CernVM
SoFTware Appliance

Virtualization R&D: Activities and Perspectives

Predrag Buncic
Carlos Aguado Sanchez
Jakob Blomer
Artem Harutyunyan

CernVM
SoFTware Appliance

Overview

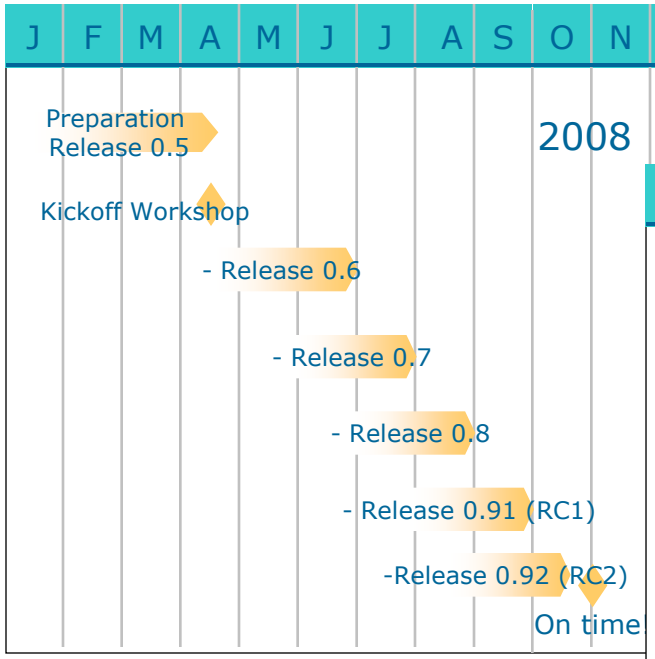
- Release Status
  - CernVM 2.0 main features
  - Release process
  - CernVM FS v2
- Scaling up CernVM
  - Building CDN for CernVM
- Supporting Infrastructure
- CernVM & Clouds
- Performance issues
- From R&D to service
- Conclusions

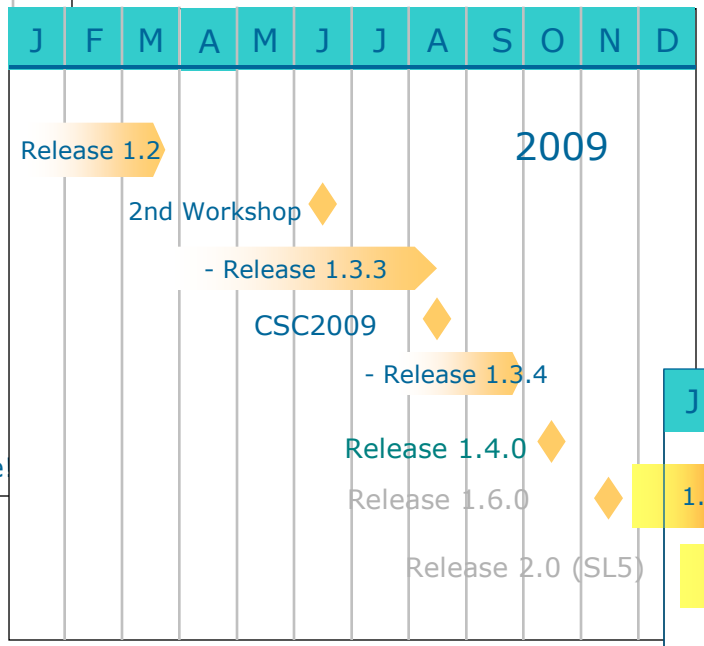
CernVM
SoFTware Appliance

Release Status

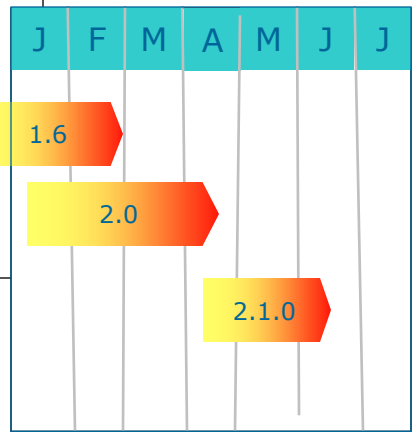
Release Status

- Semi production operation in 2009
  - Final CernVM 1.6 release 28/02/2010
  - Aiming for release in March





2.0.5 RC1 released 17/04
2.0,7 RC2 released 06/05
2.1.0 released 10/06

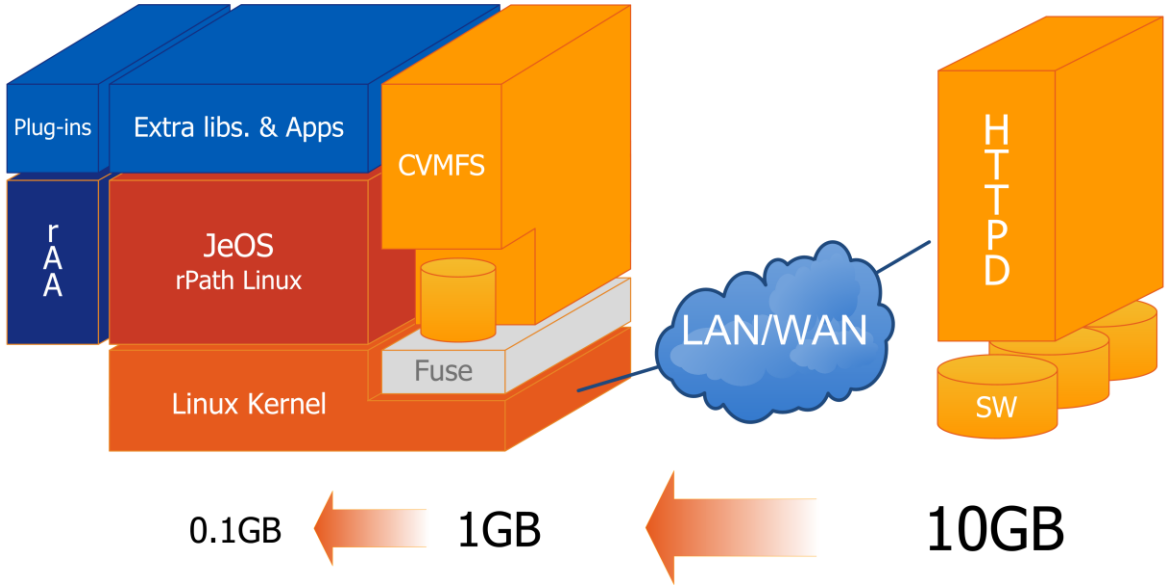


- Quick development cycles
  - In close collaboration with experiments
  - Very good feedback from enthusiastic users
- Modest manpower at CERN
  - 1 Fellow & doctoral student

CernVM 2.x main features

- OS
  - Now fully based on SL5 delivered by rPath (imported RPMs are automatically kept up to date with upstream distribution)
  - New appliance UI based or rPath Appliance Platform Agent 3.0
  - Xfce window manager for desktop version (smaller memory footprint)
- CernVM FS v2
  - Transparent file compression
  - Integrity checks using checksums, signed file catalog
  - Catalogs are given time to live, which allows for automatic updates
  - Supports nested catalogs
  - Supports failover and load balancing mechanisms for chain of forward/reverse proxy servers
  - Supports managed file cache with quotas
  - Can be deployed on physical machines (RPMs are provided)
  - Used beyond software delivery (ATLAS Conditions Data)

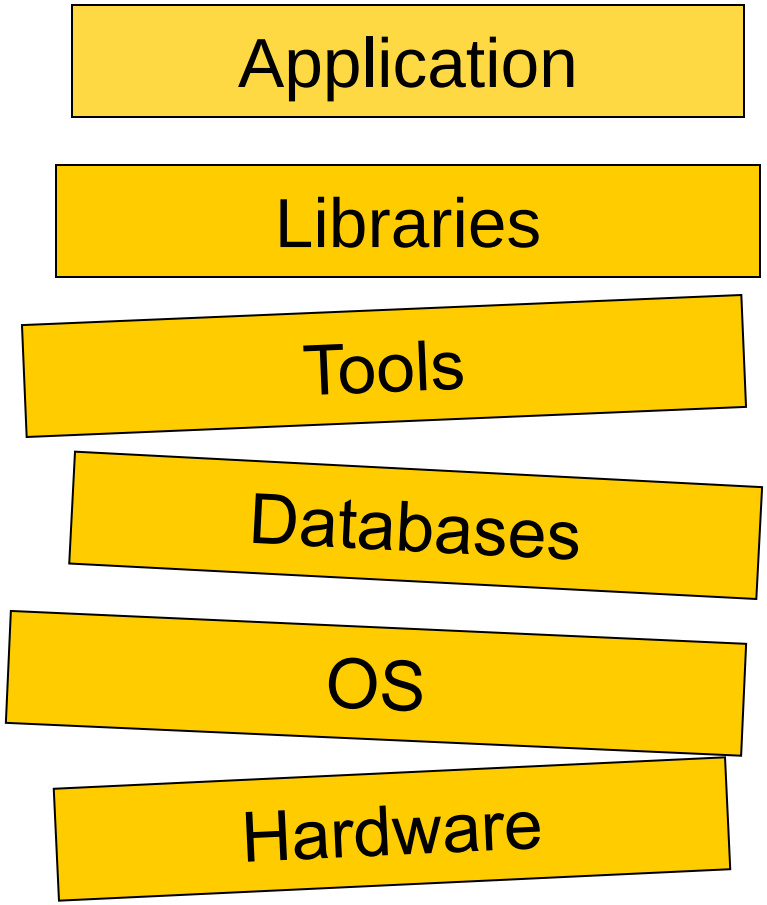
CernVM Model



1. Minimal Operating System (common platform) sufficient to satisfy the most basic use cases of LHC experiments
2. File system based on HTTP protocol and optimized for software distribution using aggressive caching and capable of off-line operations
3. Appliance Agent providing a simple Web UI for configuration and maintinance

CernVM
SoFTware Appliance

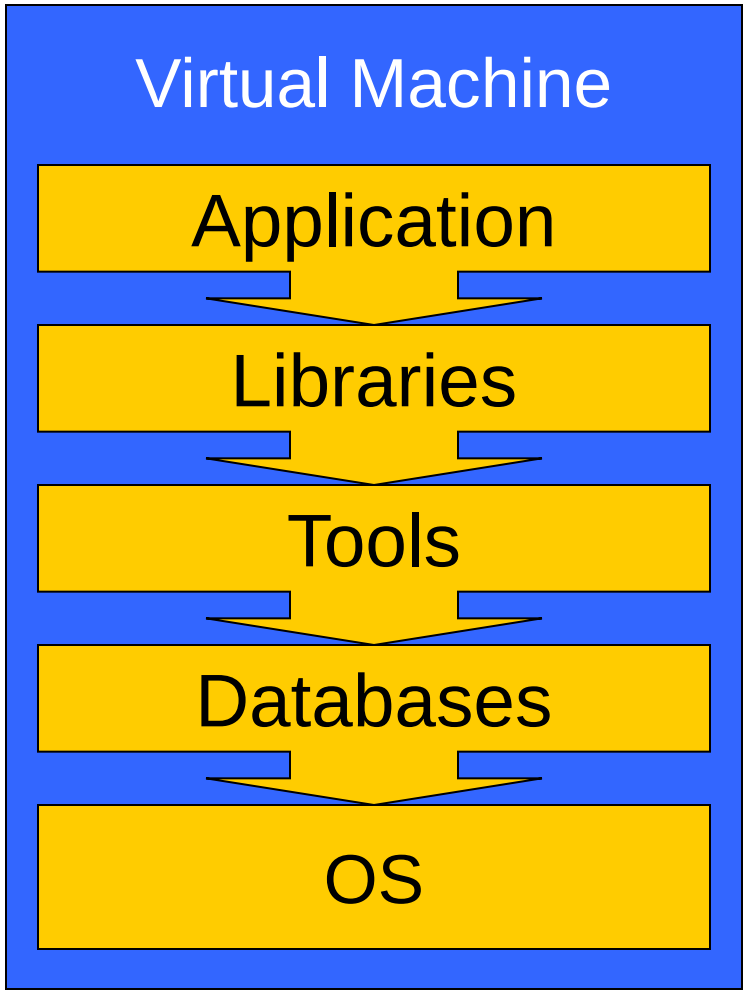
Horizontal Integration



- Traditional model
  - Horizontal layers
  - Independently developed
  - Maintained by the different groups
  - Different lifecycle
- Application is deployed on top of the stack
  - Breaks if any layer changes
  - Needs to be certified every time when something changes
  - Results in deployment and support nightmare

