High Availability using virtualization

Federico Calzolari

Scuola Normale Superiore - INFN Pisa

SCUOLA NORMALE
SUPERIORE

Aims and Requirements

Aims

- zero cost High availability service

Requirements

- full exploitation of virtual environment features

Outline

- **High Availability** definition and measure
- **Virtualization** definition and features
- **Scenario**
  - Grid data center
- **Infrastructure**
  - Preboot eXecution Environment PXE
  - Storage: from NAS to SAN
- **Solutions**
  - High availability using virtualization
  - Redundancy in virtual environments
  - Physical to Virtual migration
- **Experimental data**
  - Operation in a real crash example
- **Spin-off**
  - Host on-demand and Cloud computing

Abstract

High availability has always been one of the main problems for a data center. Till now high availability was achieved by host per host redundancy, a highly expensive method in terms of hardware and human costs. A new approach to the problem can be offered by virtualization.

Using virtualization, it is possible to achieve a redundancy system for all the services running on a data center. This new approach to high availability allows the running virtual machines to be distributed over a small number of servers, by exploiting the features of the virtualization layer: start, stop and move virtual machines between physical hosts.

The 3RC system is based on a finite state machine, providing the possibility to restart each virtual machine over any physical host, or reinstall it from scratch. A complete infrastructure has been developed to install operating system and middleware in a few minutes. To virtualize the main servers of a data center, a new procedure has been developed to migrate physical to virtual hosts.

The whole Grid data center SNS-PISA is running at the moment in virtual environment under the high availability system.

High availability definition

- **High Availability**
  - system design protocol that ensures a certain degree of operational continuity during a given period.
- **Fault Tolerance**
  - property that enables a system to continue operating properly in the event of the failure of some of its components.
- **Data Reliability - Redundancy**
  - property of some disk arrays which provides fault tolerance [no data lost in case of disk failure].

supplied by:

- **Load Balancing**
  - technique to spread work between many computers, processes, disks or other resources.
- **Failover**
  - capability to automatically switch over to a redundant or standby computer server, system, or network.

High availability features and measure

High availability features

- User does not have to care about how/where to access services/data
- Reduce downtime to a minimum

High availability measure

- Availability is described in "number of nines"; the number N of nines describes a system available a fraction A of the time

  $N = - \log_{10} (1 - A)$

- Availability is usually expressed as a percentage of uptime in one year:
  - 99.9% ▸ downtime 8.76 hours / year [my target]
  - 99.99% ▸ downtime 52.6 minutes / year
  - 99.999% ▸ downtime 5.26 minutes / year [telecommunications]

Virtualization definition

Virtualization

- Abstraction of computer resources
- Abstraction layer that allows each physical server to run one or more virtual servers, decoupling operating system and applications from the underlying physical server.

Virtualization benefits

- 1 service/host:

  split a multi processor server into more independent virtual hosts

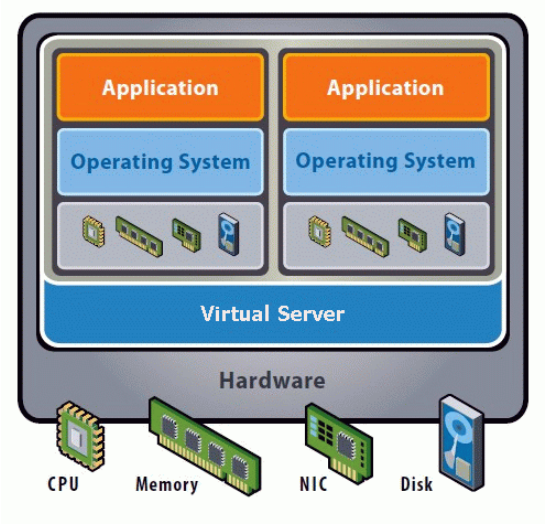
supplied by:

- VMware: NOT open source, but free version [my choice]
- Xen: open source, free, virtualization and para-virtualization, Kernel patch
- KVM: future?

