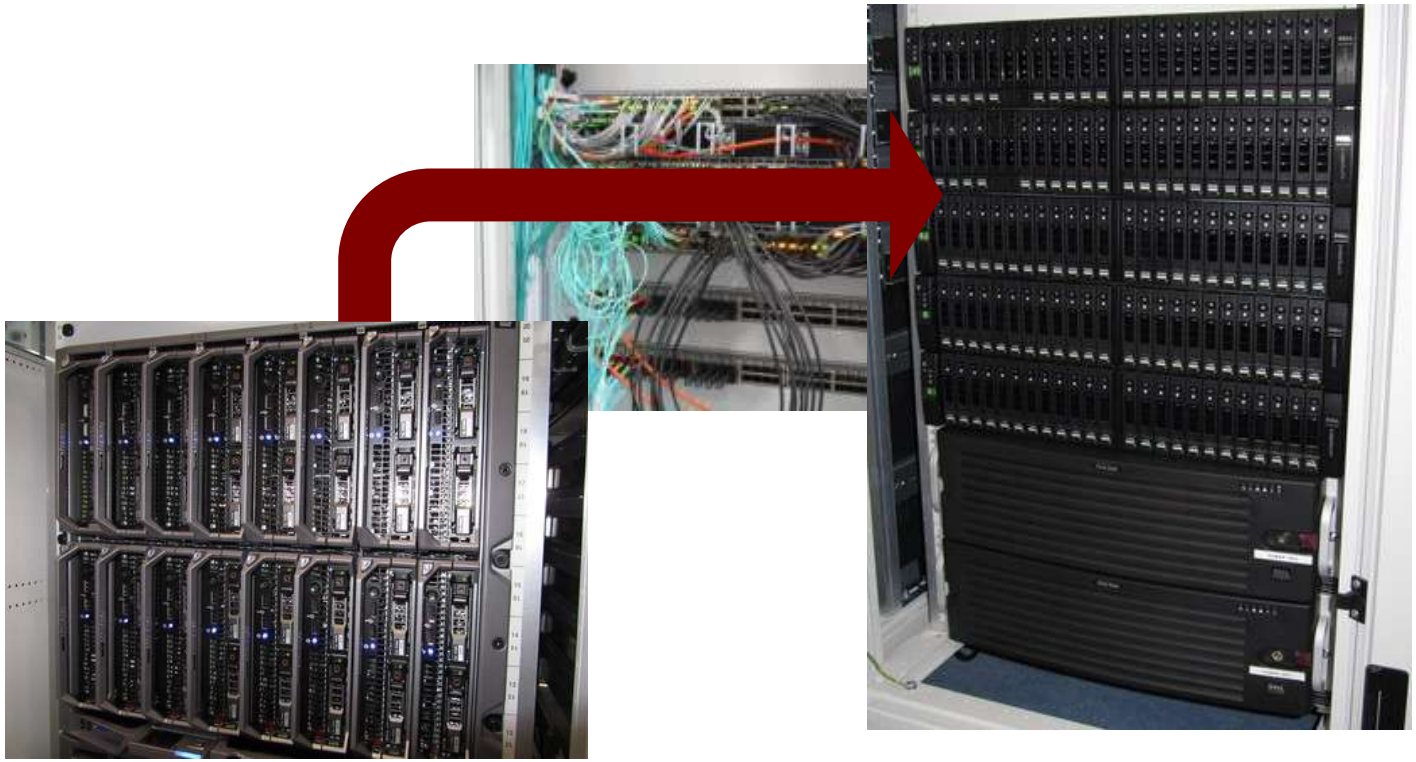
Getting Started with Sint & Piet

Status as per 4 April 2012



David Groep
Nikhef
Amsterdam
PDP & Grid

Elements of NDPF Service Cluster

- GSP Piet "MTDR24": 16 blade servers
- Sint SAN storage "MTDR23":50 TByte
- Existing set of Generics 2008/A
  - few of these will become a security VM cluster
  - 1-2 others become test systems
- Existing blades "bl0"
  - 1-5, 13,14 will be upgraded to dual FC SAN
  - 10G stacked switches replace pass-through (Thu.)
- Installnet switch (still now: deelmlx)
  - Seperate installation network ve11 @1+10G
  - IPMI management LAN ve4 for managementonly

