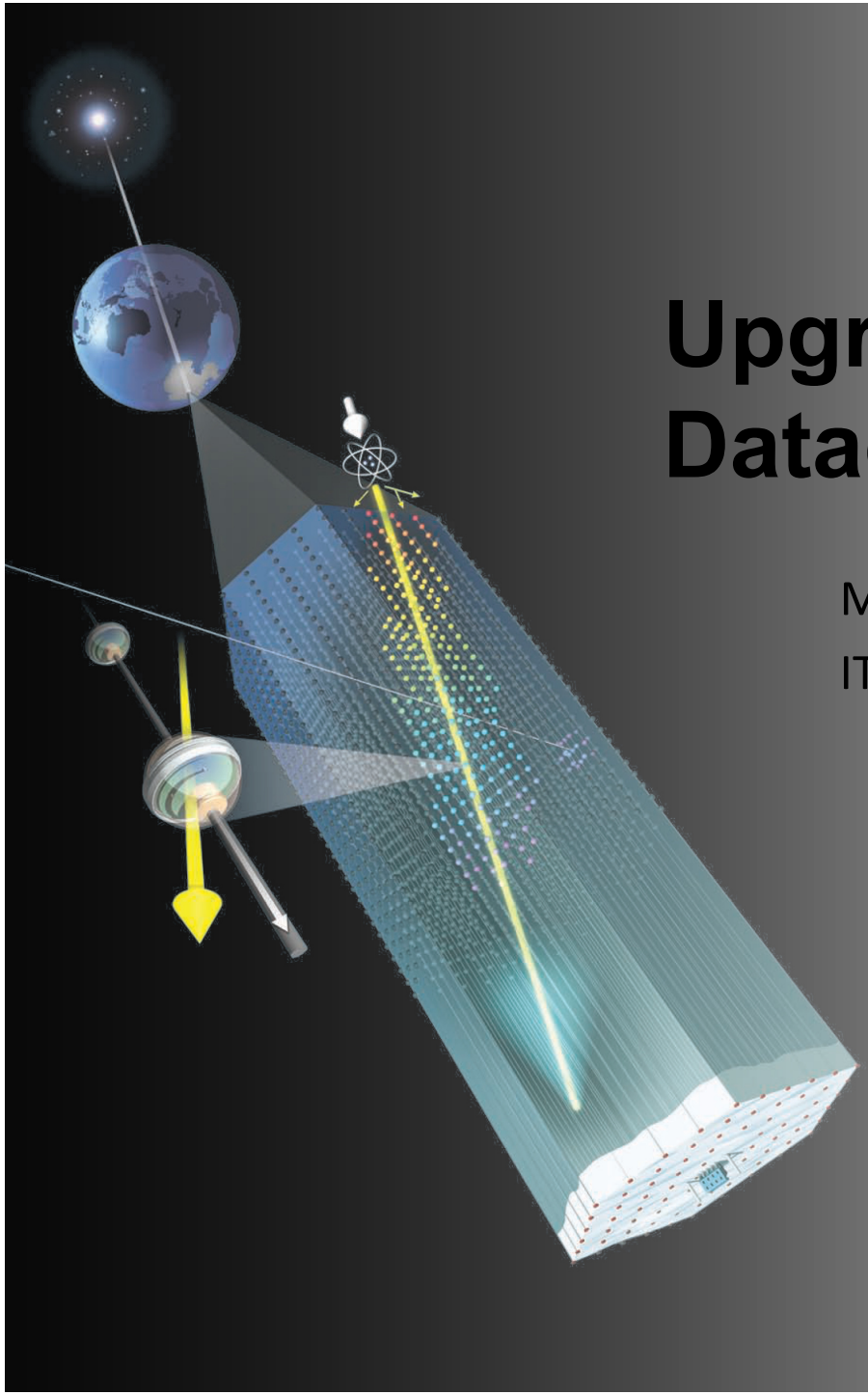
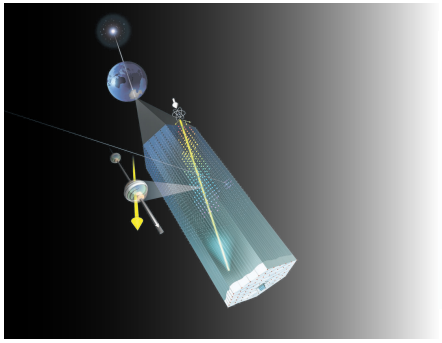