CernVM
SoFTware Appliance

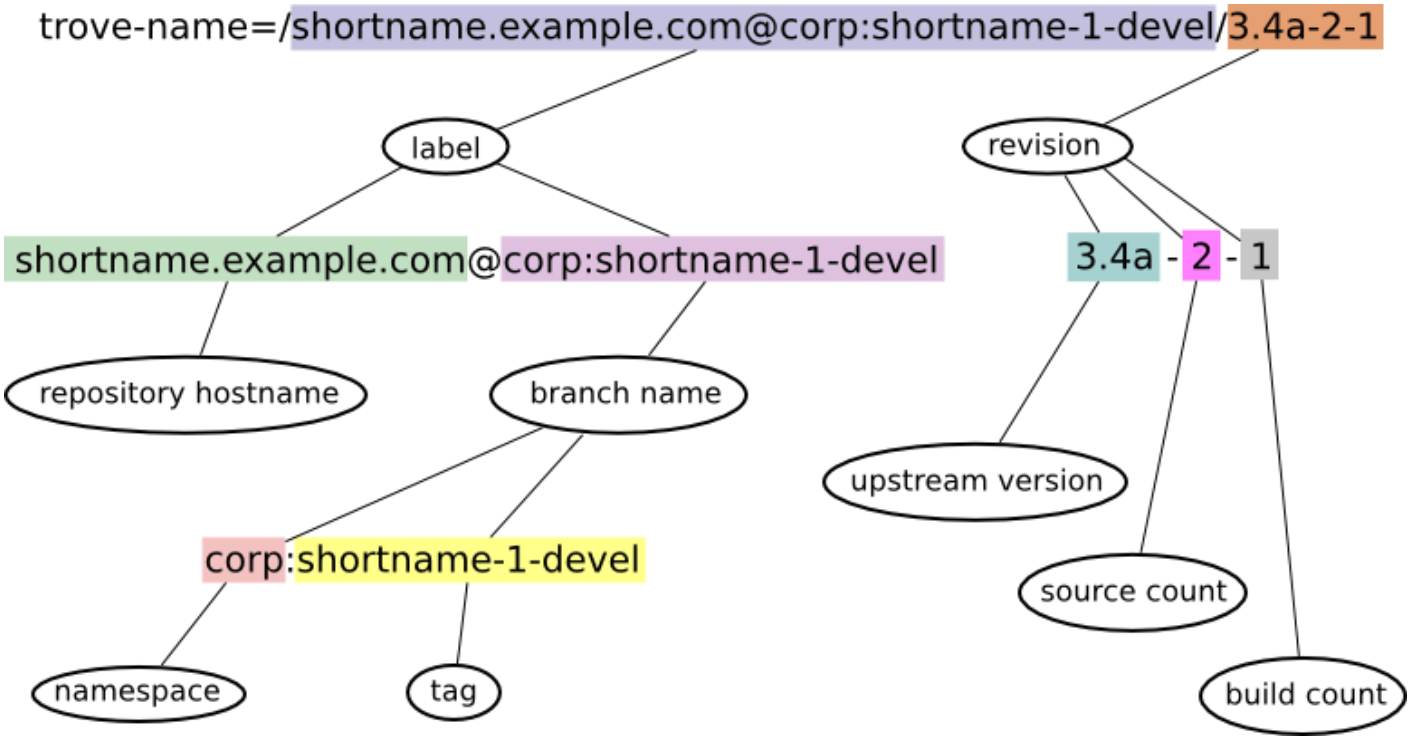
Vertical Integration



- Application driven approach
  - Analyzing application requirements and dependencies at build time
  - Adding required tools and libraries
  - Adding minimal OS and dependencies and bundling all this into Virtual Machine image
- Virtual Machine images are versioned just like the applications
  - Assuring accountability to mitigate possible negative aspects of newly acquired application freedom

CernVM
SoFTware Appliance

Conary Package Manager

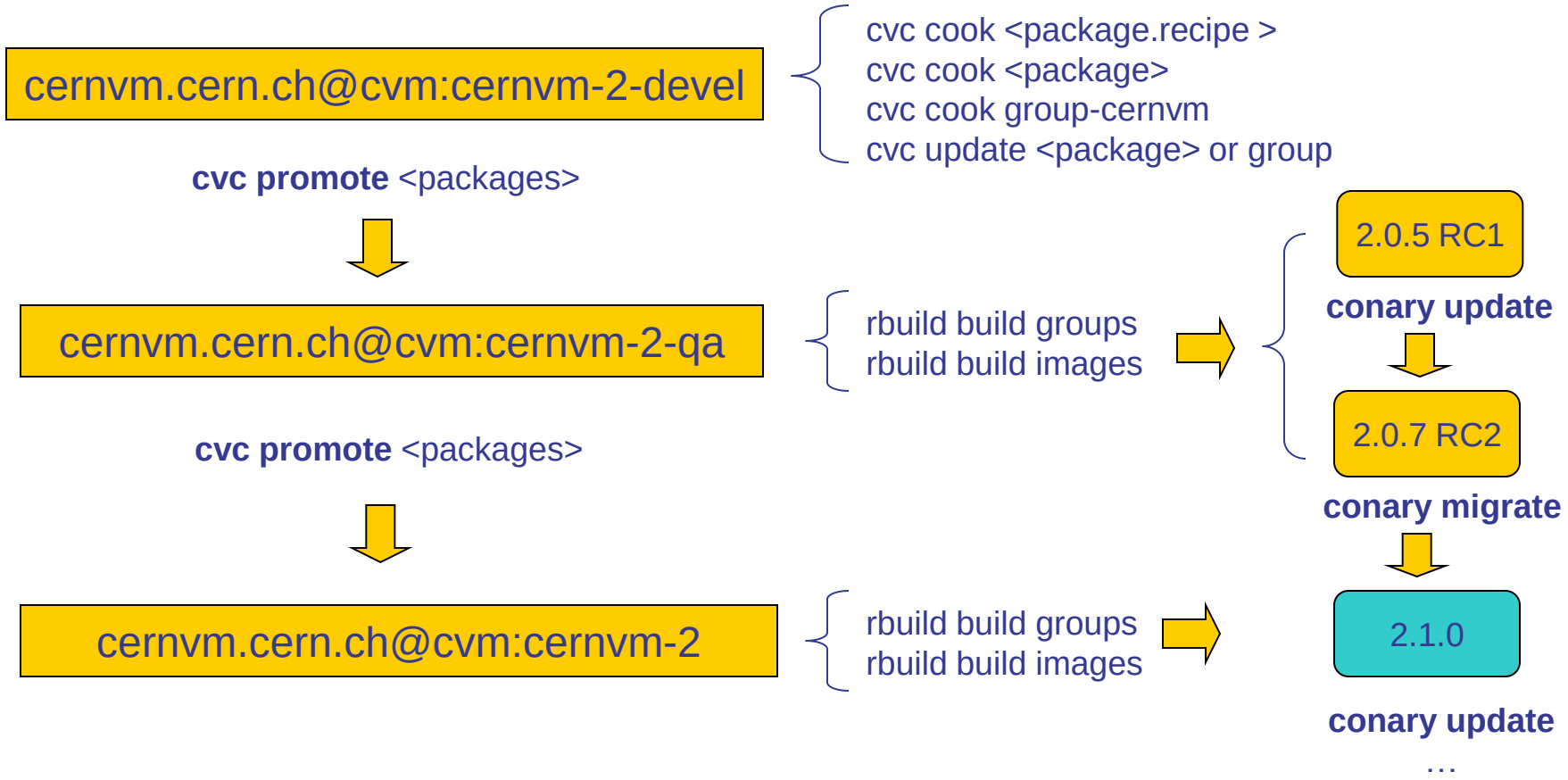


Fully accountable, every component on the system is versioned (database)

Allows updates, rollbacks and can reproduce exact system configuration at any given time

Can use multiple repositories (private and public) with signed packages

CernVM Release Process



CernVM
SoFTware Appliance

Scientific Linux

rpm import

cernvm.cern.ch@cvm:cernvm-2

scientific.rpath.org@rpath:sl-5

ruild build images

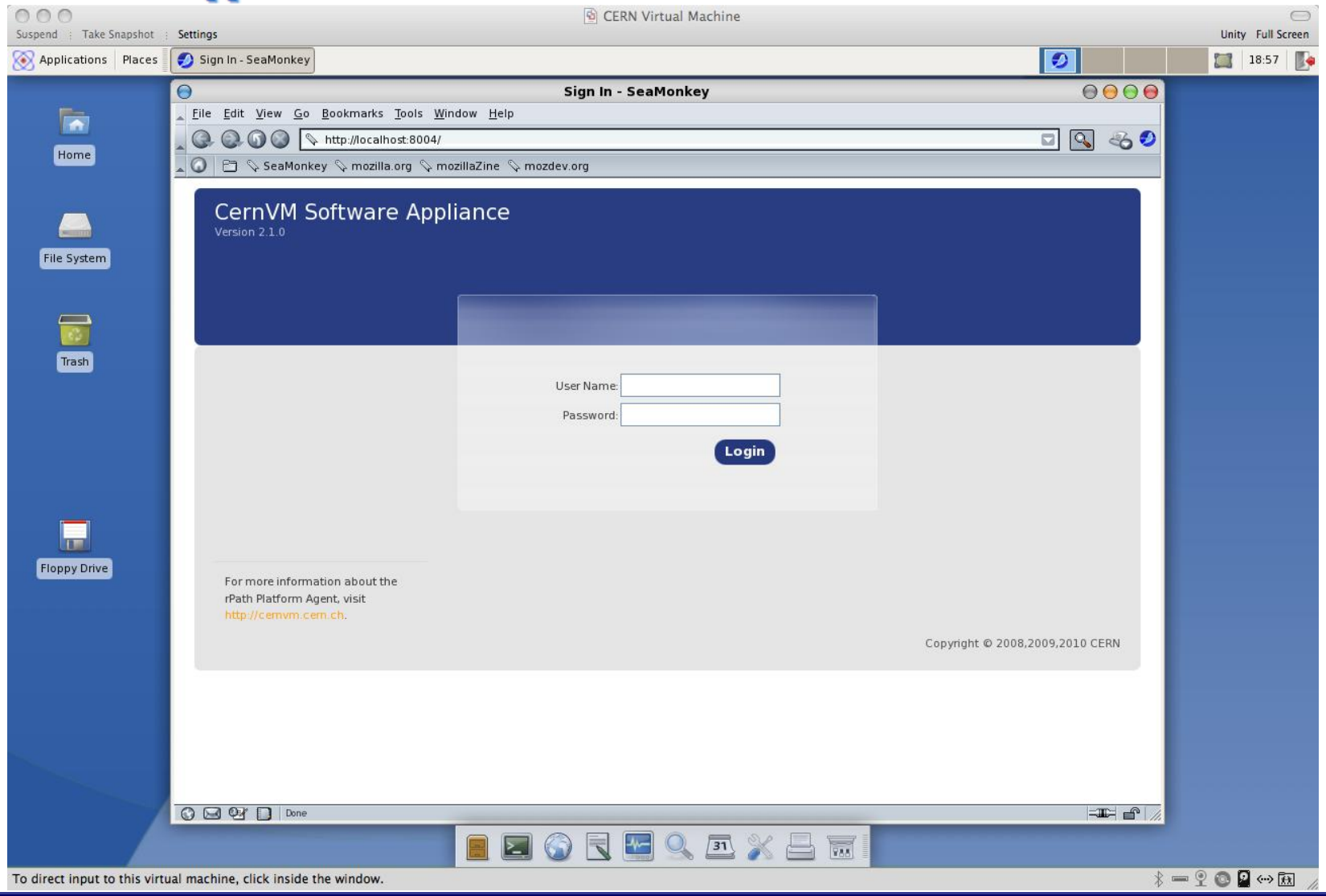
CERN Virtual Machine

Release: CERN Virtual Machine, version 2.1.0

Release created Sun, 13 Jun 2010 16:52:56 CEST
Release updated Mon, 14 Jun 2010 15:39:43 CEST
Release published Mon, 14 Jun 2010 15:39:54 CEST