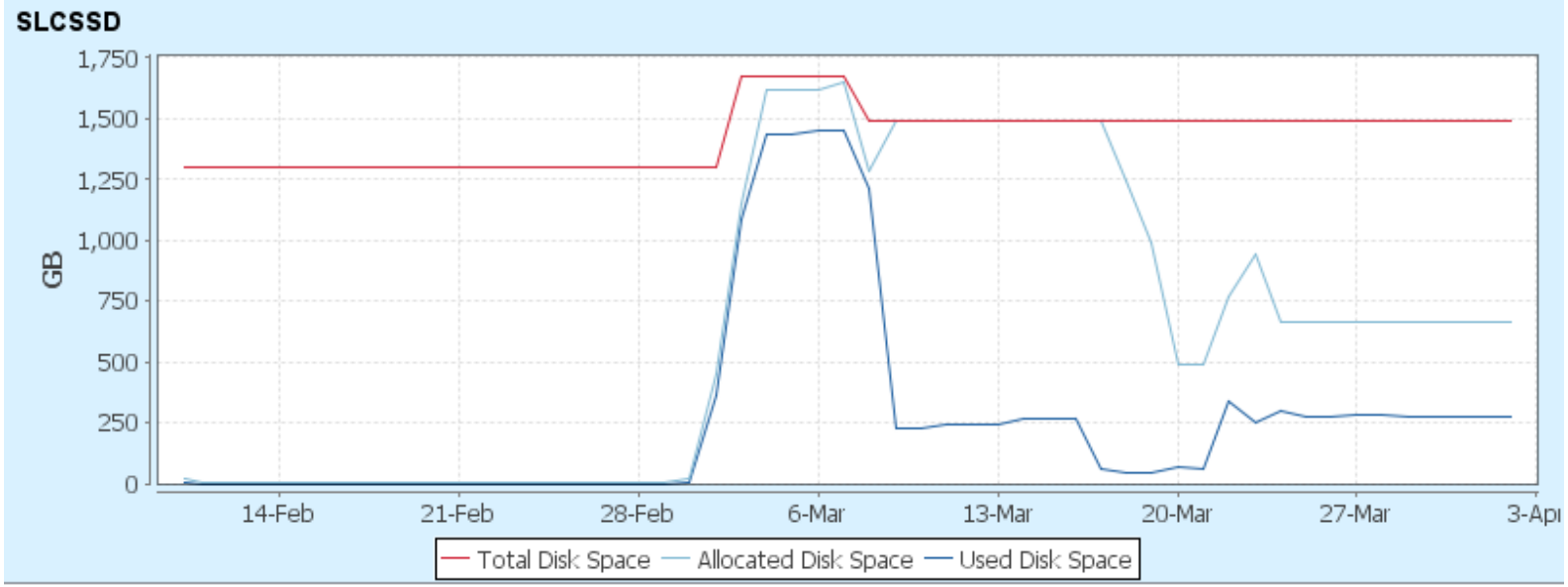
David Groep
Nikhef
Amsterdam
PDP & Grid

Storage

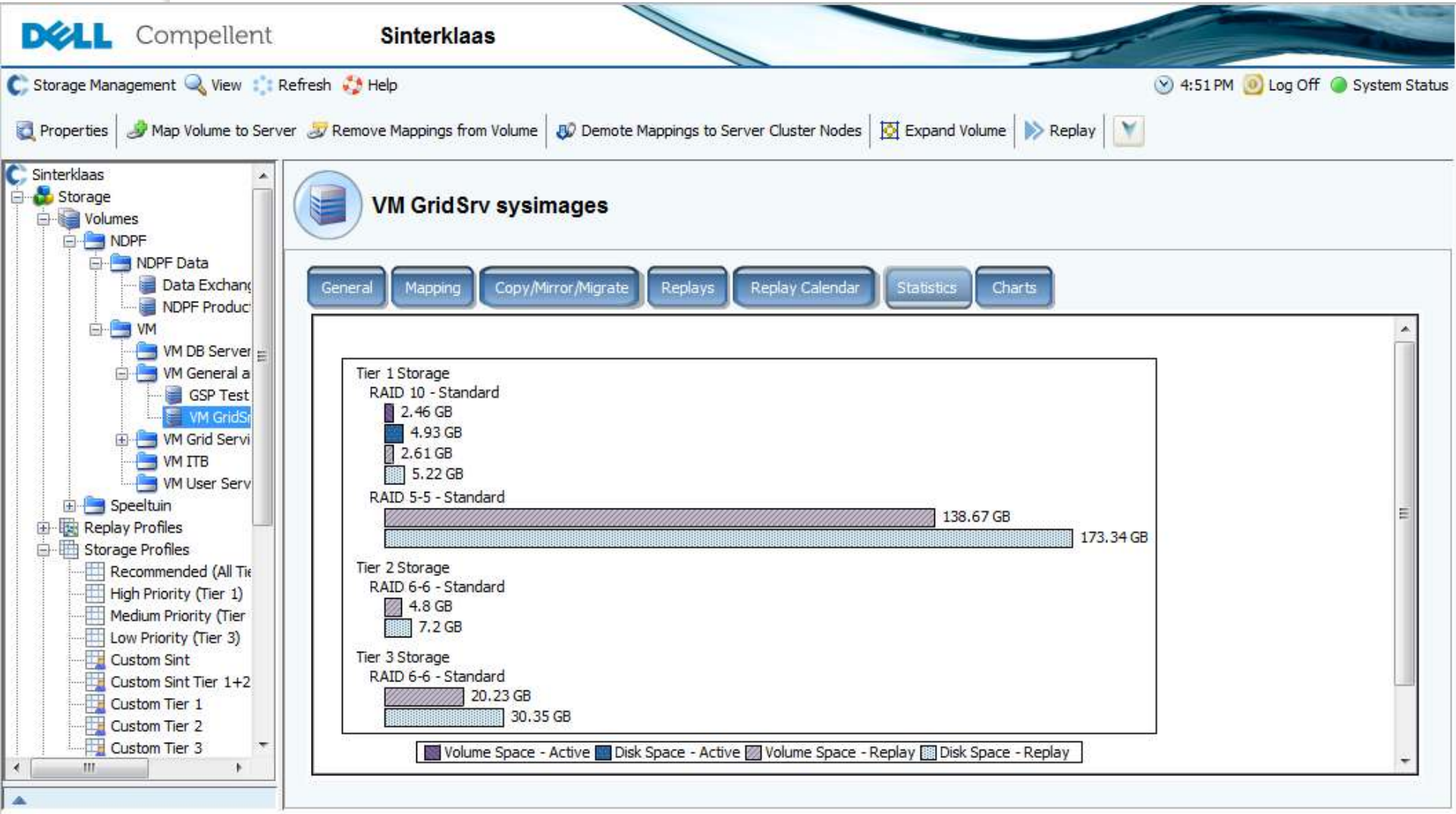
- Dell Compellent Storage Center (6.0.3)
  - sint.ipmi.nikhef.nl collective management
  - username: Admin
- 78 TByte gross capacity
  - auto-tiering ESSD, 10kRPMSAS, 7k2RPM N/L-SAS
  - Default RAID 6-6 on Tier 2&3
  - Default RAID-10 and 5-5 on Tier 1
  - Effective capacity ~ 50 TByte