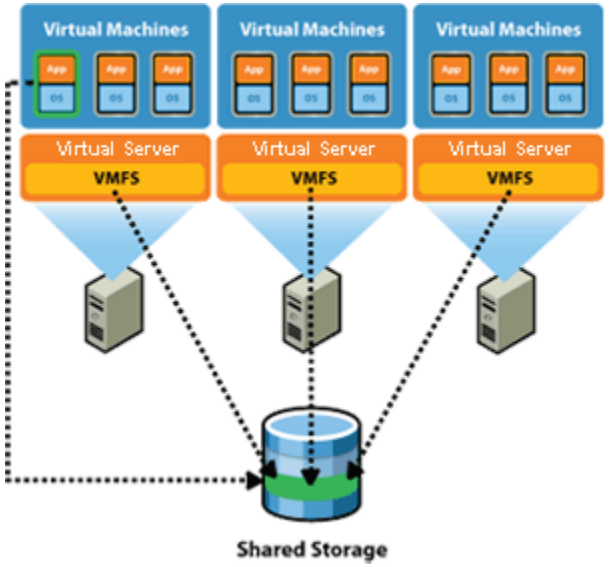
Virtualization features

What can Virtualization do?

- A single server can host multiple Virtual machines, each one providing a specific service.
- More servers can share a common external filesystem to ease virtual disk (VMFS) moving.



Virtualized architecture



Shared Storage

Why Virtualization?

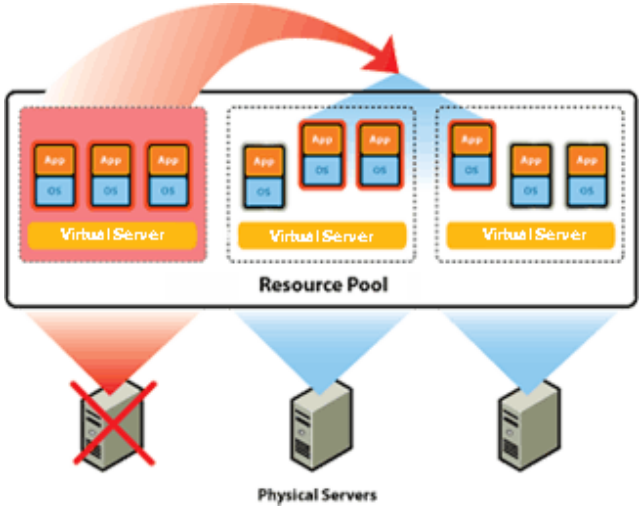
Virtualized High availability

- decouple hardware from software
- suspend/recover virtual machines
- virtual machines migration
- increase server density
- better control and manageability

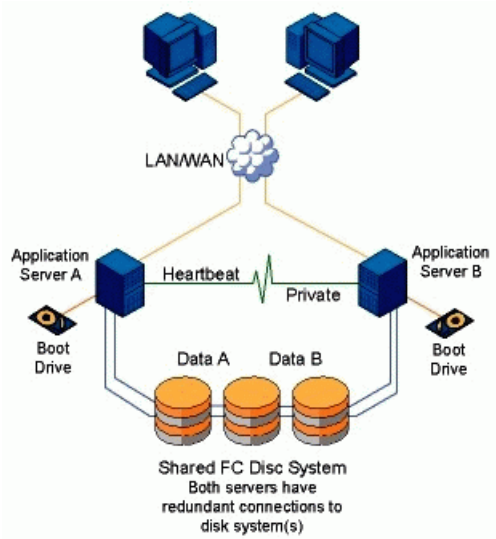
Heartbeat High availability

- host per host redundancy
- double cost for
  - **hardware**
  - **configuration**

Virtualized solution



Heartbeat Classical solution



Scenario

Grid Data Center

- 1 + Computing element: communication between farm and external (gateway)
- 1 + Storage element: disk server with SRM features
- 1 Batch Queuing System master
- 1 Monitoring service
- 1 BDII: Berkeley Database Information Index (Information provider)
- 5 Services: specific Virtual Organization applications
- 1 + User Interface: user access to Grid
- 1 Cache proxy server: Squid
- N Worker nodes: computational nodes

