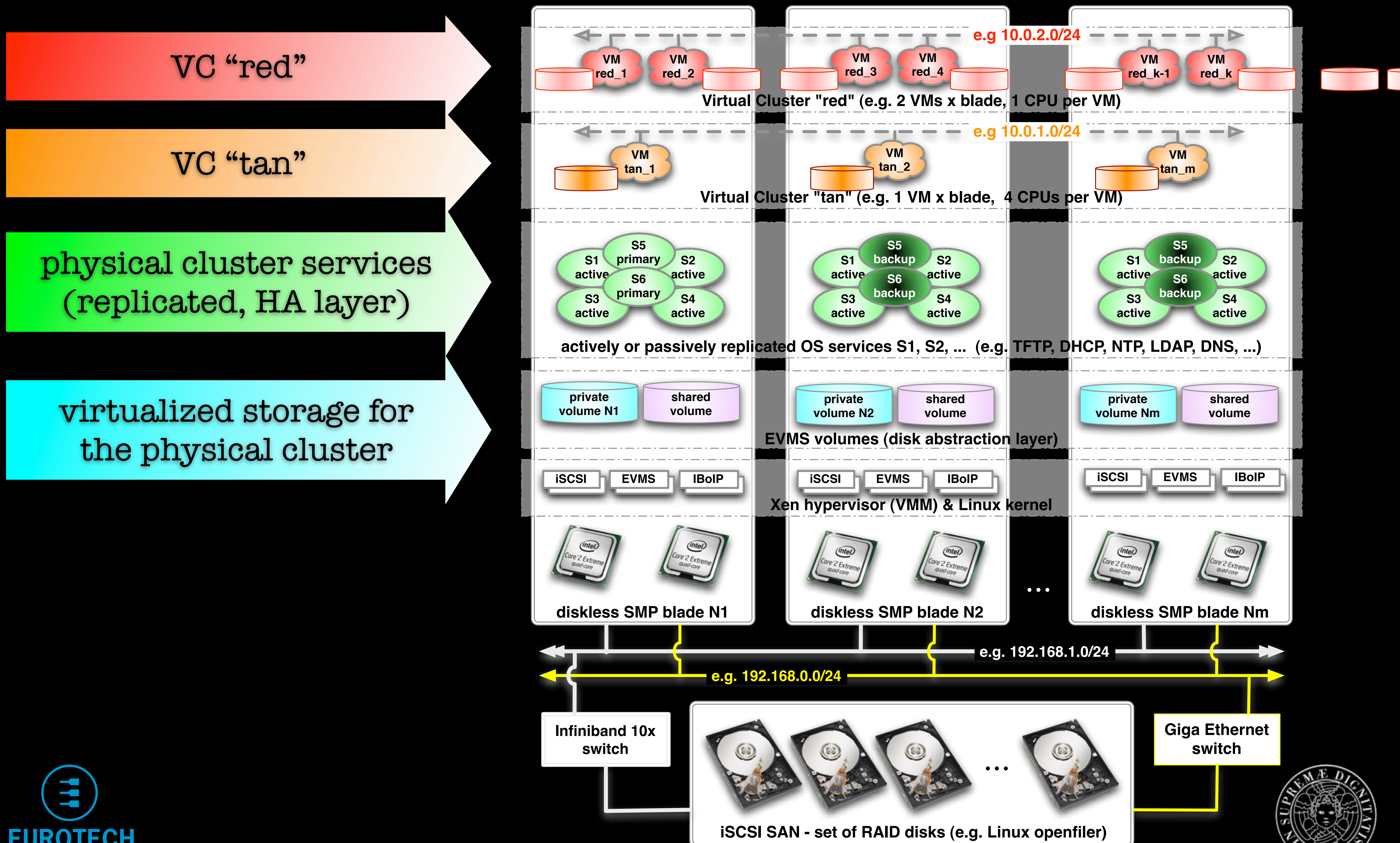
Virtual Cluster "tan"
2VMs x 2VCPUs
10.0.1.0/24

Virtual Cluster "pink"
4VMs x 4VCPUs
10.0.0.0/24

Virtual Cluster "green"
4VMs x 1VCPUs
10.0.3.0/24

Physical Cluster + external SAN
InfiniBand + Ethernet
4 Nodes x 4 CPUs
Cluster InfiniBand 192.0.0.0/24
Cluster Ethernet 192.0.1.0/24
Internet Gateway 131.1.7.6

Big Picture



EUROTECH GROUP

High Availability

by way of active and passive replication

High availability

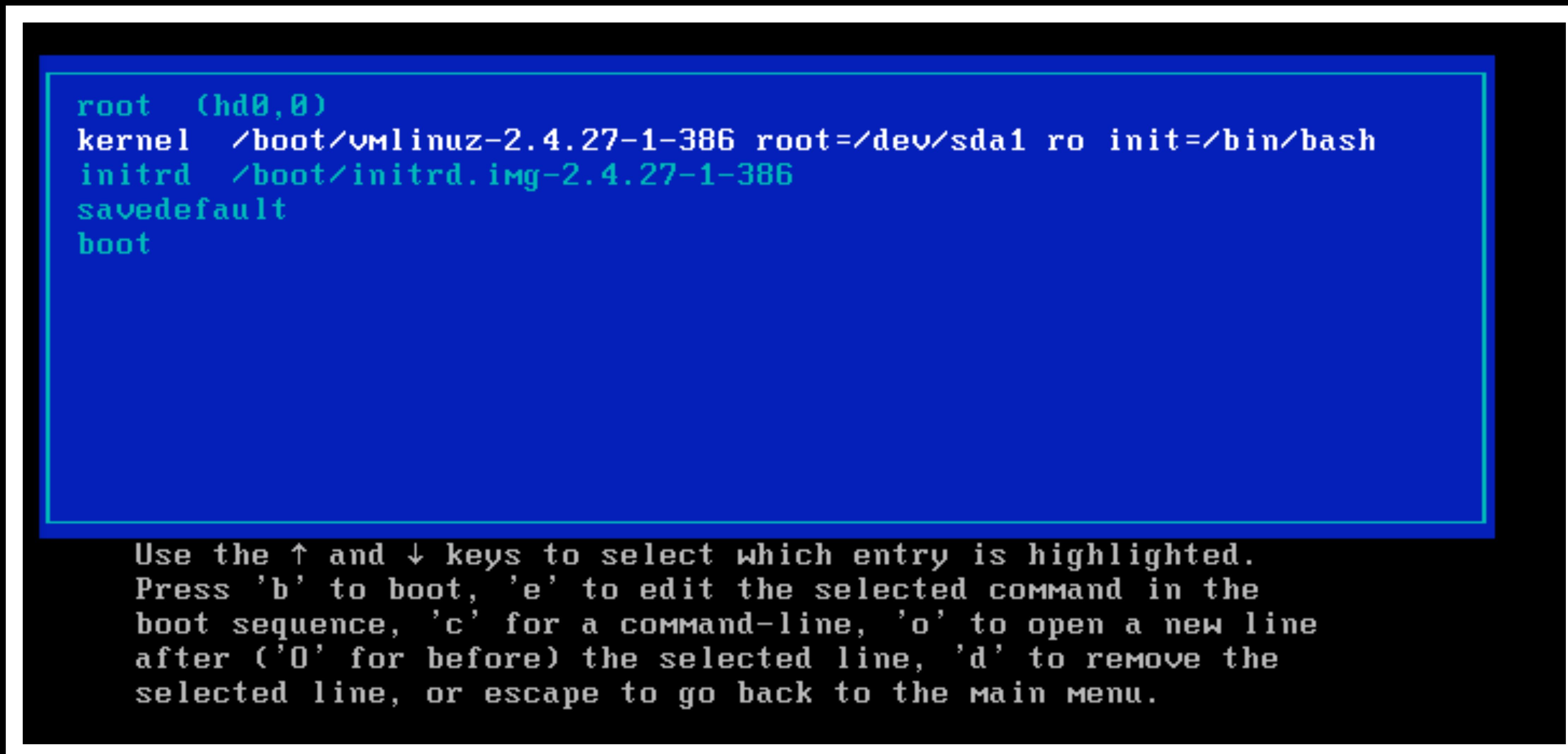
- 24/7 cluster availability
  - to be not confused with application-level fault tolerance … here we would like to ensure that the cluster survive, not its applications
- High-availability means redundancy
  - robust hardware
    - e.g. 5 power supplies, 4 independent network switches, …
    - iSCSI-over-Infiniband and Fiber channels to storage
    - RAID storage
  - service replication
    - all nodes are identical, i.e. no master node
    - all essential services are replicated on all nodes
    - each node can be hot-swapped, switched on/off with no impact on cluster availability and stability

EUROTECH GROUP

How to replicate services (sample)

Service	FT model	Notes
DHCP	active	Pre-defined map between IP and MAC
TFTP	active	All copies provide the same image
NTP	active	Pre-defined external NTPD fallback via GW
IB manager	active	Stateless service
DNS	active	Cache-only
LDAP	service-specific	Service-specific master redundancy
IP GW	passive	Heartbeat with IP takeover (via IP aliasing)
Mail	node-oriented	Local node and relays via DNS
SSH/SCP	node-oriented	Pre-defined keys
NFS	node-oriented	Pre-defined configuration
SMB/CIFS	node-oriented	Pre-defined configuration

