LHC Cloud Computing with CernVM

Ben Segal / CERN (b.segal@cern.ch)

Predrag Buncic / CERN (predrag.buncic@cern.ch)

and:

David Garcia Quintas, Jakob Blomer, Pere Mato, Carlos Aguado Sanchez / CERN
Artem Harutyunyan / Yerevan Physics Institute
Jarno Rantala / Tampere University of Technology
David Weir / Imperial College, London
Yao Yushu / Lawrence Berkeley Laboratory

ACAT 2010, Jaipur, India

February 22-27, 2010

CernVM Background

- **Over the past couple of years, the industry has redefined the meaning of some familiar computing terms**
  - Shift from glorious ideas of a large public infrastructure and common middleware (“Grids”) towards end-to-end custom solutions and private corporate grids
- **New buzzwords**
  - Amazon Elastic Computing Cloud (EC2)
    - Everything is for rent (CPU, Storage, Network, Accounting)
  - Blue Cloud (IBM) is coming
  - Software as a Service (SaaS)
  - Google App Engine
  - Virtual Software Appliances and JeOS
- **In all these cases, *virtualization* emerged as a key enabling technology, and is supported by computer manufacturers**
  - Multiple cores
  - Hardware virtualization (Intel VT, AMD-V)

CernVM Motivation

- **Software @ LHC Experiment(s)**
  - Millions of lines of code
  - Complicated software installation/update/configuration procedure, different from experiment to experiment
  - Only a tiny portion of it is really used at runtime in most cases
  - Often incompatible or lagging behind OS versions on desktop/laptop
- **Multi core CPUs with hardware support for virtualization**
  - Making laptop/desktop ever more powerful and underutilised
- **Using virtualization and extra cores to get extra comfort**
  - Zero effort to install, maintain and keep up to date the experiment software
  - Reduce the cost of software development by reducing the number of compiler-platform combinations
  - Decouple application lifecycle from evolution of system infrastructure

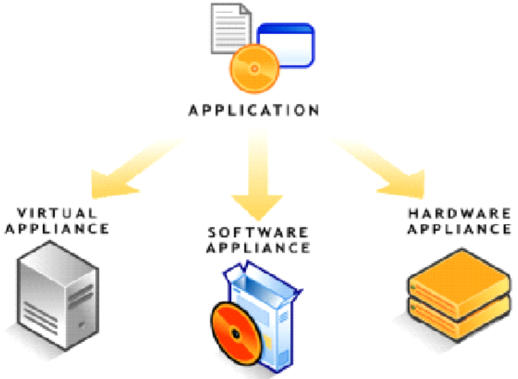
How do we do this?

- **Build a “thin” Virtual Software Appliance for use by the LHC experiments**
- This appliance should
  - provide a complete, portable and easy to configure user environment for developing and running LHC data analysis locally and on the Grid
  - be independent of physical software and hardware platforms (Linux, Windows, MacOS)
- This should minimize the number of platforms (compiler-OS combinations) on which experiment software needs to be supported and tested, thus reducing the overall cost of LHC software maintenance
- All this is to be done
  - in collaboration with the LHC experiments and OpenLab
  - By reusing existing solutions where possible

Key Building Blocks

- **rBuilder from rPath (www.rpath.org)**
  - A tool to build VM images for various virtualization platforms
- **rPath Linux 1**
  - Slim Linux OS binary compatible with Red Hat / SLC4
- **rAA - rPath Linux Appliance Agent**
  - Web user interface
  - XMLRPC API
    - Can be fully customized and extended by means of plugins (#401)
- **CVMFS - CernVM file system**
  - Read-only file system optimized for software distribution
    - Aggressive caching
  - Operational in offline mode
    - For as long as you stay within the cache