# Upgrading the 222 Datacenter in 2011

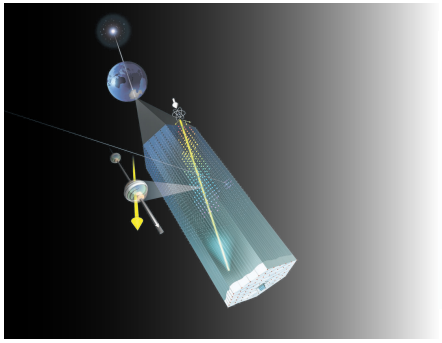
Martin Merck, UW-Madison

IT Lunch Mar. 3rd 2011



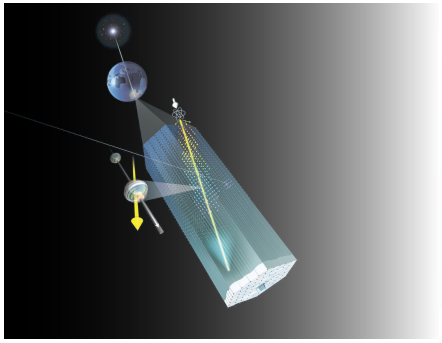
# Goals

- Replace aging hardware
  - Old DL380 G3/G4 and DL385 hardware
  - Setup very organic
  - Don't meet performance needs for DBS, etc.
- Make system more fault tolerant
  - Redundant network connectivity
  - VMs with check pointing
  - VM migration to different hardware
- Make systems more manageable
  - Configuration management
- Make space for expansions
  - Free rackspace to allow storage expansion



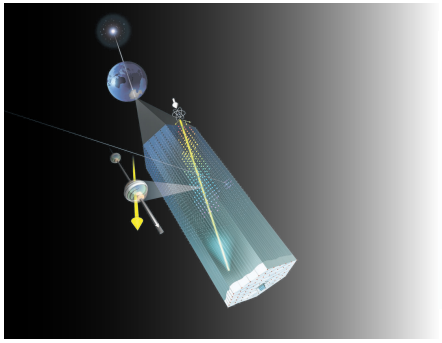
# Upgrades planned for 2011

- 10 Gbit Ethernet Backbone in 222 Datacenter
  - 10 Gbit row end switches to connect all switch uplinks
  - Allow future upgrade to 10GBit Ethernet on storage servers
  - Make redundant paths to the network backbone
- Upgrade of npx3 cluster
  - Double cluster capacity by adding 50 nodes for a total of 600 cores
  - Currently 1U DELL R410 best price but still looking into Blades
- Extend DataWarehouse
  - Add a new Hitachi AMS 2500 but already with 7 Trays (full rack)
  - 336 disks (2TB) or 673 TB (~ 612TiB) raw
  - 33 RAID6 arrays, 66 LUNs (7.75 TB). Usable 511 TB (~465 TiB)