EUROTECH GROUP



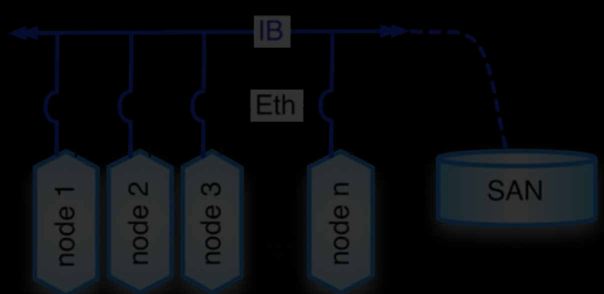
A novel boot sequence

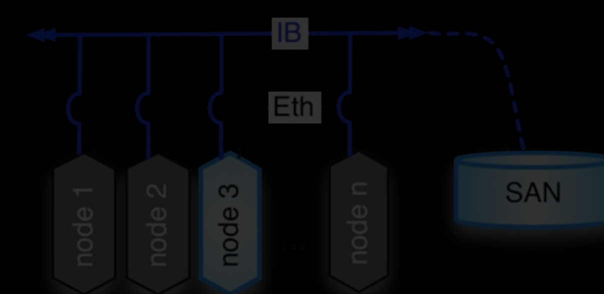
to support master-less clusters

Install without a master

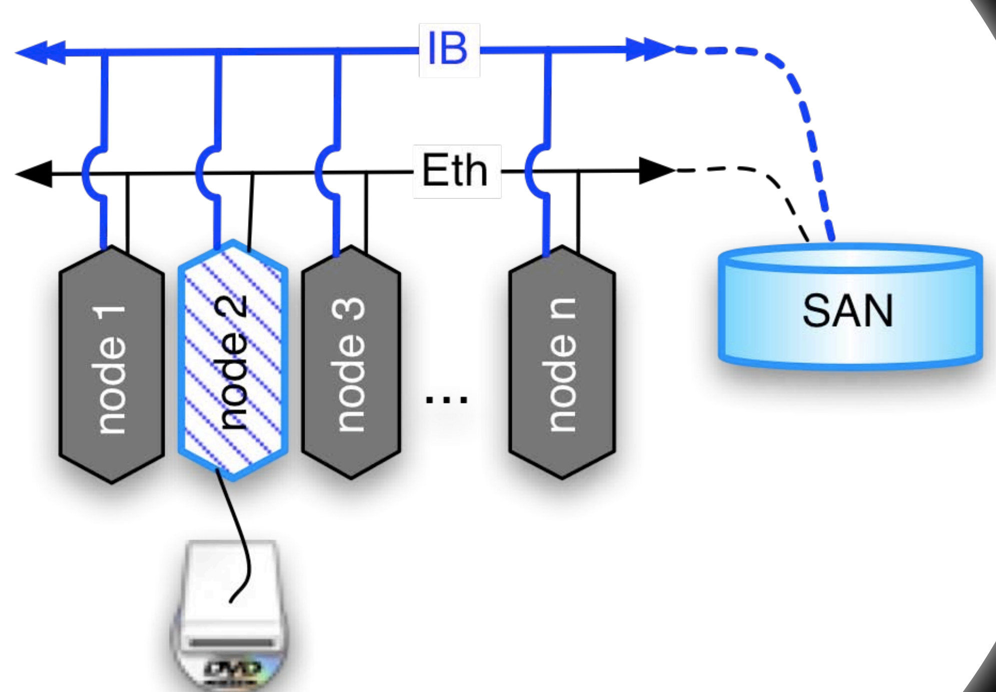
- How to install a cluster without a master?
  - the process should begin somehow
- Our solution: metamaster
  - one of the node behave transiently as a master, start the process, then become a standard node
- At the end of the installation all nodes are identical

Install without a master

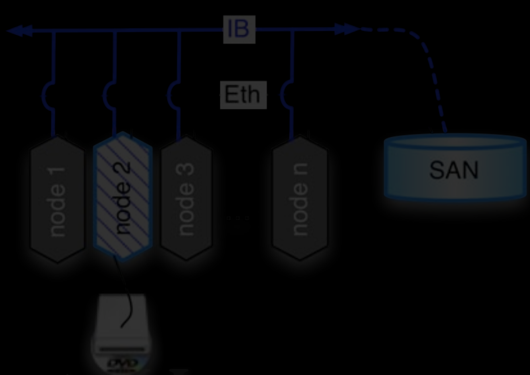


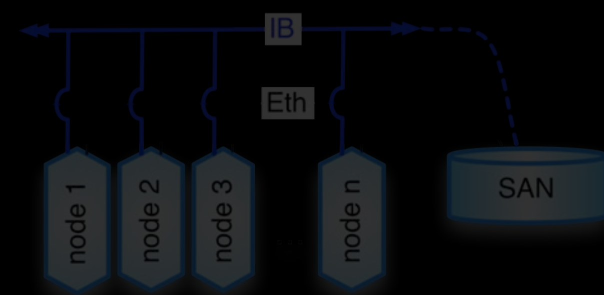


- How to install … aster?
  - the proces…
- Our solu…
  - one of … art the process,
    then b…
- At the e… e identical

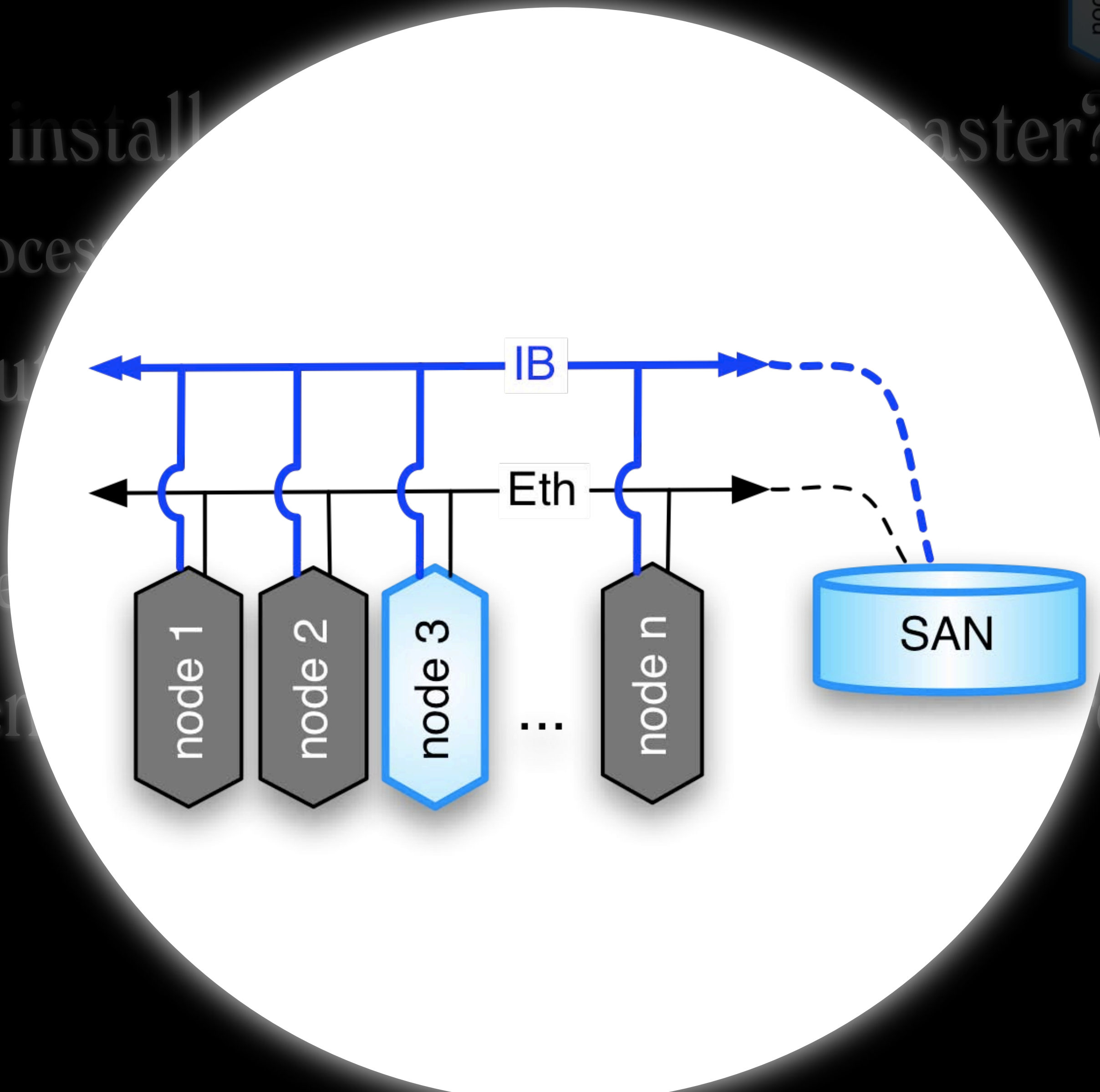


Install without a master

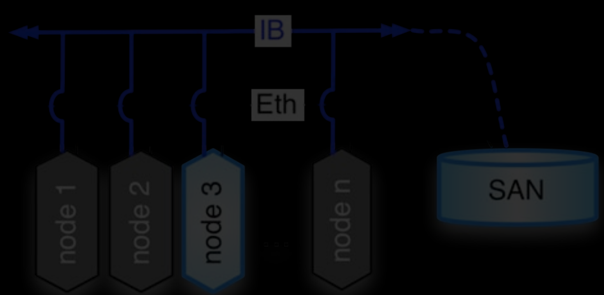


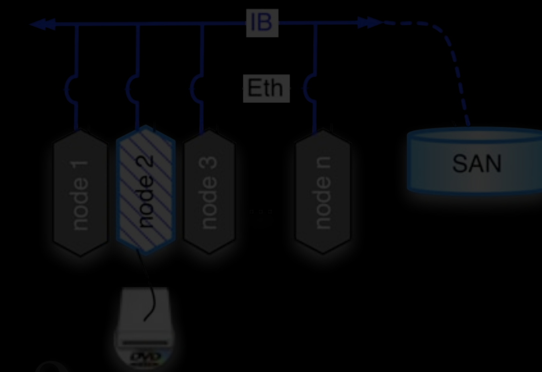


- How to install ... aster?
  - the proces...
- Our solu...
  - one of ... art the process,
    then b...
- At the e... e identical

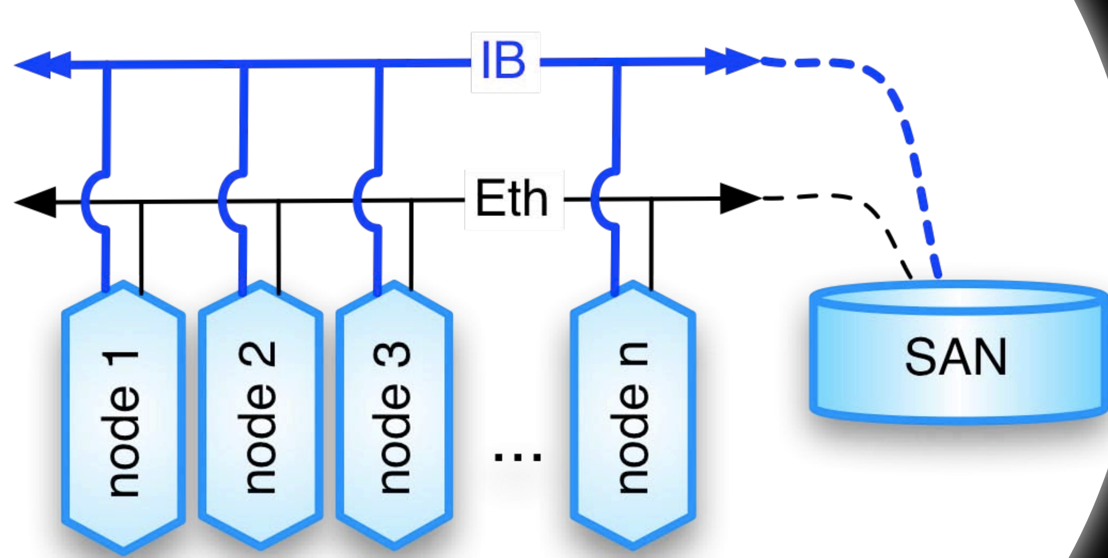


Install without a master





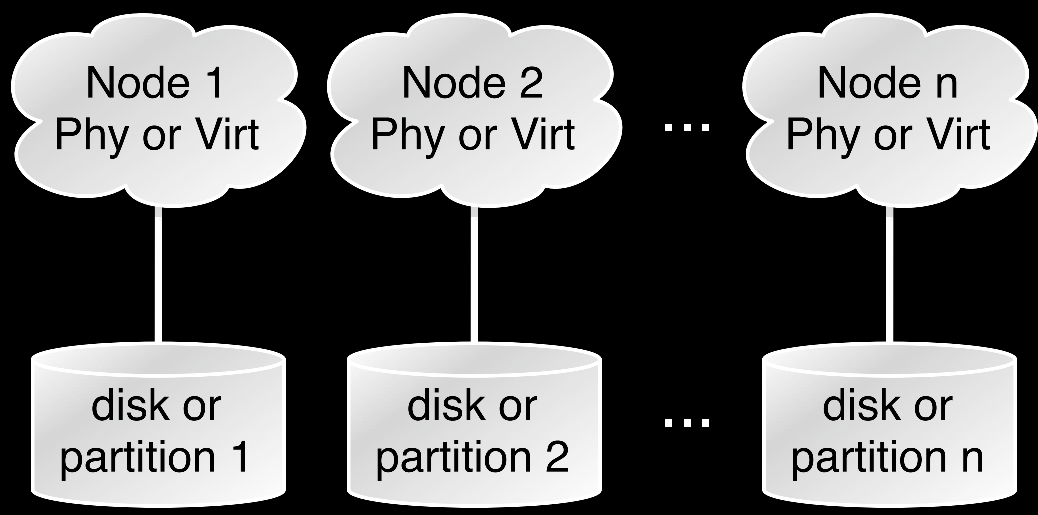
- How to install ... aster?
  - the proces ...
- Our solu ...
  - one of ... art the process, then be ...
- At the en ... e identical