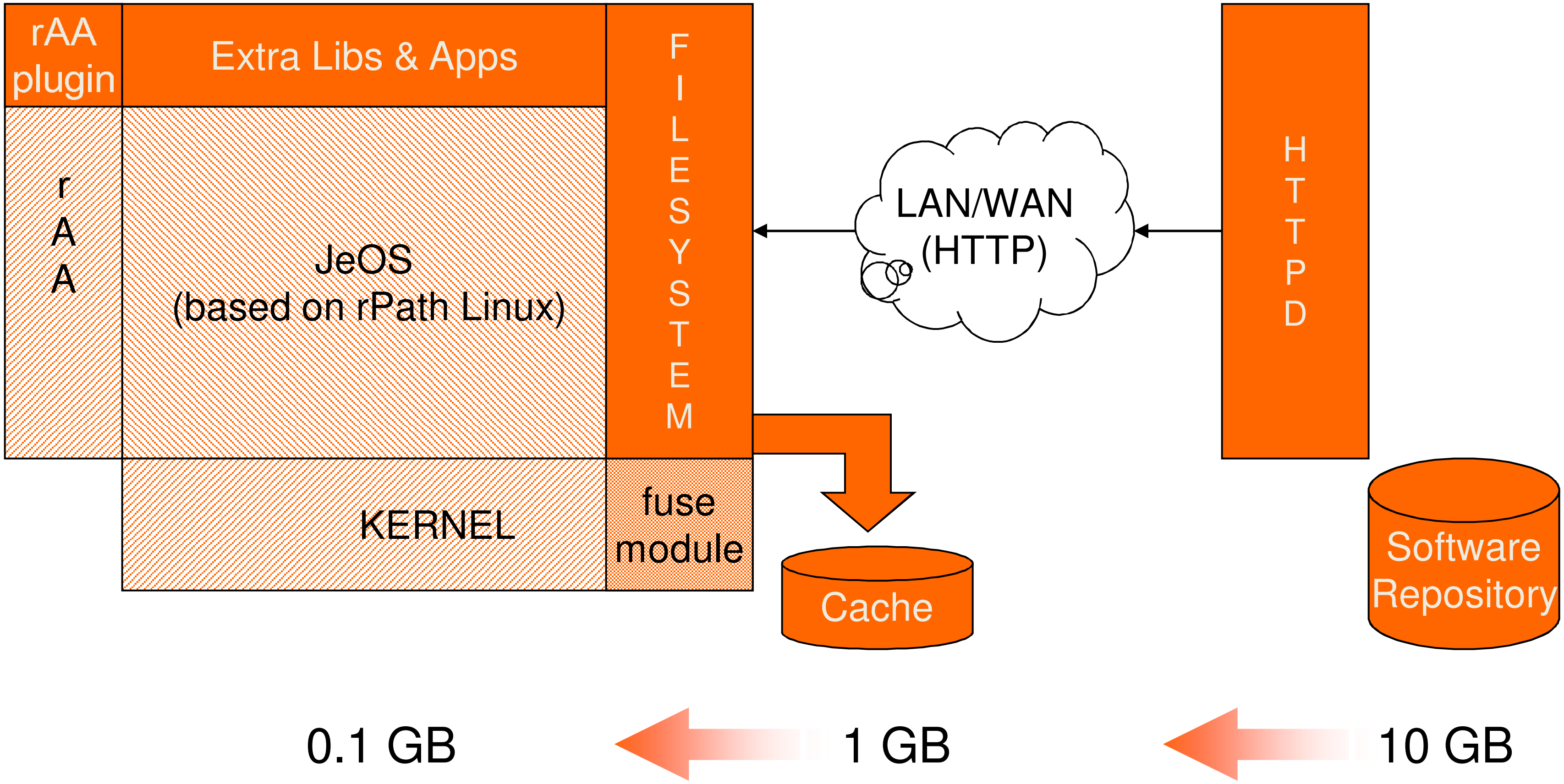
□



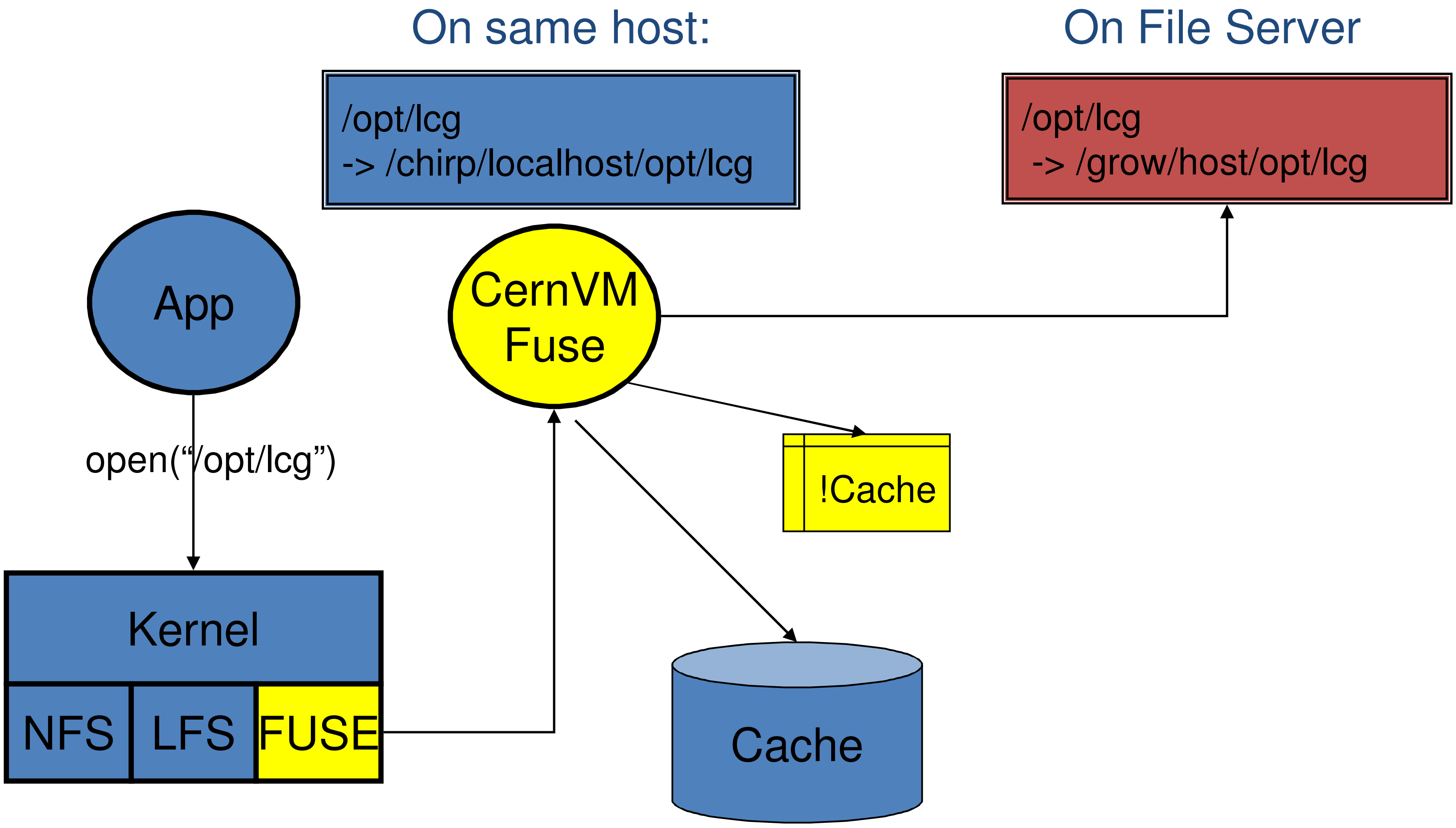
Build types

- Installable CD/DVD
- Stub Image
- Raw File System Image
- Netboot Image
- Compressed Tar File
- Demo CD/DVD (Live CD/DVD)
- Raw Hard Disk Image
- VMware ® Virtual Appliance
- VMware ® ESX Server Virtual Appliance
- Microsoft ® VHD Virtual Appliance
- Xen Enterprise Virtual Appliance
- Virtual Iron Virtual Appliance
- Parallels Virtual Appliance
- Amazon Machine Image
- Update CD/DVD
- Appliance Installable ISO
- Sun Virtual Box Image