David Groep
Nikhef
Amsterdam
PDP & Grid

Auto-tiering effect on used storage



David Groep
Nikhef
Amsterdam
PDP & Grid


DELL Compellent
Sinterklaas
Storage Management
View
Refresh
Help
4:51 PM
Log Off
System Status
Properties
Map Volume to Server
Remove Mappings from Volume
Demote Mappings to Server Cluster Nodes
Expand Volume
Replay
Sinterklaas
Storage
Volumes
NDPF
NDPF Data
Data Exchang
NDPF Produc
VM
VM DB Server
VM General a
GSP Test
VM GridSr
VM Grid Servi
VM ITB
VM User Serv
Speeltuin
Replay Profiles
Storage Profiles
Recommended (All Tie
High Priority (Tier 1)
Medium Priority (Tier
Low Priority (Tier 3)
Custom Sint
Custom Sint Tier 1+2
Custom Tier 1
Custom Tier 2
Custom Tier 3
VM GridSrv sysimages
General
Mapping
Copy/Mirror/Migrate
Replays
Replay Calendar
Statistics
Charts
Tier 1 Storage
RAID 10 - Standard
2.46 GB
4.93 GB
2.61 GB
5.22 GB
RAID 5-5 - Standard
138.67 GB
173.34 GB
Tier 2 Storage
RAID 6-6 - Standard
4.8 GB
7.2 GB
Tier 3 Storage
RAID 6-6 - Standard
20.23 GB
30.35 GB
Volume Space - Active
Disk Space - Active
Volume Space - Replay
Disk Space - Replay

NIKHEF

Caveats

- Tier migration only works on replay volumes
  - If you do not configure replay, data stays in primary storage most, usually on Tier-1
- Choose a 'custom' template for all volumes
  - for most volumes 'Custom Sint' is OK
  - databases and high-IOPS: 'Custom Sint Tier 1+2 HS'

David Groep
Nikhef
Amsterdam
PDP & Grid

LUNs for VM hosting

- You can (and should) host more than one VM system disk on a LUN
  - But not too many since performance (IO queues) are per-LUN
  - Allocation is sparse, so '0's take no space
  - On re-writing old VM images, space is not reclaimed
- Put transient VMs in a dedicated pool LUN
  - LUN can be removed after completion
- Guideline: 10-15 VMs per LUN
- Separate LUNs for Data, databases and $HOME