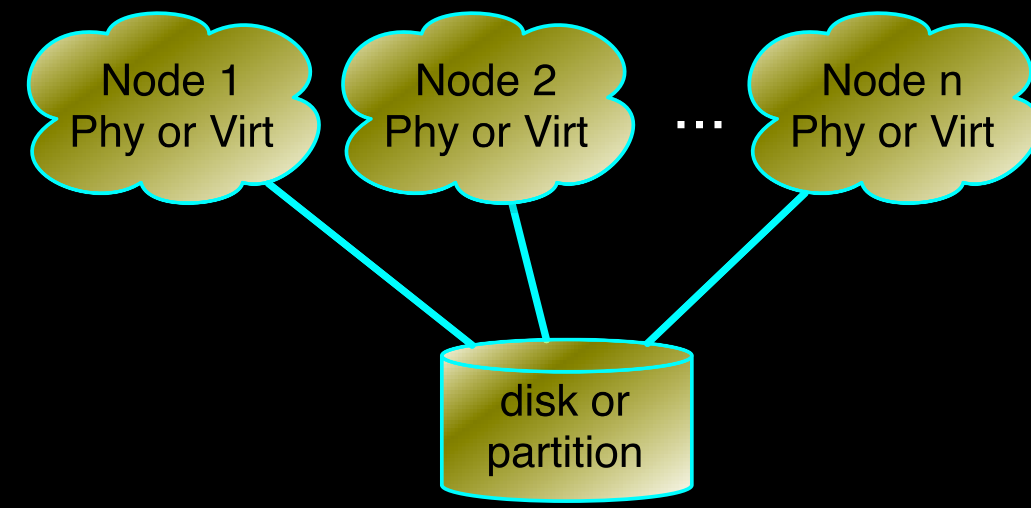


Storage virtualization

an efficient, constant time-space solution
for physical and virtual clusters

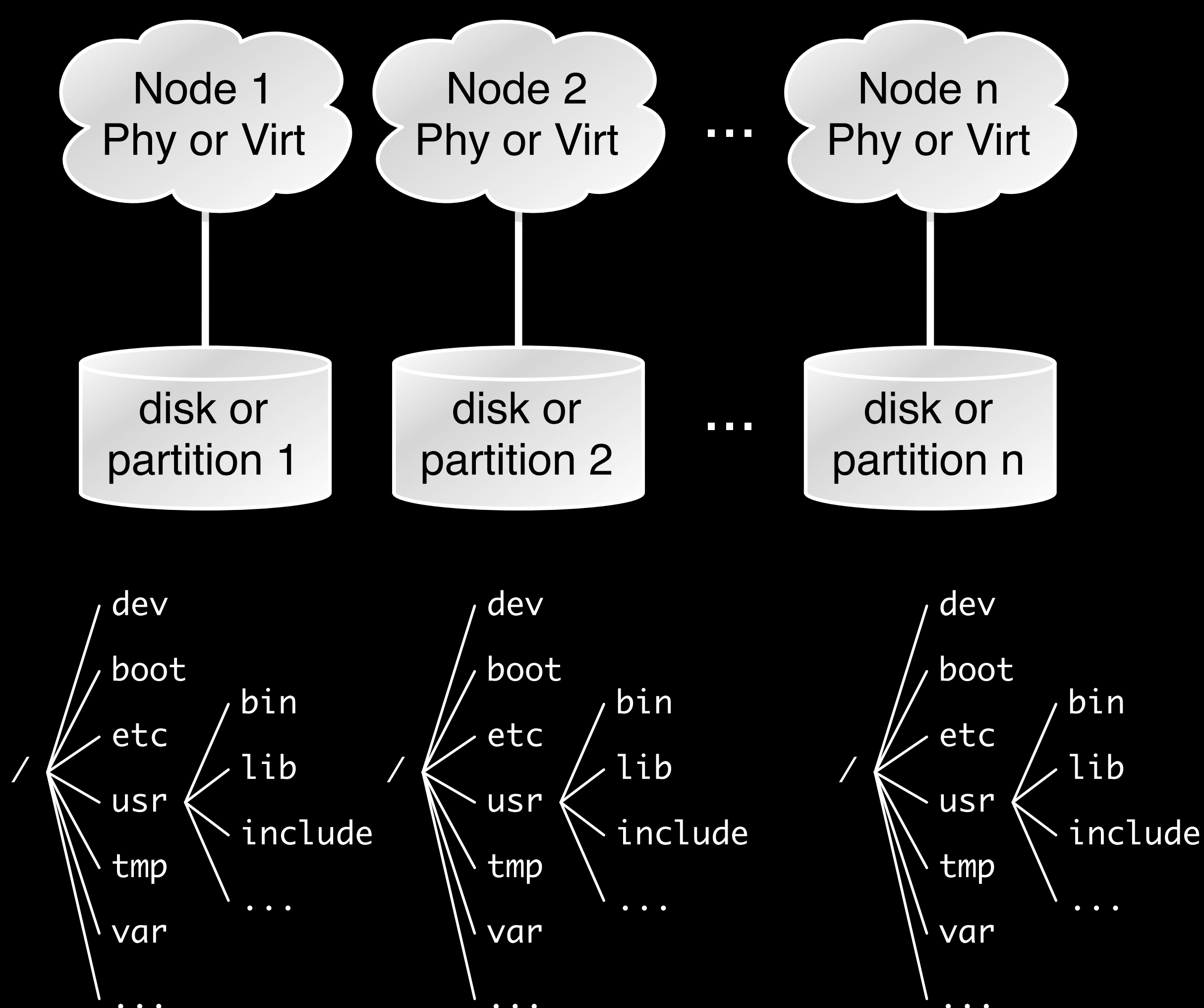
Single copy OR full replication?