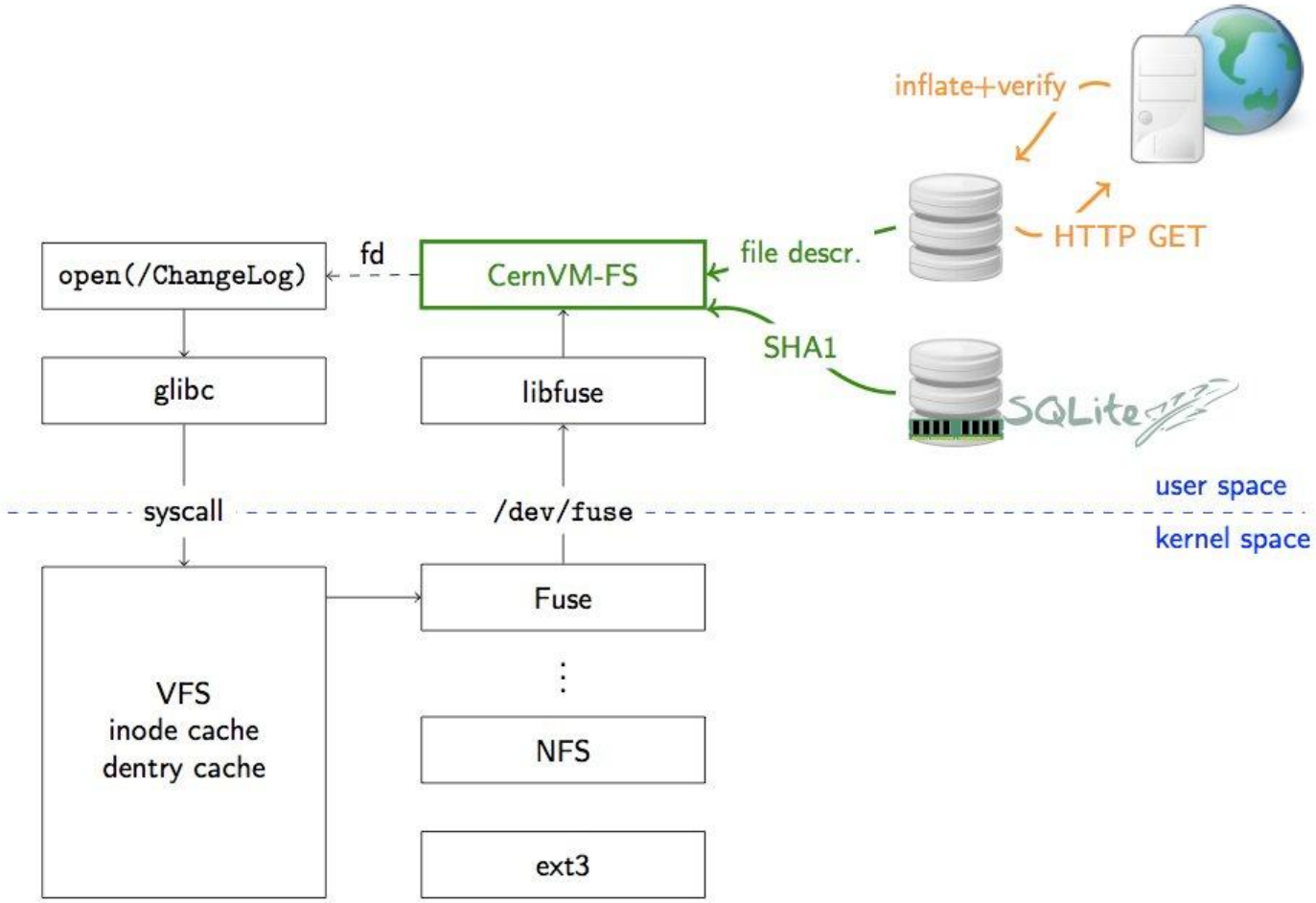
VMware, VirtualBox, QEMU, KVM,
Parallels, HyperV, Xen
x86, x86_64
…

Is this a Mac or a PC?



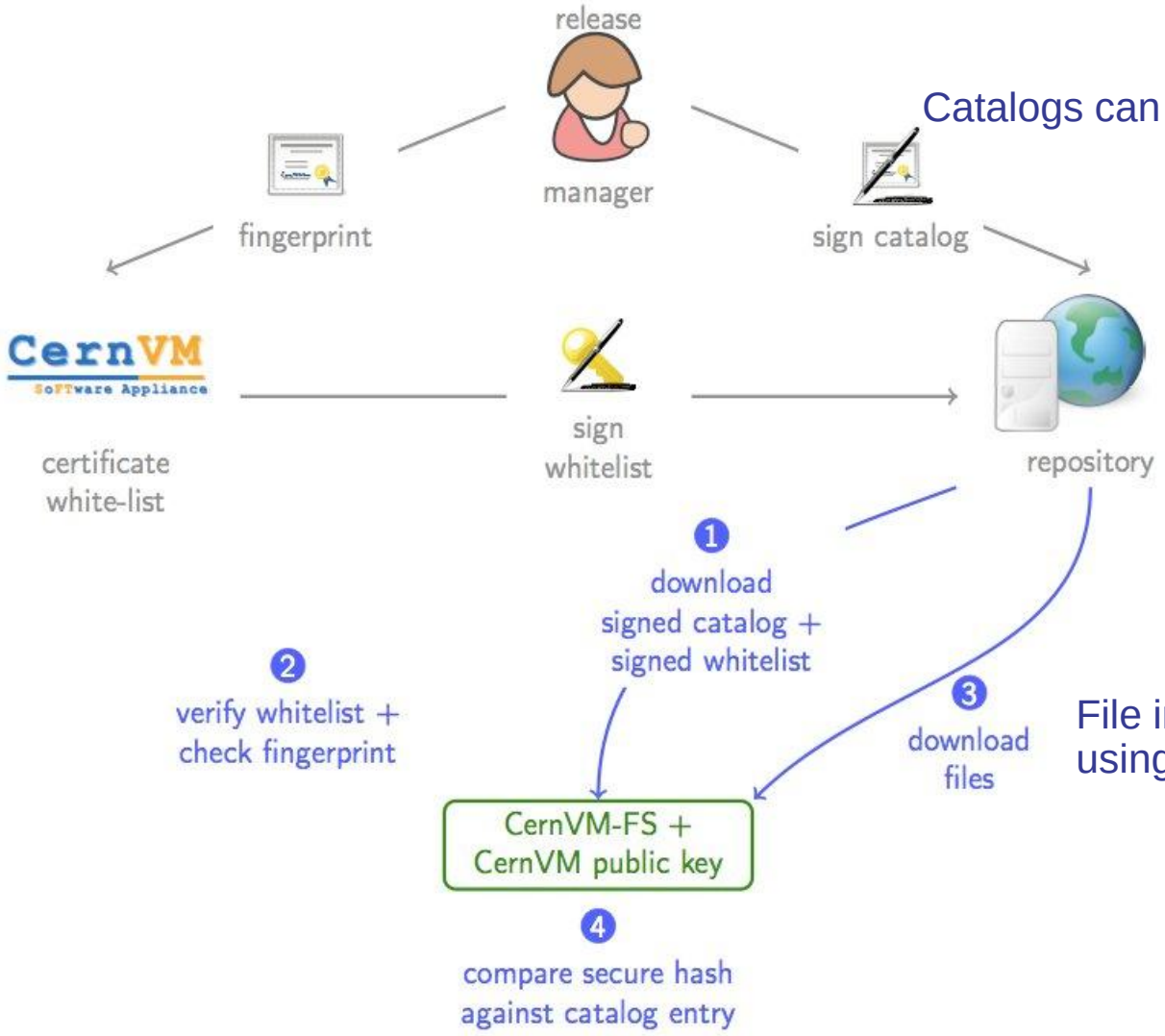
CernVM
SoFTware Appliance

CVMFS 0.2 (v2)