Assigning LUNs to hosts

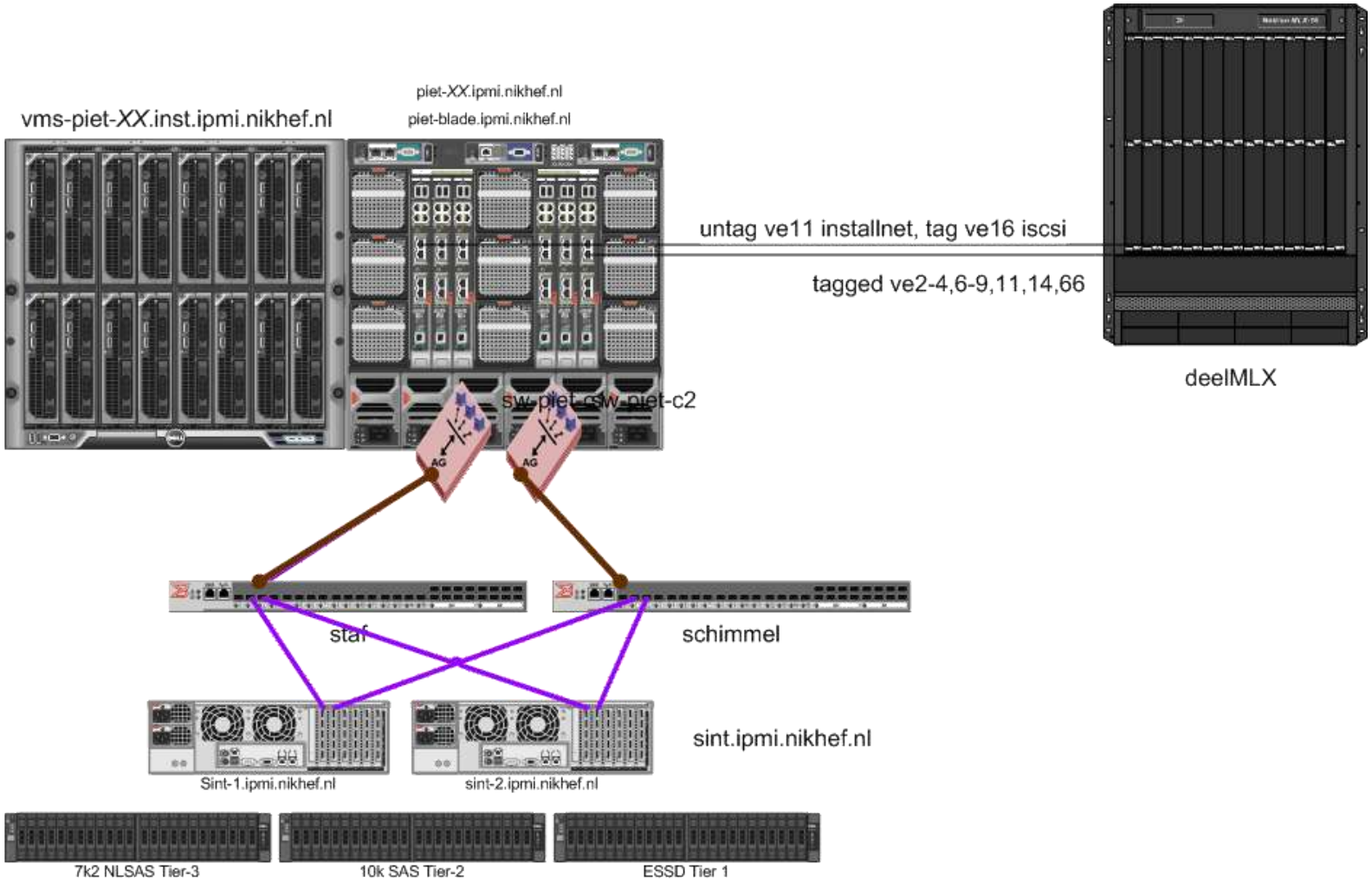
- Xen clusters are *Server Clusters* which share use of each LUN
  - this 'server cluster' must be defined as such on the Compellent, or Sint will warn for inconsistencies
  - XCP/XenServer "Piet" with the dual FC paths uses EL5/Xen5.x multi-path "MP" IO
  - Generics 2008/A iSCSI only has a single path, so uses XenServer 5.x IO
- You *cannot* attach LUNs to a single server in a XCP cluster
- You *cannot* share LUNs without using LVM

David Groep
Nikhef
Amsterdam
PDP & Grid

Linking storage and compute

- We have a dual-redundant FC mash-up
- Key FC concepts:
  - "WWN": world-wide name, identifies an FC end-point or port on a FC card
  - "zone": a group of server end-points that can see eachother
  - 'soft zoning': a zone based on WWNs
  - 'hard zoning': zone based on physical ports
  - "alias": a friendly name for a WWN
  - license: *everything* is licensed!
- FC: switches v.s. 'access gateways'

David Groep
Nikhef
Amsterdam
PDP & Grid



David Groep
Nikhef
Amsterdam
PDP & Grid

Staf and Schimmel

- Brocase SAN switches containing zone def
  - two independent switches for each path
  - cross-connect to the Compellent
  - server ports (and AGw's) connect to one of them, and servers have two FC ports for MPIO
  - effectively, each host sees a LUN 4 times

- Zones:
  - Z_Sint_Piet01: contains both "Sint"s and all Piet's
  - Z_Sint_R710: contains both "Sint"s and Achtbaan

David Groep
Nikhef
Amsterdam
PDP & Grid

Useful commands

- alicreate "aliName","member[; member...]“
  - name aliases after hostname and switch fabric
  - alicreate “A_Piet03_C1” “20:01:24:b6:fd:be:23:e3” *on switch staf.ipmi.nikhef.nl*
- zoneadd "zoneName", "member[; member]“
  - add named aliases (please!) to a zone
  - zoneadd “Z_Sint_Piet01”,“A_Piet03_C1; A_Piet_04_C1” *on switch staf*
- on schimmel it would look like
  - zoneadd “Z_Sint_Piet01”,“A_Piet03_C2”