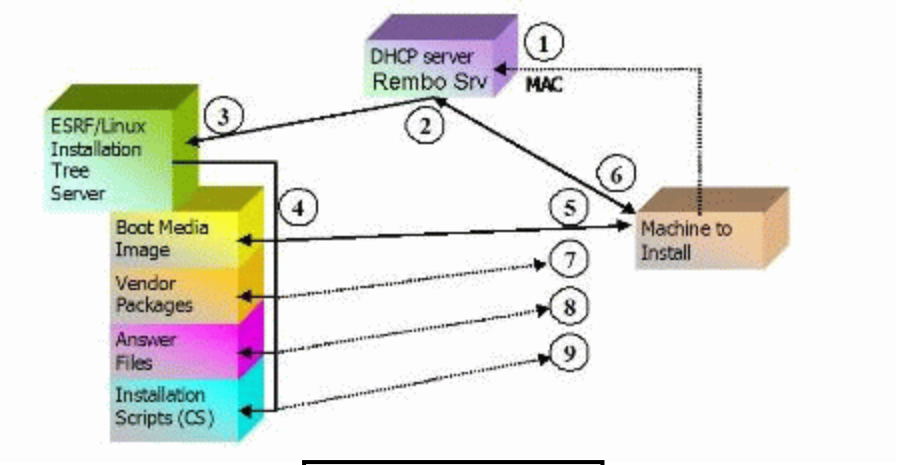
What is necessary to grant service?

- ALL but Worker nodes (~ 20 hosts)

Infrastructure - PXE

How to provide an automatic host installation?

- DHCP
- DNS HINFO (Host Info) = host_type
- PXE - TFTP
- HTTP



PXE architecture

- **INFN-PISA** EGEE Grid node: 2000 CPU, 500 TB disk
- **SNS-PISA** EGEE Grid node: small, testbed
- **CNR-ISTI** EGEE Grid node: Pre Production Service

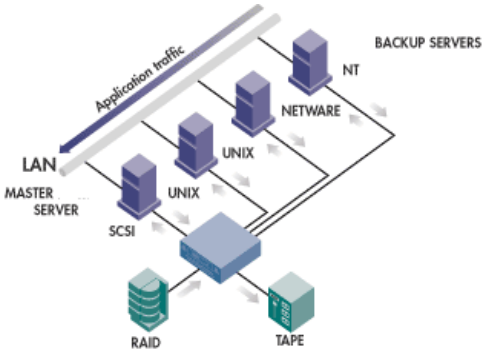
to manage up to 2000 virtual machines/disks simultaneously:

- 16 Gb/s aggregate bandwidth

Infrastructure - Storage

Storage solutions

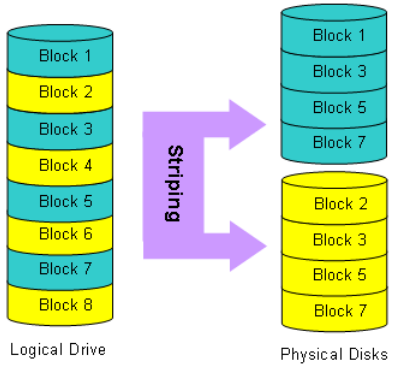
- DAS Direct Attached Storage
- NAS Network Attached Storage
- SAN Storage Area Network