More details in Jakob's talk on Tuesday afternoon…

Security



Catalogs can be signed with X.509 certificate

Access control (requiring users to register their VMs) could be implemented

File integrity is verified on download using SHA1 checksum

CernVM
SoFTware Appliance

Scaling up CernVM

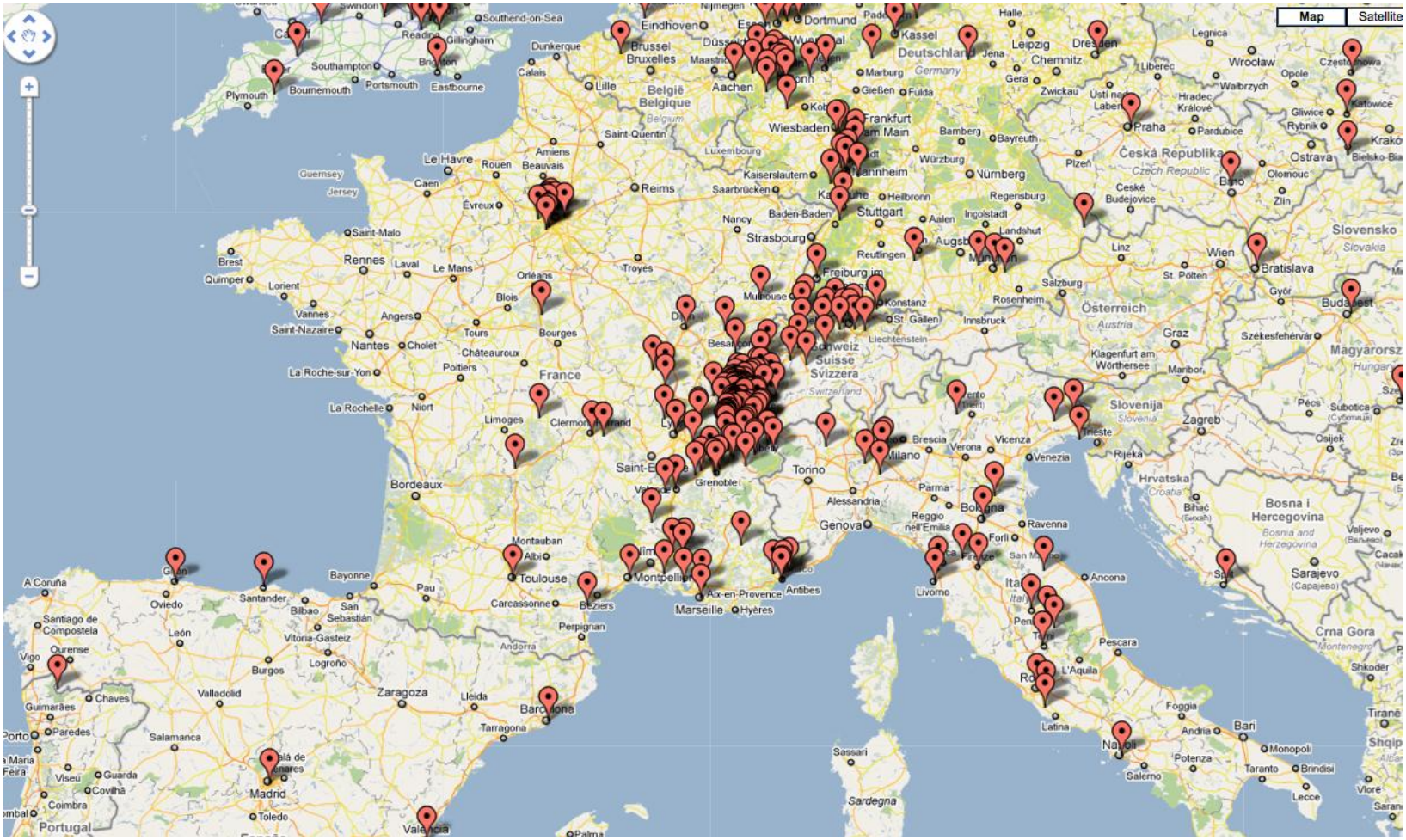
CernVM
SoFTware Appliance

CernVM Users

~2200 different IP addresses

CernVM
SoFTware Appliance

CernVM Users - Europe

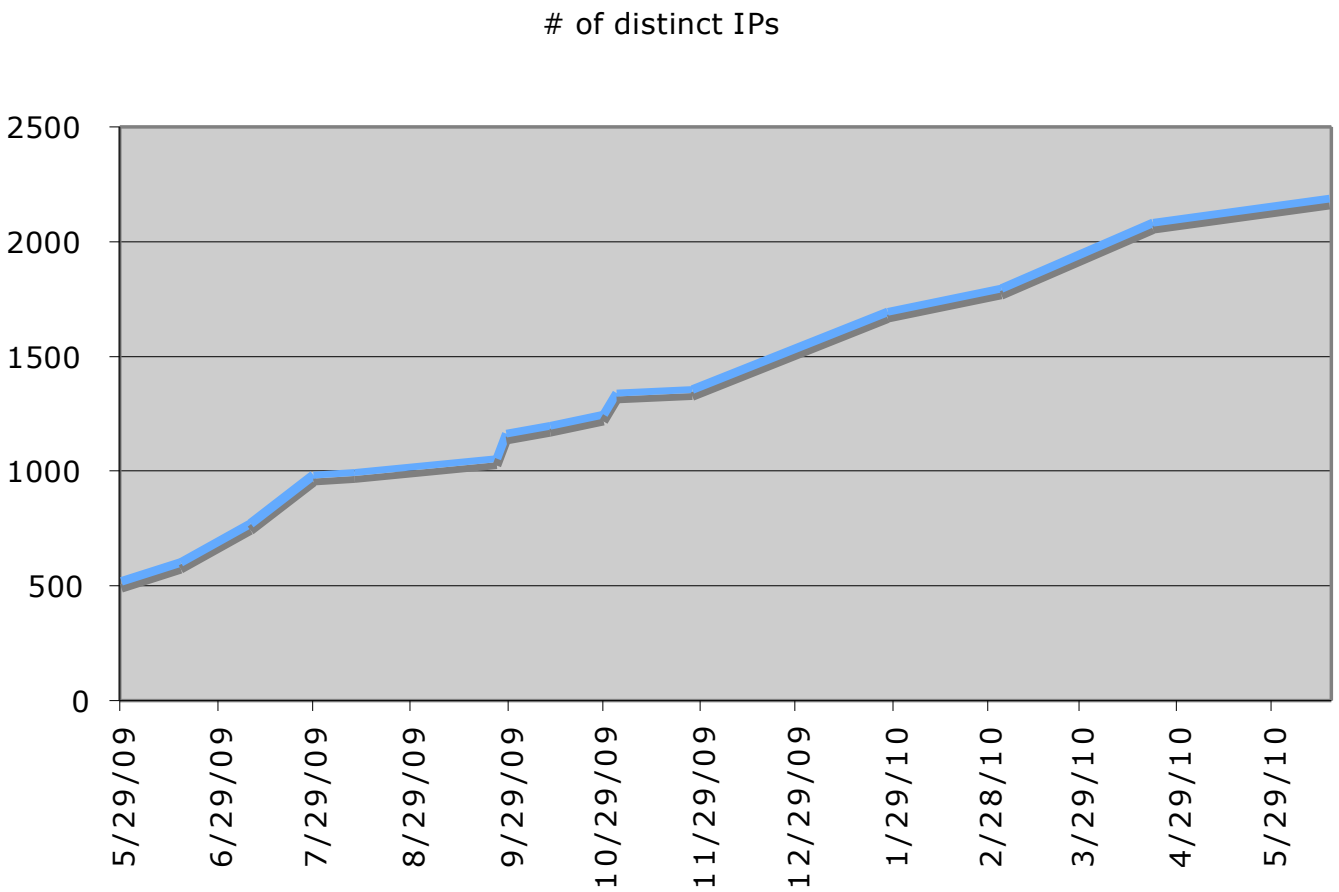


CernVM
SoFTware Appliance

CernVM Users - US

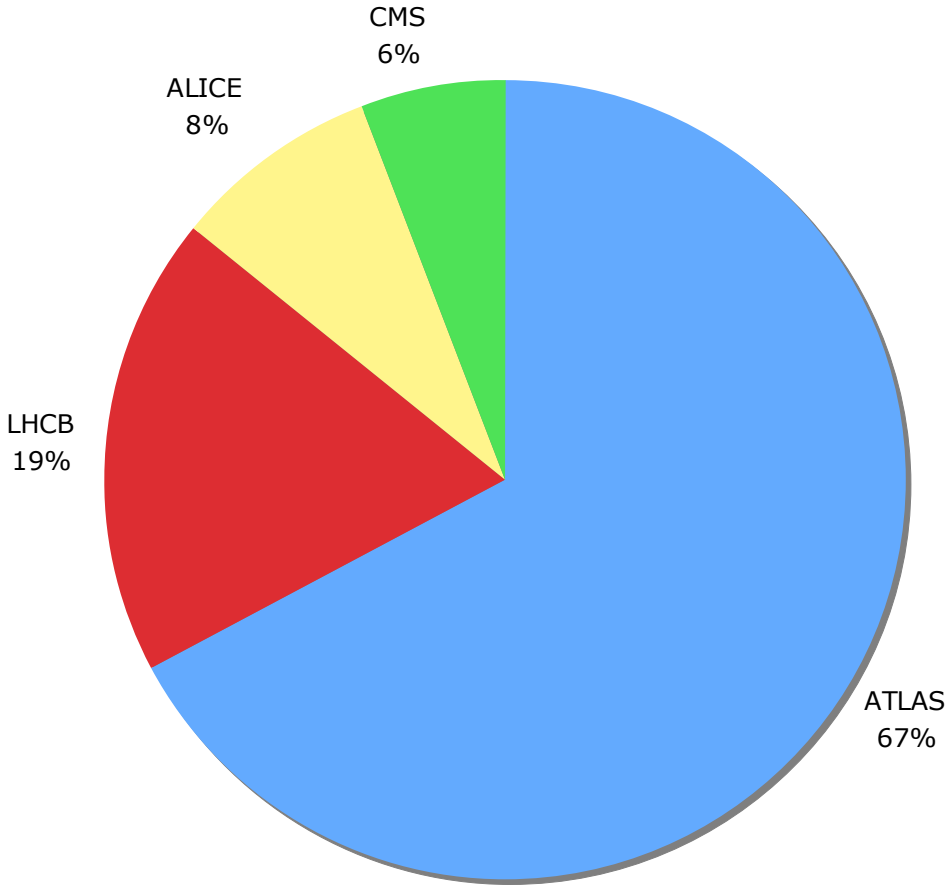
CernVM Users - CERN

Evolution over past year



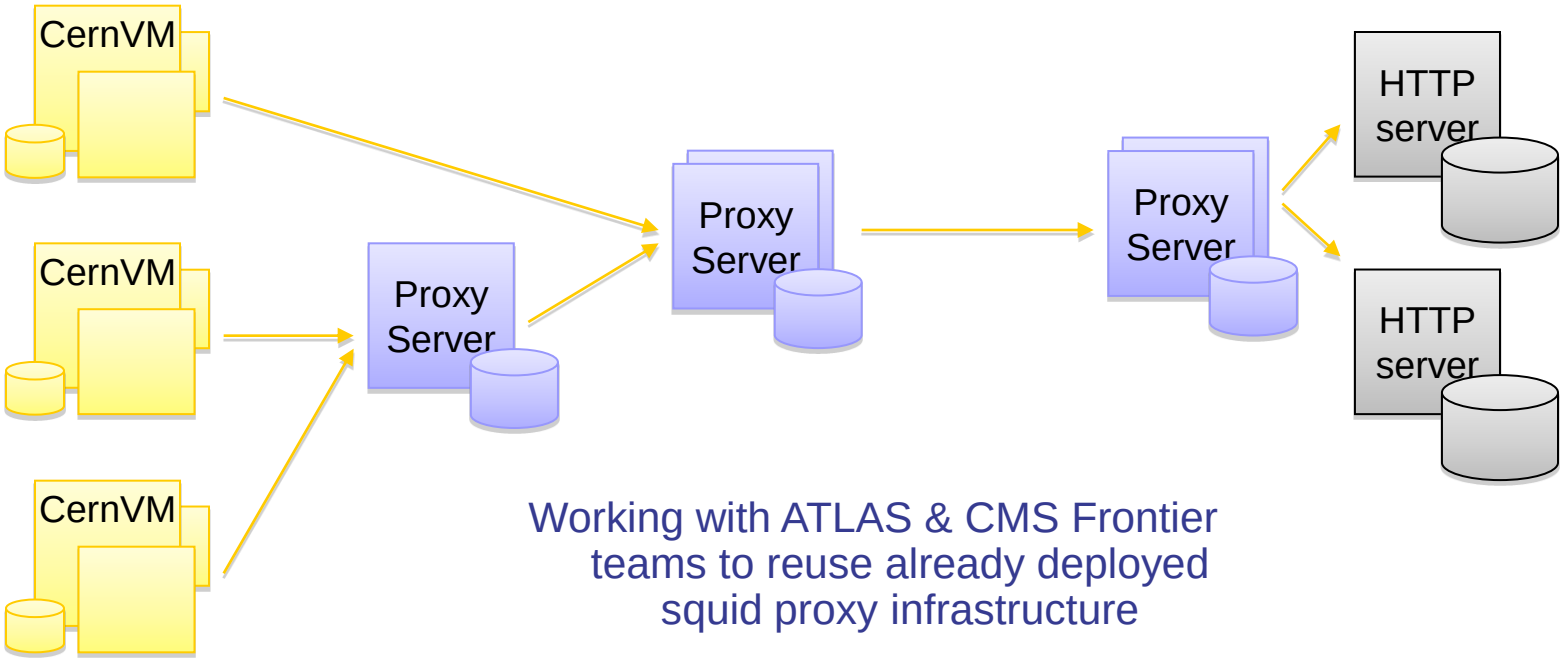
CernVM
SoFTware Appliance

Users by experiment



Scaling CVMFS

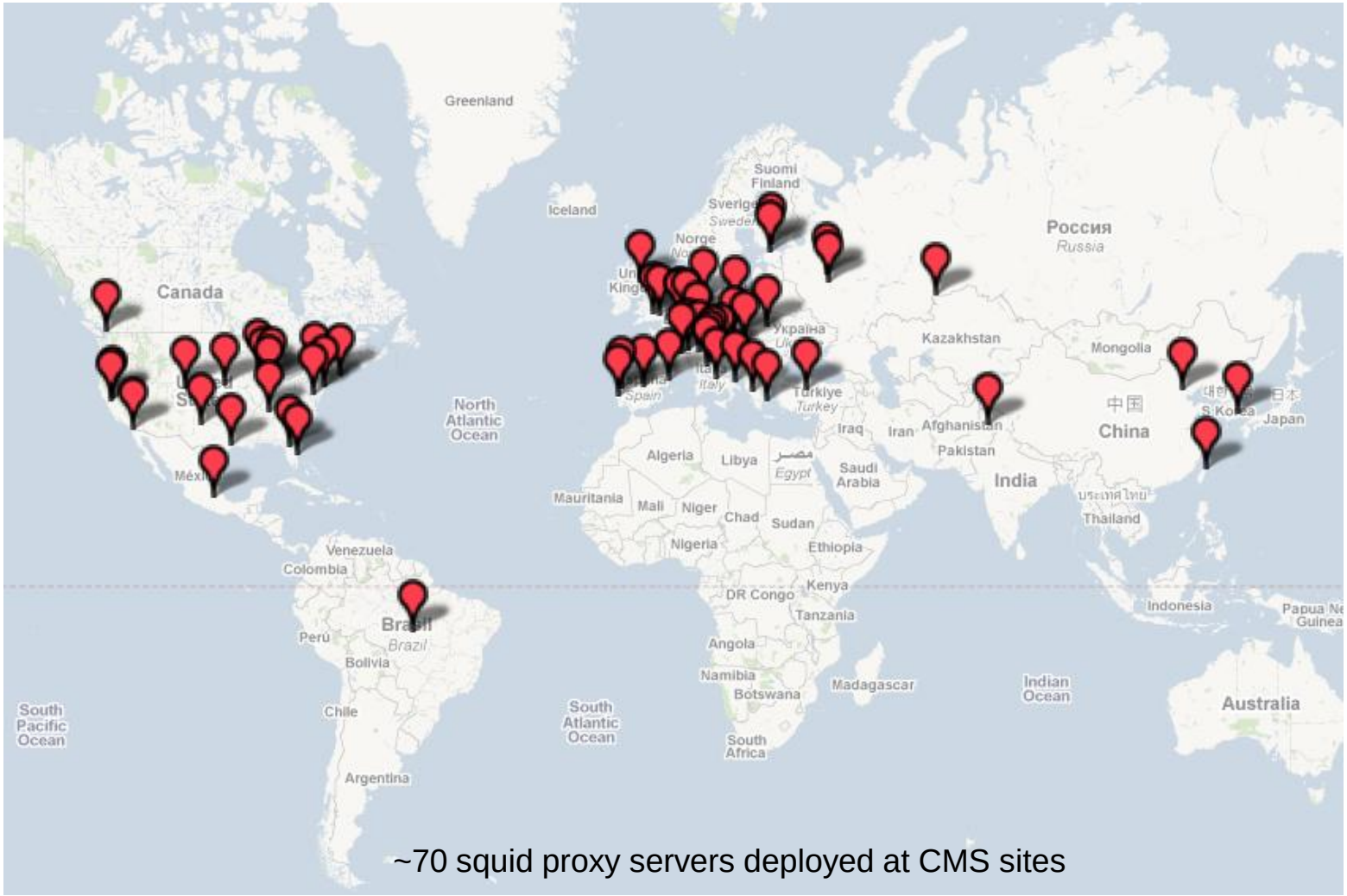
Proxy and slave servers could be deployed on strategic locations to reduce latency and provide redundancy



