IT-infrastructure in Zeuthen.

Andreas Haupt
DESY - DV
DESY Computing-Seminar
Hamburg, 6th June 2011

HELMHOLTZ ASSOCIATION

DESY

Outline

- DESY in Zeuthen – an overview
- The computing centre in Zeuthen
- Network & communication infrastructure
- Computing & storage infrastructure
- HA storage with DATAcore SANmelody
- Operating systems
- Monitoring
- Other services

DESY in Zeuthen – an overview

DESY in Zeuthen – an overview

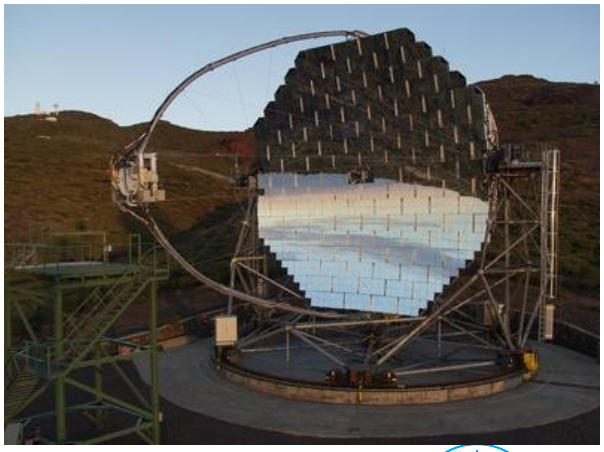
- Former “Institute High Energy Physics” of Academy of Science
- Became site of DESY in 1992
- Employees: 220
  - ~120 permanent staff (half of them are scientists)
  - ~20 PhDs, ~20 PostDocs, students, ...
  - ~20 apprentices
- More than 700 active accounts

DESY in Zeuthen – Astroparticlephysics Activities

- Participation in the Cherenkov Telescope Array Project
  - Evaluation of ACS (Alma Control Software) for the Array Observation Center
  - Taking part in Monte Carlo Production

- Computing support for the IceCube Experiment
  - Zeuthen is European IceCube Data Center (Tier1)
  - Disaster Recovery Center for IceCube
  - IceCube is now using the (WLCG) Grid Infrastructure

- Analysis of MAGIC (Gamma Ray Telescope) data
  - Using noticeable parts of mass storage and farm nodes
- THAT (astroparticle theory)
  - Simulations on parallel cluster



DESY

DESY in Zeuthen – LHC activities

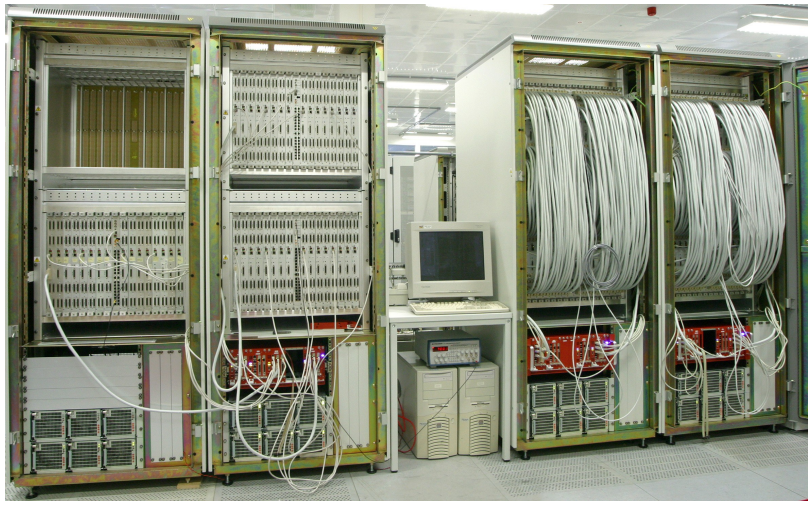
- Zeuthen has active Atlas & CMS groups
- Active contributions in central services
  - operation of DDM / computing frameworks
  - Monitoring frameworks
- Storage & Computing resources
- Remote control room established for ATLAS

DESY in Zeuthen – PITZ

- Photo Injector Test-facility Zeuthen (Electron gun development, test & optimization facility for FLASH & the planned XFEL)
  - Sparc-CPU VME crates running Solaris as control systems
  - Logging, control, DAQ hardware on standard DELL systems
  - Simulation / analysis on batch farm

DESY in Zeuthen – Lattice QCD

- Large parallel lattice calculations
- Large lattice samples stored on Lustre
- Long term history and experience in parallel computing
  - 90s: IBM SP1/SP2 parallel computers
  - APE development, QPACE & parallel clusters
  - See also: Peter Wegner's talk on 4th July
- APEnext switched off on April 30th
  - New parallel cluster as replacement
- Grid activity: ILDG (International Lattice Data Grid)
  - Distribution of large data samples via standard grid tools (SRM, GridFTP)
  - Metadata service to search for "special" datasets