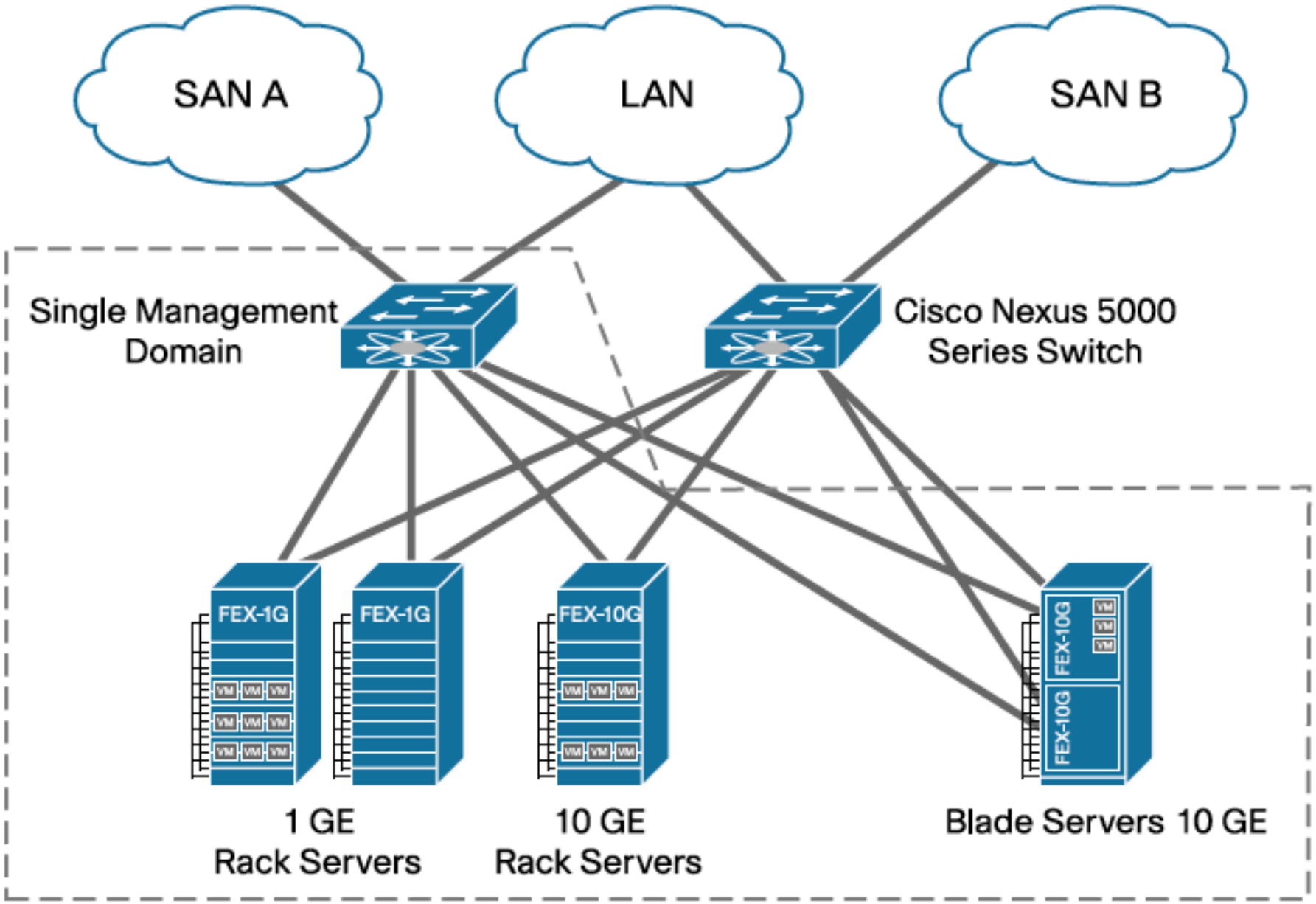
# Upgrades planned for 2011

- Extend npx2-uwa with hardware from pole
  - At least add 5 HP blade servers (80 cores)
  - Maybe add some G6
- Create GLustreFS based storage cluster /data/uwa
  - Use HP DL380 G6 and 4 x Promise + 2 x XRaid from pole as nodes
  - Add pyrite / Digidata array (was used as GulsterFS testbed)
  - Add retired Promise and possibly EONStore from 222
  - Expected capacity (~ 60 TB)
- Update and consolidate services
  - Consolidate on ~14 Servers (50-100 VMs) on KVM
  - Upgrade/Migrate to Scientific Linux 6.0
  - Introduce configuration management

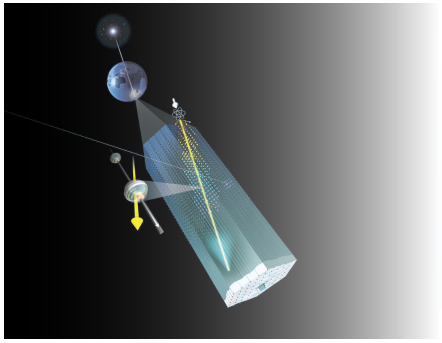


# Network Backbone upgrade

- Use CISCO Nexsus 5000 / 2000 series
  - 2 Nexus 5548 as backbone connected redundantly to 6509 core switch
  - Managed as single entity, 960GBit/s switching
  - 6-10 Nexus 2000 fabric extenders connected crosswise to Nexus 5500 switches (Gives 20 Gbit/s uplinks and failover)
  - Servers cross connected to 2 fabric extenders (FEX).
    Allows failover but currently no aggregation of bandwith
- Possibility of FCoE connection of Lustre servers
  - One connection for Ethernet and fiberchannel
  - Switch can act as FC switch ad uplink native (1/2/4/8 Gbit/s) FC
  - Needs special CISCO Ethernet cards in servers
  - Needs full outage to recable FC network

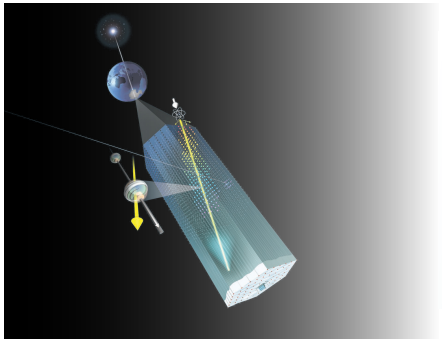



10 Gigabit Ethernet, Data Center Ethernet and Fibre Channel over Ethernet (FCoE)



# Service consolidation

- Goals
  - Save rackspace
  - Upgrade to SL 6.0 (yum)
  - Increase manageability (nagios/ganglia/OM/Puppet)
  - Fault tolerant (rapid recovery from failures, no automatic failover)
- Requirements for VM solution
  - Check pointing of VM state
  - Migration to alternative hardware (requires shared storage)
  - Resource (memory/CPU) over commit
  - Easy resource reallocation (dynamic?)
  - GUI management/monitoring
  - Direct hardware allocation (LUNs, network)



