Plug-and-play Virtual Appliance Clusters Running Hadoop

Dr. Renato Figueiredo
ACIS Lab - University of Florida

Introduction

- You have so far learned about how to use Hadoop clusters
- Up to now, you have used resources configured by others

- In this lecture you will learn about ways of deploying *your own* software stack using *virtual appliances*
- And we will overview a system that makes for simple configuration of groups of virtual appliances – i.e. *virtual clusters*

Objectives

- Concepts you will learn:
  - What is a virtual appliance?
  - What is a GroupVPN?
  - What is a virtual cluster?

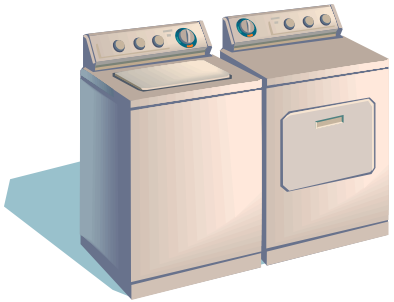
- Demonstrations, software that you will be able to take and follow on your own
  - Deploy your Hadoop cluster (and beyond)
    - On clouds – e.g. FutureGrid, EC2, private cloud
    - On your own local resources – desktops
    - Even across institutions

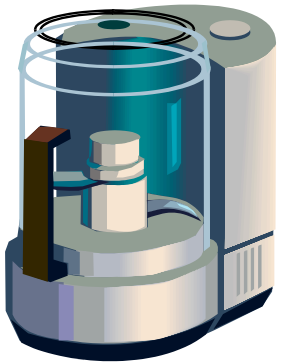
Outline

- Virtual appliances and the Grid appliance
- GroupVPN – easy to use, social VPNs
- Case study and demonstration: creating your own Hadoop cluster
  - Local resources
  - Cloud resources
  - Across providers

What is an appliance?

- Physical appliances
  - Webster – “an instrument or device designed for a particular use or function”





What is an appliance?

- Hardware/software appliances
  - TV receiver + computer + hard disk + Linux + user interface

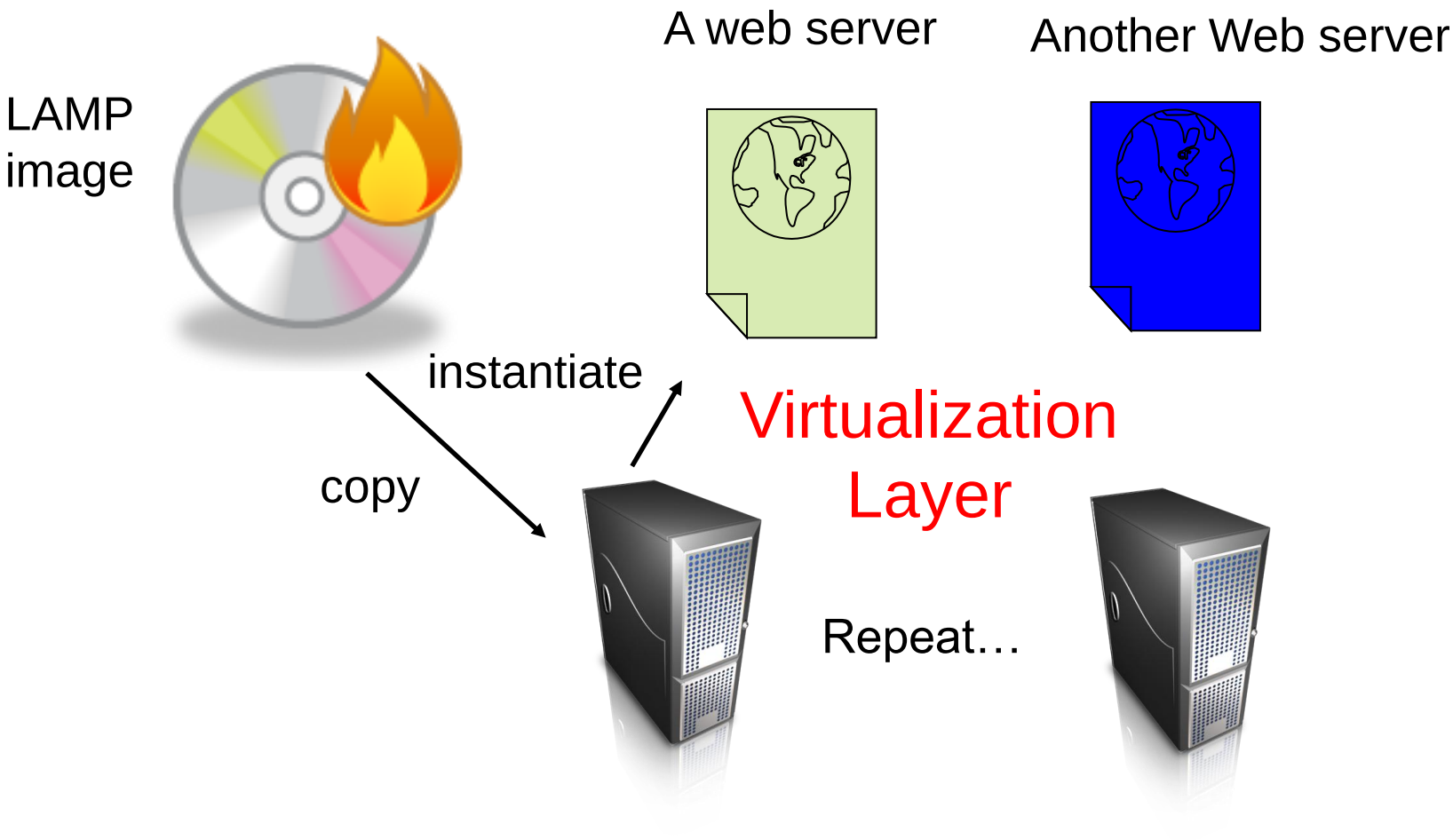
  - Computer + network interfaces + FreeBSD + user interface