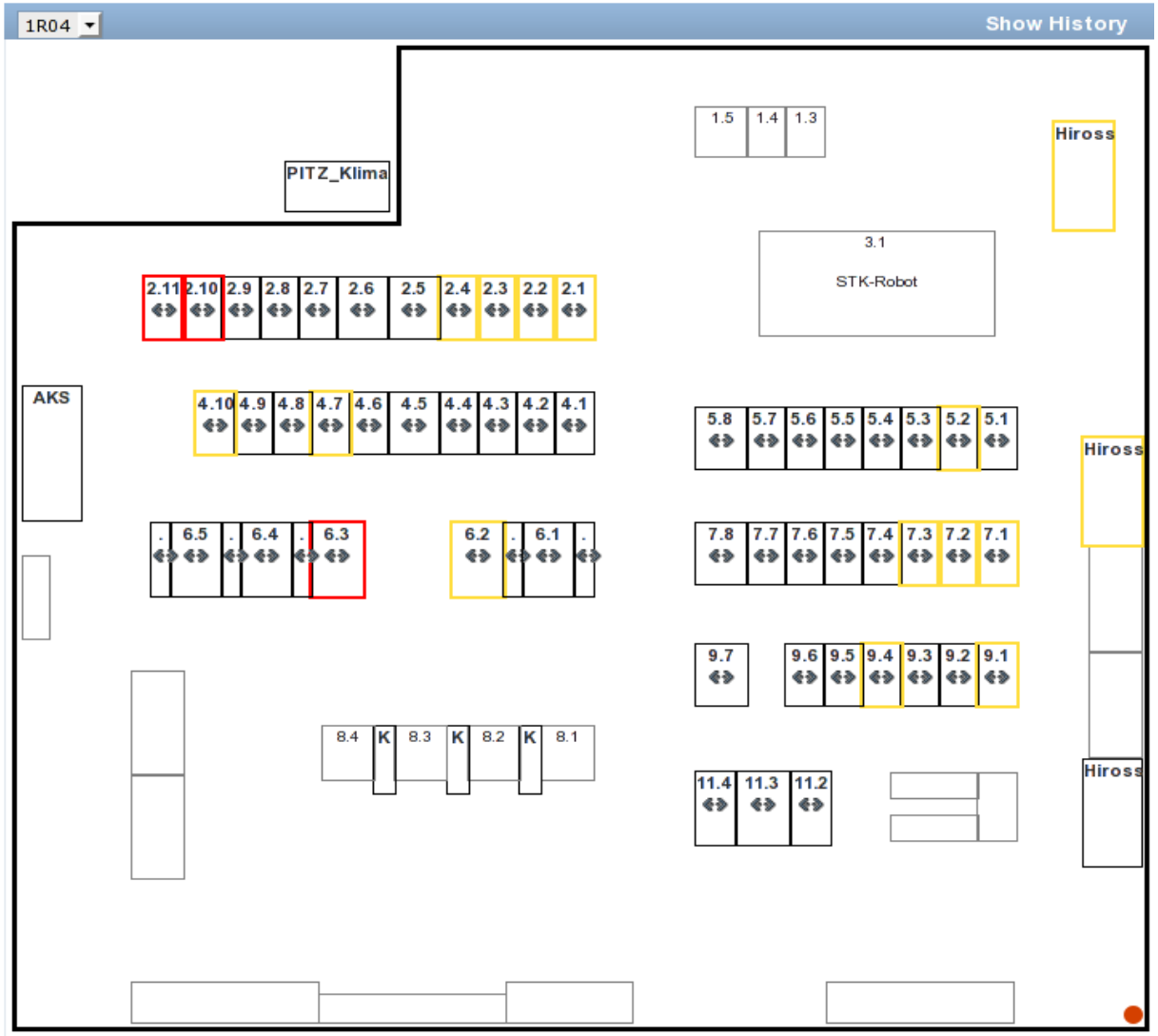
The computing centre in Zeuthen

The computing centre in Zeuthen – an overview

- 19 people
  - Taking care of any computing / telecommunication related work (there is no computing manpower in the experiments)
- UPS:
  - 3 systems running in parallel with 160 kW capacity each
  - → 1 UPS running as redundancy buffer
- Air conditioning system:
  - 5 water-cooled Rittal racks with 20 kW cooling power each
  - Air conditioning with heat exchanger: 130 kW cooling power
  - Air conditioning with cooled water: 100 kW cooling power
  - In summer months we clearly see we reached the limit of the current air cooling capacity
- Currently more than 900 active server / storage systems
  - ~240 kW sustained power consumption

The computing centre in Zeuthen – computing room 1

- ~250 m²
- Almost 50 racks
  - Vendor: Knürr
  - 5 Rittal water-cooled racks
  - 6 rows



The computing centre in Zeuthen – future enhancements

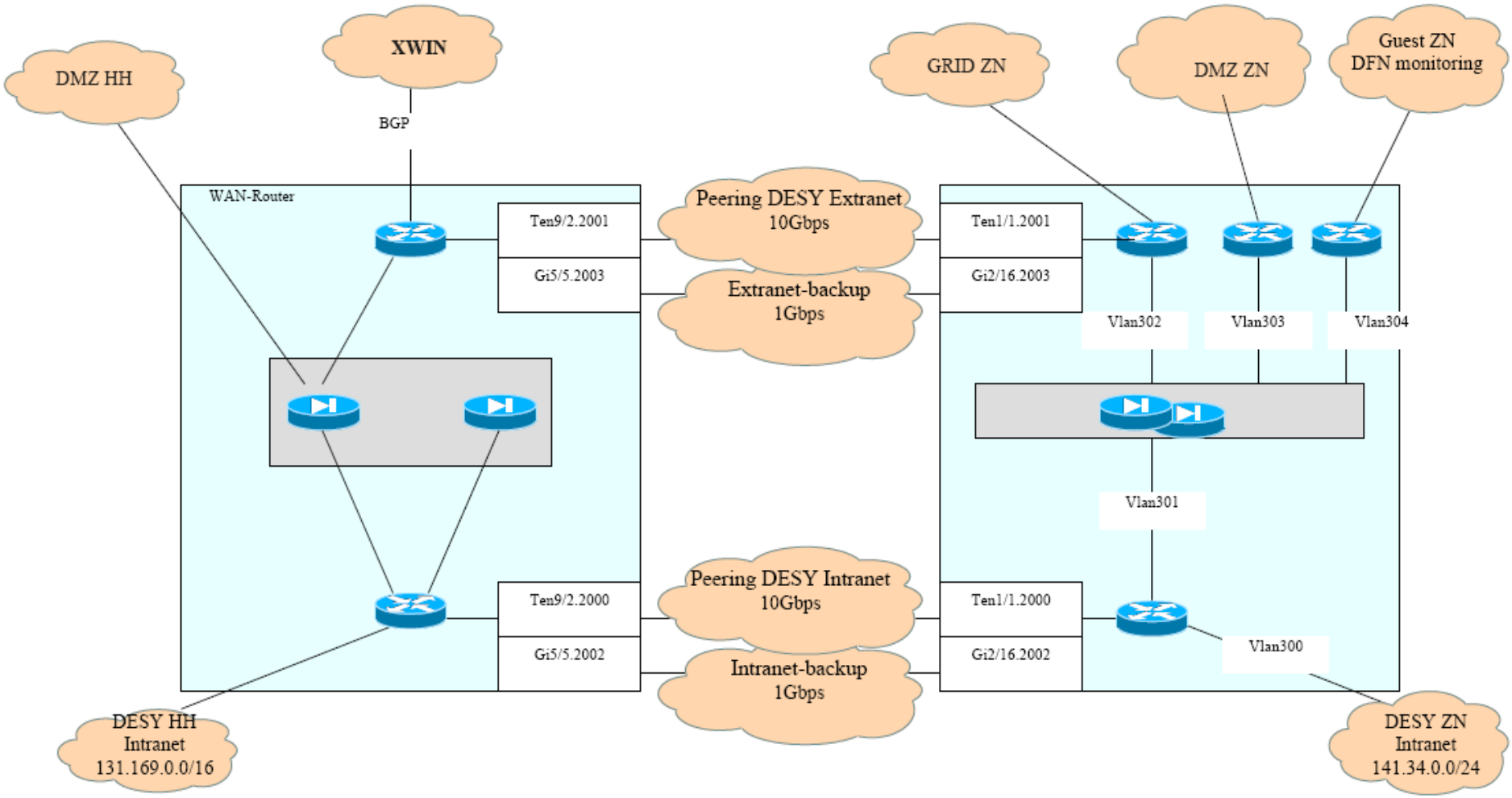
- New computing room:
  - 11 x 15 m (165 m²)
  - 5 rack rows