“Thin” Software Appliance

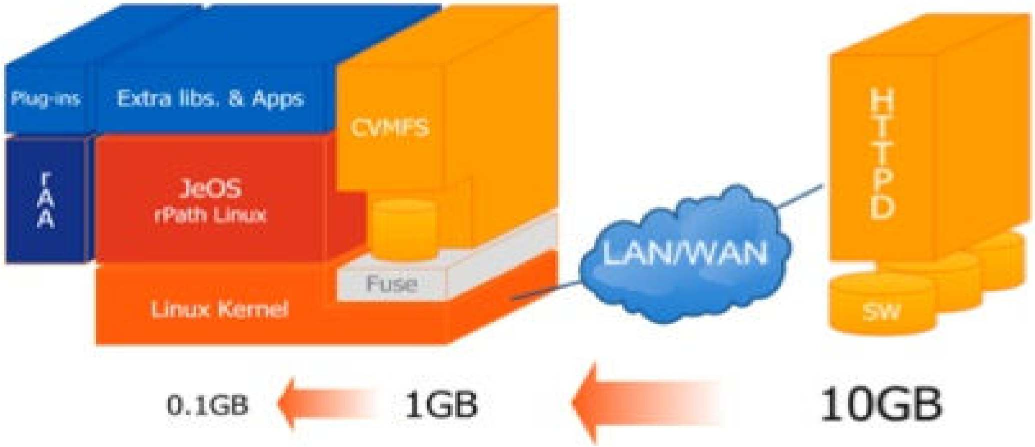


CernVM File System

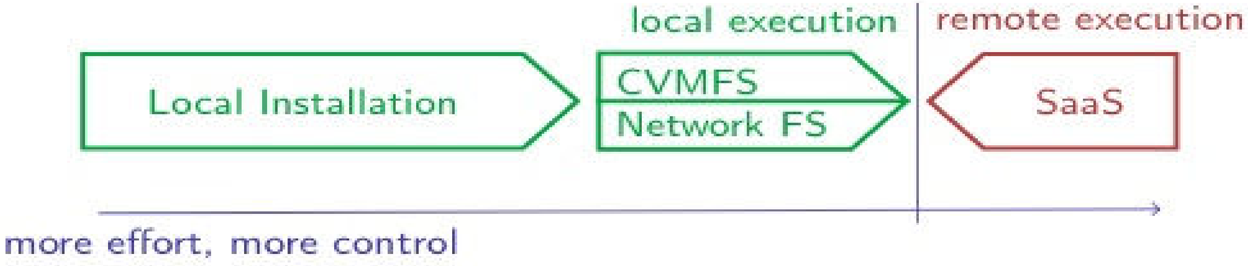


CernVM File System

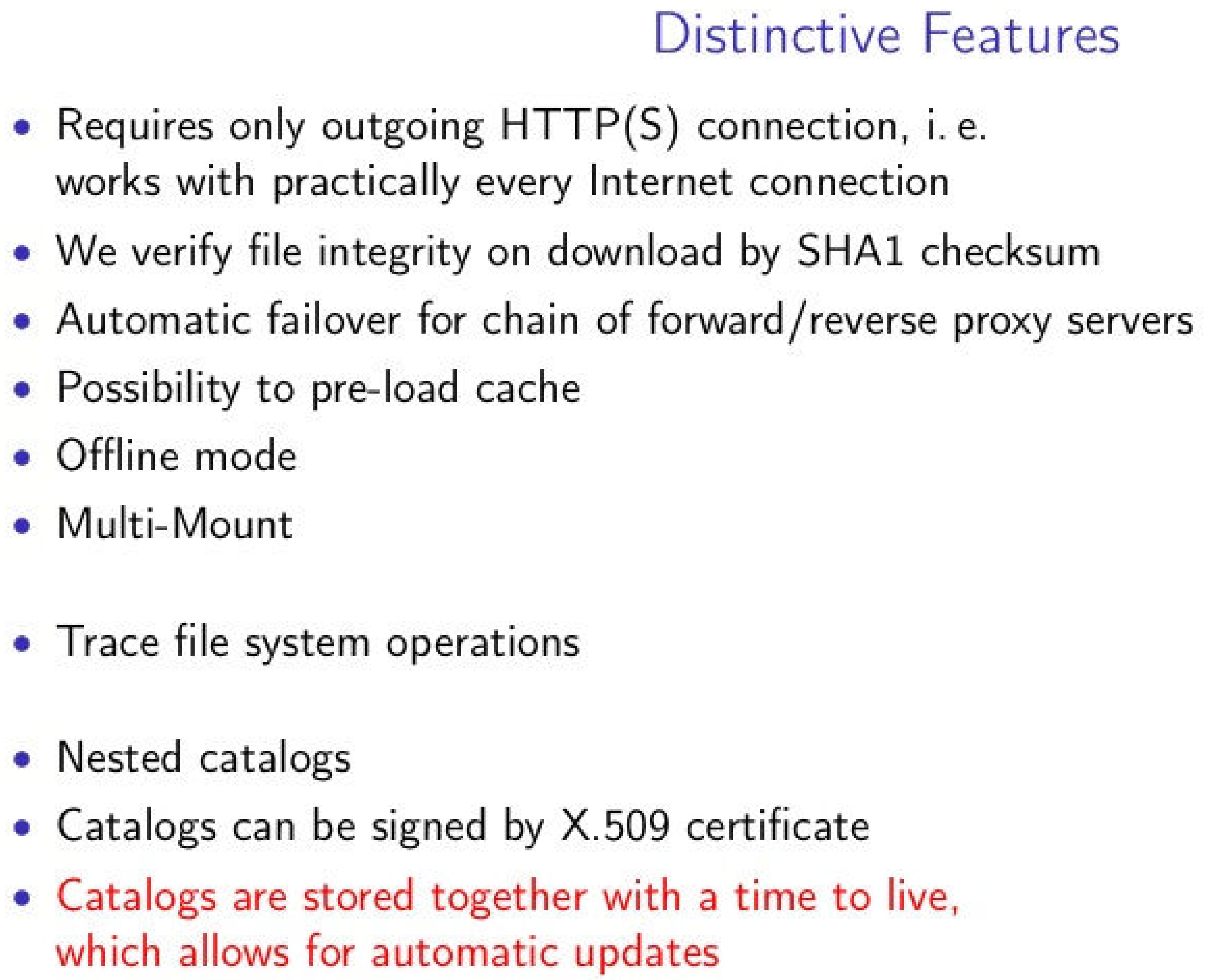
Software Distribution for Virtual Machines



- “Ready to run” binaries, i. e. `/mnt/cvmfs` mirrors destination of `make install`
- Read-only, public files



CernVM File System



Distinctive Features

- Requires only outgoing HTTP(S) connection, i. e. works with practically every Internet connection
- We verify file integrity on download by SHA1 checksum
- Automatic failover for chain of forward/reverse proxy servers
- Possibility to pre-load cache
- Offline mode
- Multi-Mount
- Trace file system operations
- Nested catalogs
- Catalogs can be signed by X.509 certificate
- Catalogs are stored together with a time to live, which allows for automatic updates

Building LHC Computing Clouds

- For example, the LHC Cloud worker nodes may be:

Tier 2, private cluster, or Amazon EC2 nodes, etc.