What is a virtual appliance?

- An appliance that packages software and configuration needed for a particular purpose into a virtual machine “image”
- The virtual appliance has no hardware – just software and configuration
- The image is a (big) file
- It can be *instantiated* on hardware

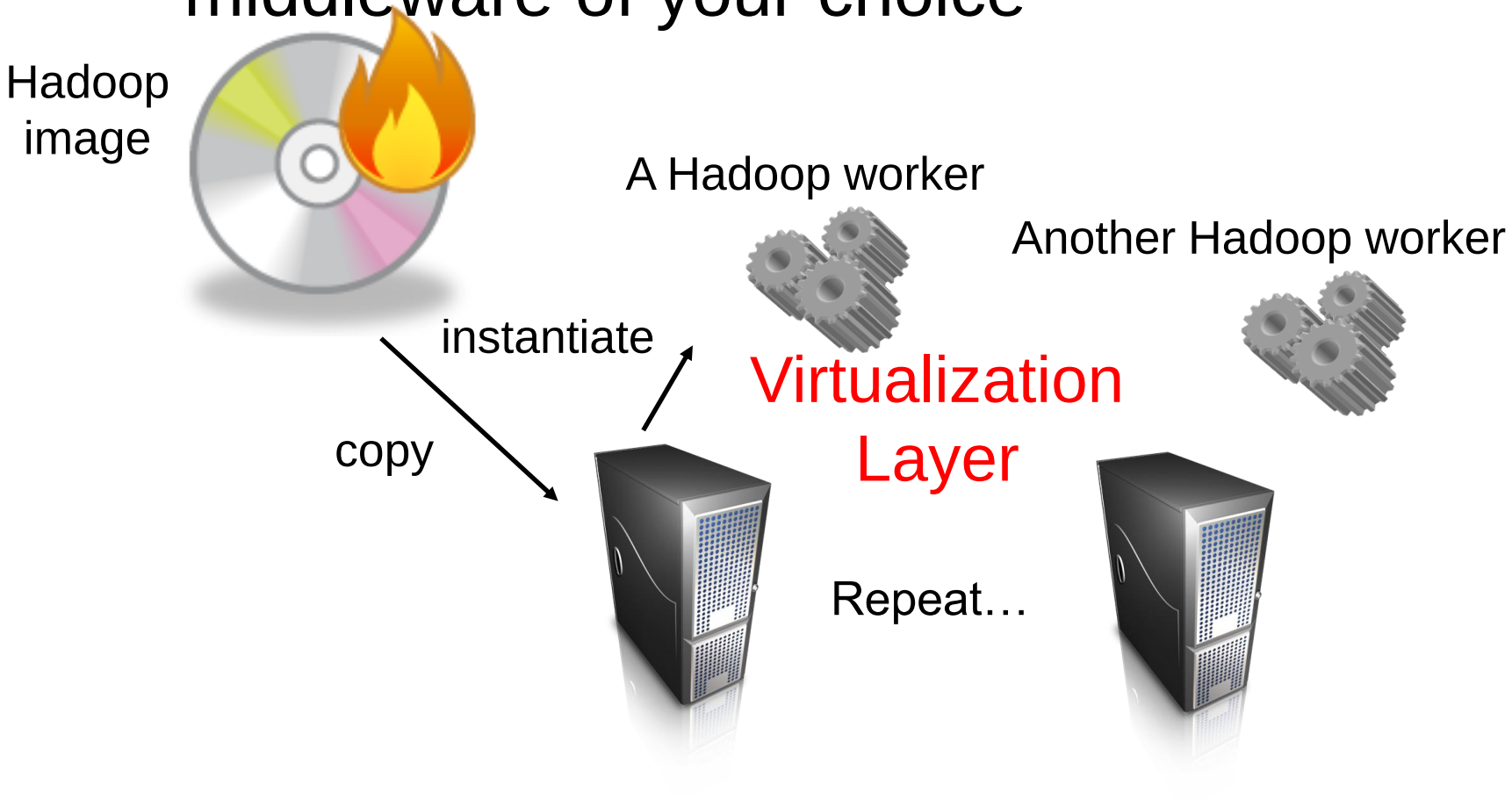
Virtual appliance example

- Linux + Apache + MySQL + PHP



We were talking about Hadoop?

- Replace Apache, MySQL, PHP with the middleware of your choice



What about the network?