The computing centre in Zeuthen – future enhancements

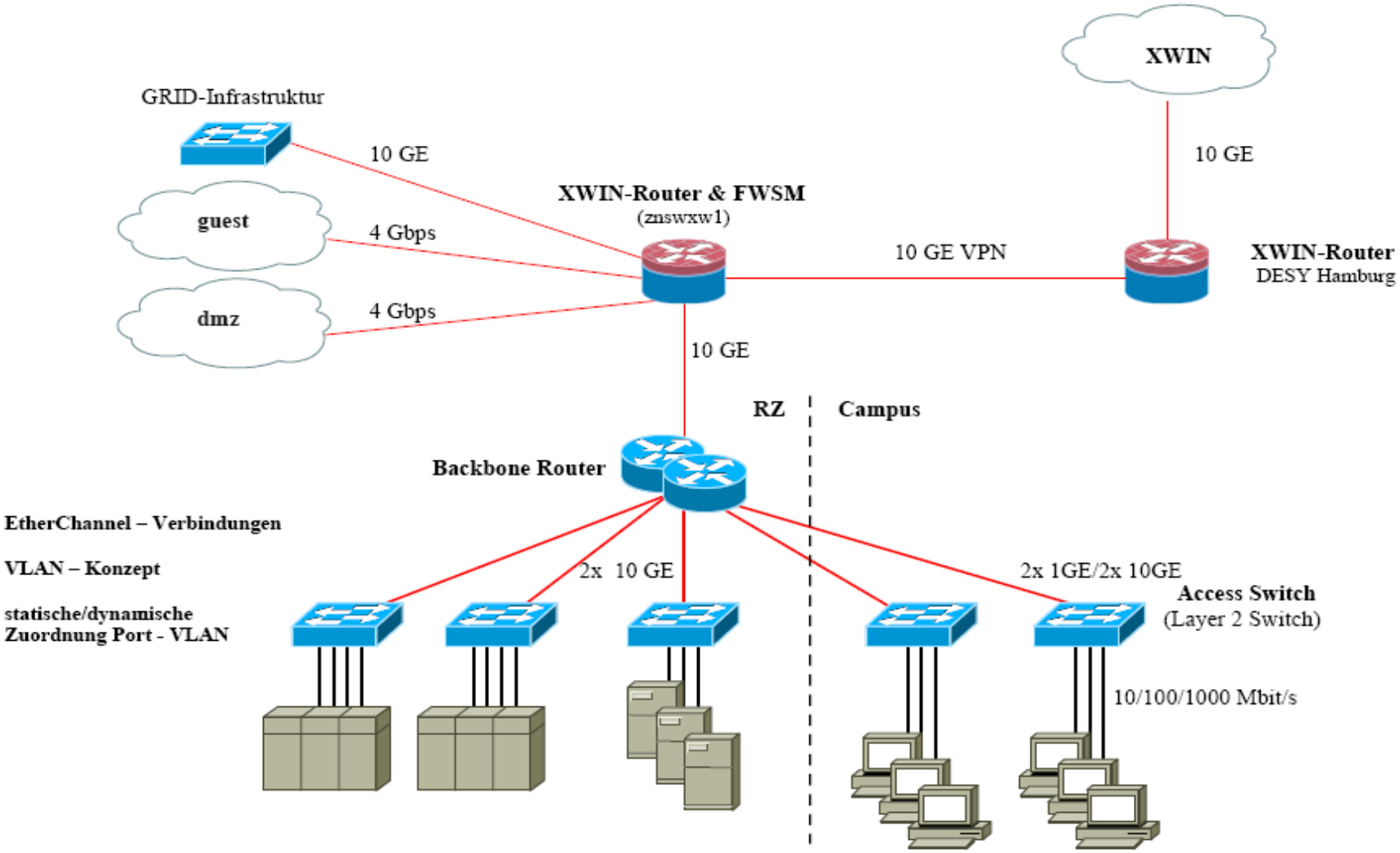
- New air conditioning system
  - Financed with money from KPII
  - Capacity: up to 500 kW for both rooms
  - Free cooling area on top of the lecture-hall building

Network & communication infrastructure

Network infrastructure – connection Hamburg - Zeuthen



Network infrastructure – overview Zeuthen



Network infrastructure – hardware in use

- In the computing centre:

Cisco Catalyst 6509	WAN-Router, integrierte Firewall (10GE VPN Verbindung nach DESY Hamburg)
Cisco Nexus 7010	Router für Netze der zentralen Systeme im Rechenzentrum
Cisco Catalyst 6509	Router für über den Campus verteilte Netze
Cisco Catalyst 6509	L2-Switche für Anschluß Server-Systeme mit in der Regel 2x 1Gbps gebondet
Arista 7140T	L2-Switch (40-Port 1/10GE) für Serveranbindung mit 10Gbps
Cisco Catalyst 4000/2950	L2-Switche für die Aufnahme Management-Interface der Serversysteme (1Gbps)
Cyclades Terminalserver	Aufnahme ser. Consolen aller Netzwerkkomponenten

- L2 switches on the Zeuthen campus
  - Cisco Catalyst 4500, Cisco Catalyst 3750, Cisco Catalyst 2950
- Summed up:
  - 3 routers
  - 6 L2 switches for servers (~900 active ports)
  - 6 L2 switches for server management interfaces (~480 active ports)
  - 40x L2 switches for systems on the Zeuthen campus (~2100 active ports)

Network infrastructure

- Wireless LAN:
  - 2x Cisco WLAN Controller 4402 (for redundancy)
  - 50 Cisco access points 1230/1240 (802.1b/g)
  - Authentication for SSIDs “DESY” & “eduroam” via Radius servers in Hamburg
- Management tools:
  - Spectrum – Network management
  - NTop – network traffic control
  - NetDisco – inventory, topology
  - KI – access / management of serial console interfaces
  - Wireless Control System (WCS) – management software for all WLAN components (configuration, administration, reporting)

Communication infrastructure

- Phone system: Avaya Integral-33
  - Analogue / digital / IP phones
  - Link to Hamburg Integral-55 via 10GE VPN connection
- 9 Cisco conference phones
  - Integrated into Hamburg IP-phone infrastructure
- 7 Tandberg video conferencing systems
  - Management software: “Cisco Telepresence Management Suite” (used DESY-wide)