Working with ATLAS & CMS Frontier teams to reuse already deployed squid proxy infrastructure

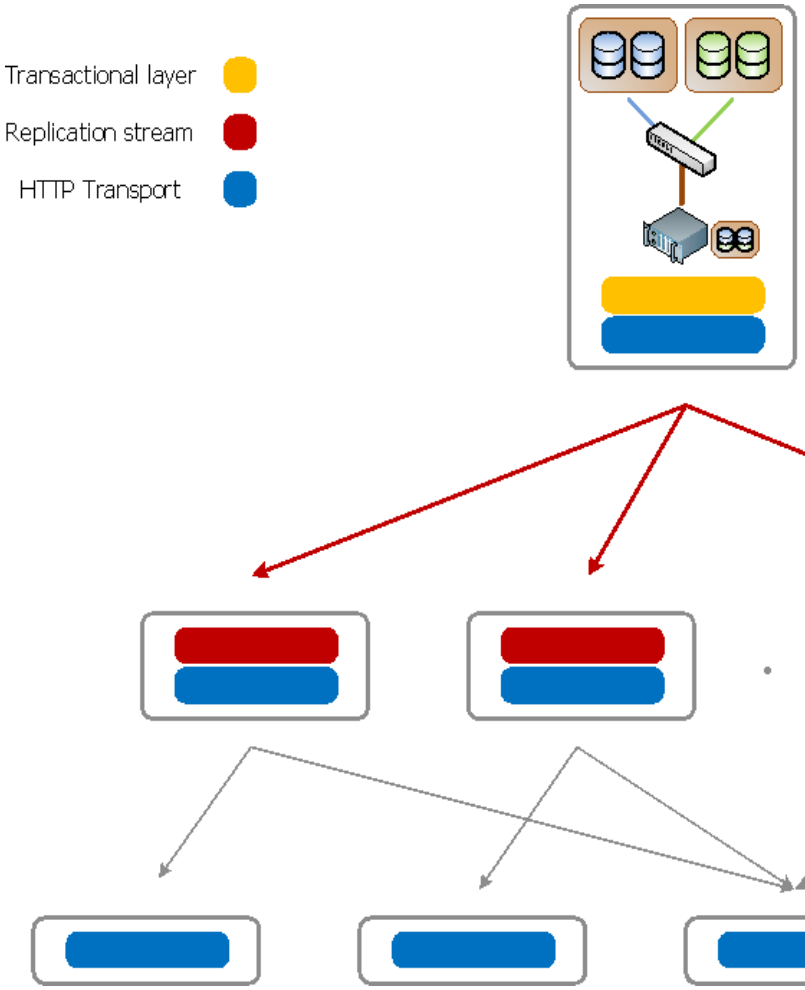
CernVM
SoFTware Appliance

C(ernVM)DN



~70 squid proxy servers deployed at CMS sites

C(ernVM)DN + SimpleCDN



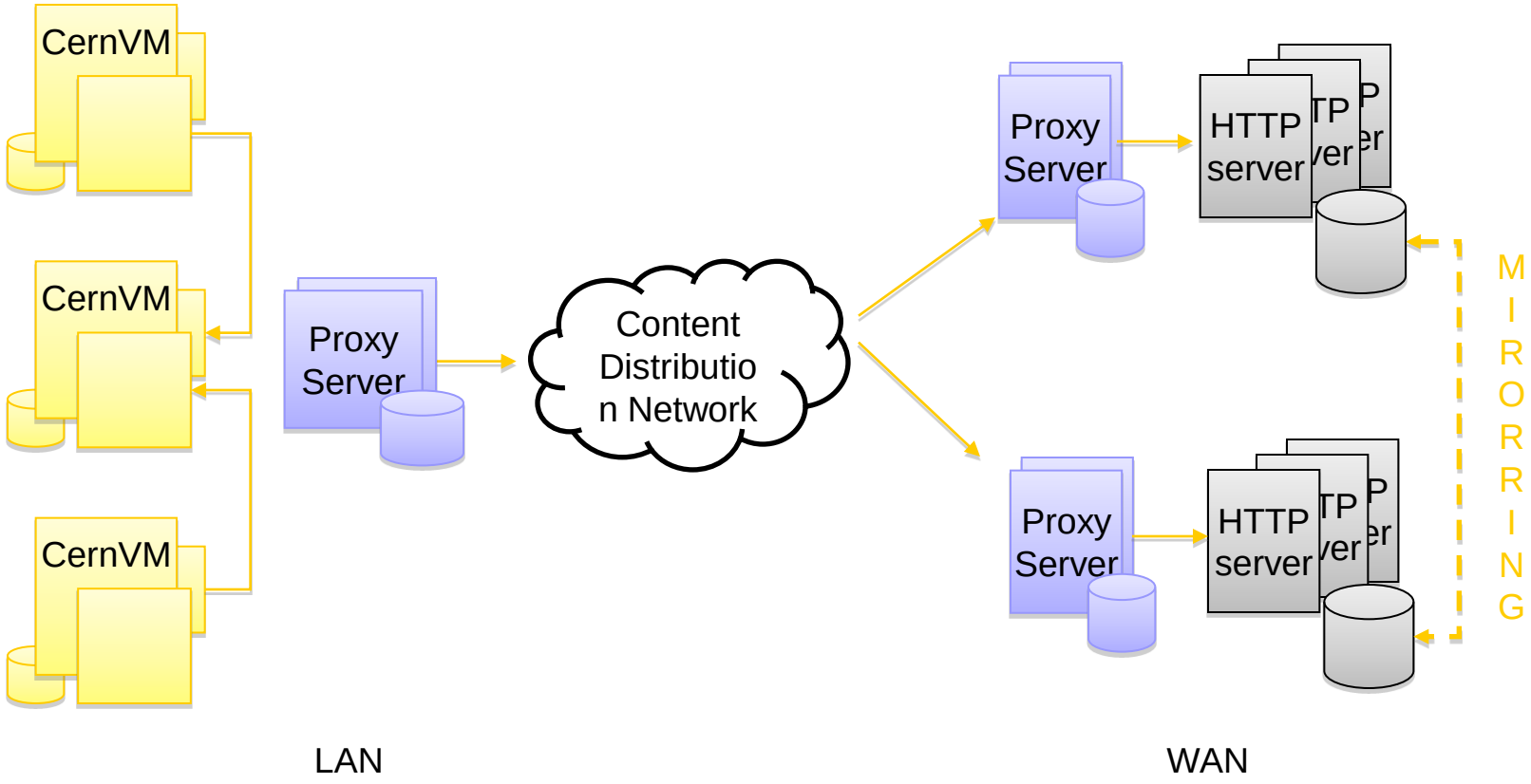
Transactional layer:
• Where releases are created and published by repository managers
• Can act as a proxy for off-site repositories (calibration files, non-CERN experiments)

SimpleCDN
12 edge locations, has access to over 1+ Tbps of regional and international fiber connections through peering

Replication layer (streams):
• Offline: ZFS streams, rsync, etc.
• Online: squid cache manager

Edge of the CDN
User addressable via central configuration or DNS

Final Goal (CVMFS v3)



CROWD: P2P like mechanism for discovery of nearby CernVMs and cache sharing between them. No need to manually setup proxy servers (but they could still be used where exist)

Use Content Delivery Network to remove a single point of failure and fully mirror the central distribution to at least one more site.