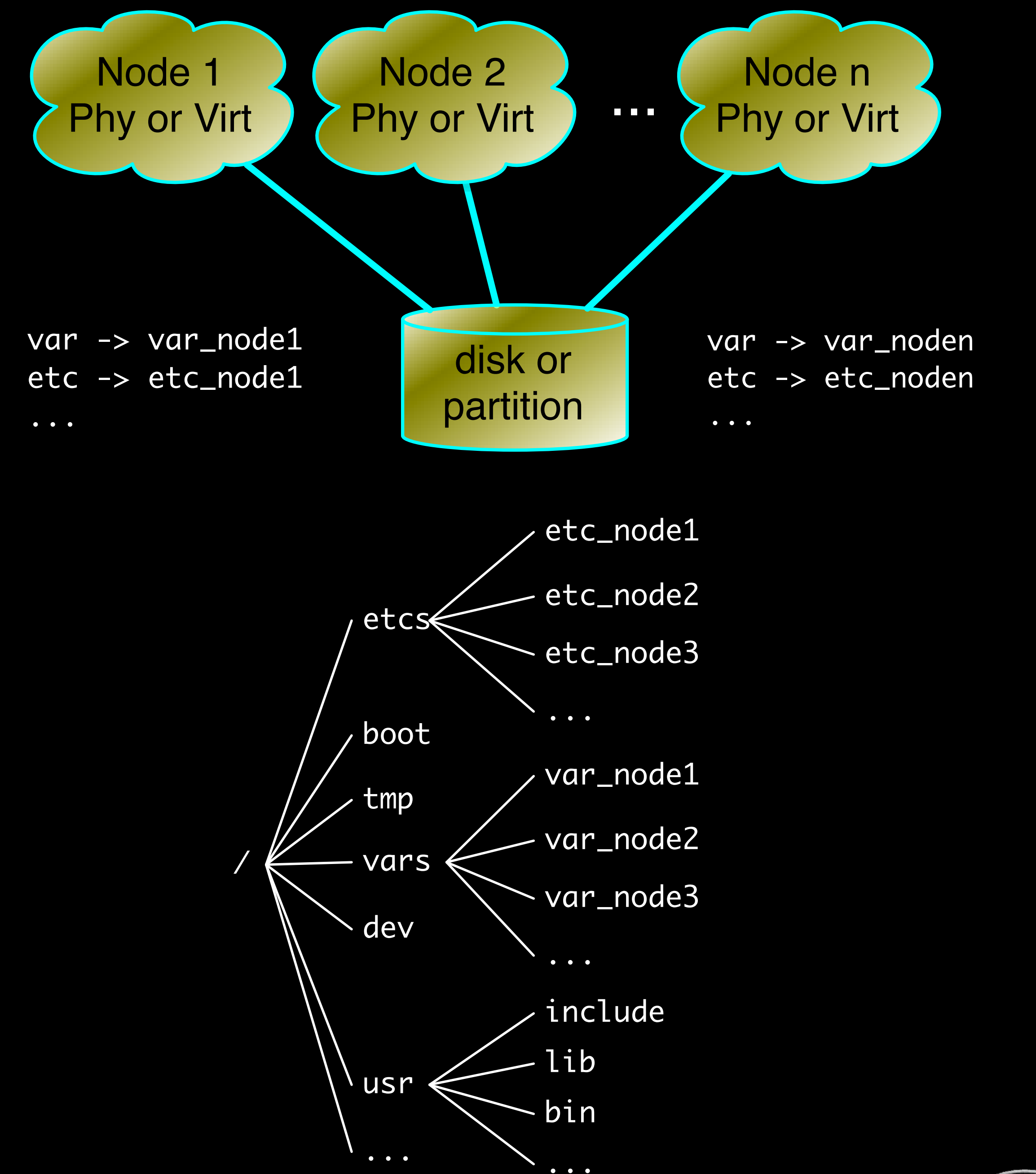




EUROTECH GROUP

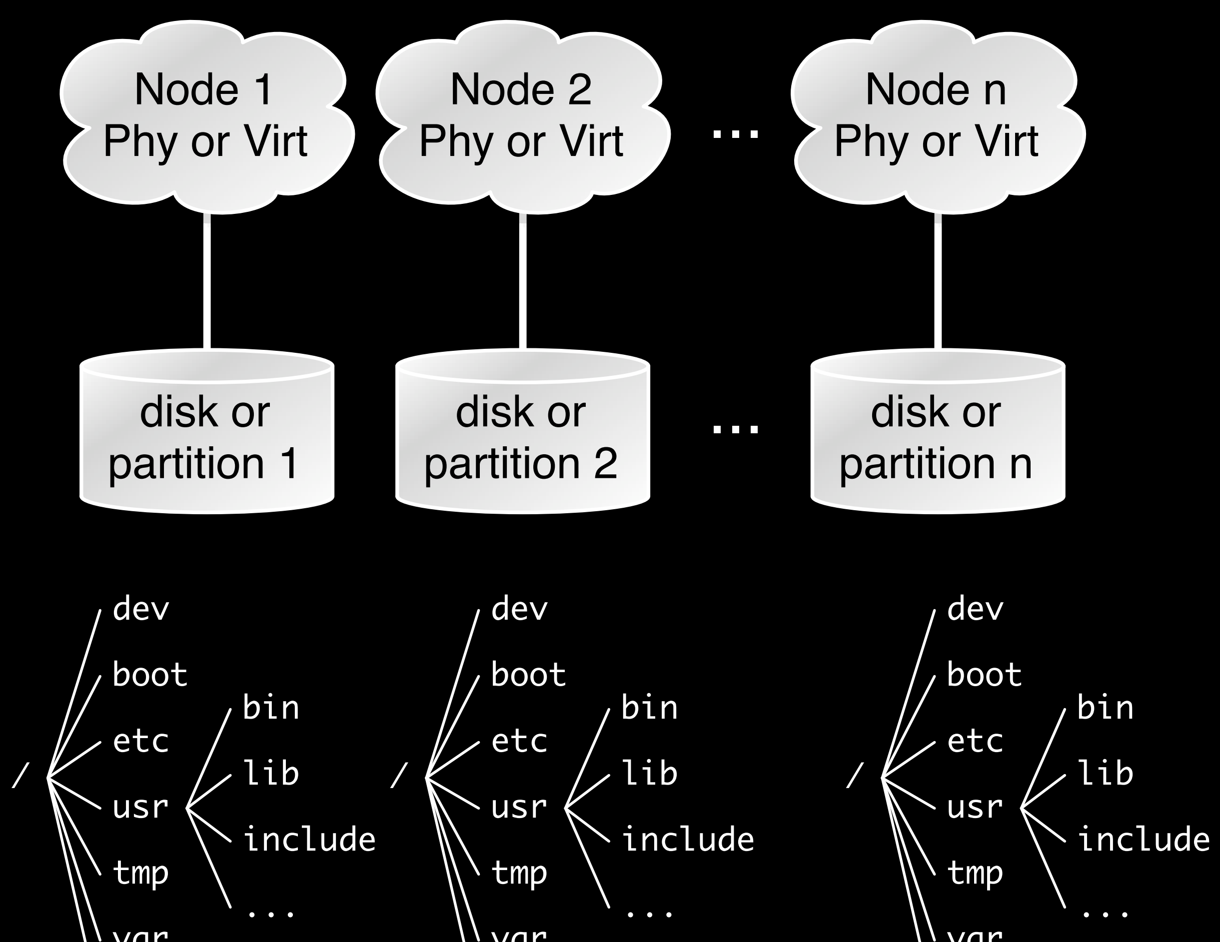
Single copy OR full replication?

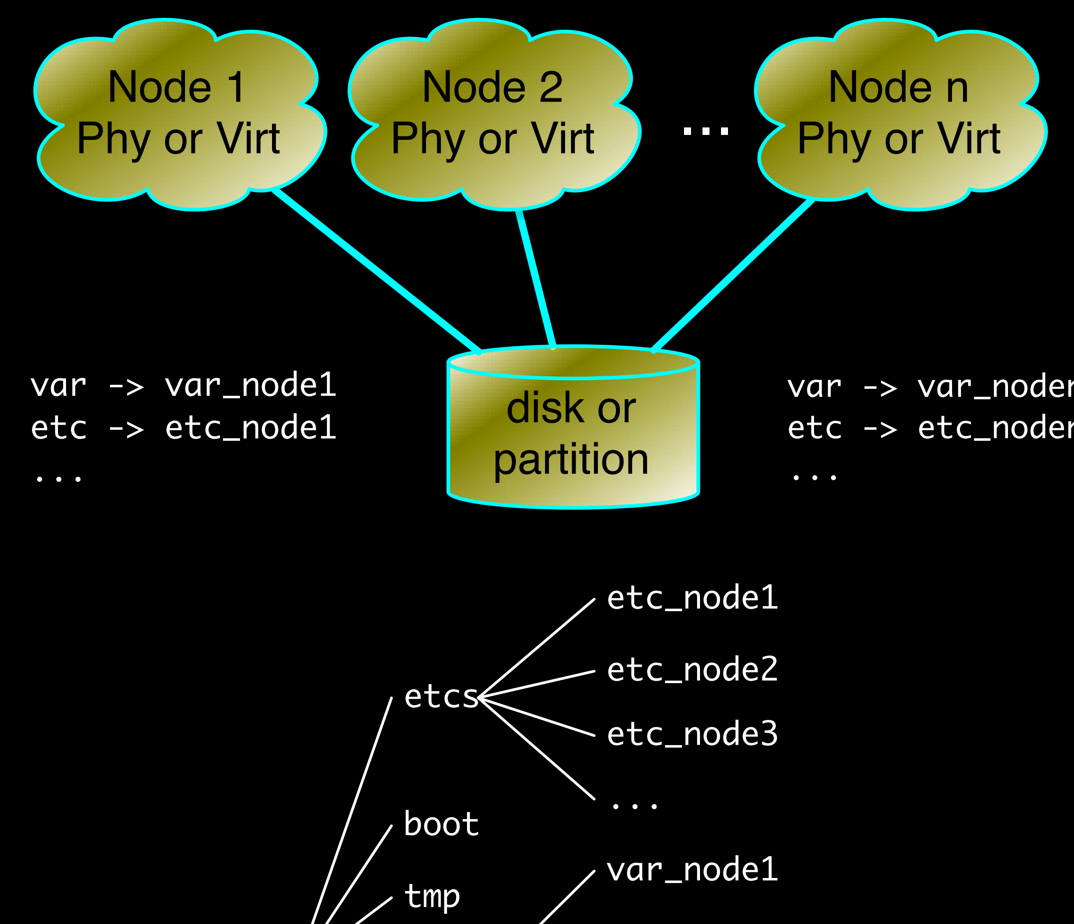




EUROTECH
GROUP

Single copy OR full replication?





- Each node (physical or virtual) has its own copy of the whole disk
- Transparent, easy to build and update
- OS does not need customization
- Inefficient in time and space - O(n*size). Identical OS files are replicated

- Each node (physical or virtual) share a disk (a file system, actually)
- Not transparent, complex to build and update
- OS does need customization
- Efficient in time and space - O(size). OS files are not replicated

Physical and virtual storage wish list

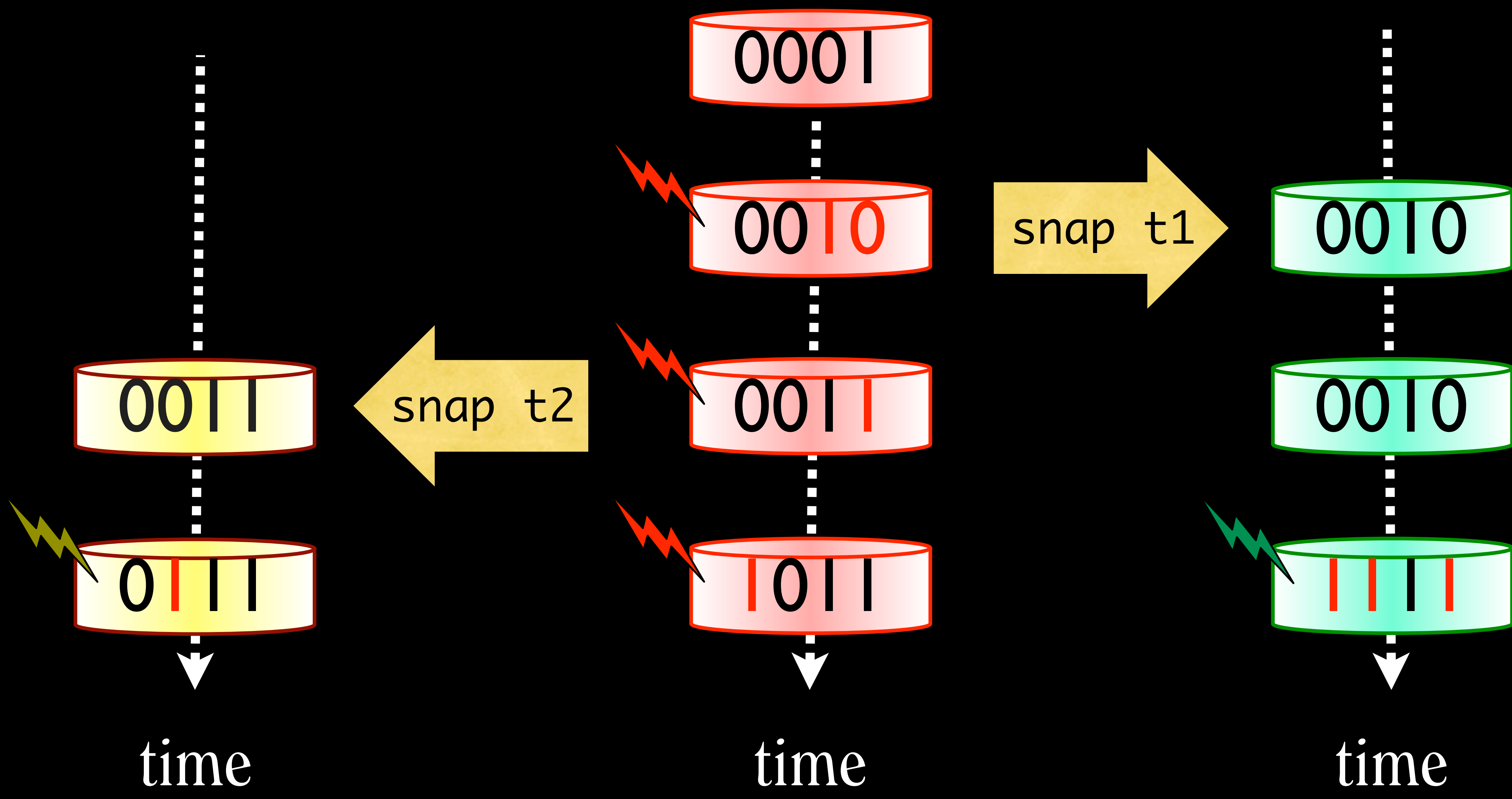
- Transparent, flexible, efficient (time and space)
  - independent from the running OS
  - creation/destruction of volumes should be dynamic
  - trivial solution not suitable
    - e.g. 50 nodes x 10 GB x 100MB/s = ~ 2 hours (optimistic forecast)
    - destroy system stability during operation due to high I/O pressure
    - e.g. 50 nodes x 5 GB = 100 GB just for OS files