# Proxmox GUI samples

You are logged in as 'root'

proxmox

Home | Logout — Proxmox Virtual Environment 1.5 — www.proxmox.com

**VM Manager**
- Virtual Machines
- Appliance Templates
- ISO Images

**Configuration**
- System
- Storage
- Backup

**Administration**
- Server
- Logs
- Cluster

## Proxmox Virtual Environment

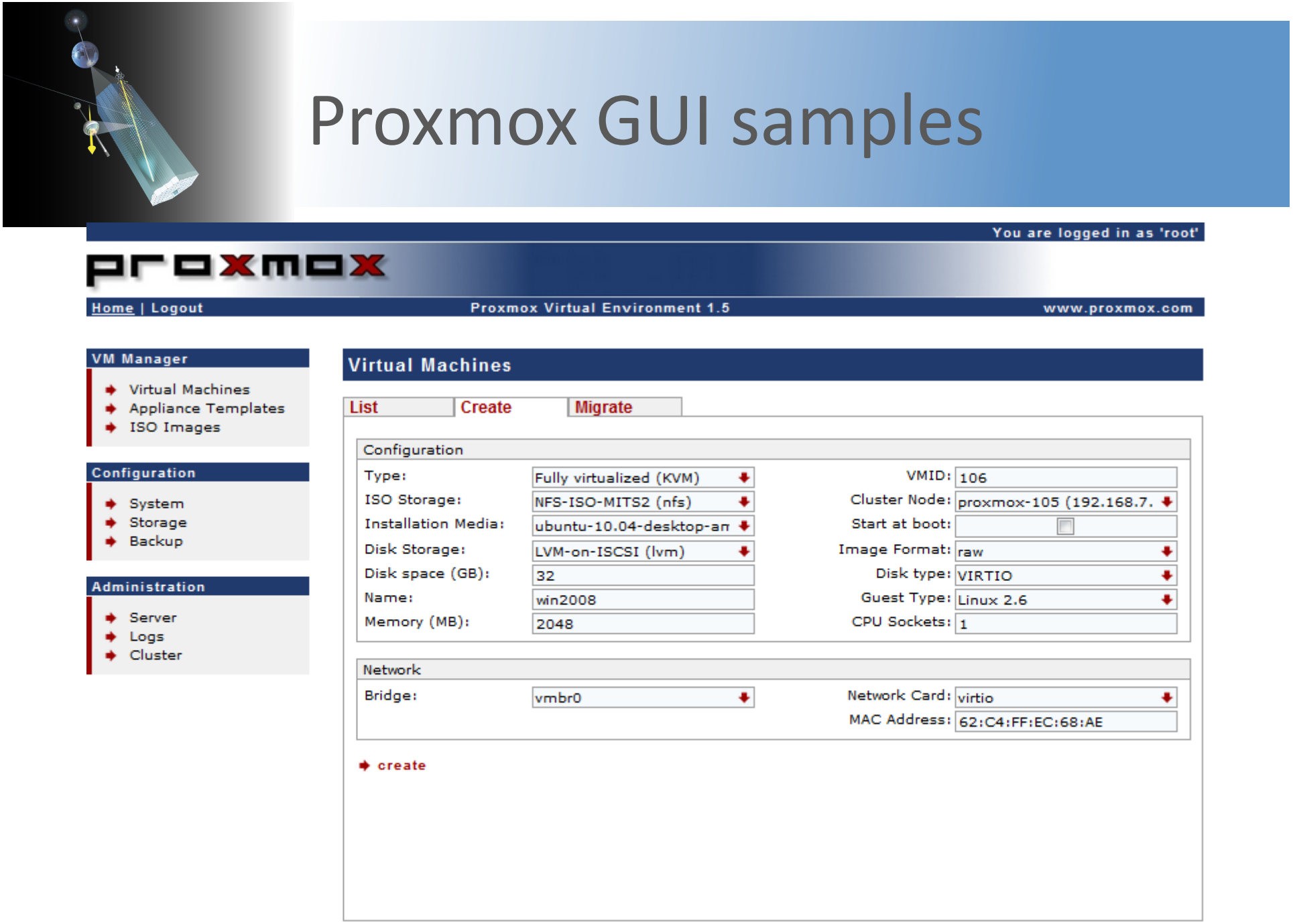
**Welcome** to the Proxmox Virtual Environment!