Computing & storage infrastructure

Computing infrastructure

Batch Farm
926 Cores

Parallel Cluster
1024 Cores, IB

Hamburg 350 km

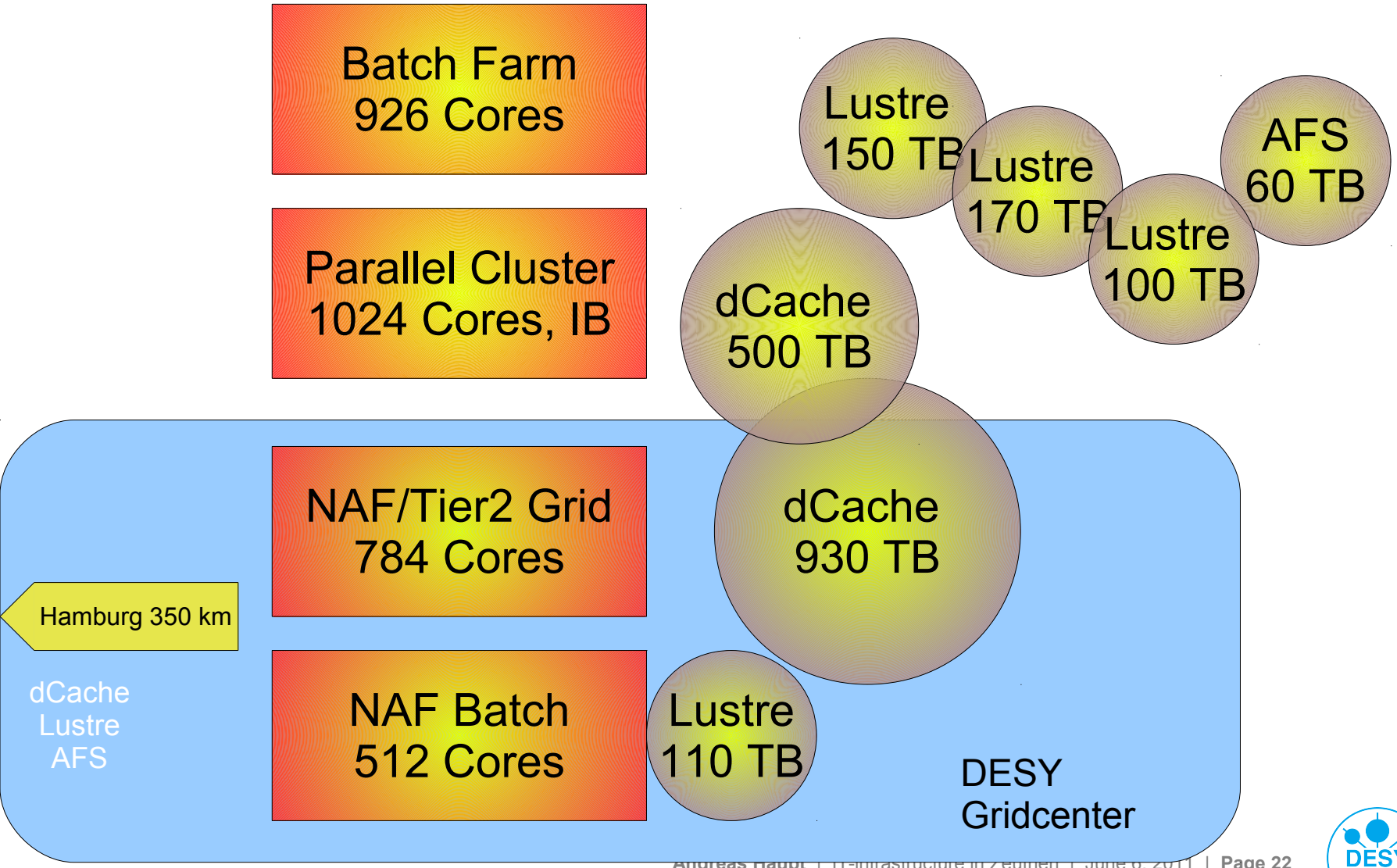
NAF/Tier2 Grid
784 Cores

NAF Batch
512 Cores

WLCG Tier2 centre for
ATLAS, CMS, LHCb
+
Grid resources for other VOs
+
Terascale Alliance
National Analysis Facility for
LHC/ILC physics

DESY

Computing & Storage infrastructure



Used hardware

- Vendor for most systems: DELL
  - DELL one winner of the DESY-wide “server advertisement”
- Hardware is running smoothly
  - DIMMs (DDR3, FB) affected by ECC problems rather often
  - Some rather seldom disk losses
- Same hardware is used for many services
  - Guarantees a big spare part pool when systems have finished their “main lifetime” with vendor hardware support
- Typical hardware support for purchased systems: 3 years
  - Replacement of defect parts via “on-site service” on the next business day
  - Some “very important” systems with replacement service within 4 hours
- Concentrate on “commodity” hardware
  - No real vendor dependency
  - Can buy systems “on demand” in small chunks

Used hardware – typical examples

R510 – storage brick
Or virtualisation node

R610 - service node

R610 + MD1000 JBOD
Storage brick

M610 – compute / wgs nodes

Hardware provisioning workflow

- Most of the hardware provisioning done via “Abrufaufträge”
  - DESY contract with DELL
- Purchased systems typically ready within 2-4 weeks (from ordering an offer with its specific configuration to system's pre-production)
- Workflow:
  1. With the official order at DELL the system is completely registered in our VAMOS database in status “ordered”
  2. Some days before systems arrive we get a mail with system's data (S/N, MAC address) → data copied into VAMOS
  3. When system arrives, status is changed to production and built into racks
  4. System ready to be installed (all needed information already in place as “asset management” and “configuration” database are identical)
  5. System comes up as configured in VAMOS

Computing infrastructure – Local batch farm

- Local batch system with currently installed 928 cores
  - Users: Icecube, CTA, PITZ, Theory, ...
- Average farm node:
  - DELL blade system
  - 3-4 GB RAM / cpu core
  - 2x 2,5” SAS disks (Raid0)
  - At least 20GB scratch disk space in $TMPDIR
  - SL5
- Grid Engine 6.2u5 

  - Last open source version provided by SUN
  - Zeuthen traditionally used SGE and its predecessor Codine since the mid 90s
  - Plan to purchase support contract from Univa – current maintainer of the open source version

DESY

Computing infrastructure – Parallel cluster

- 9 Dell M1000e blade chassis (8 in batch system) with each:
  - 16 Dell PowerEdge M610 with 2x Xeon X5560 (2.8 GHz), 24GB RAM each
  - QDR Infiniband (1x Mellanox switch per Chassis)
- Altogether 13,5 Tflops peak performance
- Chassis interconnect via Infiniband
  - Fat tree with 4 QDR Mellanox switches
- Grid Engine 6.2u5 batch system
  - Tight integration of OpenMPI 1.4, Mvapich2 1.6 (also linked against Intel-Compiler)
  - SL5 (managed like all other systems in the computing centre)
- 93TB Lustre connected with the cluster via DDR-Infiniband
- Users: Alpha, NIC, THAT, FLA

DESY

Grid Computing

- EGI / WLCG site DESY-ZN (part of the “DESY Gridcenter”)
  - Tier2 for Atlas, LHCb
  - Icecube / CTA
  - Additional VOs: Hera (hone, zeus, hermes), ILC/Calice
  - ILDG
- Cream-CE & lcg-CE in front of a 784 core (8,6 k HepSpec06) Torque batch system
  - NFS server with 96GB RAM to serve VO-directories
- 4 storage elements
  - Atlas: 750TB dCache
  - LHCb 180TB dCache
  - HSM-connected, grid-enabled dCache (500TB online) – mainly QCDSF, Icecube
  - BeStMan SE as grid gatekeeper to Lustre instances for Icecube / CTA
- Top-Level BDII, Squid for Atlas, ...