EUROTECH GROUP

Peculiarities of VC storage

- The nodes of a VC are homogenous (same OS)
  - 99% of OS-related files are identical in all VMs
  - no reason to have heterogeneous nodes in VC, since we can have many heterogenous VCs
- Keep these files in single copy
  - the solution, to be transparent, should not exhibit this to the nodes (both physical and virtual)
  - can be done exploiting snapshots
    - can be hardly used “as is”

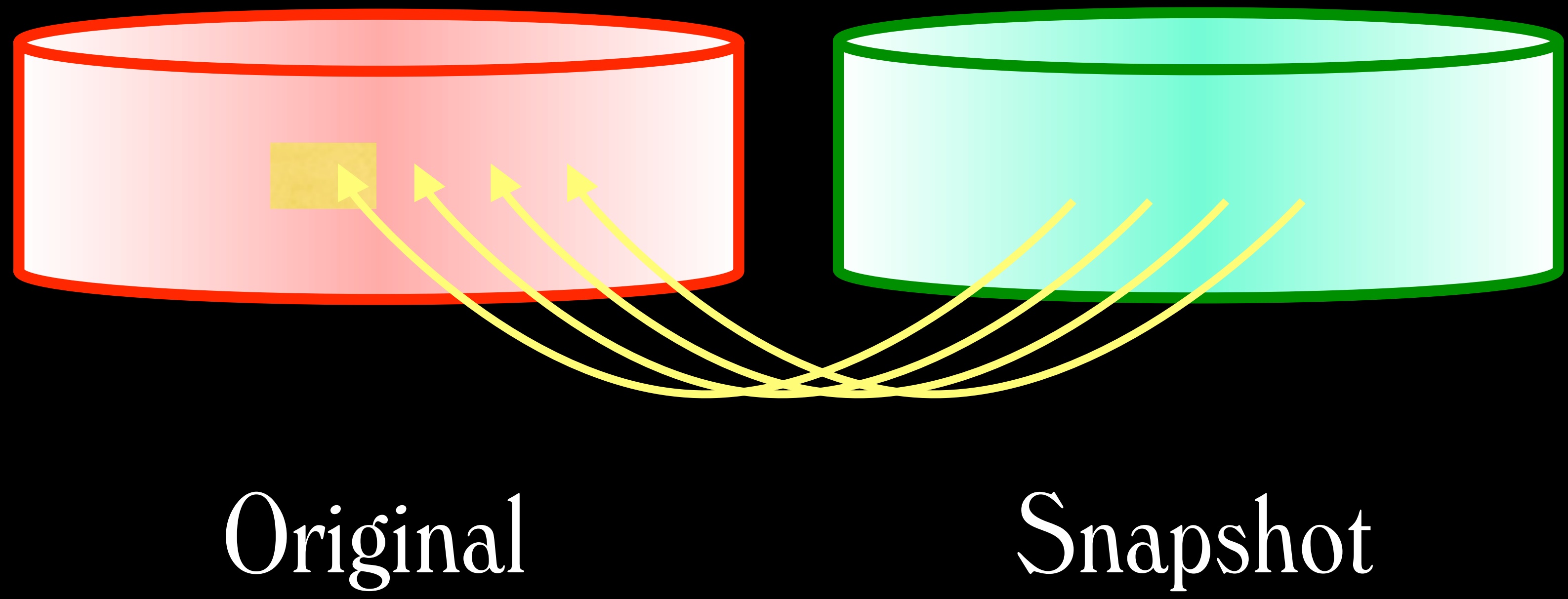
Understanding snapshots



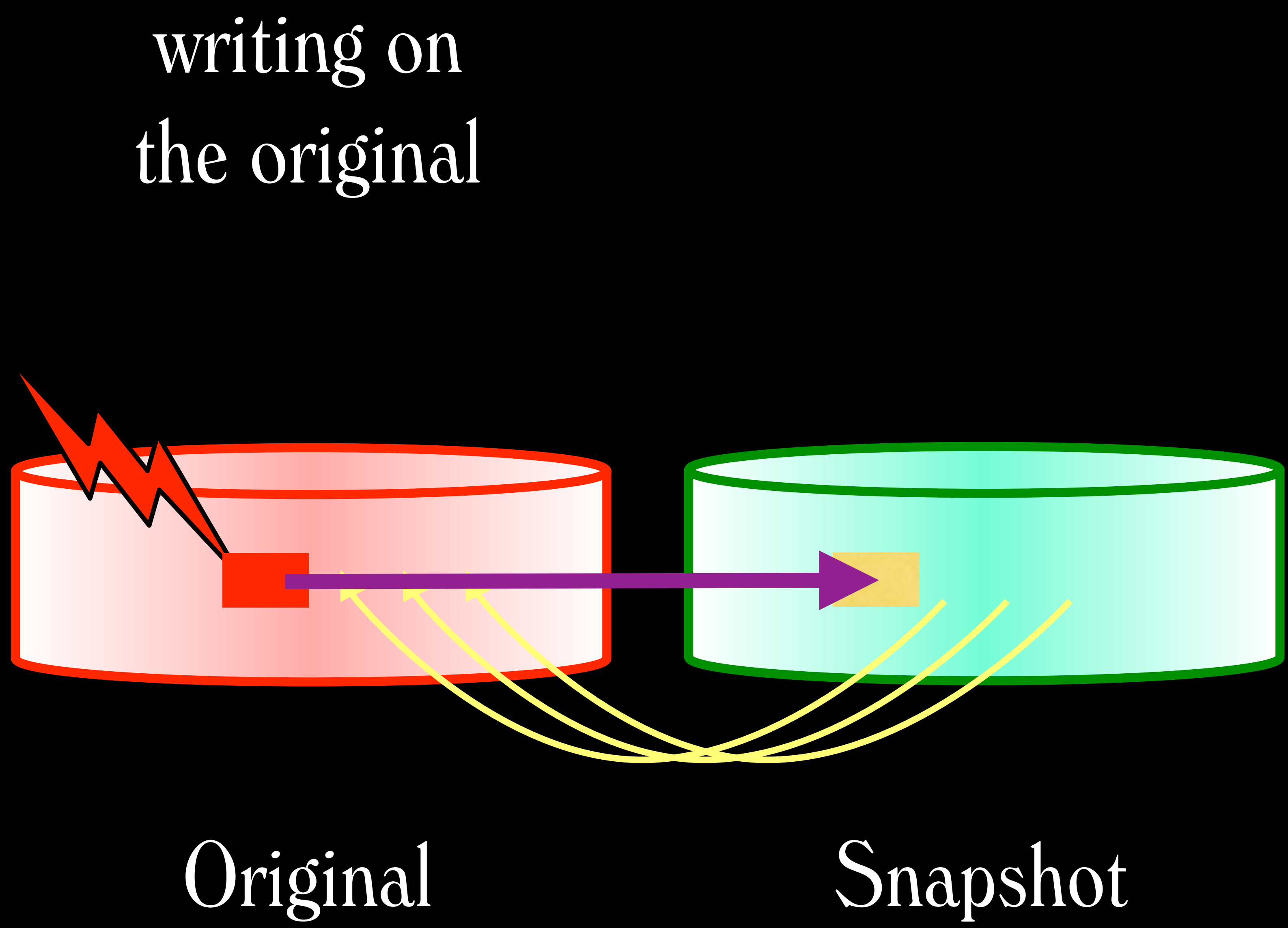
Snapshots

- Usually used for online backups
  - both original and snapshot can be written
  - FS-neutral, transparent
- Supported in standard tools (e.g. LVM, EVMS)
  - in Linux implemented via dm_snapshot module (device mapper)
  - used also in other systems, such as WinXP system recovery machinery
- Can be implemented in several ways
  - copy-on-write, redirect-on-write, split-mirror, ...
- They have been used also to store/share VM images
  - for a single machine, not for clusters ...