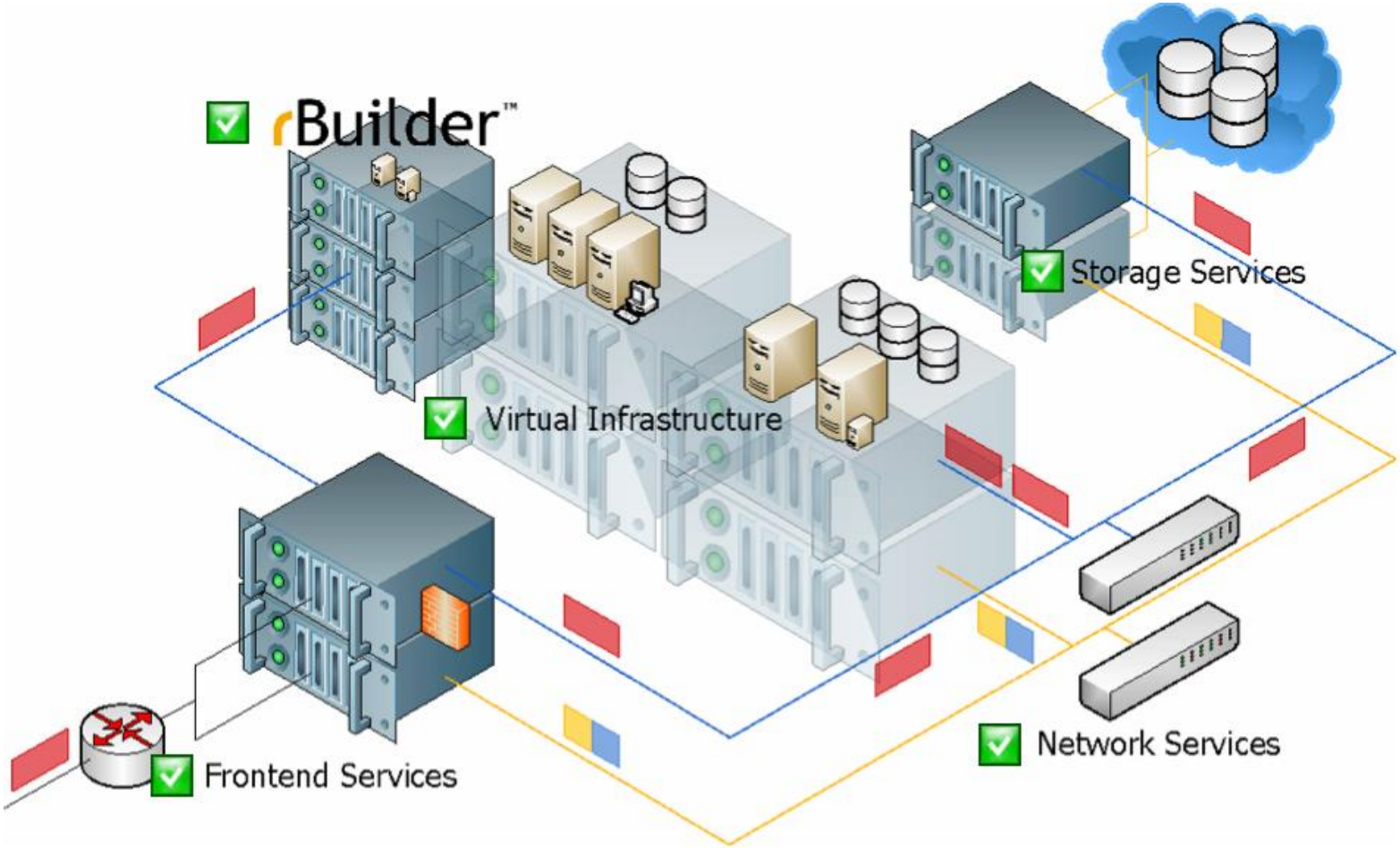
CernVM

SoFTware Appliance

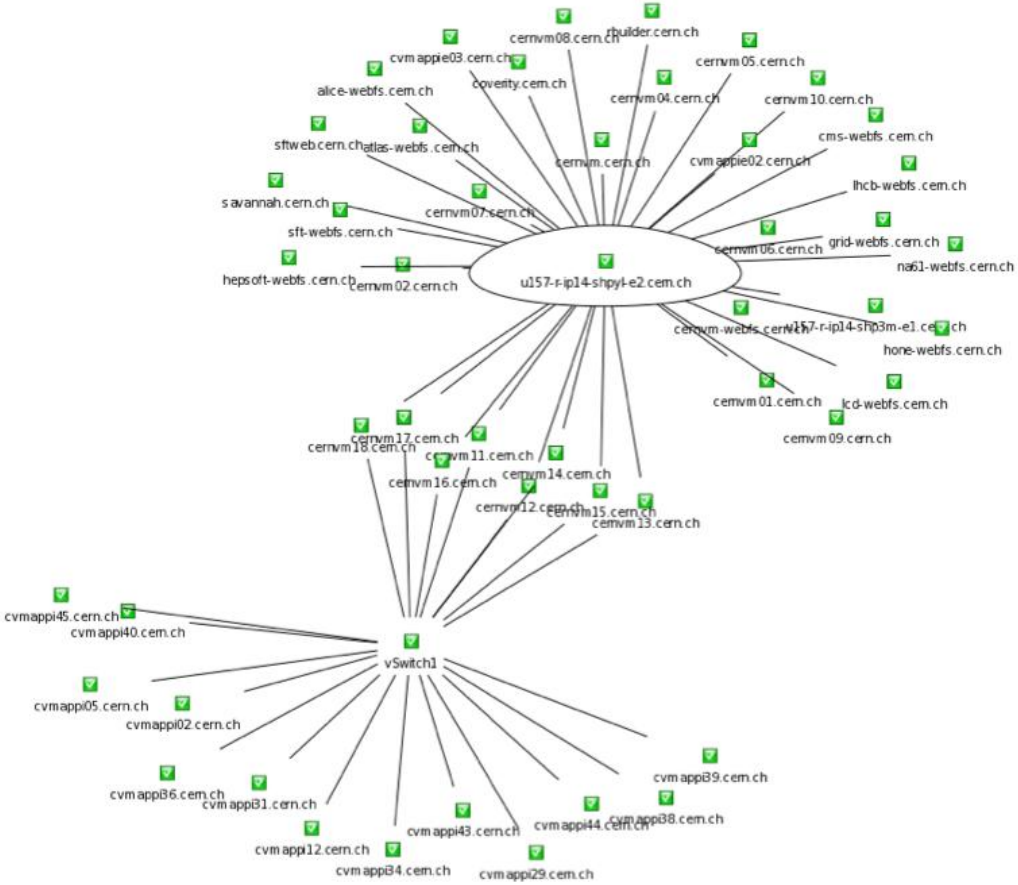
Supporting Infrastructure

CernVM
SoFTware Appliance

Virtualized Infrastructure



Currently Supported CVMFS Repositories

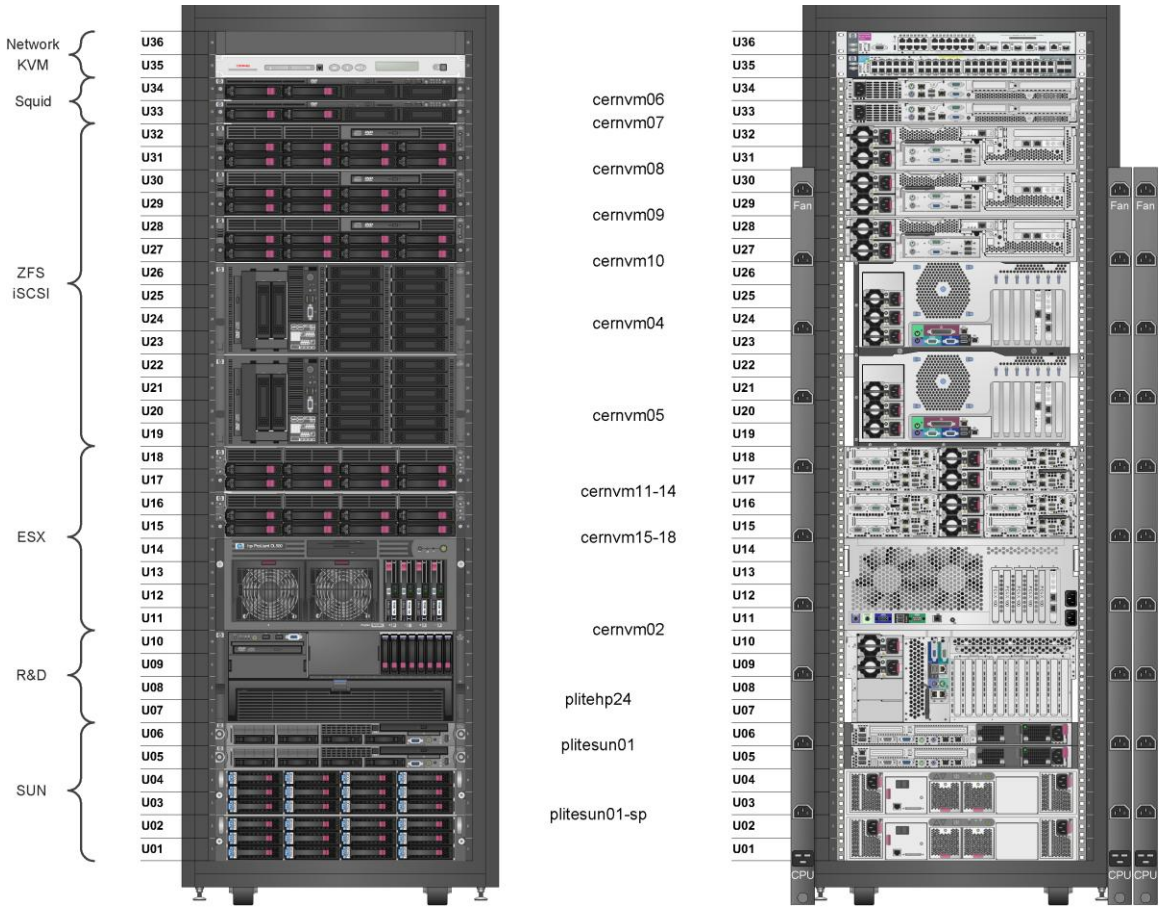


CernVM Services Network Map

- Experiments
  - ATLAS
  - ALICE
  - LHCb
  - CMS
  - NA61
  - LCD
  - H1
- Projects
  - HEPSOFT (TH)
  - SFT (AA externals)

CernVM
SoFTware Appliance

CernVM Servers



2 x 4 x HP DL170 G6 = 64 cores in 4U
192 GB, 3 GB/core, 20kCHF

1 x HP DL 580 = 24 cores in 4U
48 GB, 2 GB/core, 30kCHF

CernVM
SoFTware Appliance

CernVM & Clouds

CernVM
SoFTware Appliance

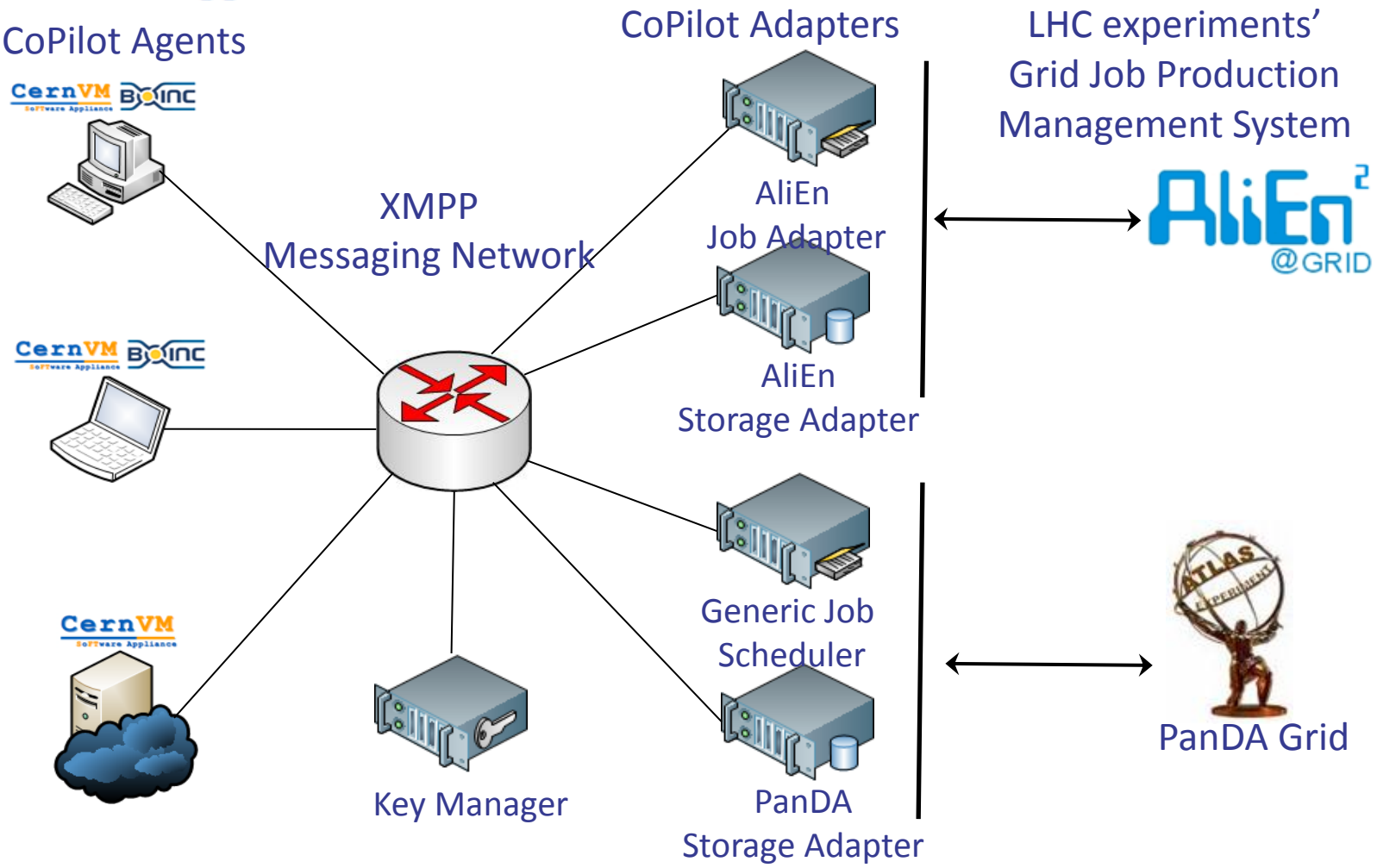
Follow up from last year's workshop

1. Provide infrastructure in CERN's computer centre for the preparation of CernVM images and the Virtual Organization's application software delivery to them. CernVM images are generated by tools of the CernVM project, http://cern.ch/cernvm, which provides a virtual software appliance for developing and running LHC data analysis
2. Include the capability to run CernVM images in CERN's virtualized batch initiative.

CernVM@Cloud

- CernVM as job hosting environment on Cloud
  - Ideally, users would like to run their applications on the Grid (or Cloud) infrastructure in exactly the same conditions in which they were originally developed
  - Software can be efficiently installed using CVMFS
  - HTTP proxy assures very fast access to software even if VM cache is cleared
  - Multiple VM instance can share memory
  - VMware, KVM
- CernVM already provides development environment and can be deployed on Cloud (EC2, Nimbus, OpenNebula*)
  - One image supports all four LHC experiments
  - Easily extensible to other communities

CernVM CoPilot Architecture



Message driven system can be scaled up in case of high load by just adding new Adapter instances

CoPilot for PanDA

CernVM SoFTware Appliance

Generic Job Scheduler

PanDA Server

0. Request a job for execution

1. Instruct CoPilot Agent to start PanDA pilot

2. Request a job for execution

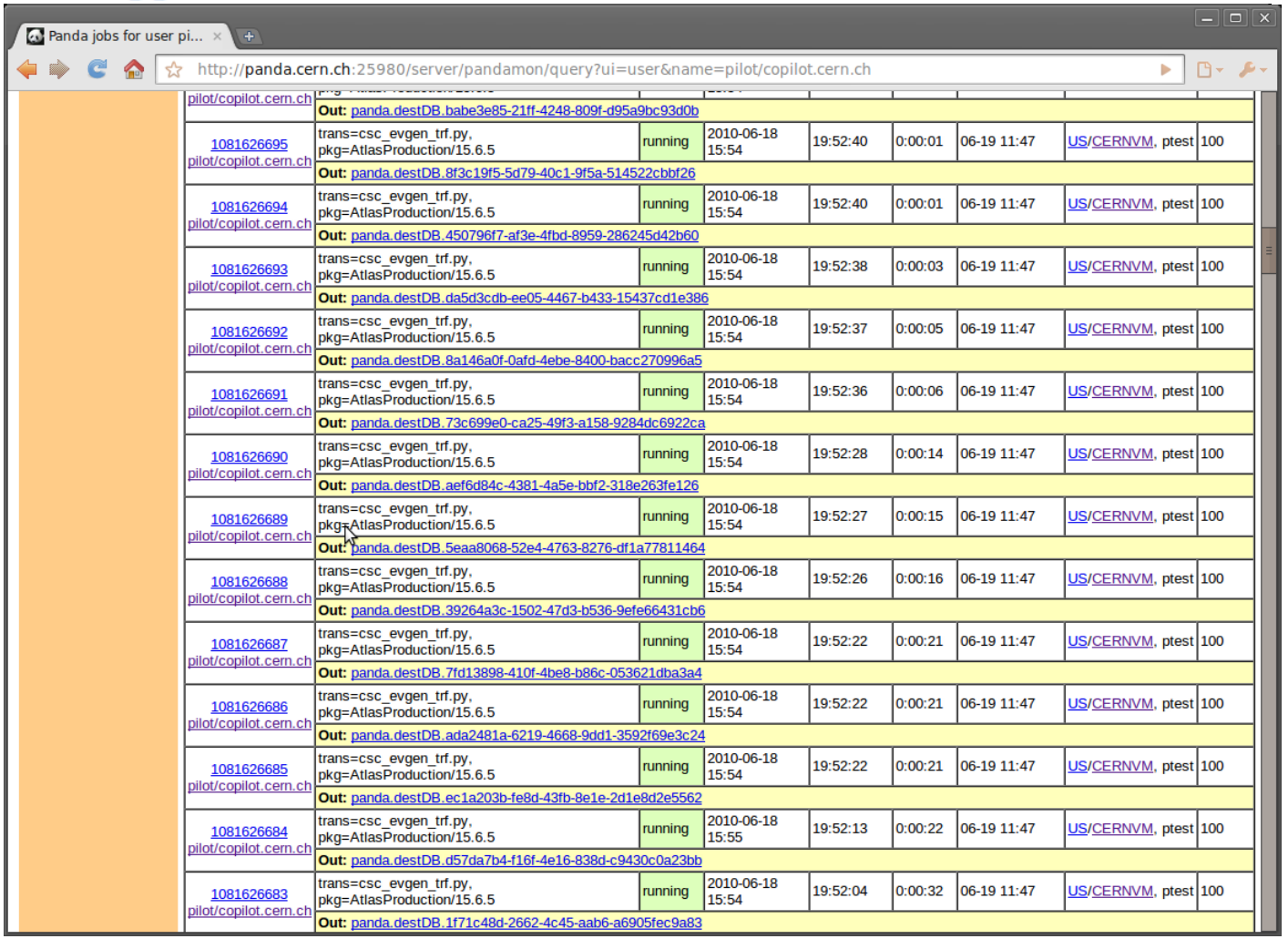
ATLAS Storage Adapter

3. Upload the output files once the job is done

4. Register output files, mark the job as completed

Key Manager service is used to secure the communication of the CoPilot components