DESY

NAF

- Joint effort between DESY departments IT & DV
  - Installation / configuration
- Distributed: resources in Hamburg & Zeuthen
- Zeuthen resources:
  - 512 cpu cores (batch, interactive workgroup servers)
  - 128 cpu cores as NAF share in the grid site “DESY-ZN”
  - ~110TB Lustre (Atlas & LHCb)
  - 12TB AFS
  - Infrastructure services for redundancy (AFS database server, Kerberos – KDC, Syslog)

Storage infrastructure

- dCache
- Lustre
- AFS
- Tape Library (TSM, OSM)

AFS

- AFS cell ifh.de
- Versions in use on servers: 1.4.12, 1.4.14
- 3 database servers
- Currently 26 fileserver in production
  - ~ 60 TB
- Servers running SL5 x86_64
  - Ext3 filesystem
- Home directories, group space, scratch space, ...
- Management mostly automated
  - Space management by “group administrators” using afs_admin
  - Regular backups, restore as a “self service” for users
  - Database of all “mountpoint – volume pairs” created nightly

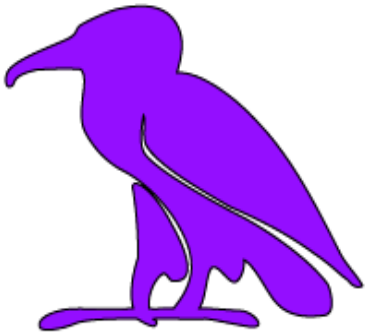
DESY

Lustre

- Currently ~420TB installed spread over 5 instances
  - ~25 OSS (file servers)
- Version: 1.8.5
- Smooth, stable production
- “Grid enabled”
  - Access via grid tools provided by BeStMan SRM & Globus GridFTP server
- Self-written monitoring tool:
  - Send reports about “misbehaving” users to experiment's computing contact persons (used space, avg. file size)

dCache

- 3 production instances running in Zeuthen
  - "old", first dCache instance with enabled HSM connected to tape library (500 TB online)
    - still running version 1.8 – plan to migrate to 1.9.12
  - ATLAS instance (750TB)
    - running "old" golden release version 1.9.5
  - LHCb instance (180TB)
    - running "old" golden release version 1.9.5
- All instances "Grid enabled"
  - SRM, GridFTP
- Currently ~ 90 file servers (pool nodes) active



Tape library

- SL 8500 robot
- 3000 slots
  - Up to 4 PB
  - LTO2 – LTO4 tapes (> 50% LTO3, 16% LTO4)
  - Currently 16% “free”
- OSM (HSM component for dCache)
  - 2 LTO3, 4 LTO4 drives
  - mover nodes: SUN nodes (x86) running Solaris
- TSM (backup)
  - 3 LTO3 drives for TSM6 connected to DELL R510, SL5
  - 3 LTO3 drives for old TSM5 – will be removed if backup data is “aged out”
  - 4 LTO3 drives via switch connected to Hamburg backup instance (used to duplicate Hamburg backup data)