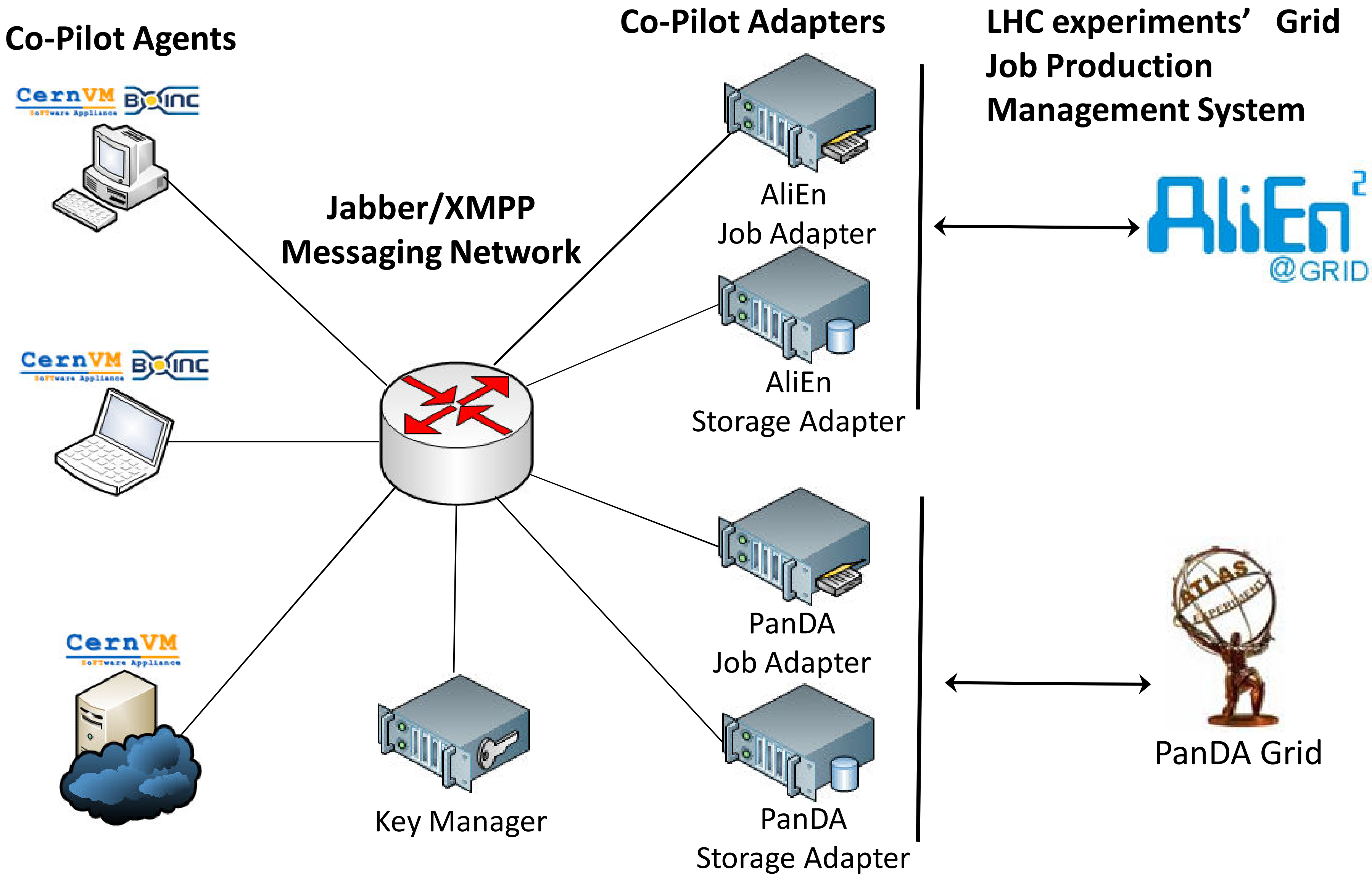
By running CernVM on these nodes:

- Solves porting problem to all client platforms (Windows, Mac, Linux):
- Solves image size and update problems
- Solves job production interface problems