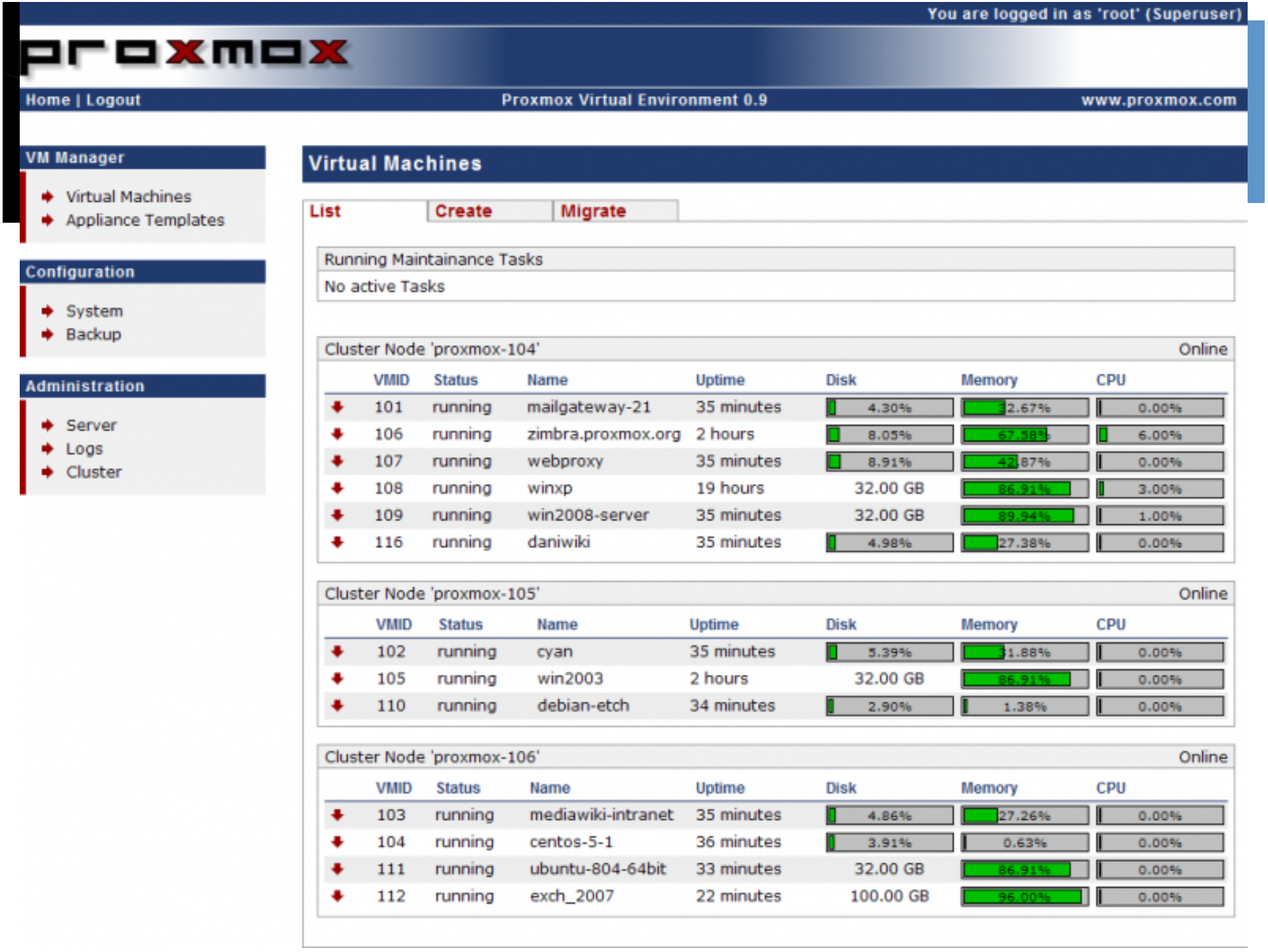
For more information please visit our homepage at www.proxmox.com

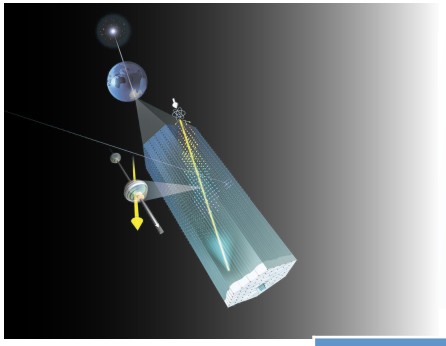
| Hostname | IP Address | Role | State | Uptime | Load | CPU | IODelay | Memory | Disk |
|---|---|---|---|---|---|---|---|---|---|
| proxmox-105 | 192.168.7.105 | Master | active | 00:08 | 0.01 | 0% | 0% | 3% | 11% |
| proxmox-106 | 192.168.7.106 | Node | active | 00:04 | 0.01 | 0% | 0% | 3% | 2% |

| Local System Status ('proxmox-105') | Online |
|---|---|
| Uptime | 17:52:16 up 00:08, load average: 0.01, 0.04, 0.02 |
| CPU(s) | 4 x Intel(R) Xeon(R) CPU X3220 @ 2.40GHz |
| CPU Utilization | 0.40% |
| IO Delays | 0.00% |
| Physical Memory (7.79GB/245MB) | 245MB |
| Swap Space (7.00GB/0KB) | 0KB |
| HD Space root (94.49GB/10.04GB) | 11.20% |
| Version (package/version/build) | pve-manager/1.5/4728 |
| Kernel Version | Linux 2.6.24-11-pve #1 SMP PREEMPT Fri May 14 09:28:08 CEST 2010 |