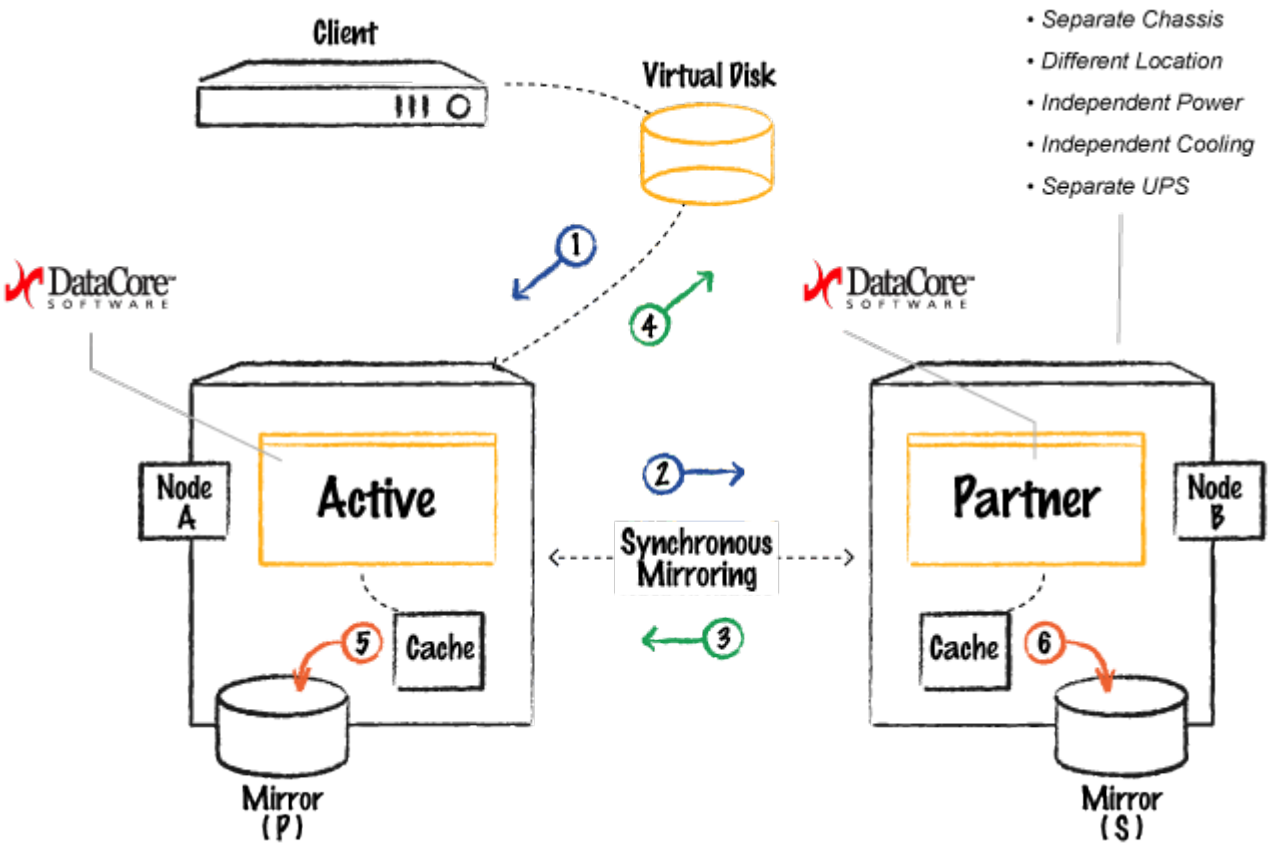
HA storage with DATAcore SANmelody

HA storage with DATAcore SANmelody – Features

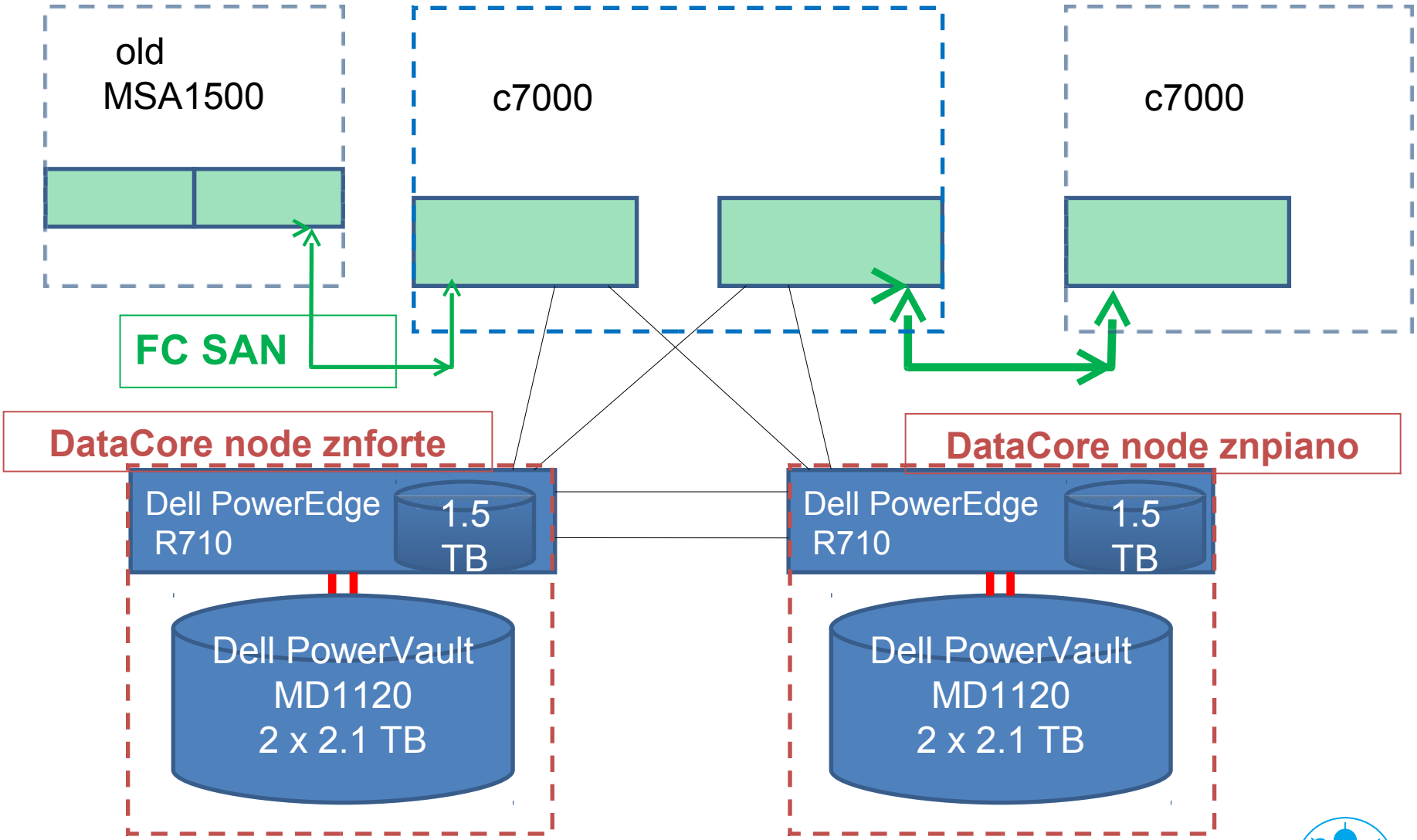
- Virtual disk pooling
- Synchronous mirroring
  - 2 nodes - High Availability
- High-speed caching (1TB cache per node)
- Load balancing
- Thin provisioning
  - oversubscription - disk space allocation when required
- RAID striping
- Online snapshot
- Remote replication (not used so far)
- Central management (nodes individually)

HA storage with DATAcore SANmelody

- Overview:



HA storage with DATAcore SANmelody – diagram



Operating systems / configuration management

Operating systems – UNIX / Linux

- Scientific Linux
  - SL4 – SL6 (work horse currently: SL5)
  - Mostly x86_64 (even on desktops)
  - Policy to upgrade every system to latest “minor release”
  - Virtualisation with XEN (part of SL5), will move to KVM (part of SL6)
  - Desktops managed by DV completely – no “root” access by users
- Solaris
  - Only some systems left (mainly management nodes for tape library & installation service...)
  - PITZ VMEs

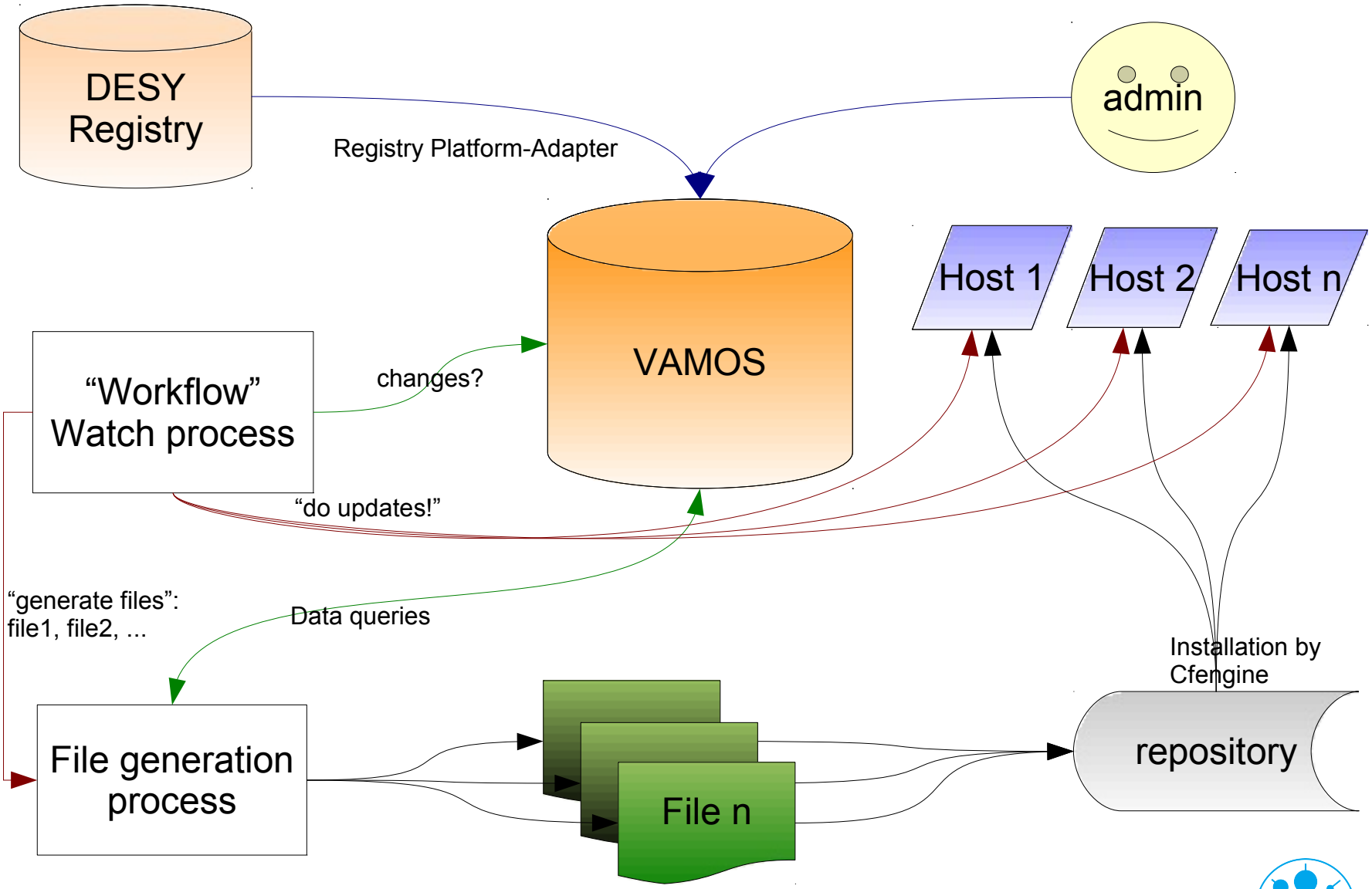
Operating systems – Windows

- Zeuthen systems integrated into the DESY-wide WIN.DESY.DE domain
  - Installed using “local” infrastructure
  - Replica of “Active Directory” server in Zeuthen
- Mostly desktop PCs
  - More than 200 systems
  - Some DAQ systems
- File service for Zeuthen users (see SANmelody slides)
- Terminal service
- Work on NetInstall packages