- Multiple Web servers might be completely independent from each other
- Hadoop workers are not
  - Need to communicate and coordinate with each other
  - Each worker needs an IP address, uses TCP/IP sockets
- Cluster middleware stacks assume a collection of machines, typically on a LAN (Local Area Network)

Enter virtual networks

NOWs, COWs

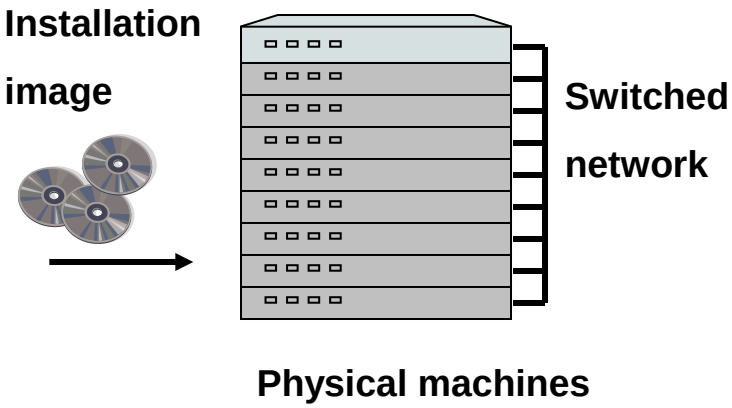
- **Local-area**
- **Physical machines**
- **Self-organizing switching**

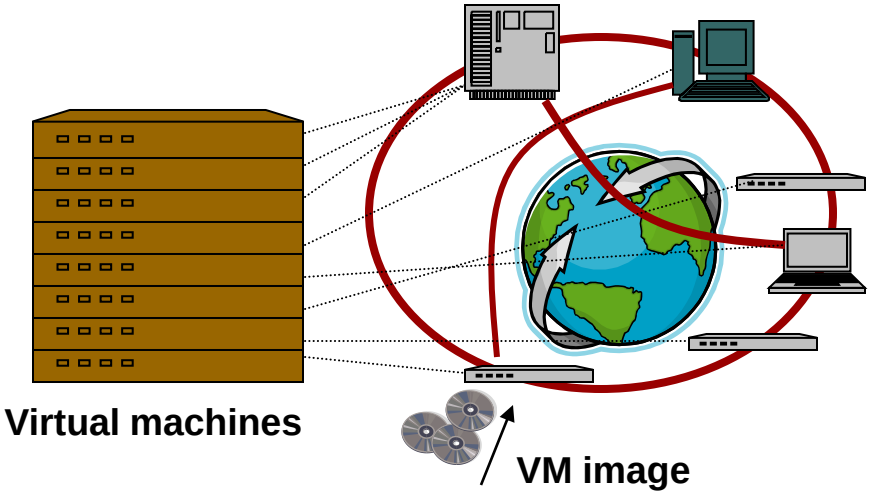
(e.g. Ethernet spanning tree)