# Proxmox GUI samples

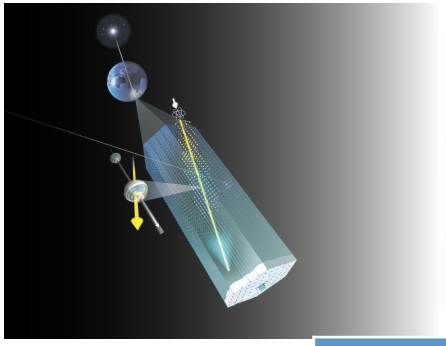



You are logged in as 'root' (Superuser)
proxmox
Home | Logout
Proxmox Virtual Environment 0.9
www.proxmox.com
VM Manager
Virtual Machines
Appliance Templates
Configuration
System
Backup
Administration
Server
Logs
Cluster
Virtual Machines
List
Create
Migrate
Running Maintainance Tasks
No active Tasks
Cluster Node 'proxmox-104'
Online
VMID Status Name Uptime Disk Memory CPU
101 running mailgateway-21 35 minutes 4.30% 32.67% 0.00%
106 running zimbra.proxmox.org 2 hours 8.05% 67.58% 6.00%
107 running webproxy 35 minutes 8.91% 42.87% 0.00%
108 running winxp 19 hours 32.00 GB 86.91% 3.00%
109 running win2008-server 35 minutes 32.00 GB 89.94% 1.00%
116 running daniwiki 35 minutes 4.98% 27.38% 0.00%
Cluster Node 'proxmox-105'
Online
VMID Status Name Uptime Disk Memory CPU
102 running cyan 35 minutes 5.39% 31.88% 0.00%
105 running win2003 2 hours 32.00 GB 86.91% 0.00%
110 running debian-etch 34 minutes 2.90% 1.38% 0.00%
Cluster Node 'proxmox-106'
Online
VMID Status Name Uptime Disk Memory CPU
103 running mediawiki-intranet 35 minutes 4.86% 27.26% 0.00%
104 running centos-5-1 36 minutes 3.91% 0.63% 0.00%
111 running ubuntu-804-64bit 33 minutes 32.00 GB 86.91% 0.00%
112 running exch_2007 22 minutes 100.00 GB 96.00% 0.00%



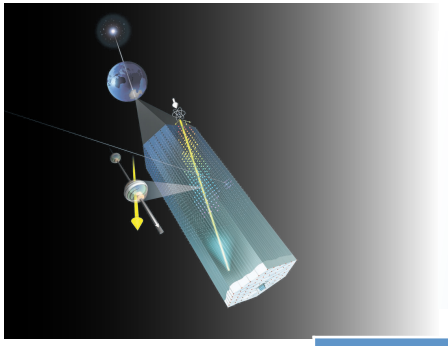
# DMZ services on VMs

| VM | Service |
|---|---|
| flounder | RT |
| barramundi | Gamma Ray follow up |
| shad | Webmail |
| goby | Mantis |
| sardella | Optical Follow up |
| herring | mailman (lists) |
| bonefish | DOMprod CGI? |
| cuttlefish | external DNS |
| cerulean | Drill software ? |
| mukte | driller WWW |
| tan | - |
| yellow (old) | - |
| shiner | svn |
| anchovy | dm-ice (wiki / doc-db) |
| zebrafish | racktables / netdisco |
| whiting | ARA (wiki / doc-db / mailman) |
| boxfish | software |
| croaker | blog |
| mullet | - |
| dorado | plasma-astro (wiki) |



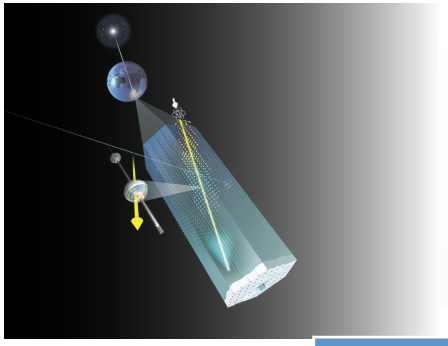
# DMZ services on hosts

| Host | Service |
|---|---|
| pub1/pub2 | Access |
| pub-nx | Access |
| pubara | Access |
| rhn | RedHat Network |
| lackey | external WWW |
| bright | internal WWW |
| x2100 | ganglia |
| ocher | wiki / indigo (events) |
| bianco | docushare |
| snow | gallery |
| brick | Data Warehouse WWW |
| cygnus | i3live WWW |
| i3live-2 | i3live devel |
| yellow | i3moni |
| skynet | ITS |
| dbs2 | I3OmDb |
| dbs4 | SimprodDB |
| aspen/t1000/giskard | GridFTP |



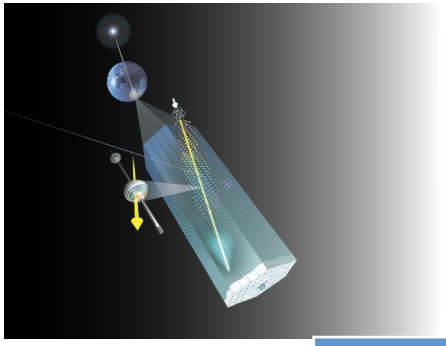
# 223 services on hosts

| Host | Service |
|---|---|
| slate | Web development |
| guppy1 | Users |
| guppy2 | Users |
| guppy64a | Users |
| cobalt1 | Users |
| cobalt2 | Users |
| cobalt3 | Users |
| cobalt64a | Users |
| Retrospect | Backup |