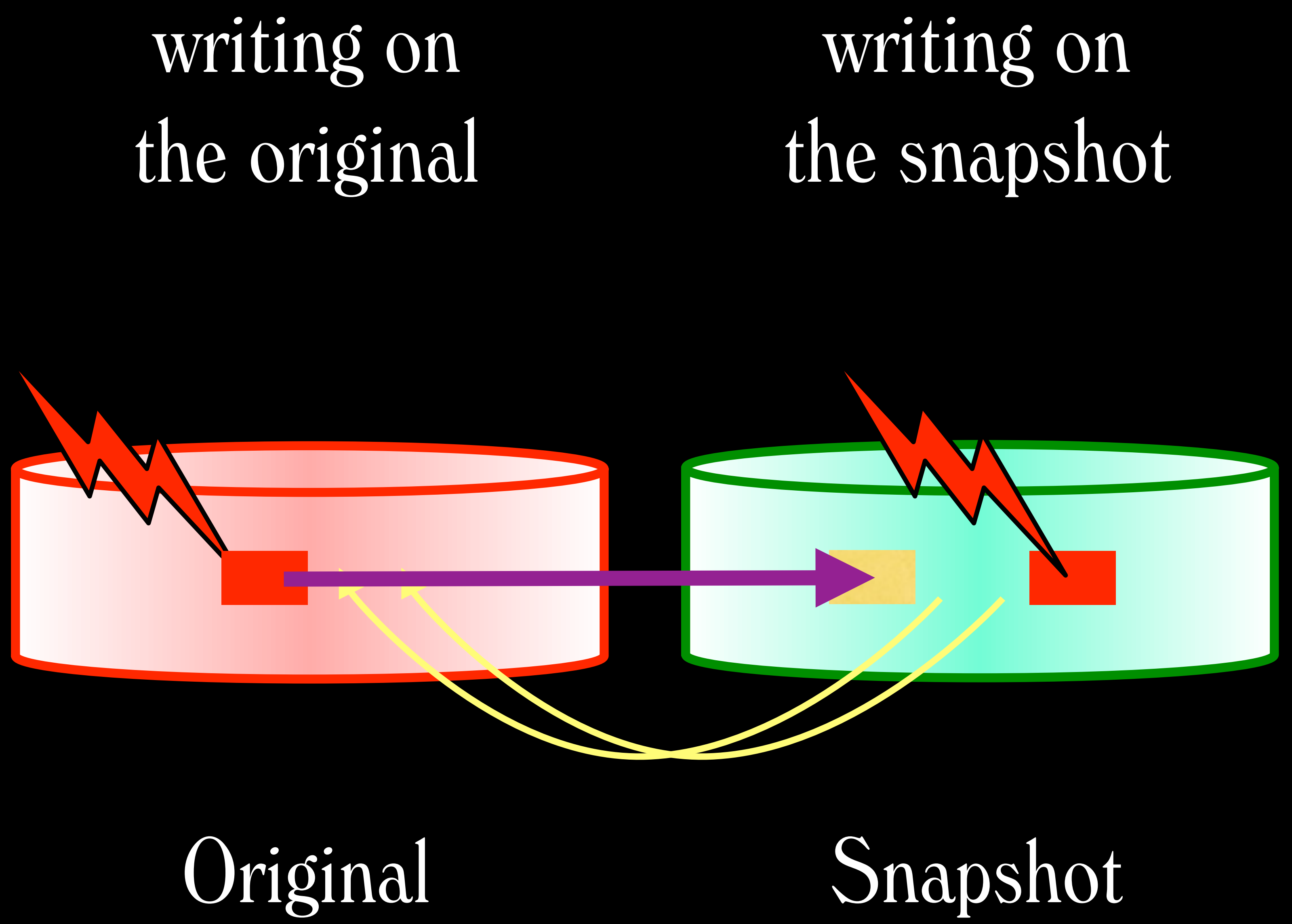
Copy-on-write



Copy-on-write



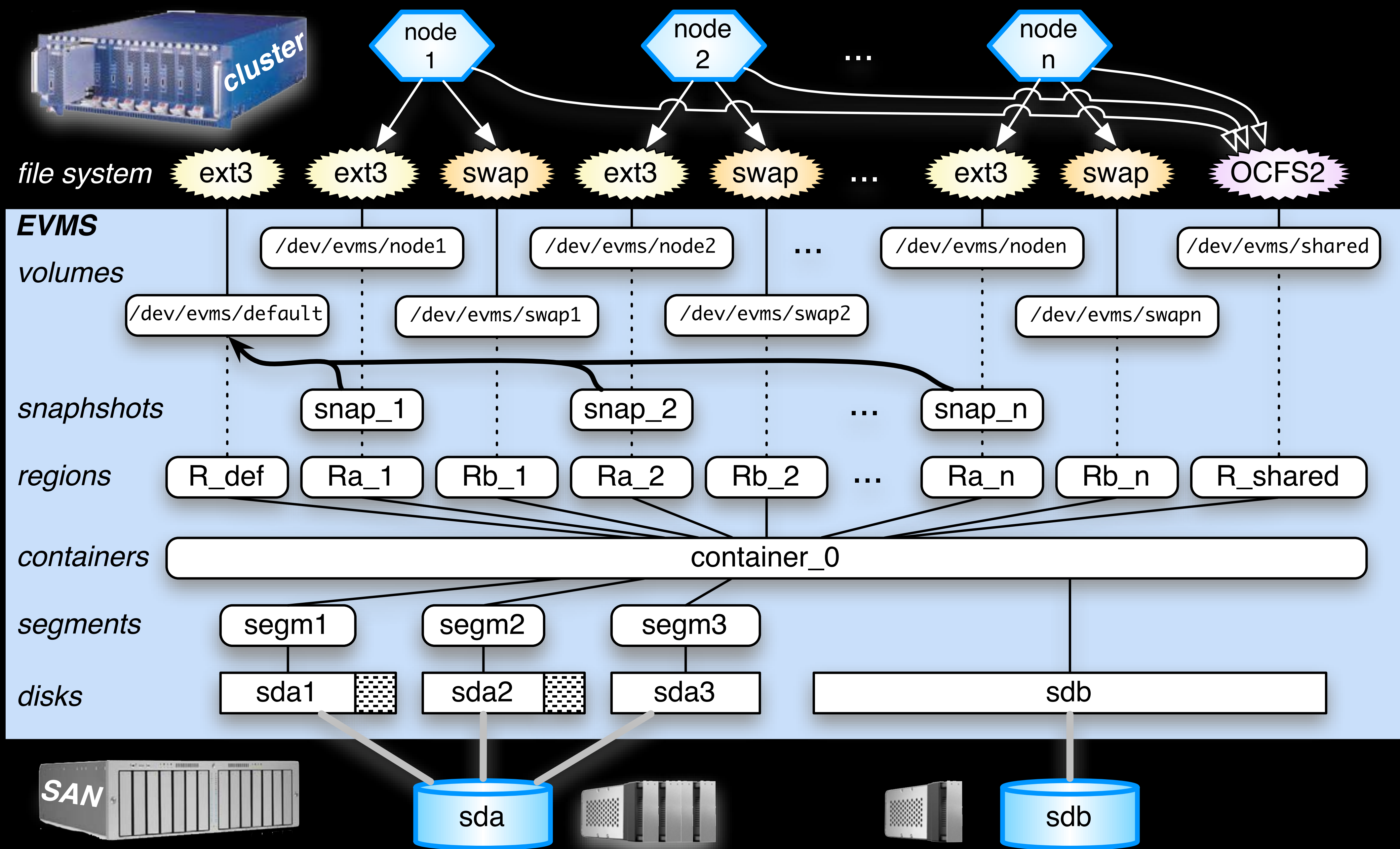
Copy-on-write



Concurrent Snapshots

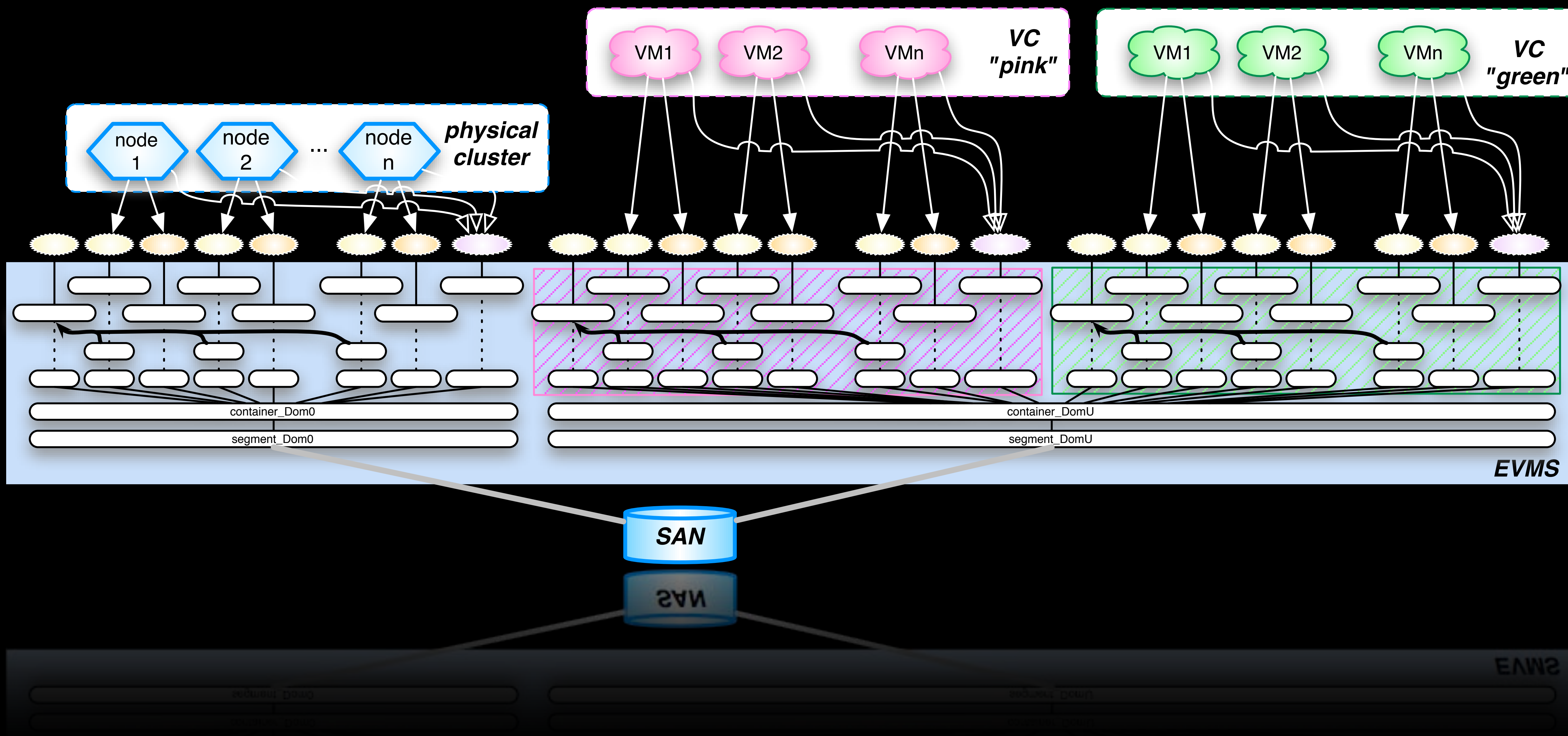
- Snapshots are not designed for parallel systems
  - if used with standard semantics all snapshots should remain active in all nodes (even those not accessing them)
  - they are buffered in kernel space, thus consume kernel memory
    - space linear in the number of snapshots system-wide
- VirtuaLinux introduces and uses a novel semantics
  - relax the standard semantics maintaining the correctness
    - mark as read-only the blocks that will not change in the original (e.g. OS files)
    - enable the deactivation of not used snapshots
      - correct provided the original is read-only
      - standard semantics cannot enforce it since has no way to mark as read-only
    - implemented as EVMS plugins, no kernel space changes

VirtuaLinux storage virtualization
(physical cluster)



VirtuaLinux storage virtualization

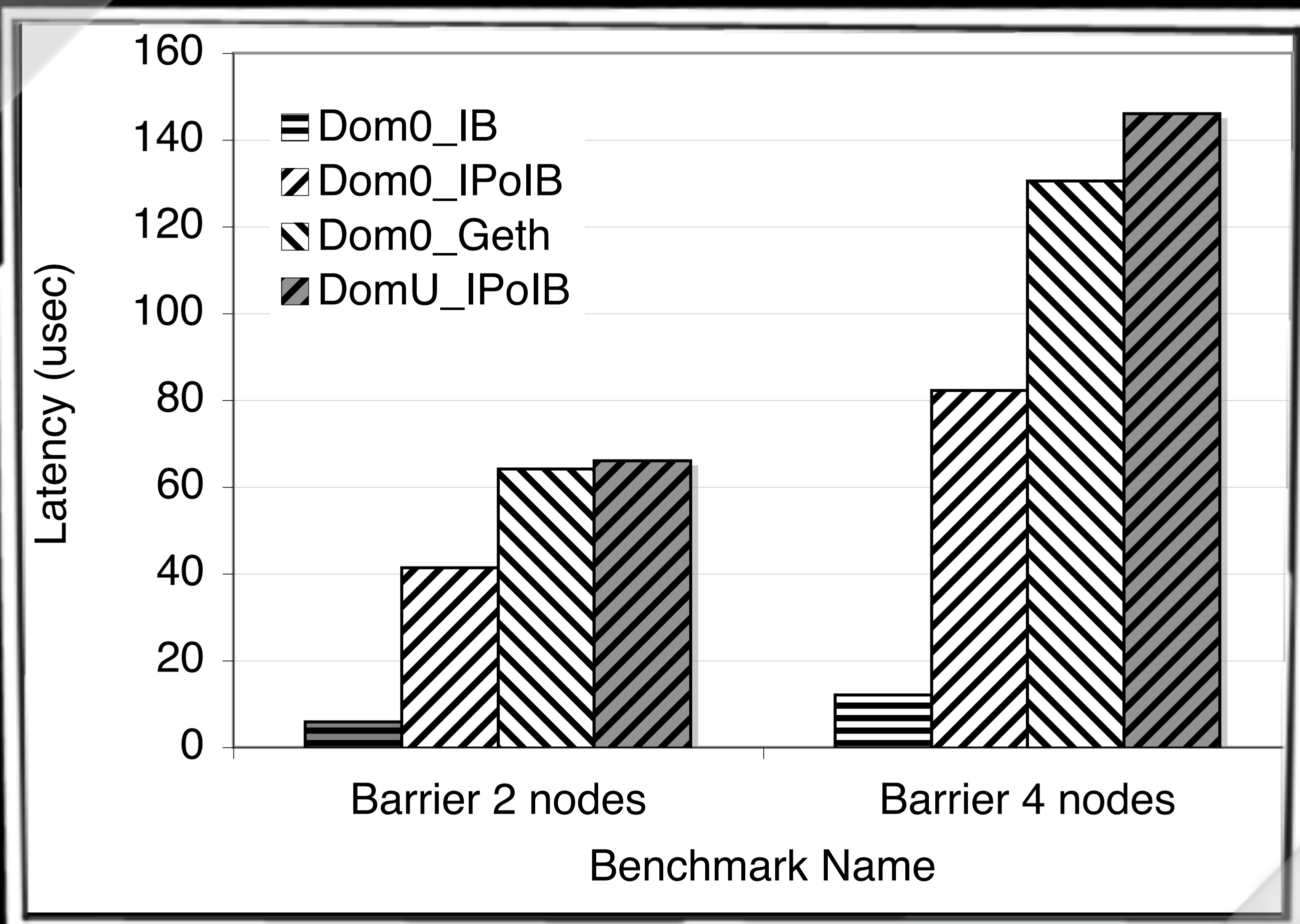
(virtual clusters)