- All done without changing physicists' code or procedures

- How is it done ? … using "CoPilot" system …

CernVM Co-Pilot architecture

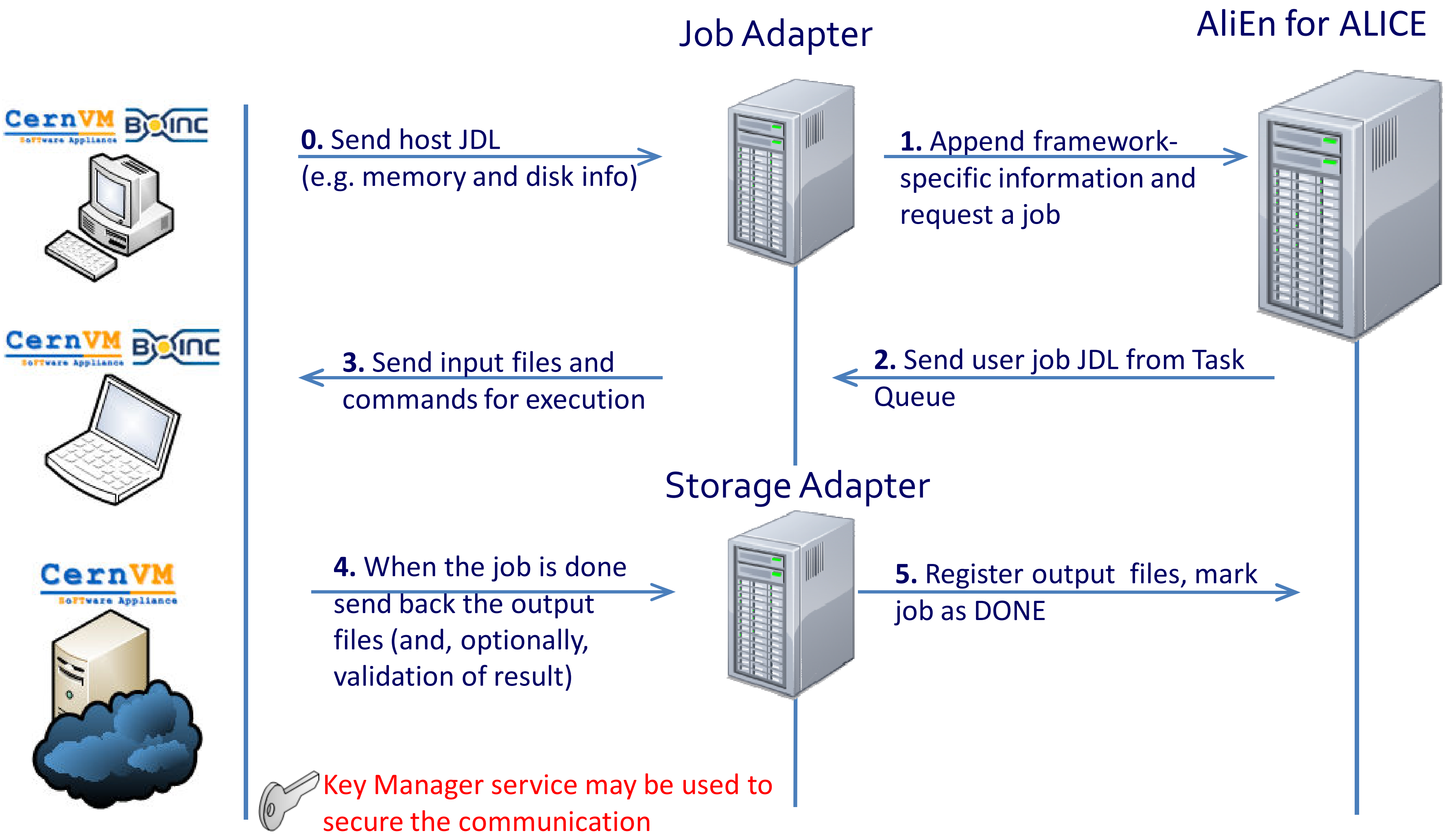


The use of Jabber/XMPP allows to scale the system in case of high load by just adding new Adapter instances

The work of CernVM Co-Pilot Adapters

- Job Adapter
  - Receives job execution requests from an Agent
  - Contacts Grid services (e.g. AliEn/PanDA/Dirac) and gets a job for execution
  - Gets the necessary input files from the Grid file catalogue
  - Makes input files available for the Agent via Chirp
  - Instructs Agent to download the input
  - Instructs Agent to execute job command (e.g. ‘root myMacro.C’)
- Storage Adapter
  - Receives job completion request
  - Provides Agent with the Chirp directory to upload job output
  - Puts the job output to the Grid file catalogue
  - Contacts Grid services and sets the final status of the job
- **NOTE: Grid credentials are handled by Adapters, not sent to Agents**

CernVM Co-Pilot job execution (ALICE)


CernVM Software Appliance
BOINC
CernVM Software Appliance
BOINC
CernVM Software Appliance

Job Adapter

AliEn for ALICE

0. Send host JDL (e.g. memory and disk info)

1. Append framework-specific information and request a job

2. Send user job JDL from Task Queue

3. Send input files and commands for execution

Storage Adapter

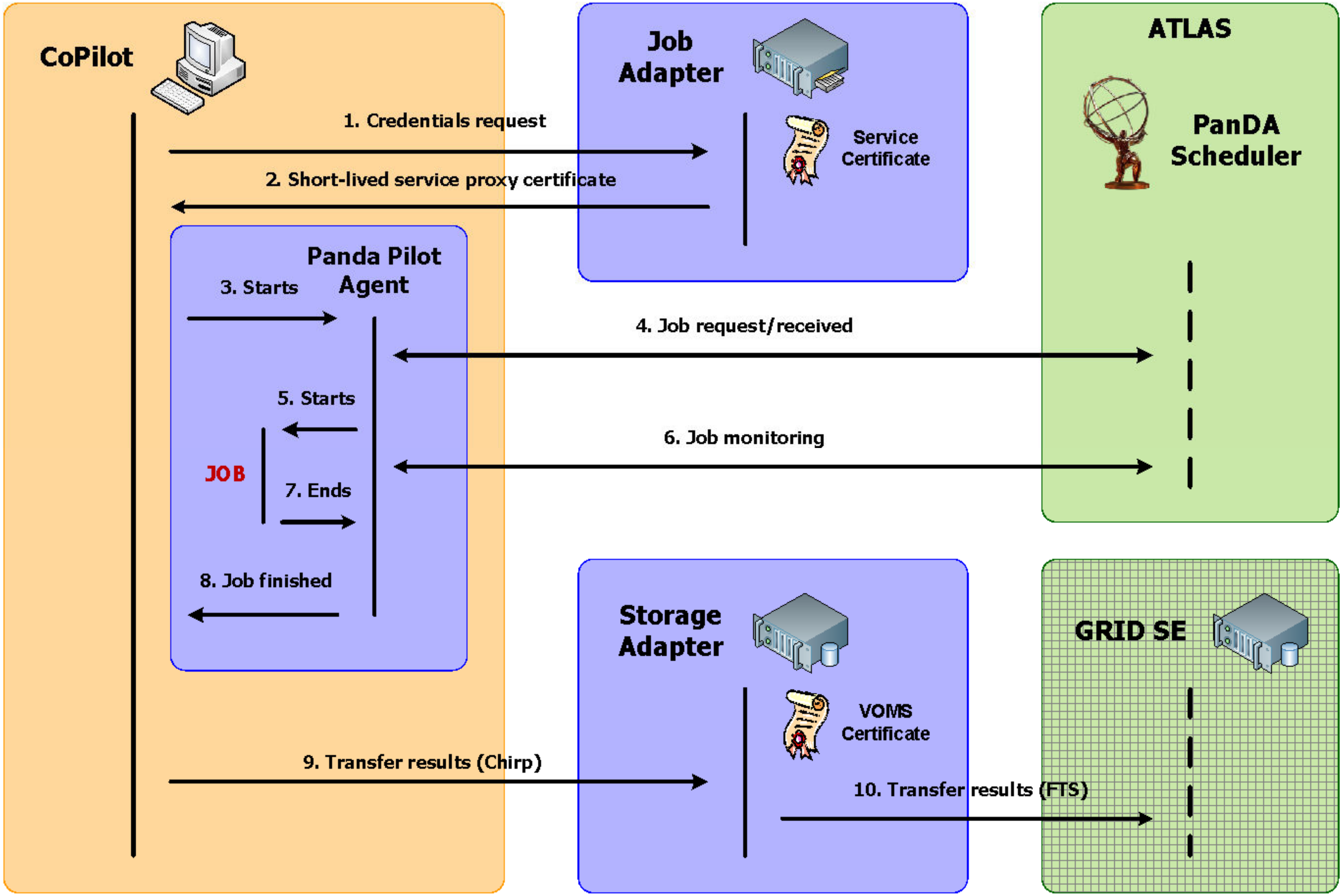
4. When the job is done send back the output files (and, optionally, validation of result)