Configuration management – VAMOS

- Central management database
  - Account / personal information (fed by DESY Registry via a “Platform Adapter”)
  - Host (incl. asset) information (services to be configured, packages to be installed, ...)
  - Network information
  - Supports inheritance of attributes
- Cfengine used to bring host schema in database onto the real system
- Watched for changes by a daemon
  - Starts a configured “workflow” process based on the changes
- All typical configuration files generated out of it
  - /etc/passwd, /etc/group, /etc/netgroup, /etc/hosts, ssh_known_hosts, ...
  - Name server configuration, DHCP server configuration, ...
  - Automatic creation of accounts in AFS, Kerberos, Gridengine, ...
  - Automatic creation of host-keys for ssh, Kerberos, ...

Configuration management – VAMOS



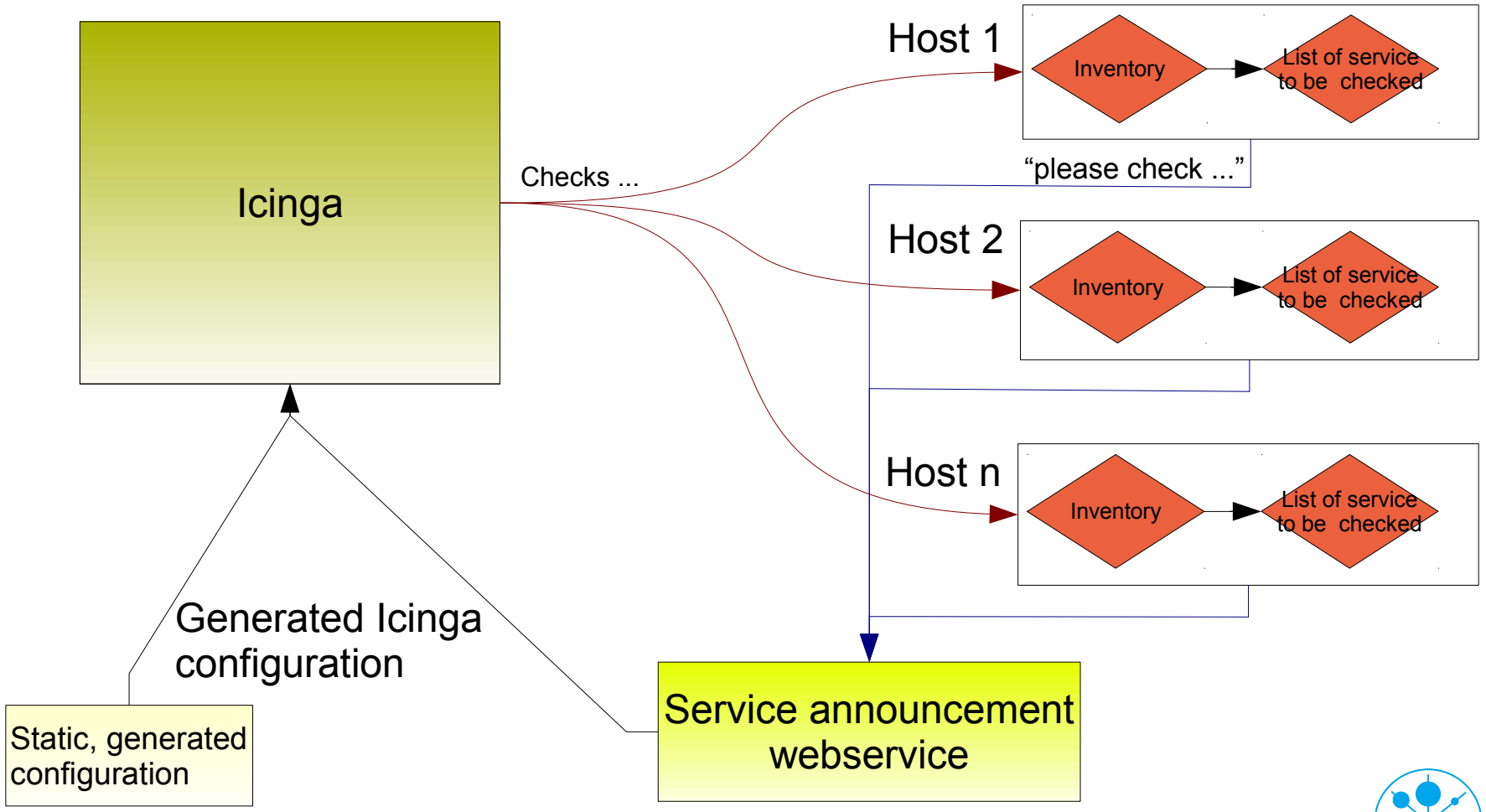
Monitoring

Monitoring – Icinga

- Icinga 1.2.1
- Only little configuration changes needed for successful migration from Nagios 3
  - Fork needed as Nagios core developer denied further real “open source” progress
- Monitoring of 1100 hosts, printers and other devices like UPS, air conditioning system, ...
- All in all more than 12000 services monitored
  - Developed service check plugins for e.g. “system event log”, NIC bonding status, certificate lifetime, hanging fileservices, ...
- Configuration (to be monitored services) created automatically
  - Systems request service monitoring via a webservice
  - “special” systems like UPS configured statically
- Hardware monitoring:
  - Evaluation of SNMP traps
  - Evaluation of “system event log” via IPMI

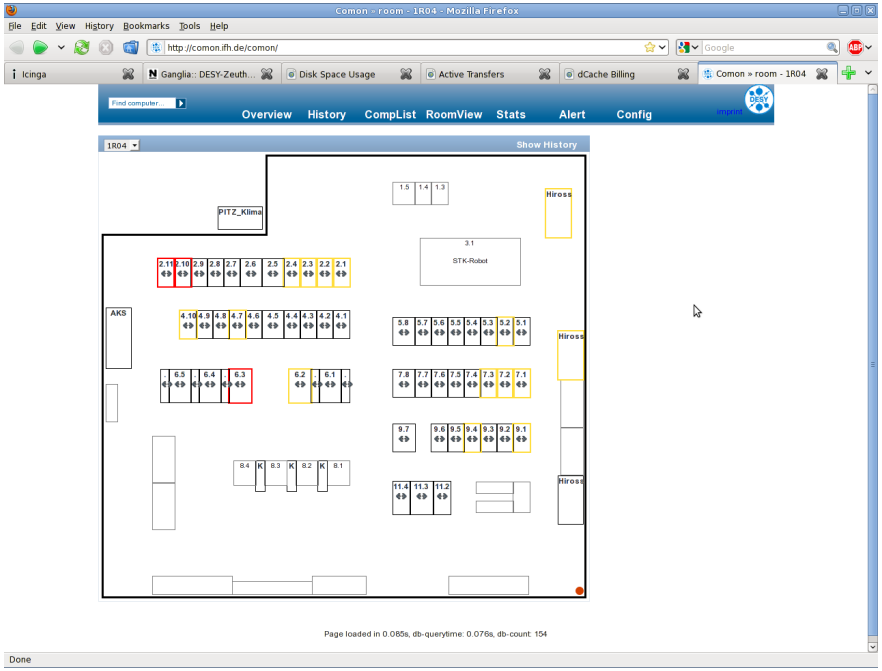
Monitoring – Icinga

- Automatic announcements of services to be checked:

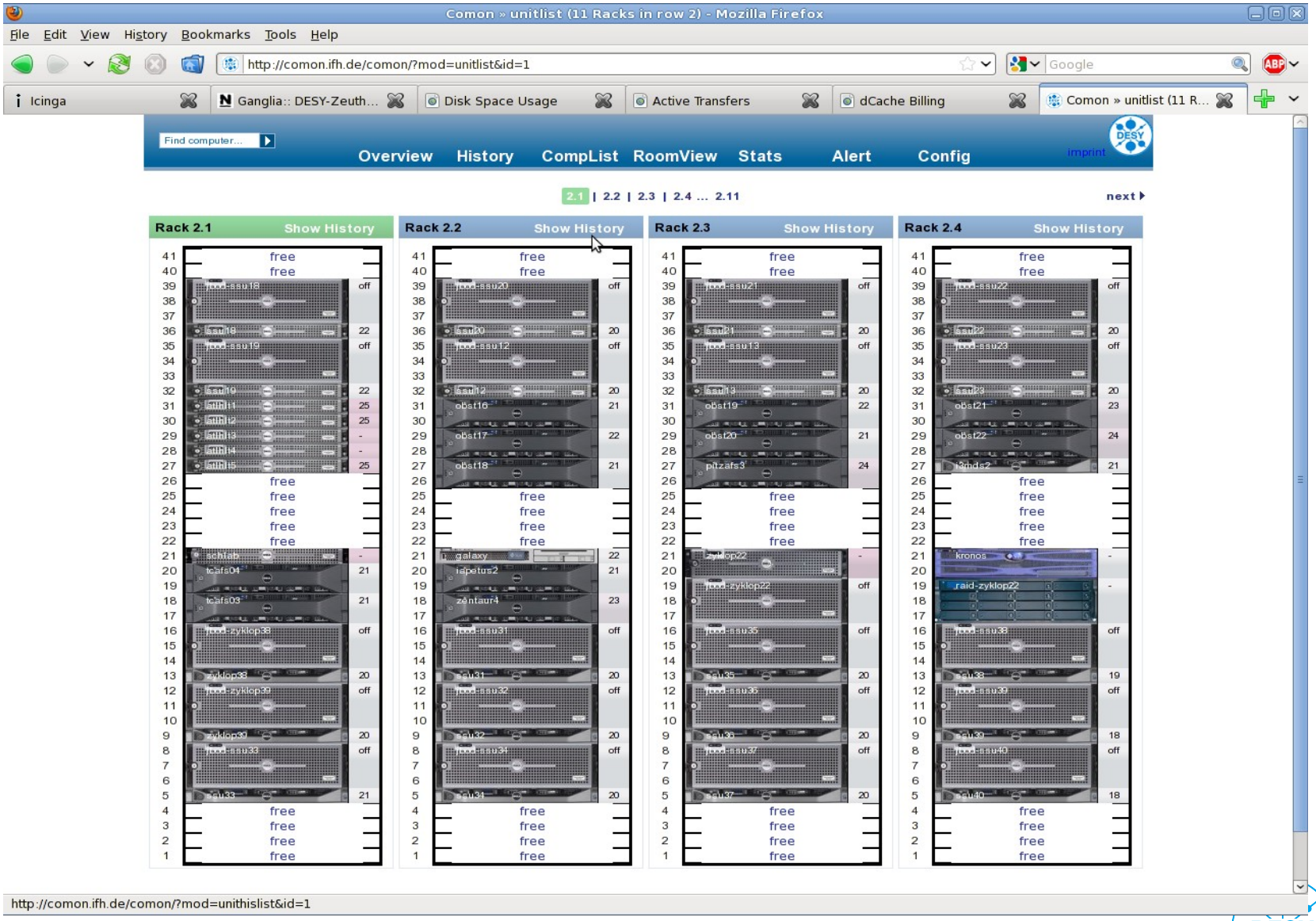


Monitoring – Comon

- Overview of the computing centre room(s)
  - “Where is host x ?”
- Temperature monitoring
- Air conditioning system & servers provide their ambient temperature reading to Comon
  - Website to visualize this
- Recognize “hot” and “cold” zones
- Alarming in case of air conditioning malfunction



Monitoring – Comon



Other services

Other services – Printing

- ~ 40 printers / “multi-function devices” on the campus
  - Vendor mostly HP / RICOH
- Print service via CUPS & Samba
  - Problems with some printer drivers after Samba 3.3.10
  - Separate print server for 64bit Windows systems
  - Planned: Windows print server - Windows clients don't use Samba any more (will have advantages as well as disadvantages)
- Future: CUPS service on SL6 (version 1.4.x)
  - Windows Printserver for Windows clients – forwards print requests to Cups

Other services

- Subversion 

  - Centrally managed repositories for all DESY groups
  - 4 global admin (2 in HH, 2 in ZN), many repository admins
  - https://svnsrv.desy.de
  - Currently ~120 repositories, 450 developers
  - WebGUIs exists for tasks like creating new repositories, adding users, browsing repositories, ...
  - Authentication scheme allows participation without DESY account
- IMAP service for users in Zeuthen
  - Dovecot 1.2, migration to 2.x planned 

  - Server-side scripting using sieve (filtering, vacation messages, ...) - a management WebGUI exists
- Wiki farm 
  - Moin Moin
  - Atlas, CTA, Icecube, PITZ, DV, ... (altogether currently 19 wikis)

Other services – User support

- Single queue in the DESY Request Tracker - uco-zn@desy.de
  - No hierarchic (1st, 2nd level) support structure
  - All colleagues providing user services are subscribed to the queue and take over appropriate tickets
  - Direct communication with users
- Short-term meetings with experiment representatives
  - Procurement plans
  - Things that went wrong and should be improved
  - “keep in touch”

Summary

- The Zeuthen computing centre is equipped with hardware that
  - meets the CPU and storage needs of the research groups and projects working at the Zeuthen campus
  - gets renewed regularly to keep up with the increasing demands
- The Zeuthen computing centre provides excellent resources for the experiments achieved by
  - highly standardized hardware
  - highly automated workflows for deploying hardware and software
  - efficiently configured services with minimal or even no human interaction
  - timely reactions to user requests
- The topic “Scientific Computing” was not covered in detail
  - See upcoming talk by Peter Wegner on 4th July 2011