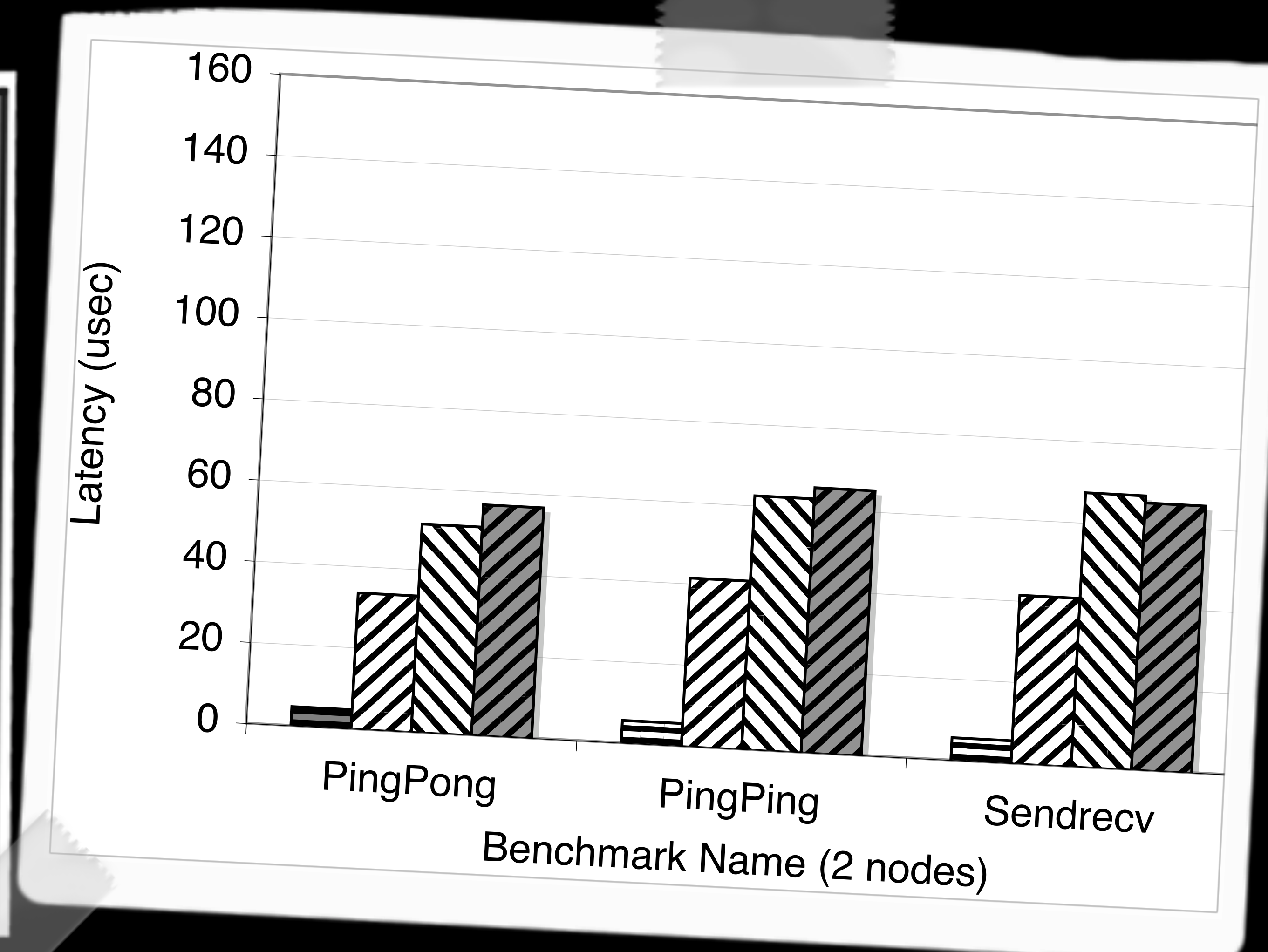


Experiments

Vcpu, Vnetwork and Vstorage overhead

Communication Latency

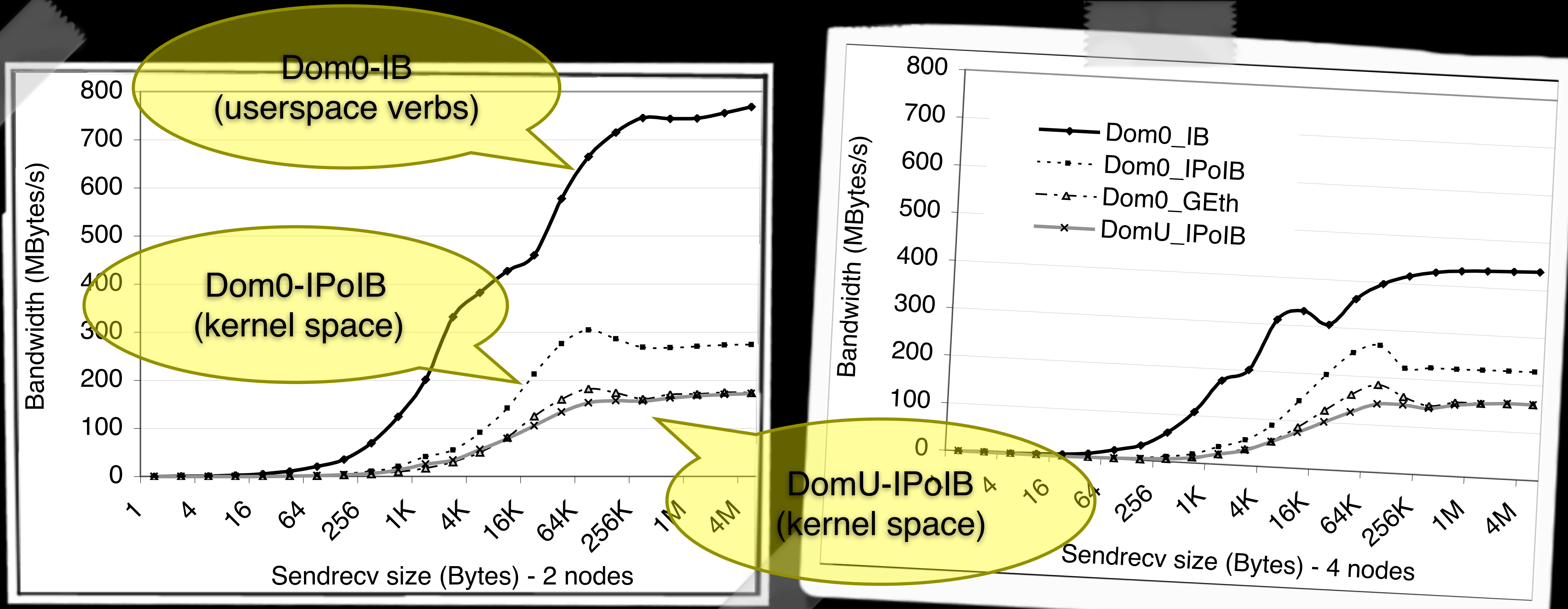




Dom0_IB	Ubuntu Dom0, Infiniband user-space verbs (MPI-gen2)
Dom0_IPoIB	Ubuntu Dom0, Infiniband IPoverIB (MPI-TCP)
Dom0_Geth	Ubuntu Dom0, Giga-Ethernet (MPI-TCP)
DomU_IPoIB	Ubuntu DomU, virtual net on top of Infiniband IPoverIB (MPI-TCP)

EUROTECH GROUP

Communication Bandwidth



Dom0_IB	Ubuntu Dom0, Infiniband user-space verbs (MPI-gen2)
Dom0_IPoIB	Ubuntu Dom0, Infiniband IPoverIB (MPI-TCP)
Dom0_Geth	Ubuntu Dom0, Giga-Ethernet (MPI-TCP)
DomU_IPoIB	Ubuntu DomU, virtual net on top of Infiniband IPoverIB (MPI-TCP)

EUROTECH GROUP

Performance (CPU & SO)

Micro-benchmark	Ub-Dom0 *vs* CentOS	Ub-DomU *vs* CentOS	Ub-DomU *vs* Ub-Dom0
Simple syscall	+667%	+726%	+7%
Simple open/close	+36%	+34%	-2%
Select on 500 tcp fd's	+51%	+51%	0%
Signal handler overhead	+112%	+127%	+7%
Protection fault	+246%	+293%	+13%
Pipe latency	+115%	+31%	-40%
Process fork+execve	+143%	+119%	-10%
float mul	∼0%	∼0%	∼0%
float div	∼0%	∼0%	∼0%
double mul	∼0%	∼0%	∼0%
double div	∼0%	∼0%	∼0%
RPC/udp latency localhost	+35%	-7%	-31%
RPC/tcp latency localhost	+35%	-5%	-30%
TCP/IP conn. to localhost	+32%	+3%	-22%
Pipe bandwidth	-38%	+51%	+144%

EUROTECH GROUP

Virtual Storage Performance

Additional layer on top of iSCSI	read	write	rewrite
none (reference raw iSCSI access)	60	88	30
EVMS standard volume	66	89	32
EVMS snap, fresh files	63	88	31
EVMS snap, files existing on original	**63**	**7**	**31**

VirtuaLinux 1.1 (multi tier)