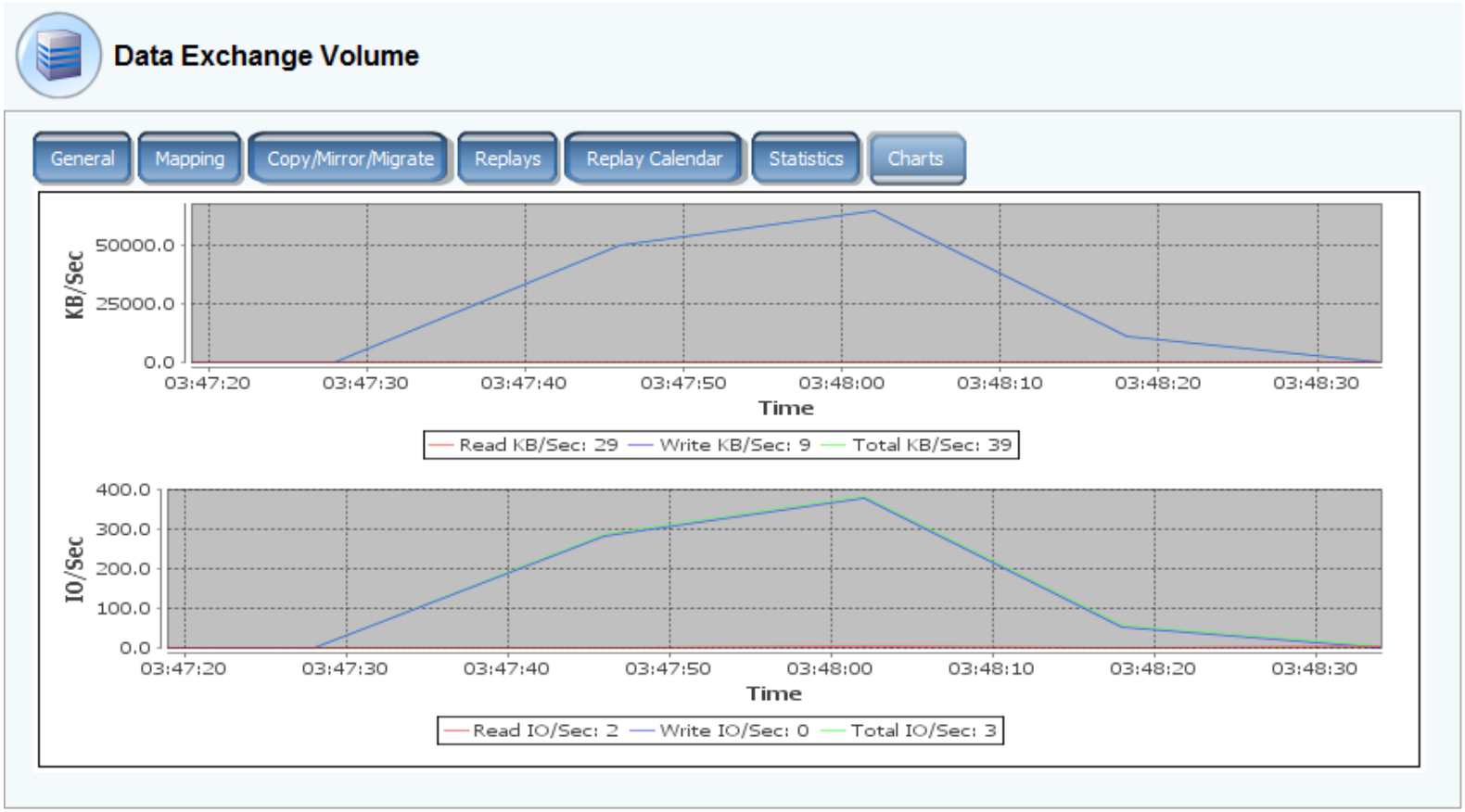
David Groep
Nikhef
Amsterdam
PDP & Grid

FC Configurations

- changes in the CLI are the *defined configuration*, not yet the *effective configuration* settings
- Save the configured settings
  - `cfgsave`
- Enable the configured settings
  - `cfgenable "C_Staf"`
- There can be multiple configuration, but for the time being that would just be confusing
- Note that a single WWN can be a member of many zones

David Groep
Nikhef
Amsterdam
PDP & Grid

iSCSI – for Generics and more



David Groep
Nikhef
Amsterdam
PDP & Grid

GSP Piet

- 16 M610 blades,
  - 96 GByte RAM, 2x600 GB SAS disk in RAID-1 config
  - 2x6 cores with HT
  - eth0+1: 1 Gbps, switch fabric A
  - eth2+3: 10 Gbps, switch fabric B
  - dual FC: 8 Gbps, switch fabric C
- Enclosure: piet-blade.ipmi.nikhef.nl
- Switches: sw-piet-<fabric><1|2>.ipmi.nikhef.nl
- Blade DRAC: piet-<01..16>.ipmi.nikhef.nl