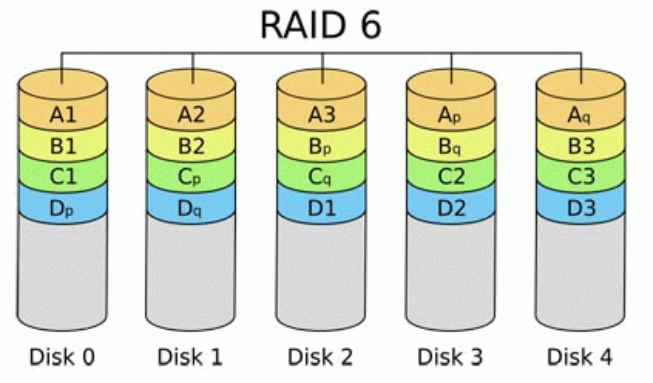
Storage architecture

Requirement: reliable storage

- RAID Redundant Array of Independent Disks
- DRBD Distributed Replicated Block Device - Mirror over Network



Data Striping



RAID 6

A new approach to High availability

RELAXED High availability

- A "relaxed" High availability service is a system able to restore any previously running application in less than 10 minutes from the crash time.
- A relaxed system may ensure the application redundancy required in the greater part of cases.

How can a Relaxed High availability service be achieved?

- Virtual machines are highly portable between computers.
- A virtual machine can pause operation, be moved or copied to another physical computer, and there resume execution exactly where it left off.

Hysteresis

Tendency of a system to respond differently to the same stimulus depending on the initial state of the system.

definition by Claudia Guida, Molecular Biologist @IEO Milan

3RC Project: 3 Re Cycle

Finite state machine with hysteresis

- Reboot
- Restart
- Reinstall

Each physical host can backup all the others

3RC logo

Requirements

- redundant controller [shared]
- reliable storage
  - SAN or NAS via FC or NFS
  - RAID over network: DRBD

Goals

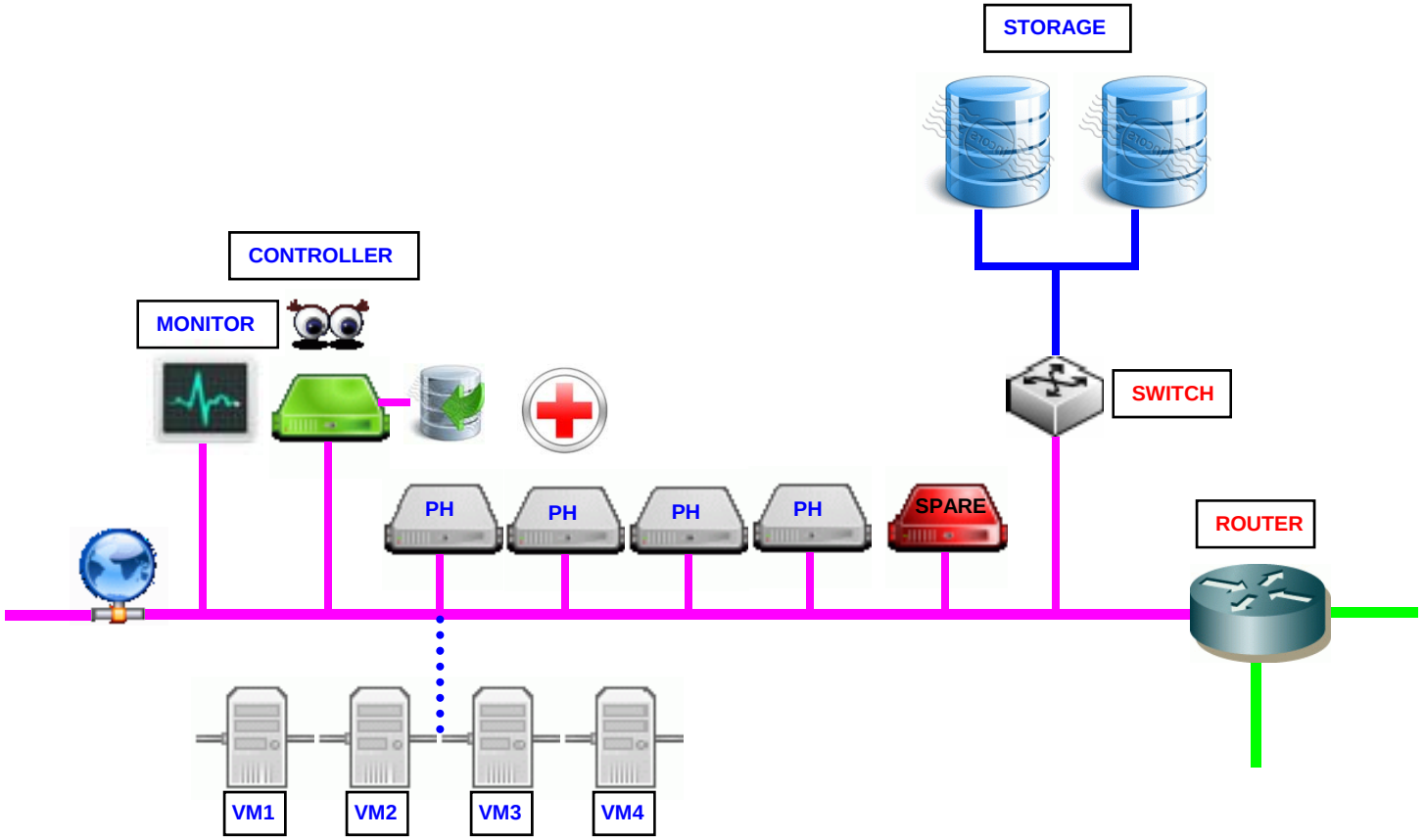
- relaxed High Availability: recovery time < 10 min
- backup solution ONLY @disaster_time