5. Register output files, mark job as DONE

Key Manager service may be used to secure the communication

For the detailed description of the communication protocol please see:
https://cernvm.cern.ch/project/trac/cernvm/wiki/CoPilotProtocol

CernVM Co-Pilot job execution (ATLAS)



For the detailed description of the communication protocol please see:
https://cernvm.cern.ch/project/trac/cernvm/wiki/CoPilotProtocol

Building a Volunteer Cloud

- **In this case, the Cloud worker nodes are:**

  Volunteer PC's running CernVM…

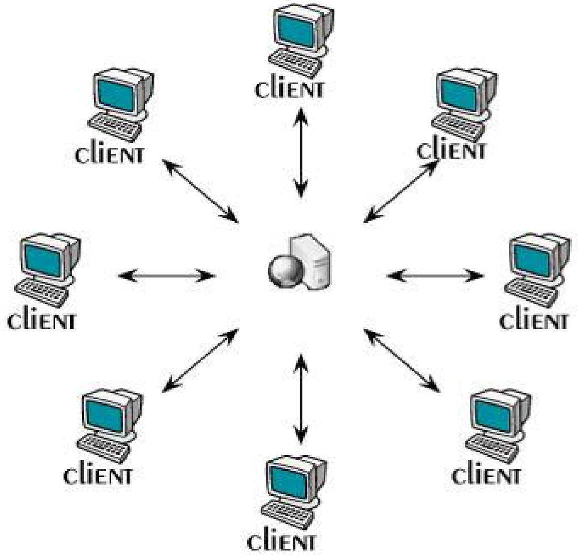
- **Solves porting problem to all client platforms (Windows, Mac, Linux):**
- **Solves image size problem**
- **Solves job production interface problem**

- **All done without changing existing BOINC infrastructure (client or server side)**
- **All done without changing physicists' code or procedures**

- **How is it done ? … with CoPilot and BOINC …**

What is BOINC?

- "**B**erkeley **O**pen **I**nfrastructure for **N**etwork **C**omputing"
- **Software platform for distributed computing using volunteered computer resources**
- **http://boinc.berkeley.edu**
- **Uses a volunteer PC's unused CPU cycles to analyse scientific data**
- **Client-server architecture**
- **Free and Open-source**
- **Also handles DESKTOP GRIDS**



Some volunteer computing projects

SCIENCE
SETI@home (BOINC)
evolution@home
eOn
climateprediction.net (BOINC)
Muon1
LHC@home (BOINC)
Einstein@Home(BOINC)
BBC Climate Change
Experiment (BOINC)
Leiden Classical (BOINC)
QMC@home (BOINC)
NanoHive@Home (BOINC)
μFluids@Home (BOINC)
Spinhenge@home (BOINC)
Cosmology@Home (BOINC)
PS3GRID (BOINC)
Mars Clickworkers

LIFE SCIENCES
Parabon Computation
Folding@home
FightAIDS@home
Übero
Drug Design Optimization Lab (D2OL)
The Virtual Laboratory Project
Community TSC
Predictor@home (BOINC)
XGrid@Stanford
Human Proteome Folding (WCG)
CHRONOS (BOINC)
Rosetta@home (BOINC)
RALPH@home (BOINC)
SIMAP (BOINC)
malariacontrol.net (BOINC)
Help Defeat Cancer (WCG)
TANPAKU (BOINC)
Genome Comparison (WCG)
Docking@Home (BOINC)
proteins@home (BOINC)
Help Cure Muscular Dystrophy (WCG)

MATHEMATICS & CRYPTOGRAPHY
Great Internet Mersenne Prime Search
Proth Prime Search
ECMNET
Minimal Equal Sums of Like Powers
MM61 Project
3x + 1 Problem
Distributed Search for Fermat
Number Divisors
PCP@Home
Generalized Fermat Prime Search
PSearch
Seventeen or Bust
Factorizations of Cyclotomic Numbers
Goldbach Conjecture Verification
The Riesel Problem
The 3*2^n-1 Search
NFSNET
Search for Multifactorial Primes
15k Prime Search
ElevenSmooth
Riesel Sieve
The Prime Sierpinski Project
P.I.E.S. - Prime Internet Eisenstein Search
Factors of k*2^n±1
XYYXF
12121 Search
2721 Search
Operation Billion Digits
SIGPS
Primesearch

INTERNET PERFORMANCE
Gómez Performance ($)
Network Peer
NETI@home
dCrawl
DIMES
Red Library DLV
Majestic-12
Boitho
PeerFactor
DepSpid
Pingdom GIGRIB
Project Neuron(BOINC)

ECONOMICS
MoneyBee
Gstock

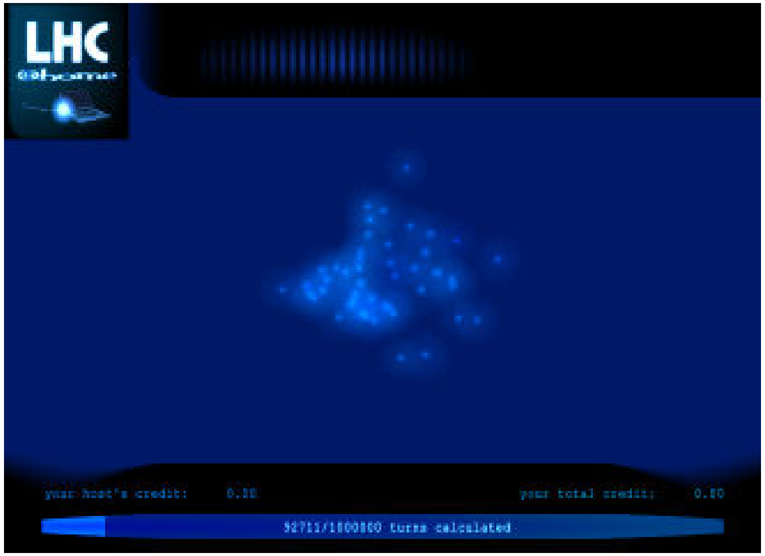
GAMES
ChessBrain
Chess960@home (BOINC)

ART
Electric sheep
Internet Movie Project
RenderFarm@home (BOINC)

Source: http://distributedcomputing.info/

The BOINC community

- Competition between individuals and teams for “credit”.
- Websites and regular updates on status of project by scientists.
- Forums for users to discuss the science behind the project.
- E.g. for LHC@home, the volunteers show great interest in CERN and the LHC.
- Supply each other with scientific information and even help debug the project.