David Groep
Nikhef
Amsterdam
PDP & Grid

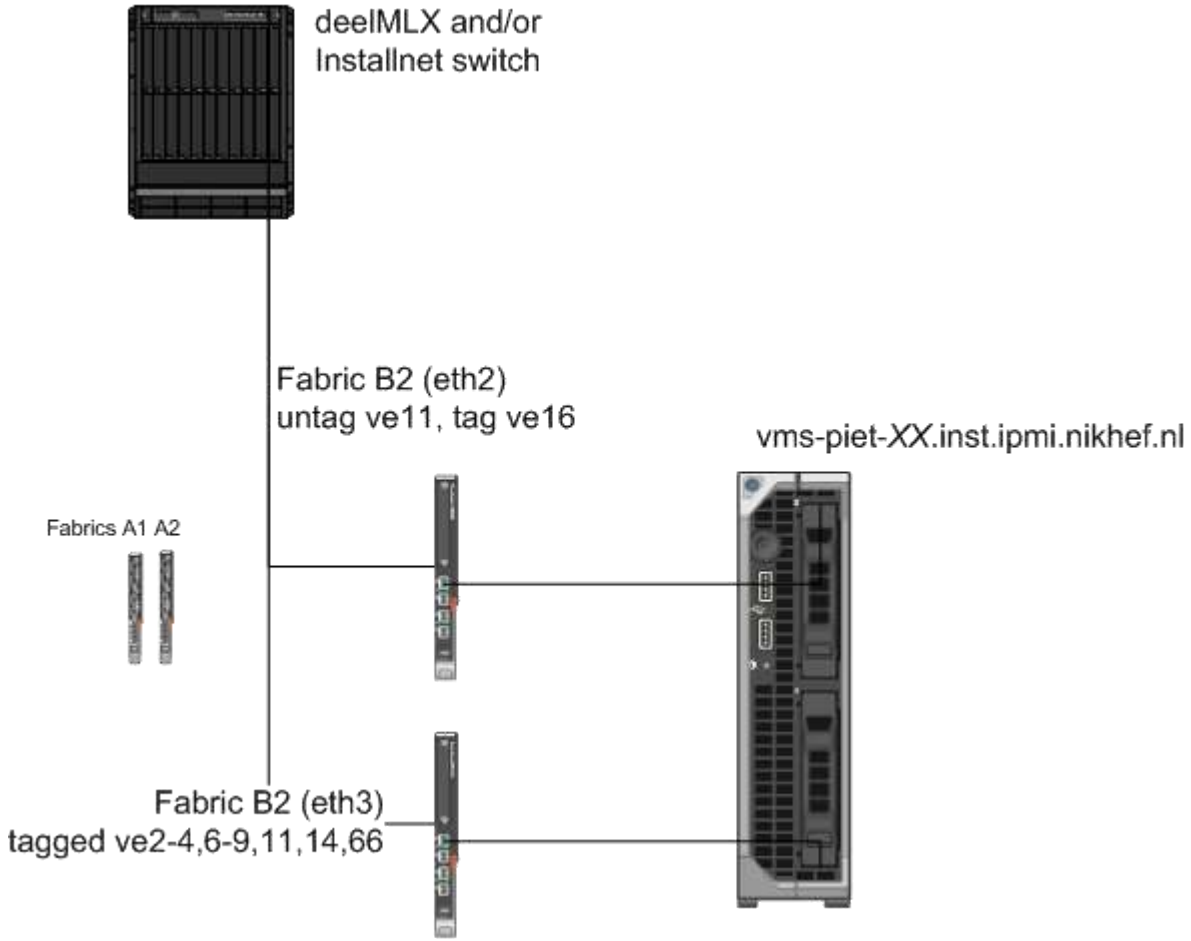
Recommended BIOS

- CPU: **virtualisation enabled** ☺
- no memory checking on boot
- boot order: disk, <rest>

- To install
  - mount the virtual CD-ROM via DRAC, via Java applet or ActiveX control
  - press F11 during boot
  - select 'Virtual CD-ROM'
  - Proceed with installation

David Groep
Nikhef
Amsterdam
PDP & Grid

Ethernet Networking Piet



David Groep
Nikhef
Amsterdam
PDP & Grid

Installation

- Follow install guidance XCP1.5beta (1.4.90)
- Hostnames
  - **vms-piet-XX.inst**.ipmi.nikhef.nl, vms-gen-XX.inst....
- Cluster master (via CNAME)
  - **pool-piet.inst**.ipmi.nikhef.nl CNAME vms-piet-16...
- static IP address (see Wiki or DNS)
  - dedicated installnet "ve11" 172.22.64.0/18
  - untagged over eth2 (eth0 for Generics 2008/A)
- DNS: boer, stal, dwalin
- NTP: stal, **salado.inst** (172.22.64.2), dwalin

David Groep
Nikhef
Amsterdam
PDP & Grid

XenServer

- XCP 1.5 looks like XenServer 6 w/o HA,WBS
- Local configuration resembles CentOS5
- Installed via CD-ROM
  - Install image ISO on stal, or download from xen.org
  - to make it look like XenServer, follow Wiki tweaks
- Fit in local 'GridSRV'-like management system
  - **run xcp-config.sh post-install script**
  - installs ssh keys of nDPFPrivilegedUsers for root
  - configure *bridged* networking & txqueue performance