Research topics

- **Monitor service**
  - check the physical/virtual hosts health status monitor
- **Remote controller**
  - perform actions over physical / virtual hosts - choice algorithm:
    - reboot
    - restart virtual machine on the same host
    - restart the whole virtual layer
    - move virtual machine to another host
    - reinstall from scratch on the same/another host - via PXE
- **Infrastructure**
  - DHCP, DNS, HTTP, PXE-TFTP
- **Storage architecture**
  - SAN, DRDB
- **Procedures**
  - physical to virtual migration

Architecture

3RC Architecture



Redundancy in virtual environment

Several redundancy strategies ▸ several availability levels

- Virtual machines on external storage
  - problems if software crashes
- Scheduled virtual machines dump: disk, ram, registers
  - dump at scheduled times ▸ recovery at time T_{n-1}
- Virtual machines with OS and MW ready to be mounted
  - virgin machine from disk copy
- Install from scratch: operating system and middleware
  - virgin machine from real installation via PXE

Recovery time

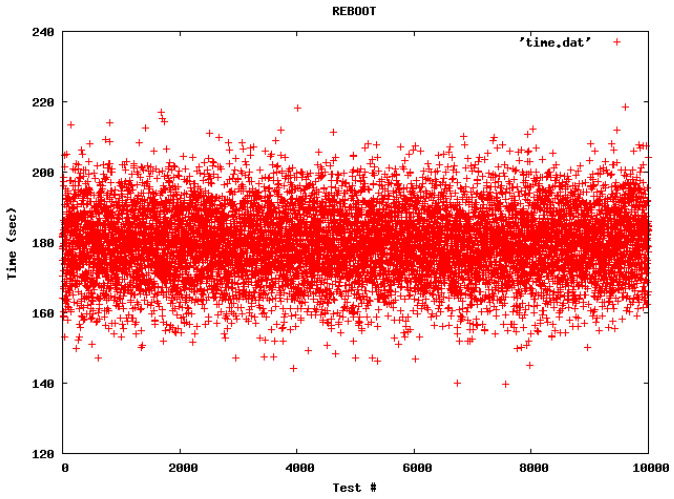
Time schedule

- monitor 70 sec ± 1
- controller 30 sec ± 30
- re-boot 80 sec ± 10 [PXE: 10 sec + boot: 70 sec]