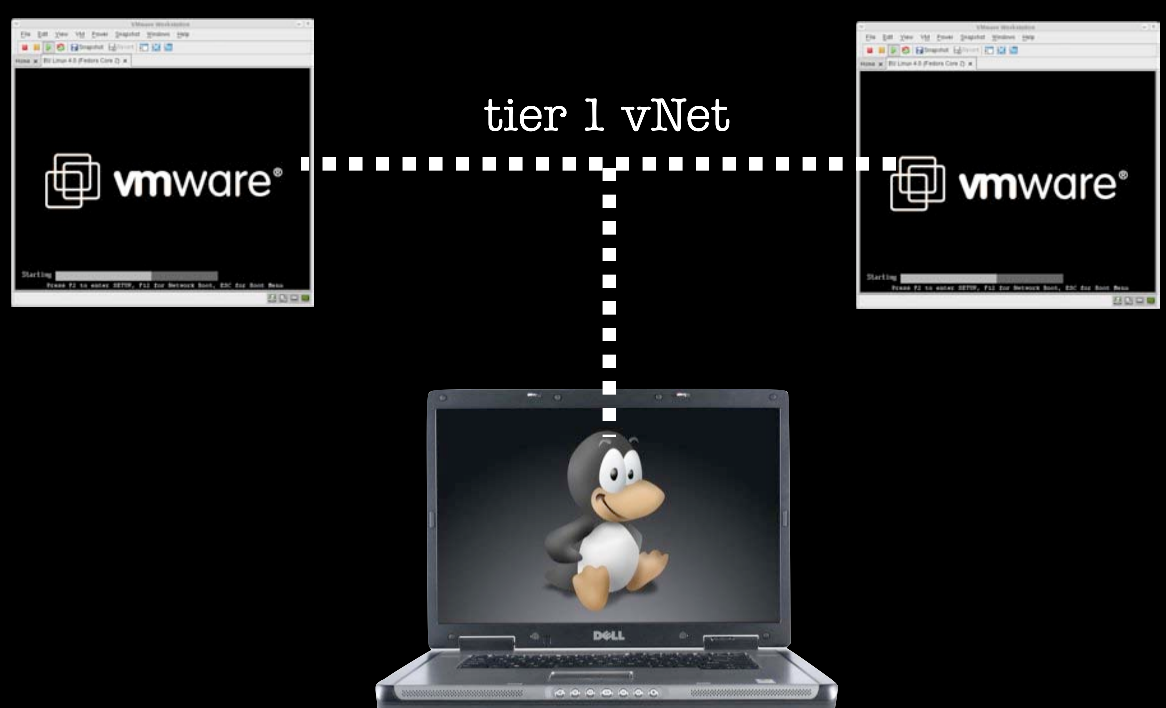
- Based on Ubuntu, kernel 2.6.19-4 (gutsy)
- Designed to support our developers working at home
  - did you have at home a cluster with a SAN?
  - the cluster is simulated by yet another level of virtualization (binary translator, e.g. VMware)
- Three-tiers (two of them virtualized)
  - tier 0, standard linux - simulates the SAN (iscsi-target)
  - tier 1, macchine VMware - simulates the physical cluster
  - tier 2, macchine Xen - nodes of virtual clusters
- Slooooow, but still, it makes the development possible
  - can be used for demo of parallel apps in conferences
    - if you have at least a 64bit core2duo Intel (≥Merom) laptop (I haven't, sorry)

EUROTECH GROUP

VirtuaLinux many-tier

tier 0
laptop standard linux
+ iSCSI target

VirtuaLinux many-tier



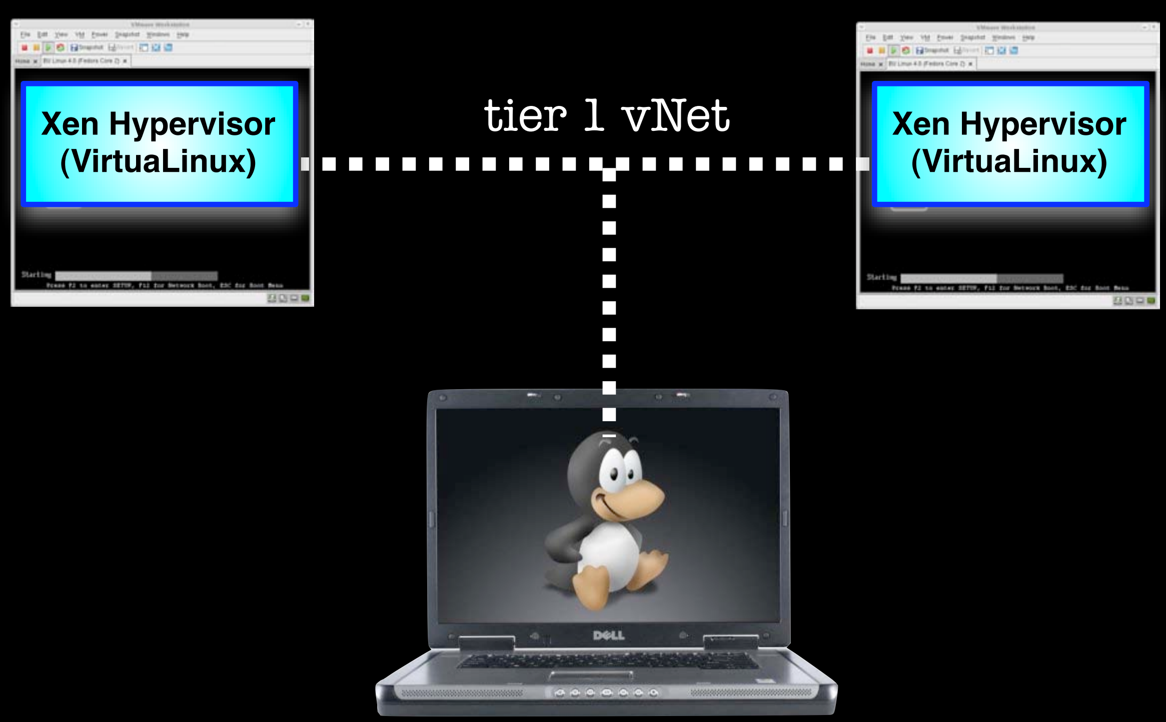
tier 1

virtual cluster
of VM diskless
(WMware o QEMU)

tier 0

laptop standard linux
+ iSCSI target

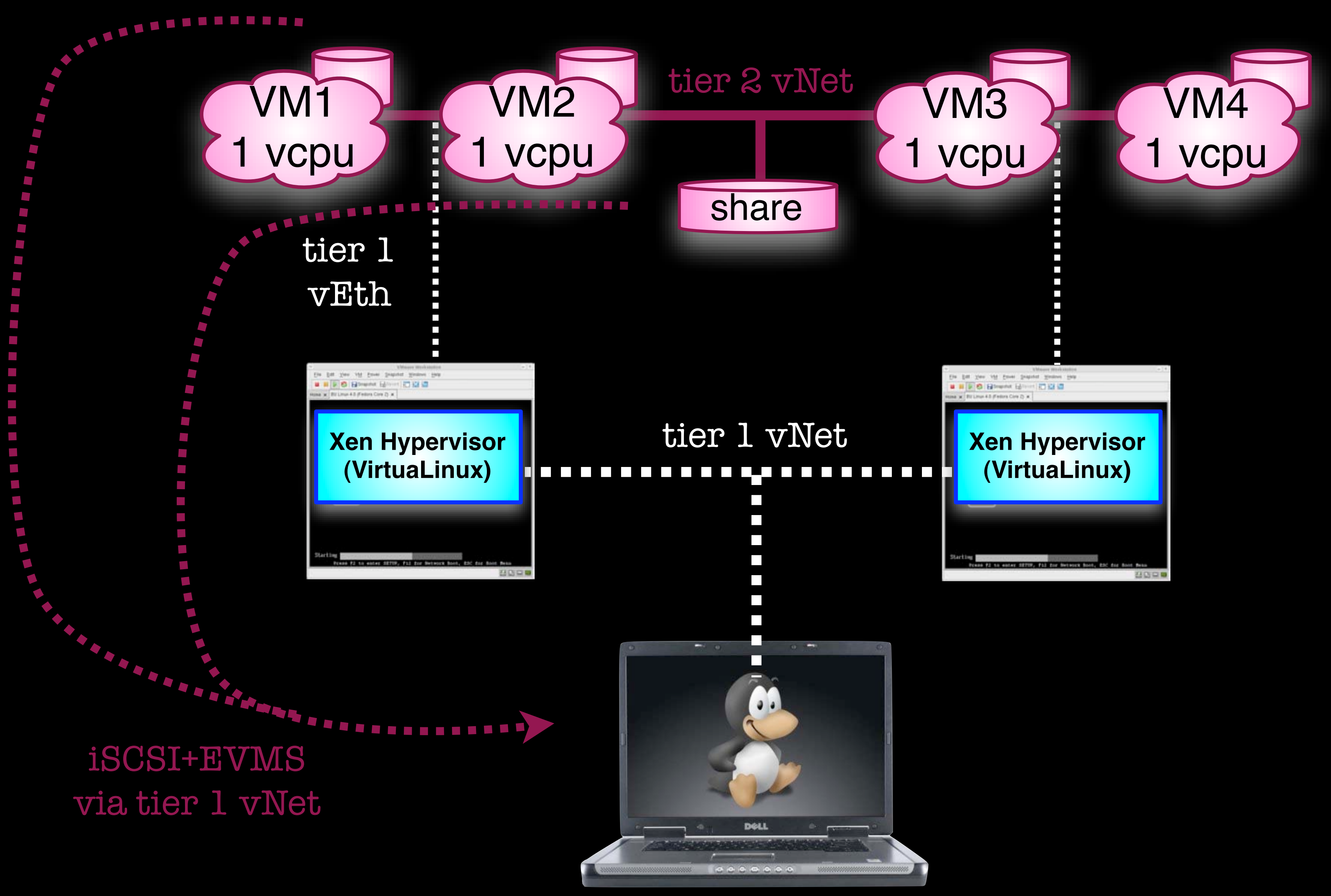
VirtuaLinux many-tier



tier 1
virtual cluster
of VM diskless
(WMware o QEMU)

tier 0
laptop standard linux
+ iSCSI target

VirtuaLinux many-tier



tier 2
VirtuaLinux
with VC (Xen)

tier 1
virtual cluster
of VM diskless
(WMware o QEMU)

tier 0
laptop standard linux
+ iSCSI target

Conclusion

- Focuses on HPC cluster for industry needs
  - reduce install and maintenance costs
  - makes it possible the consolidation and sharing
  - prevent the destruction of installation at deployment site due to weird administrator actions
- Some scientific results
  - some advance in storage virtualization
    - comparable VMware Lab Manager (not opensource)
    - performance (sometime) better than non-virtualized storage
- Some industrial results
  - currently deployed on shipped Eurotech HPC clusters
    - subsets of the whole system
  - Graphic version of VC management tools not opensource

EUROTECH GROUP

We would like to acknowledge Eurotech S.p.A. Italy and
University of Pisa for the financial support and hardware

and

all people who contributed to the development, and in particular
aldinuc, califfo, gervystar, gobex, massimot, monica_d, patton73, pierfrancesco, spinatel

Virtualinux is opensource under GPL and it is meant to be
a continually evolving experimentation framework, thus
Please do not hesitate to contact me if you would like
propose new ideas or to participate to develop our own

email: aldinuc@di.unipi.it
http://www.di.unipi.it/~aldinuc