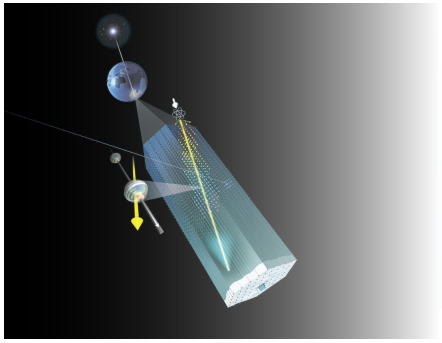
# Internal services on hosts

| Host | Service |
|---|---|
| alabaster | Nagios / MRTG |
| ebony | Splunk / syslog? / nessus? |
| ash | YUM / kickstart |
| trout | mail / postgrey / amavis |
| wabiska | LDAP / DNS / NTP |
| oken | LDAP / DNS / NTP / 222 DHCP |
| porter | Samba / FlexLM / Cups / 223 DHCP |
| perch | Admin / SNM2 / Hitrack / 118 DHCP |
| dbs1 | Web DBs / Ingest DB |
| dbs3 | Drill DB |
| marbel | sptrDownloader |
| stone | Ingest |
| bender | VMs tarpoon /sickelback / sawfish / dogfish |
| viridian | nutch search |
| mint | - |
| stucco | Altiris / SolidNet licenses |
| icepick | Cobra and OpenPlan. |
| hardcat | Hardcat Assest Management |



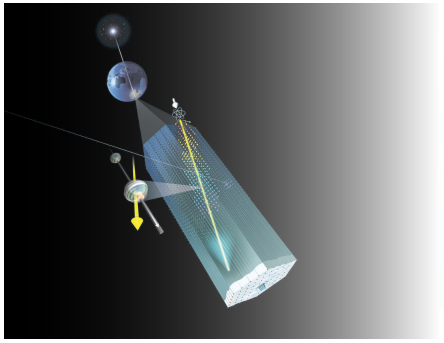
# Data Servers

| Host | Service |
|---|---|
| gibson | home |
| red | Lustre MGS |
| white / ivory / pinky | Lustre MDS |
| t101 / cerebro /wopr /gort | /data/exp OSS |
| tobor / r2d2 / wall-e / hal | /data/sim OSS |
| johnny5 / marvin / ed-209 / megatron | /data/ana OSS |
| robby /rosie | /data/user OSS |
| cylon / voltron | Lustre NFS server |
| rain / night | NFS /net/user (Apple XServe) |
| dark / ruby / cream | NFS /net/user (EonStore/Promise) |
| b782 | FTP / file access |
| chuma | Storehouse HSM |
| maloof / sam | Filetek RFS |
| klausz | HSM I/O |
| onyx | Atempo backup |
| amethyst | Atempo bachup (old) |
| charcoal | Amanda backup |
| chalk | Direct tape backup |

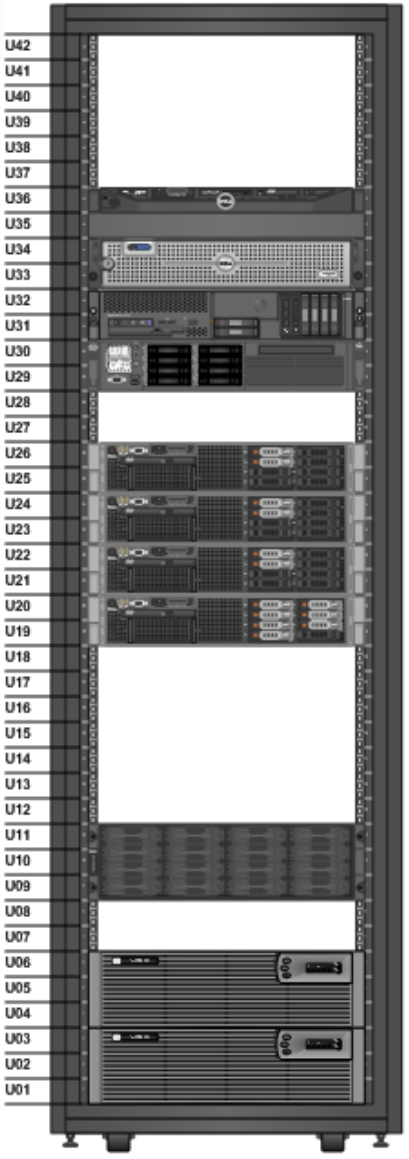


# Services to be consolidated

- Base services
  - DNS / LDAP / NTP / DHCP
  - May use backup servers instead of VM failover
- Web services
  - Currently ~24 different services
  - Some may be split and more growth in future
- Datacenter services (total of ~18)
  - Ingest / SptrDownloader …
  - Internal monitoring (nagios, ganglia, mrtg, Hitachi monitoring)
  - Split DC services (CUPS / Samba / FlexLM / puppet / syslog / mail / admin)
- Databases
  - Currently 4 (2 internal, 2 external)
  - Need to split into 8 (ARA, Auger, some WIKI stuff, own Ingest DB server)