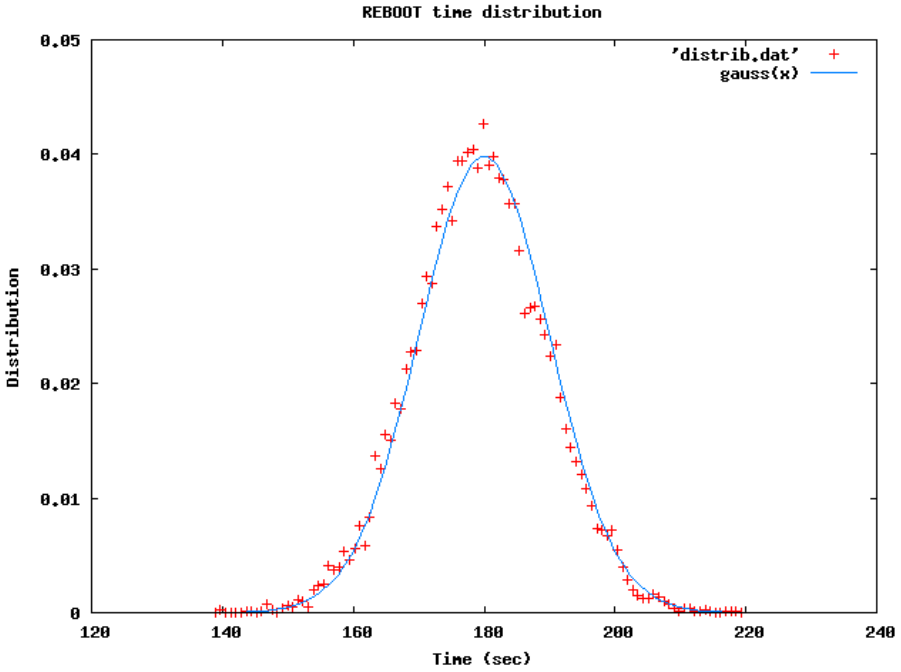
Experimental data - I

NON Destructive test

- overload
- shutdown



Recovery time - 10.000 crash test

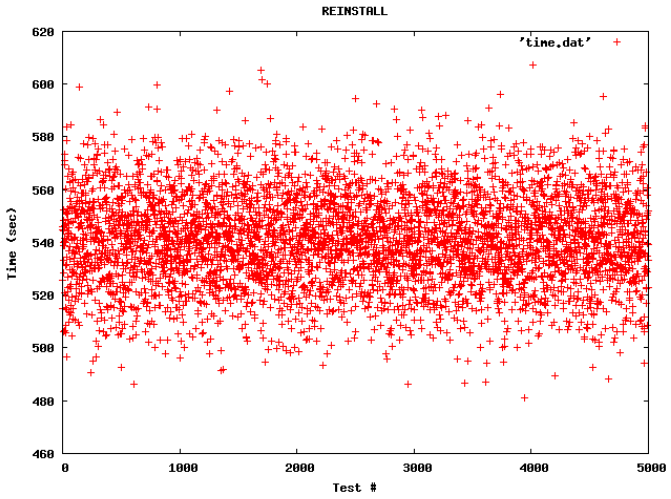


Recovery time distribution - 10.000 crash test
mean **181** sec
sigma 10 sec

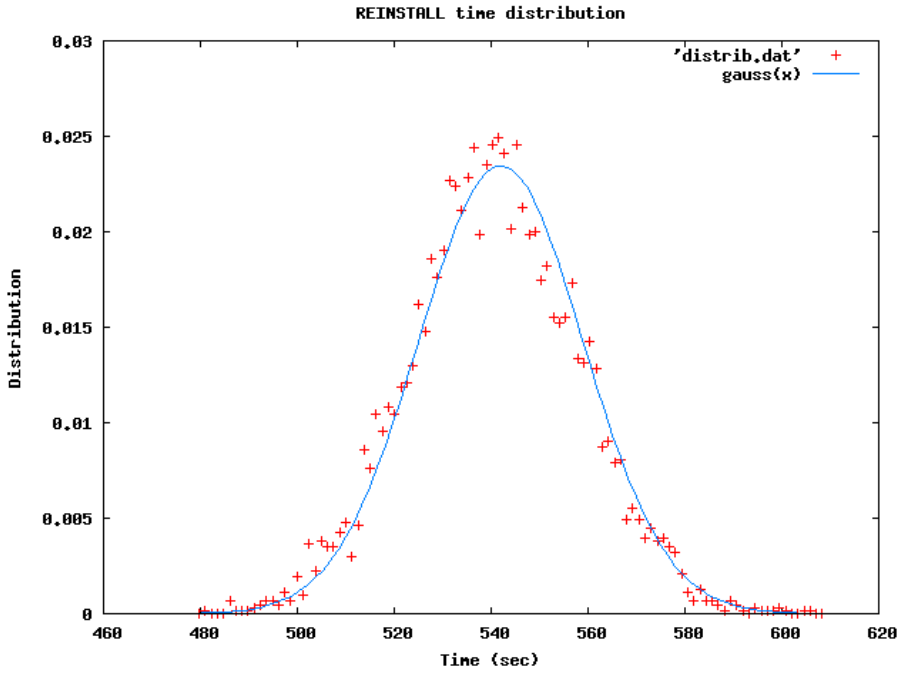
Experimental data - II

Destructive test

- rm /boot; reboot
- dd zero /sda; reboot



Reinstall time - 5.000 crash test



Reinstall time **distribution** - 5.000 crash test
mean **542** sec
sigma 17 sec