https://wiki.nikhef.nl/grid/GSP_Virtualisation_with_Xen

David Groep
Nikhef
Amsterdam
PDP & Grid

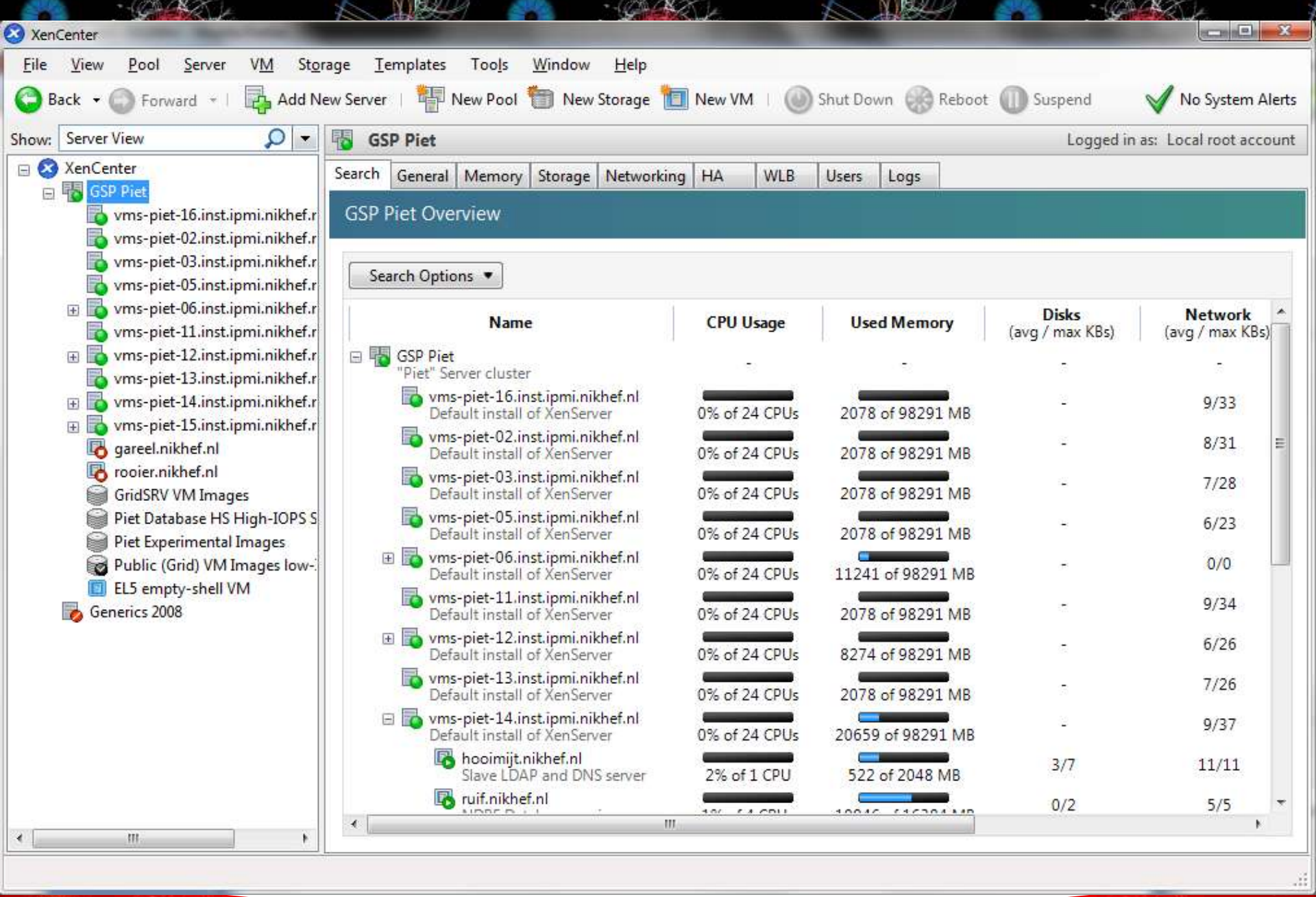
Pool Constraints

- All hosts in the pool **must have**
  - the same network configuration (devices, vlans)
  - the same FC configuration for MPIO
  - be in the same server pool on Sint
- You can't
  - copy VDIs between SRs on different pools
  - migrate VMs between pools
  - mount the same storage r/w on different pools
  - copy VDIs with the GUI (but you can with a CLI)

David Groep
Nikhef
Amsterdam
PDP & Grid

XenCenter

- Works best under Windows (sorry)
  - CLI anyway needed for advanced configuration
  - The CLI command is **xe** (see also: xe help --all)
- Configuration
  - connect to pool-piet.inst.ipmi.nikhef.nl, username: root
- Caveats:
  - no HA, so failing hardware will kill VMs
  - Live migration: right-click to move VMs
  - Maintenance mode: automatically moves VMs
  - Maintenance on pool master: trigger alternate master (and remember to update DNS CNAME please)


XenCenter
File View Pool Server VM Storage Templates Tools Window Help
Back Forward Add New Server New Pool New Storage New VM Shut Down Reboot Suspend No System Alerts
Show: Server View
XenCenter
GSP Piet
vms-piet-16.inst.ipmi.nikhef.r
vms-piet-02.inst.ipmi.nikhef.r
vms-piet-03.inst.ipmi.nikhef.r
vms-piet-05.inst.ipmi.nikhef.r
vms-piet-06.inst.ipmi.nikhef.r
vms-piet-11.inst.ipmi.nikhef.r
vms-piet-12.inst.ipmi.nikhef.r
vms-piet-13.inst.ipmi.nikhef.r
vms-piet-14.inst.ipmi.nikhef.r
vms-piet-15.inst.ipmi.nikhef.r
gareel.nikhef.nl
rooier.nikhef.nl
GridSRV VM Images
Piet Database HS High-IOPS S
Piet Experimental Images
Public (Grid) VM Images low-
EL5 empty-shell VM
Generics 2008
GSP Piet
Logged in as: Local root account
Search General Memory Storage Networking HA WLB Users Logs
GSP Piet Overview
Search Options
Name
CPU Usage
Used Memory
Disks (avg / max KBs)
Network (avg / max KBs)
GSP Piet
"Piet" Server cluster
vms-piet-16.inst.ipmi.nikhef.nl
Default install of XenServer
0% of 24 CPUs
2078 of 98291 MB
9/33
vms-piet-02.inst.ipmi.nikhef.nl
Default install of XenServer
0% of 24 CPUs
2078 of 98291 MB
8/31
vms-piet-03.inst.ipmi.nikhef.nl
Default install of XenServer
0% of 24 CPUs
2078 of 98291 MB
7/28
vms-piet-05.inst.ipmi.nikhef.nl
Default install of XenServer
0% of 24 CPUs
2078 of 98291 MB
6/23
vms-piet-06.inst.ipmi.nikhef.nl
Default install of XenServer
0% of 24 CPUs
11241 of 98291 MB
0/0
vms-piet-11.inst.ipmi.nikhef.nl
Default install of XenServer
0% of 24 CPUs
2078 of 98291 MB
9/34
vms-piet-12.inst.ipmi.nikhef.nl
Default install of XenServer
0% of 24 CPUs
8274 of 98291 MB
6/26
vms-piet-13.inst.ipmi.nikhef.nl
Default install of XenServer
0% of 24 CPUs
2078 of 98291 MB
7/26
vms-piet-14.inst.ipmi.nikhef.nl
Default install of XenServer
0% of 24 CPUs
20659 of 98291 MB
9/37
hooimijt.nikhef.nl
Slave LDAP and DNS server
2% of 1 CPU
522 of 2048 MB
3/7
11/11
ruif.nikhef.nl
0/2
5/5