PanDA on lxcloud@CERN



pilot/copilot.cern.ch	**Out:** panda.destDB.babe3e85-21ff-4248-809f-d95a9bc93d0b						
1081626695 pilot/copilot.cern.ch	trans=csc_evgen_trf.py, pkg=AtlasProduction/15.6.5	running	2010-06-18 15:54	19:52:40	0:00:01	06-19 11:47	US/CERNVM, ptest 100
	Out: panda.destDB.8f3c19f5-5d79-40c1-9f5a-514522cbbf26						
1081626694 pilot/copilot.cern.ch	trans=csc_evgen_trf.py, pkg=AtlasProduction/15.6.5	running	2010-06-18 15:54	19:52:40	0:00:01	06-19 11:47	US/CERNVM, ptest 100
	Out: panda.destDB.450796f7-af3e-4fbd-8959-286245d42b60						
1081626693 pilot/copilot.cern.ch	trans=csc_evgen_trf.py, pkg=AtlasProduction/15.6.5	running	2010-06-18 15:54	19:52:38	0:00:03	06-19 11:47	US/CERNVM, ptest 100
	Out: panda.destDB.da5d3cdb-ee05-4467-b433-15437cd1e386						
1081626692 pilot/copilot.cern.ch	trans=csc_evgen_trf.py, pkg=AtlasProduction/15.6.5	running	2010-06-18 15:54	19:52:37	0:00:05	06-19 11:47	US/CERNVM, ptest 100
	Out: panda.destDB.8a146a0f-0afd-4ebe-8400-bacc270996a5						
1081626691 pilot/copilot.cern.ch	trans=csc_evgen_trf.py, pkg=AtlasProduction/15.6.5	running	2010-06-18 15:54	19:52:36	0:00:06	06-19 11:47	US/CERNVM, ptest 100
	Out: panda.destDB.73c699e0-ca25-49f3-a158-9284dc6922ca						
1081626690 pilot/copilot.cern.ch	trans=csc_evgen_trf.py, pkg=AtlasProduction/15.6.5	running	2010-06-18 15:54	19:52:28	0:00:14	06-19 11:47	US/CERNVM, ptest 100
	Out: panda.destDB.aef6d84c-4381-4a5e-bbf2-318e263fe126						
1081626689 pilot/copilot.cern.ch	trans=csc_evgen_trf.py, pkg=AtlasProduction/15.6.5	running	2010-06-18 15:54	19:52:27	0:00:15	06-19 11:47	US/CERNVM, ptest 100
	Out: panda.destDB.5eaa8068-52e4-4763-8276-df1a77811464						
1081626688 pilot/copilot.cern.ch	trans=csc_evgen_trf.py, pkg=AtlasProduction/15.6.5	running	2010-06-18 15:54	19:52:26	0:00:16	06-19 11:47	US/CERNVM, ptest 100
	Out: panda.destDB.39264a3c-1502-47d3-b536-9efe66431cb6						
1081626687 pilot/copilot.cern.ch	trans=csc_evgen_trf.py, pkg=AtlasProduction/15.6.5	running	2010-06-18 15:54	19:52:22	0:00:21	06-19 11:47	US/CERNVM, ptest 100
	Out: panda.destDB.7fd13898-410f-4be8-b86c-053621dba3a4						
1081626686 pilot/copilot.cern.ch	trans=csc_evgen_trf.py, pkg=AtlasProduction/15.6.5	running	2010-06-18 15:54	19:52:22	0:00:21	06-19 11:47	US/CERNVM, ptest 100
	Out: panda.destDB.ada2481a-6219-4668-9dd1-3592f69e3c24						
1081626685 pilot/copilot.cern.ch	trans=csc_evgen_trf.py, pkg=AtlasProduction/15.6.5	running	2010-06-18 15:54	19:52:22	0:00:21	06-19 11:47	US/CERNVM, ptest 100
	Out: panda.destDB.ec1a203b-fe8d-43fb-8e1e-2d1e8d2e5562						
1081626684 pilot/copilot.cern.ch	trans=csc_evgen_trf.py, pkg=AtlasProduction/15.6.5	running	2010-06-18 15:55	19:52:13	0:00:22	06-19 11:47	US/CERNVM, ptest 100
	Out: panda.destDB.d57da7b4-f16f-4e16-838d-c9430c0a23bb						
1081626683 pilot/copilot.cern.ch	trans=csc_evgen_trf.py, pkg=AtlasProduction/15.6.5	running	2010-06-18 15:54	19:52:04	0:00:32	06-19 11:47	US/CERNVM, ptest 100
	Out: panda.destDB.1f71c48d-2662-4c45-aab6-a6905fec9a83						

CoPilot for AliEn

AliEn Job Adapter

AliEn Core Services

0. Request a job for execution

1. Append framework-specific information and get a job

2. Send user job JDL from Task Queue

3. Send input files and commands for execution

AliEn Storage Adapter

4. Upload the output files (and, optionally, the result of validation) once job is done

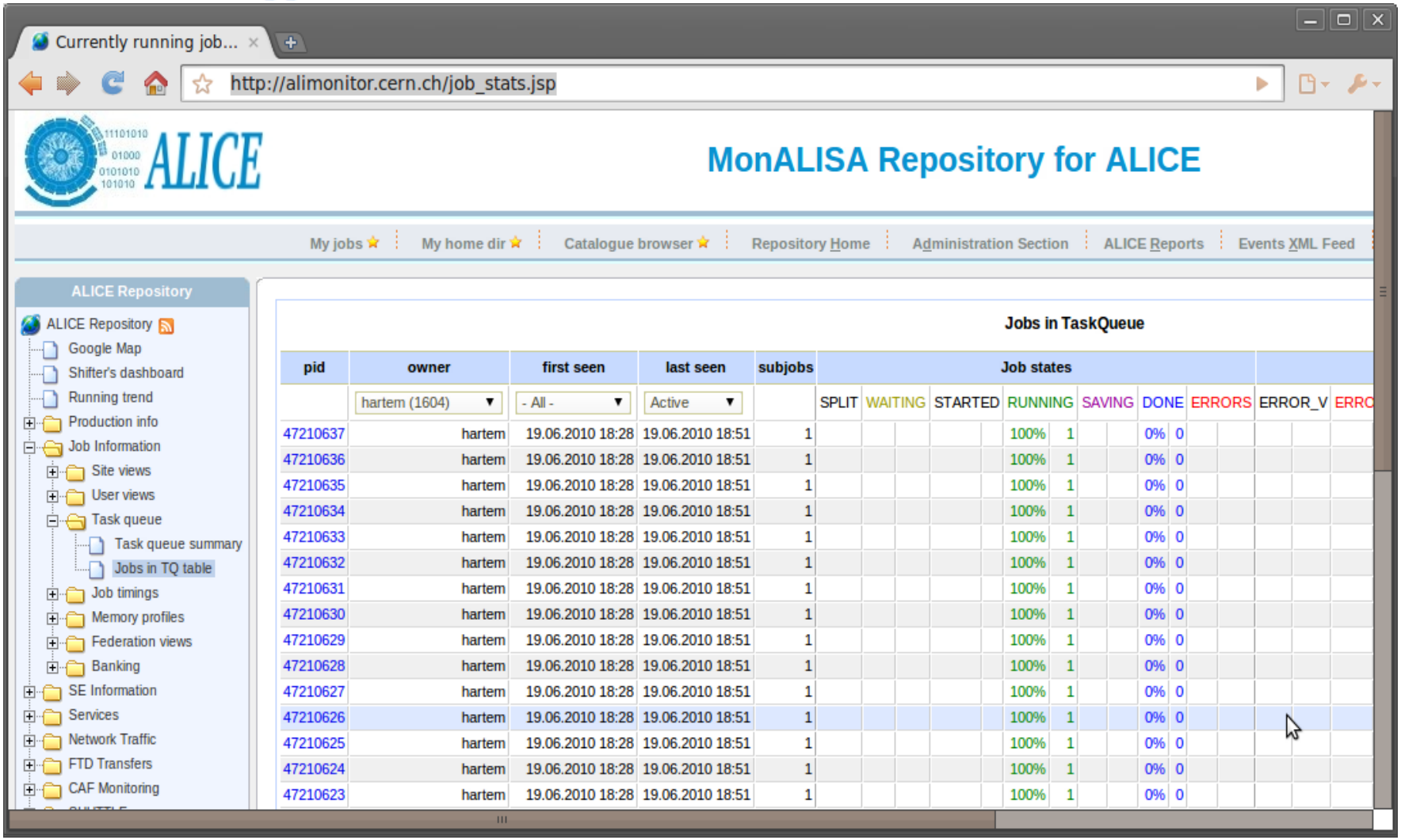
5. Register output files, mark job as DONE

Key Manager service is used to secure the communication of the CoPilot components

A. Harutyunyan, P. Buncic, T. Freeman, K. Keahey, **"Dynamic Virtual AliEn Sites on Nimbus with CernVM"**. 2010 J. Phys.: Conf. Ser. 219 072036 doi: 10.1088/1742-6596/219/7/072036