Physical to Virtual migration

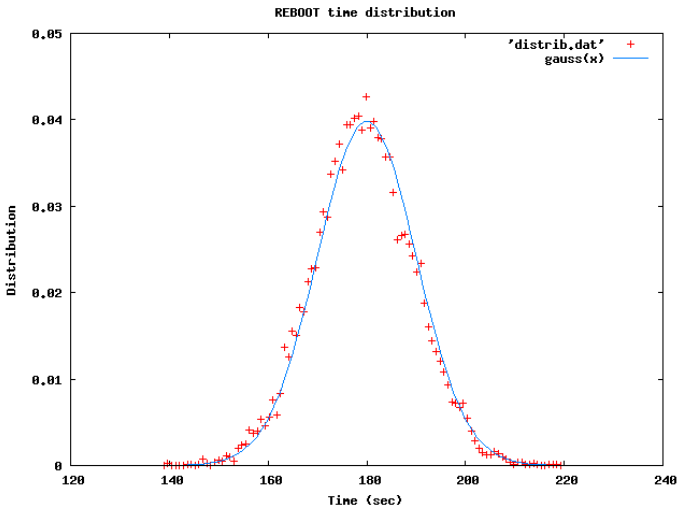
How to migrate a physical machine to a virtual machine

- physical machine RUNNING
  - create virtual disk
  - mount virtual disk with Linux live distro or Virtualization-tools
  - rsync <real> to <virtual>
  - untar <special path> [/dev]
  - grub install
  - < 20 sec downtime for switch real to virtual
- physical machine STOPPED
  - create virtual disk
  - mount virtual disk with Linux live distro or Virtualization-tools
  - dd <real> to <virtual>
  - grub install