NIKHEF

All VM provisioning

- Configure a machine with quattor
  - make sure the hardware spec is *really* what you want
  - compile the profile (the new biggerish-XML works)
  - **aii-shellfe --configure <hostname>**
  - **aii-provision-vm --n "<hostname"> -a**

```
stal:~:1030$ aii-provision-vm -n "boslook.nikhef.nl" -a
Parsing XML data from CDB, please wait ... Done.
Password:
Connecting to https://pool-piet.inst.ipmi.nikhef.nl/ as root ...
Creating new VM boslook.nikhef.nl on pool.
  SR will be Public (Grid) VM Images low-IOPS series 1
  Template used OpaqueRef:94142f41-6744-76da-881c-45570ae1938d
  OS Repository http://stal.nikhef.nl/centos/5.7/os/x86_64/
  Autoboot set to 1
  Start on ready set to no
  Creating NIC eth0 in Public Grid (194.171.97.0)
```

- After install: fixup XenTools version as per Wiki

network configuration is in .xapirc of ndpfmgr

David Groep
Nikhef
Amsterdam
PDP & Grid

Non-All

- Create host from template ("New VM")
- the install URL is like
  http://spiegel.nikhef.nl/mirror/centos/6/os/x86_64
- boot options:
  ks=<url> graphical utf8

When the install fails first time

- login to pool master
- reset the boot loader to "eliloader"

David Groep
Nikhef
Amsterdam
PDP & Grid

Tuning the VM and more

- Memory balooning can be set through the GUI
  - only after VM tools are installed
  - and fixed to proper version
- Live migration
  - again easiest from GUI
  - only within a pool
  - only to same hardware
- Across pools?
  - you cannot live migrate across pools
  - you cannot sanely share LUNs across pools

David Groep
Nikhef
Amsterdam
PDP & Grid

About disks and images

- VM disk images are stored in *Storage Repositories*
- One LUN contains one SR
- One SR can contain many disk images
- An SR is *always* LVM volume, even on local disk
- The default XCP disk image type is now VHD inside LVs, no longer raw LVM volumes
  - in you copy in disk images, create the VDI yourself with type RAW
  - when copying VDIs between SRs, they become VHD
  - you cannot copy out VHDs to raw LVs

David Groep
Nikhef
Amsterdam
PDP & Grid

Finding your SR

- An SR is hosted on a LUN and reached via
  - a local HBA for FC connected systems
  - an iSCSI HBA for Generics 2008/A
  - On FC, your LUN is visible immediately once you add the server to the server group on Sint
  - On iSCSI, you need to trigger a discover on both sides at the same time (<30sec)
- Provision the LUN on the Compellent first
- *For aii-provision-vm: making the desired SR the default SR eases creation of new VMs*

David Groep
Nikhef
Amsterdam
PDP & Grid

Import existing VM from EL5

- Create a RAW LVM

```
xe vdi-create \
     sm-config:type=raw sr-uuid={SR_UUID}   \
     name-label="My Raw LVM VDI"   \
     virtual-size={size}GiB type=user
```

- mount this LVM in the Dom0
  - a lot safer than tweaking with lvchange –ay!
  - vbd-create and vbd-plug
- Copy the contents into it (may take a while!)
  - you can try it with a live VM if the FS is stable
- **unplug please** VDI from the dom0 VBD
  - destroy the left-over vbd in the dom0
- create a VM which boots pygrub (not eliloader)

David Groep
Nikhef
Amsterdam
PDP & Grid

Backups and restore

- SRs can be detaches and attached without trouble
- ***DO NOT FORMAT AN SR***
  if you think there's useful data on it
- Attach an SR only to one pool in RW mode

David Groep
Nikhef
Amsterdam
PDP & Grid

Oops

If all is hosed recover from database dump (wiki!)

- Or, if worst comes to worst:
  - detach all SRs
  - wipe all VM hosts ("enter maintenance mode" -> "remove server from pool", or reinstall
  - create a new pool
  - re-attach the SRs, but *do not format*
  - create a bootable template
  - create VMs based on existing VDI disk images

David Groep
Nikhef
Amsterdam
PDP & Grid

En Nu Zelf ...

David Groep
Nikhef
Amsterdam
PDP & Grid