# 222: Rack B5




Front View

Base services VM01 (4VMs/8max)

Pub01

Cobalt01

Cobalt03 possible

Cobalt05 future

DMZ Services VM01 (12VMs/24max)

DMZ Services VM03 (12VMs/24max)

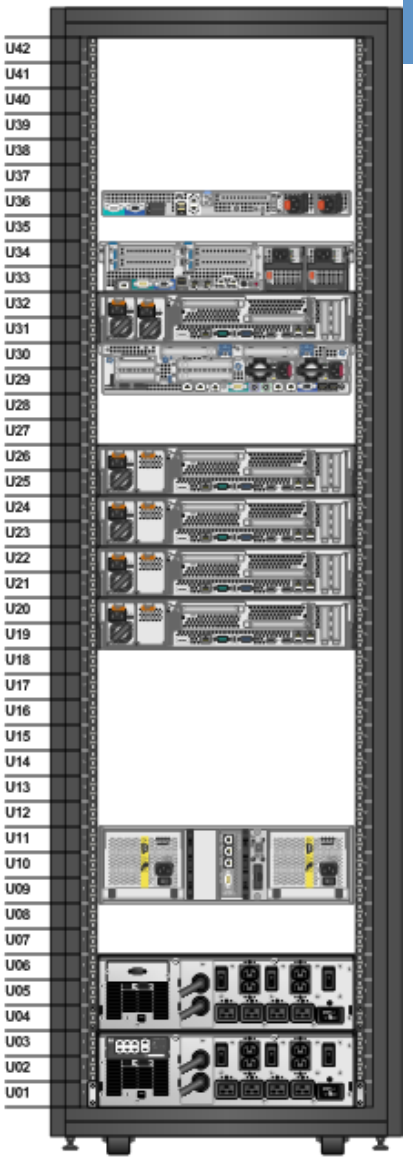
222 Services VM01 (12VMs/24max)

DB VM01 (12VMs/12max)

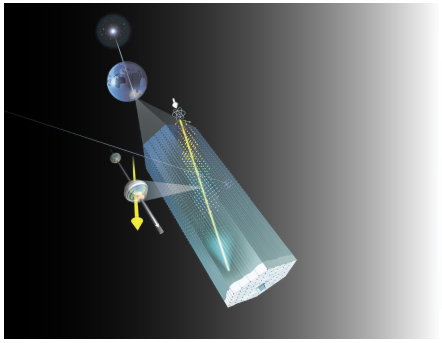
iSCSI SAN/NAS ???

222upsRB5A

222upsRB5B



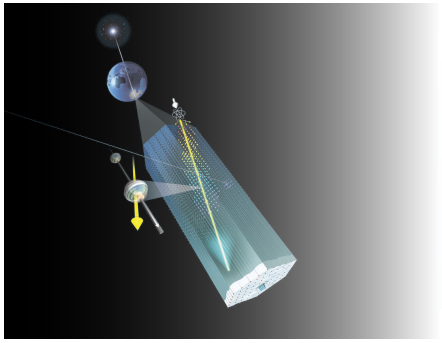

Rear View



# Datacenter layout

| Rack | Usage |
|---|---|
| RA0 | Lustre disks (Nexsan) |
| RA1 | YUM / Grid / Helpdesk / |
| RA2 | Future extension |
| RA3 | Lustre disks<br>New Hitachi AMS2500 |
| RA4 | Future extension |
| RA5 | Services Rack 2<br>Lustre disks (Nexsan + Apple) |
| RA6 | Lustre disks<br>Apple XRaid |
| RA7 | Lustre OSSs (ana / user)<br>Lustre disks (Nexsan) |
| RA8 | Lustre disks (Nexsan) |

| Rack | Usage |
|---|---|
| RB0 | Filtek HSM / Backup (Atempo +Amanda) |
| RB1 | Lustre MGS / MDS |
| RB2 | LUSTRE OSSs (sim/exp)<br>Non-Lustre /net/user |
| RB3 | npx2-uwa |
| RB4 | Lustre disks (AMS 2500) |
| RB5 | Services Rack 1 |
| RB6 | Tape Robot |
| | |
| | |



# Research topics

- VMs
  - KVM or other solutions (OpenVZ looks very interesting)
  - GUI management (lots of libvirt based solutions)
  - Live migrationz
  - Snapshot strategies
- Shared storage / dedicated storage
  - SAN / NAS solutions (NFS/ iSCSI / FC)
  - SSD performance
  - SSD access from VMs
- Configuration management
  - Puppet (currently preferred)
  - CFEngine or others
- Network redesign
  - Switching layout
  - Fibre Channel over Ethernet integration