Outcomes

- RECOVER crashed machine in **3** min
- REINSTALL broken machine in **9** min

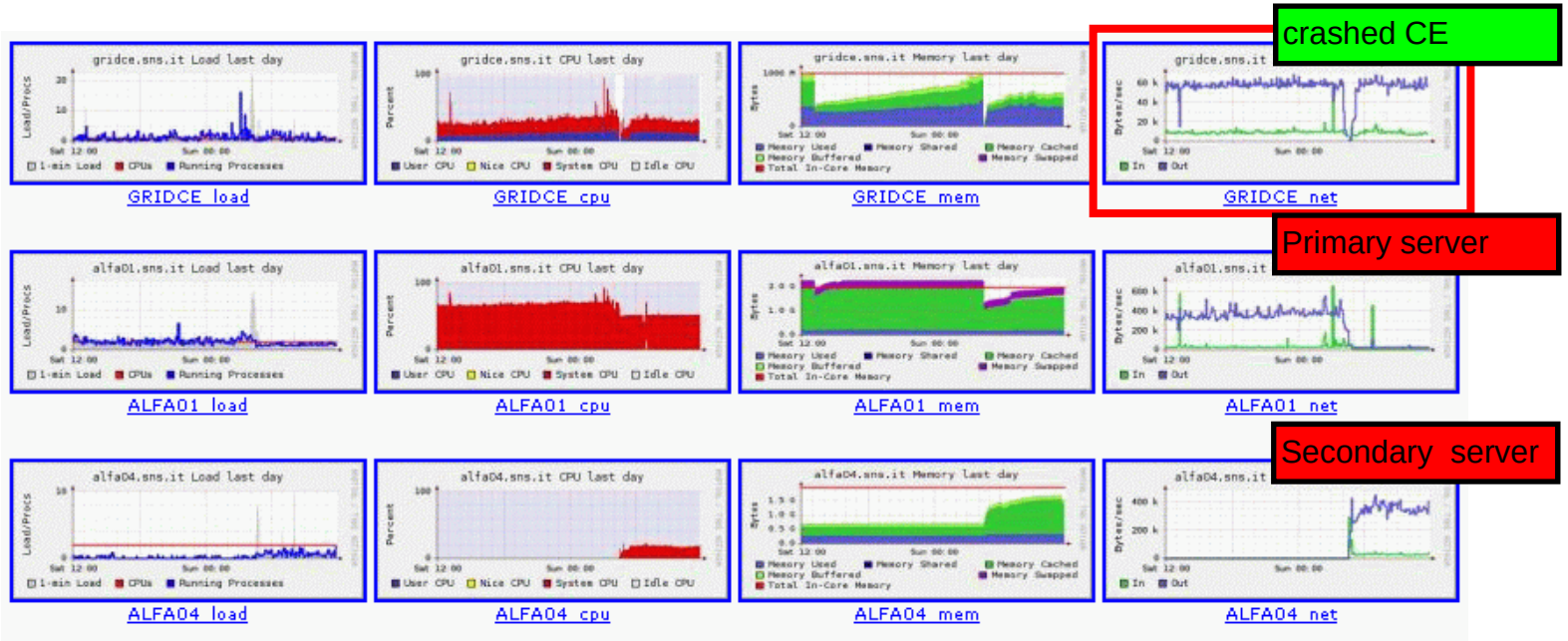
- SNS-PISA is the first EGEE/LCG Grid node
  - fully virtualized (services + WN)
  - highly available
  - NO downtime after service crash



Recovery time

Operation in a real crash example

gridce.sns.it [SNS-PISA Grid node master CE] crashes for an electrical power glitch @4:00 AM



GRIDCE crashed virtual machine
ALFA01 primary physical host
ALFA04 secondary physical host

@ crash_time the algorithm decides if restart or reinstall virtual machine over the same or another physical host

What 3RC High availability project is for

All the environments satisfied by a Relaxed High availability solution

- computing
- information
- monitoring
- users management
- GRID data center services

Note

It is important to know what a theorem states,
but it is probably more important
to know what a theorem does not state.

statement by Luigi Picasso, Theoretical Physics Professor
@University of Pisa

What 3RC High availability is NOT for

Mission critical applications

- financial transactions
- security certificates management
- real time controllers
- human health related applications

miracles [at least in the current release]

Spin-off

Host on-demand and Cloud computing

Basic concepts

- Virtualization and PXE architecture allows to bring up a server in a few minutes

Possibility to offer host on-demand

- CPU n core
- RAM n GB
- DISK n TB
- Operating System: Linux, Windows
- Middleware and Grid Applications Globus/LCG
- for T time
- at the end of time T hosts will be erased!!!

The End

Thanks

federico.calzolari@sns.it