AliEn on lxcloud@CERN

CernVM
SoFTware Appliance



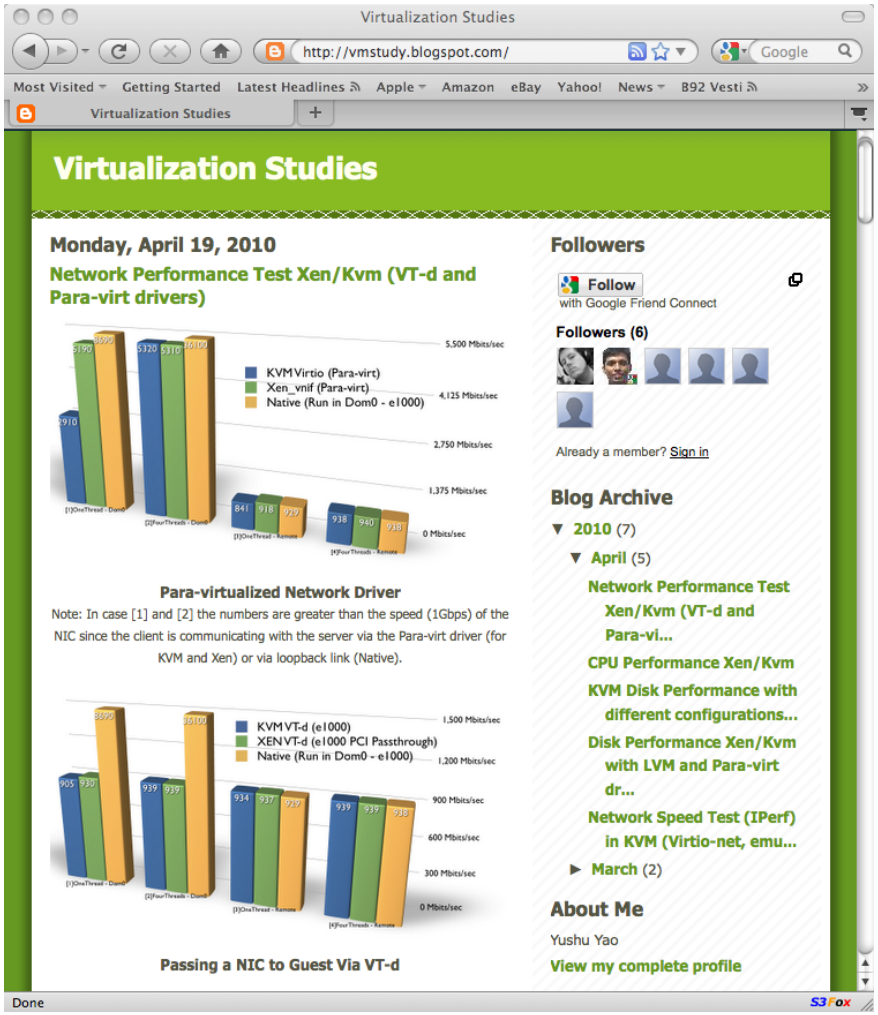
pid	owner	first seen	last seen	subjobs	Job states								
	hartem (1604)	- All -	Active		SPLIT	WAITING	STARTED	RUNNING	SAVING	DONE	ERRORS	ERROR_V	ERRO
47210637	hartem	19.06.2010 18:28	19.06.2010 18:51	1				100% 1		0% 0			
47210636	hartem	19.06.2010 18:28	19.06.2010 18:51	1				100% 1		0% 0			
47210635	hartem	19.06.2010 18:28	19.06.2010 18:51	1				100% 1		0% 0			
47210634	hartem	19.06.2010 18:28	19.06.2010 18:51	1				100% 1		0% 0			
47210633	hartem	19.06.2010 18:28	19.06.2010 18:51	1				100% 1		0% 0			
47210632	hartem	19.06.2010 18:28	19.06.2010 18:51	1				100% 1		0% 0			
47210631	hartem	19.06.2010 18:28	19.06.2010 18:51	1				100% 1		0% 0			
47210630	hartem	19.06.2010 18:28	19.06.2010 18:51	1				100% 1		0% 0			
47210629	hartem	19.06.2010 18:28	19.06.2010 18:51	1				100% 1		0% 0			
47210628	hartem	19.06.2010 18:28	19.06.2010 18:51	1				100% 1		0% 0			
47210627	hartem	19.06.2010 18:28	19.06.2010 18:51	1				100% 1		0% 0			
47210626	hartem	19.06.2010 18:28	19.06.2010 18:51	1				100% 1		0% 0			
47210625	hartem	19.06.2010 18:28	19.06.2010 18:51	1				100% 1		0% 0			
47210624	hartem	19.06.2010 18:28	19.06.2010 18:51	1				100% 1		0% 0			
47210623	hartem	19.06.2010 18:28	19.06.2010 18:51	1				100% 1		0% 0			

CernVM
SoFTware Appliance

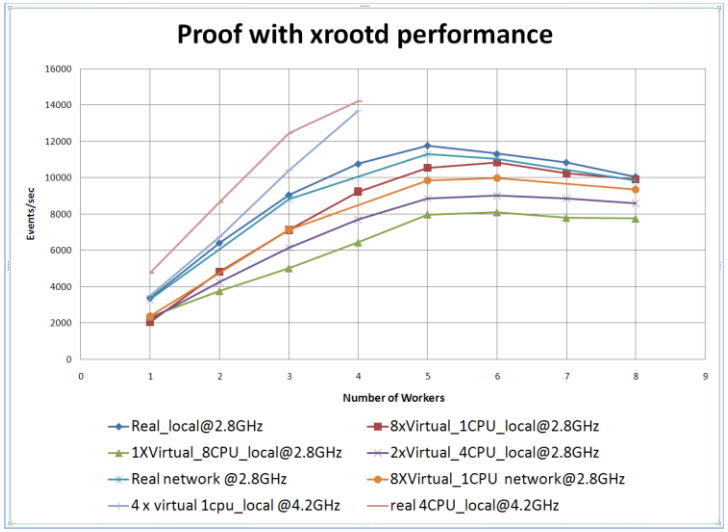
Performance

CernVM
SoFTware Appliance

Performance studies



Yushu's blog: http://vmstudy.blogspot.com

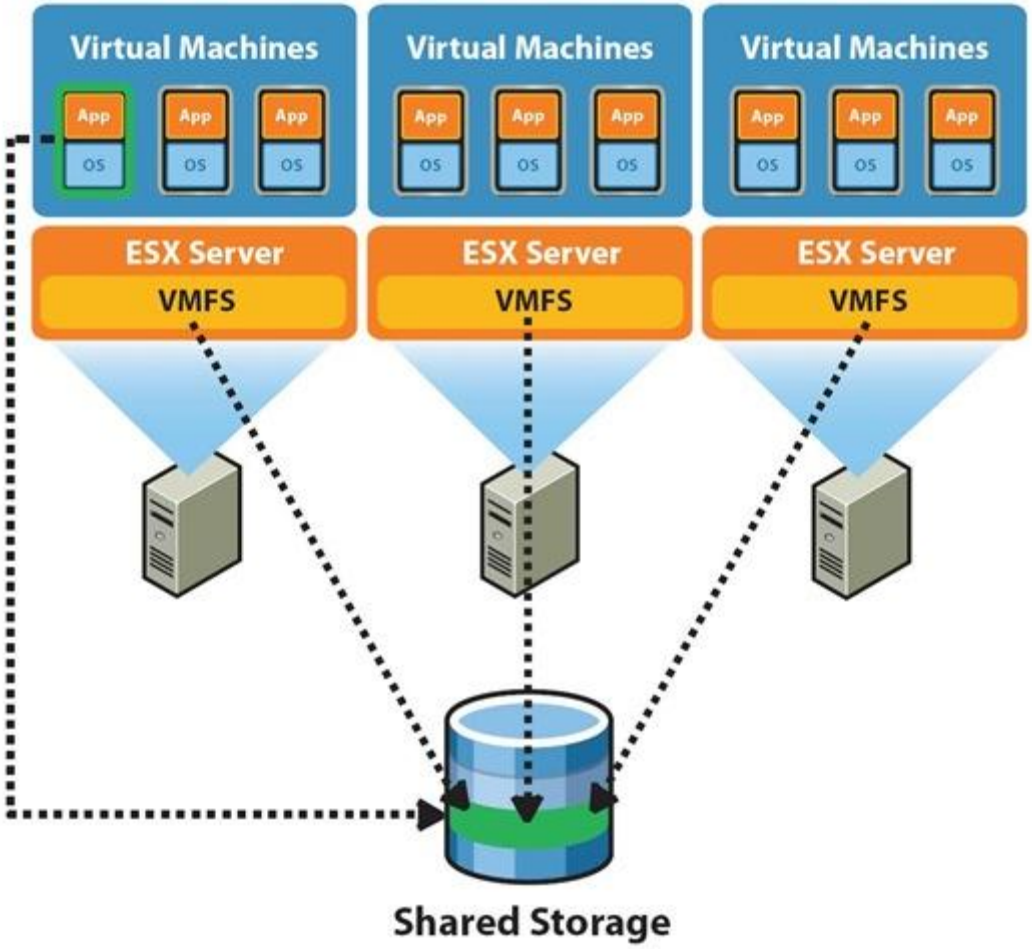


PROOF benchmarks (Waruna Fernando, Sergey Panitkin)

- Results are not yet conclusive and seem to depend on workload type or application
- Local disk I/O is general issue when comparing to native performance
  - Using paravirtualized driver helps
  - Use virtio with KVM
- Use network I/O instead of local disk
- Use PCI passthrough

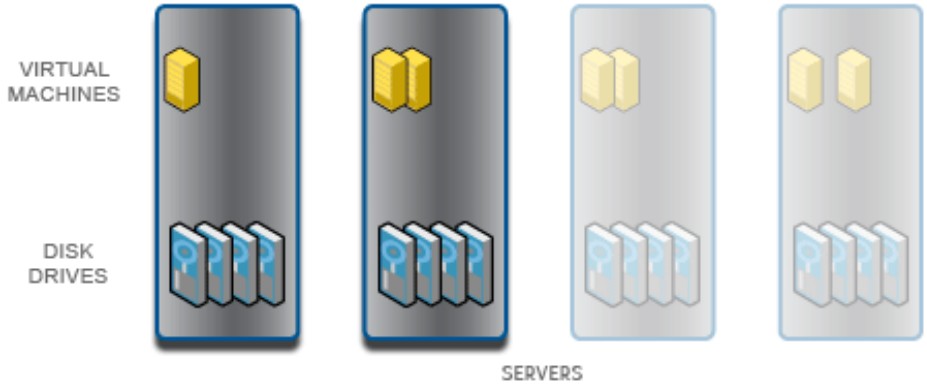
CernVM
SoFTware Appliance

Exmple: VMware Infrastructure



CernVM
SoFTware Appliance

Virtual SAN Appliance



- The LeftHand Virtual SAN Appliance (VSA™) creates a virtual storage node, including disk drives, cache, processors, and controllers, using resources that already exist inside your virtual servers.
- These nodes can be clustered together to transform existing server storage capacity into a virtual storage system, which you can then manage as a single iSCSI SAN

- Can we do the same with tools that we know and have experience with? See talk of Andreas Peters on Tuesday afternoon…

CernVM

SoFTware Appliance

From R&D to service

CernVM
SoFTware Appliance

SFT Review

- A review of the PH-SFT Group took place on 30 September 2009
- Conclusions
  - WP8 and WP9 are very important R&D projects that have succeeded in engaging the participation of the LHC experiments.
  - Both projects have made significant progress and should soon be in a position to offer services that can be adopted by the experiments.
  - We should strongly recommend that the reality of this successful approach should be considered by IT department in their future plans and thereby eliminate the potential risk of ending up with two independent incompatible and even orthogonal approaches.
- Recommendations
  - R1) understand the medium-long term objectives of the project and assuming there is take-up by the experiments prepare a plan for how the activity can be absorbed in the baseline programme of the group
  - R2) move as much as possible of the accepted back-end infrastructure to CERN IT department in order to ensure a 24x7 production service

Transition from R&D to Service

CernVM
SoFTware Appliance

Service Level Status overview

Home | Search | KPIs | Tags | Admin | Documentation | Help

CERNVM WebFS Service 24 Jun 2009 Wed 23:13:35

CERNVM WebFS Service

availability: (more)

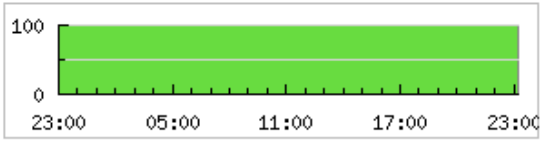
percentage: 100%

status: **available**

this service consists of:

- CERNVM WebFS ALICE Service
- CERNVM WebFS ATLAS Service
- CERNVM WebFS CMS Service
- CERNVM WebFS LHCb Service
- CERNVM WebFS Grid UI Service

availability in the last 24 hours (more):



Additional information

full name: **CERNVM WebFS Service**

group: PH-SFT

email: **cernvm.administrator@cern.ch**

web site: http://cernvm-webfs....

manager: **Carlos Aguado Sanchez**

Availability update

last update: 23:10:28, 24 Jun 2009 (3 minutes ago)

expires after: 15 minutes

rss feed with status changes

Part of (subservice of):

CERNVM Project

Admin

admin tools

CernVM
SoFTware Appliance

Conclusions

- CernVM Software Appliance
  - Used by ATLAS, LHCb and to lesser extent NA61, ALICE and CMS
  - Based on innovative second generation package manager
  - Provides versioning of every build and build products installed on the system
- Initially developed as user interface for laptop/desktop
  - Already deployable on the cloud (EC2, Nimbus)
  - Can be deployed on unmanaged infrastructure like BOINC
  - Ongoing effort to connect existing Pilot Job frameworks (AliEn, Panda, DIRAC, Condor glidein) using CoPilot framework
- Performance issues with virtualization
  - Jury is still out but It is obvious that there will be always some performance penalty
  - There are technical solutions to critical issues (paravirtualization, PCI passthrough)
  - What really counts is time-to-finish (install + develop + deploy + successfully execute + get results)
- CernVM FS (CVMFS)
  - Performs and scales very well, steps take to secure it
  - Being used not only for software distribution but also for calibration data (ATLAS)
  - Also available as standalone package outside CernVM
- Time to think about moving parts of CernVM infrastructure to IT