LHC@home screensaver

LHC@home

- Calculates stability of proton orbits in CERN’s new LHC accelerator
- System is nonlinear and unstable so numerically very sensitive. Hard to get identical results on all platforms
- About 40 000 users, 70 000 PC’s… over 1500 CPU years of processing
- Objectives: extra CPU power and raising public awareness of CERN and the LHC - both successfully achieved.
- Started as an outreach project for CERN 50th Anniversary 2004; used for Year of Physics (Einstein Year) 2005

Join LHC@home

World Year of PHYSICS 2005

CERN and Einstein? Find out about the place where energy becomes matter

What is the LHC? Discover the most powerful accelerator ever built

What is LHC@home? Help scientists at CERN to simulate particles travelling in the LHC

Why do particle physicists need accelerators? Recreating the conditions of the early universe

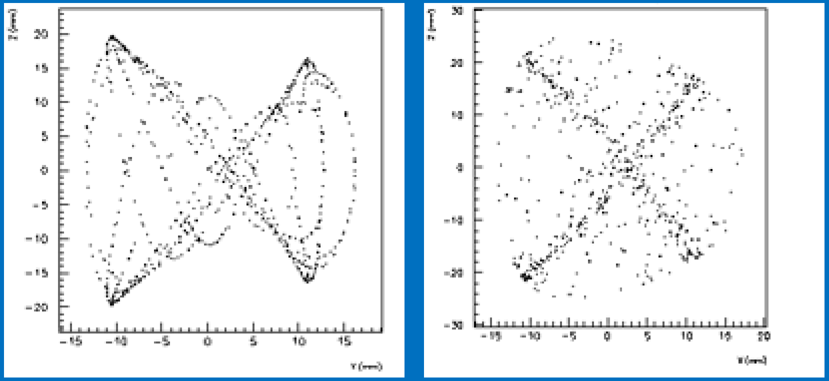
LHC@home

[illegible]

Help scientists to build the LHC!

SixTrack program

SixTrack is a Fortran program by F. Schmidt, based on DESY program
SixTrack simulates 60 particles for 100k-1M LHC orbits
Can include measured magnet parameters, beam-beam interactions
LHC@home revealed reproducibility issues, solved by E. McIntosh



Phase space images of a particle for a stable orbit (left) and unstable chaotic orbit (right).

BOINC & LHC physics code

Problems with “normal” BOINC used for LHC physics:

1) A project’s application(s) must be ported to every volunteer platform of interest: most clients run Windows, but CERN runs Scientific Linux and porting to Windows is impractical.

2) The project’s work must be fed into the BOINC server for distribution, and results must be recovered. “Job submission scripts” must be developed for this, but CERN physics experiments won’t change their current setups.

3) Job management is very primitive in BOINC, whereas physicists want to know where their jobs are and be able to manage them.

BOINC & Virtualization

CernVM and Co-Pilot allow us to solve these three problems (porting, job submission, job management) but to run CernVM guest VM’s within a BOINC host we need a cross-platform solution for:

- control of (multiple) VM’s on a host, including:
  Start|Stop|pause|resume|reset|poweroff|savestate
- command execution on guest VM’s
- file transfers from guests to host (and reverse)

BOINC & Virtualization

Details of the "VM controller" package:

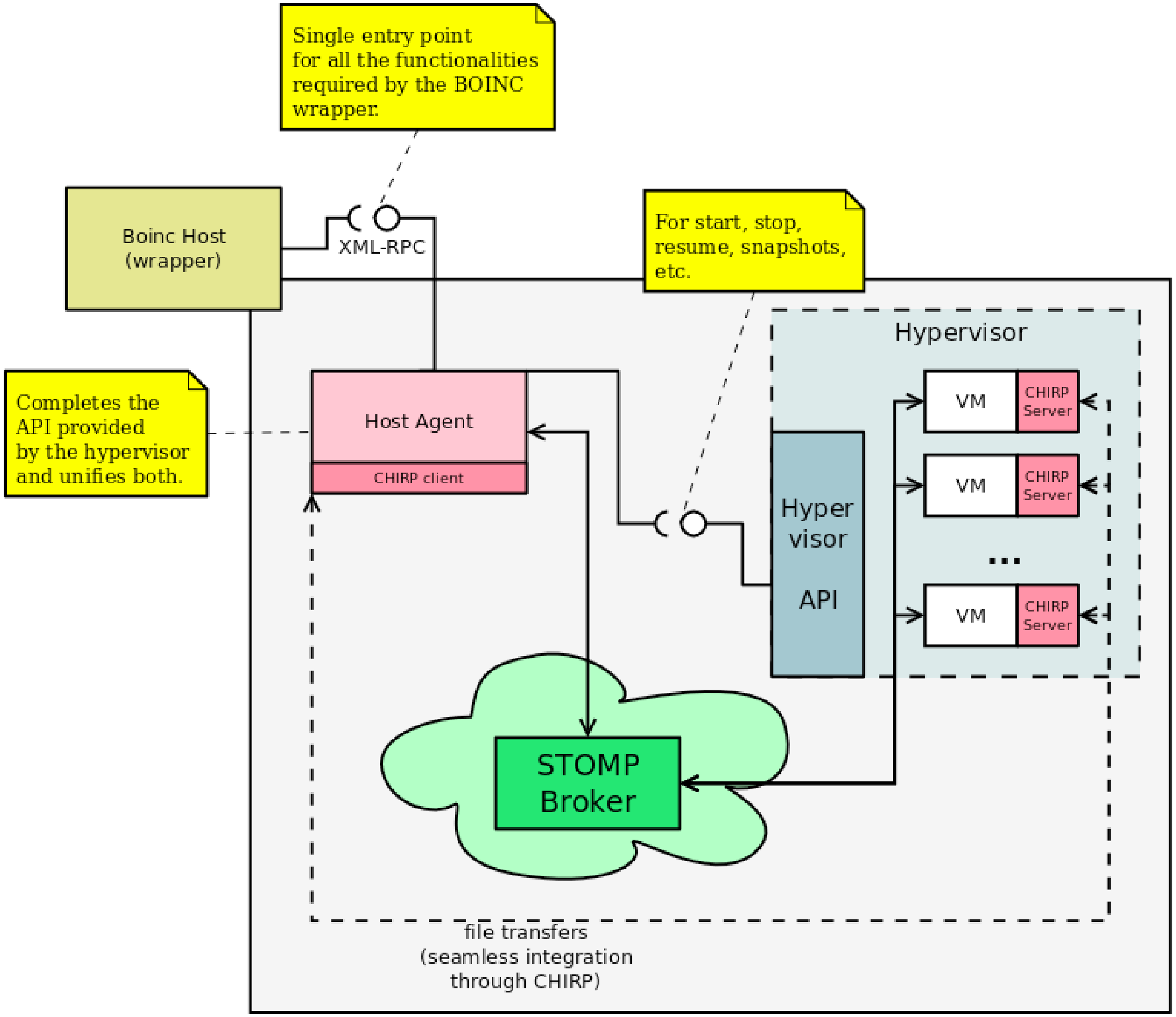
(developed by David Garcia Quintas / CERN)

- Cross-platform support - based on Python (Windows, MacOSX, Linux…).
  - Uses Python packages:

    Netifaces, Stomper, Twisted, Zope, simplejson, Chirp…
- Does asynchronous message passing between host and guest entities via a broker (e.g. ActiveMQ). Messages are XML/RPC based.
- Supports:
  - control of multiple VM's on a host, including:

    Start|Stop|pause|resume|reset|poweroff|savestate
  - command execution on guest VM's
  - file transfers from guests to host (and reverse) using Chirp

Host to VM Guest communication



Single entry point for all the functionalities required by the BOINC wrapper.

For start, stop, resume, snapshots, etc.

Completes the API provided by the hypervisor and unifies both.

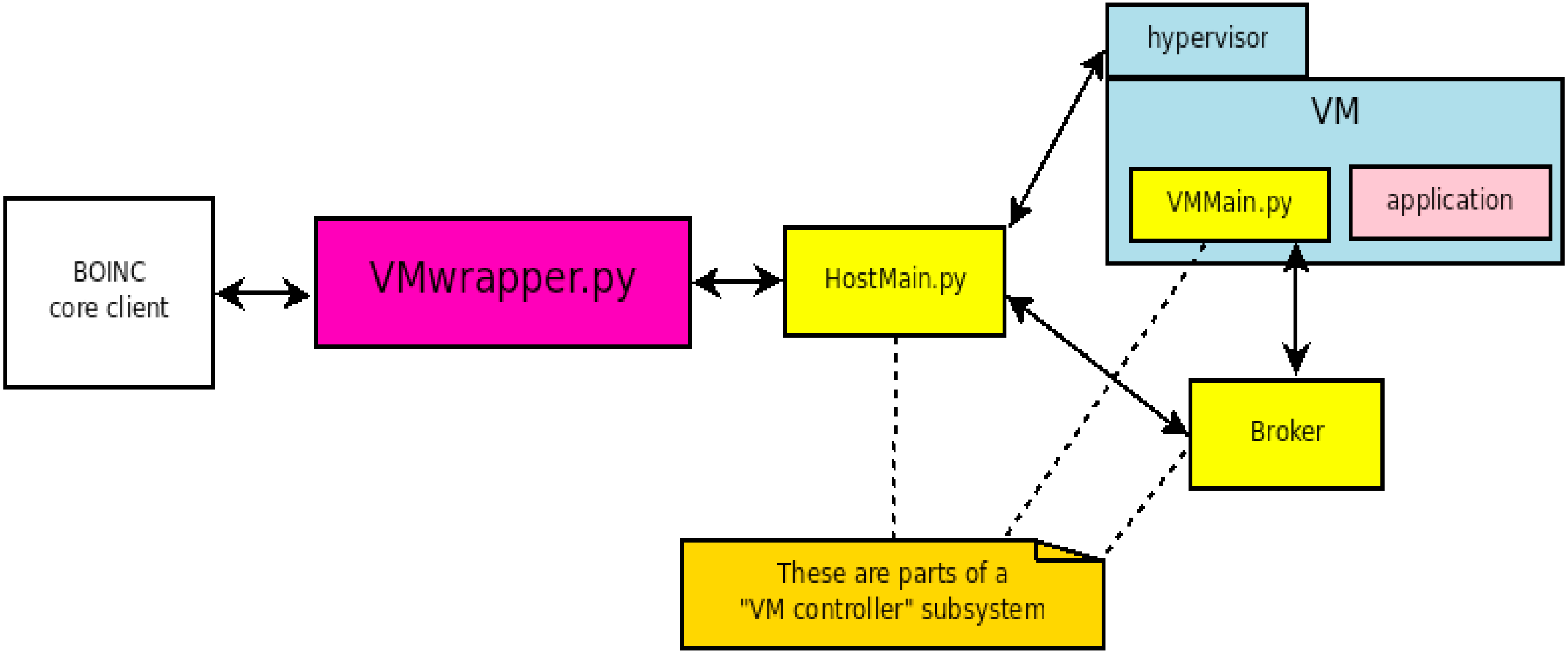
file transfers (seamless integration through CHIRP)

BOINC & Virtualization

Details of the new BOINC “Vmwrapper”:
(developed by Jarno Rantala / CERN openlab student)

- Written in Python, therefore multi-platform
- Uses “VM controller” infrastructure described above
- Back-compatible with original BOINC Wrapper
  - Supports standard BOINC job.xml files
  - For VM case, supports extra tags in the job.xml file
- Able to measure the VM guest resources and issue credit requests
  - ..including “partial credits” to allow very long-running processes/jobs

BOINC VMwrapper architecture



BOINC Virtual Cloud

Summary of the method:

- New BOINC wrapper (VMWrapper) used to start a guest Virtual machine in BOINC client PC, and execute a CernVM image.
- The CernVM image has all LHC software and CoPilot code.
- Host-to-VM communication/control provided for any BOINC PC.
- The new Vmwrapper gives BOINC client and server all the functions they need - they are unaware of VM's…
- As before, the CoPilot allows LHC job production to proceed without changes.

BOINC Virtual Cloud

Summary of results at this point:

- **Solved client application porting problem**
- **Provided host-to-VM guest communication/control**
- **The new Vmwrapper gives BOINC client all functions it needs**

- **Solutions to image size problems and physics job production interface offered by the CernVM project together with the CoPilot adapter system.**

Building a Volunteer Cloud

- **Final Summary:**

- Solved porting problem to all client platforms:
- Solved image size problem
- Solved job production interface problem

- All done without changing existing BOINC infrastructure (client or server side)
- All done without changing physicists' code or procedures

- **We have built a "Volunteer Cloud" …**