"WOWs"

- **Wide-area**
- **Virtual machines (VMs)**
- **Self-organizing overlay**

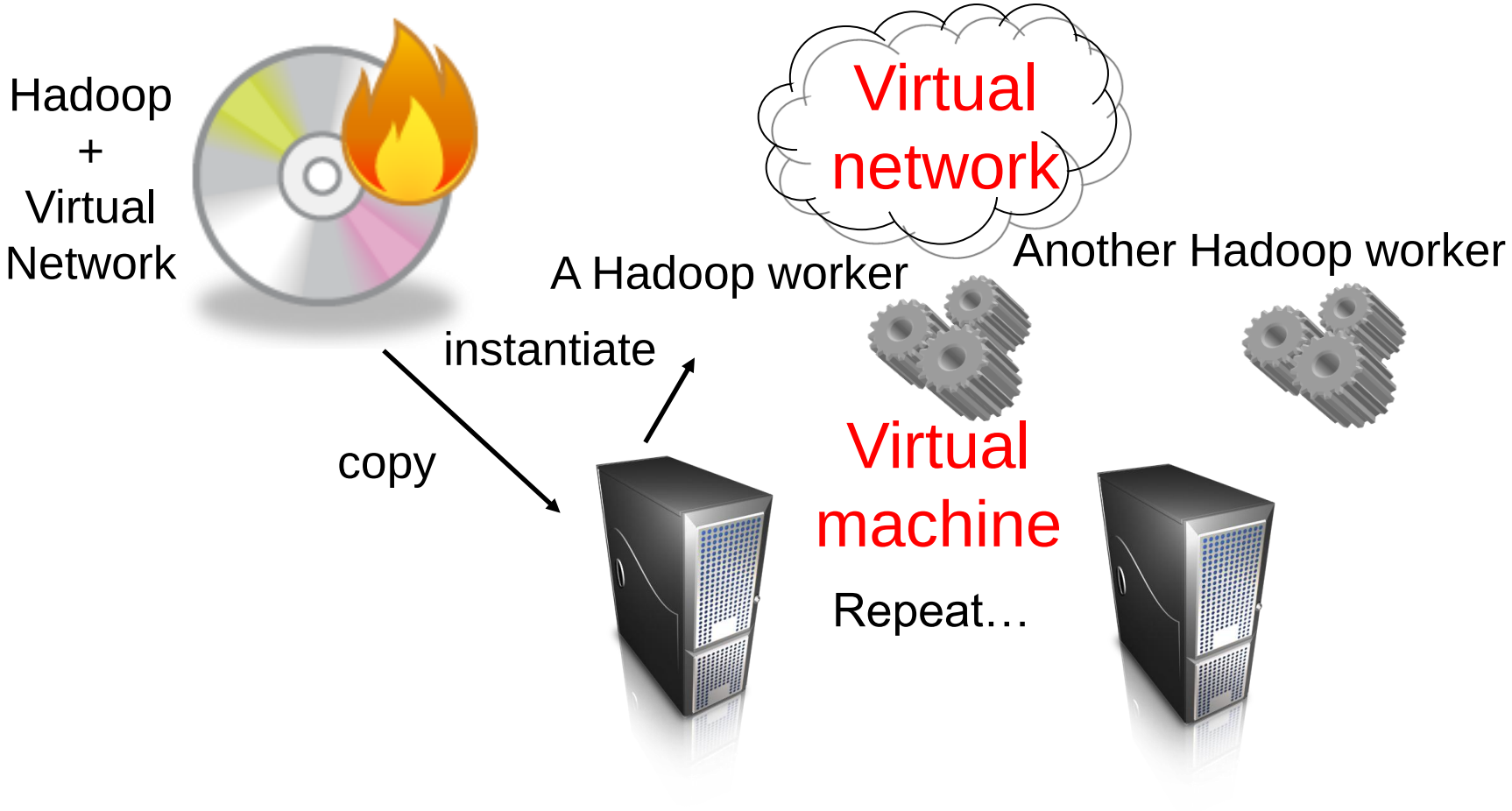
IP tunnels, P2P routing



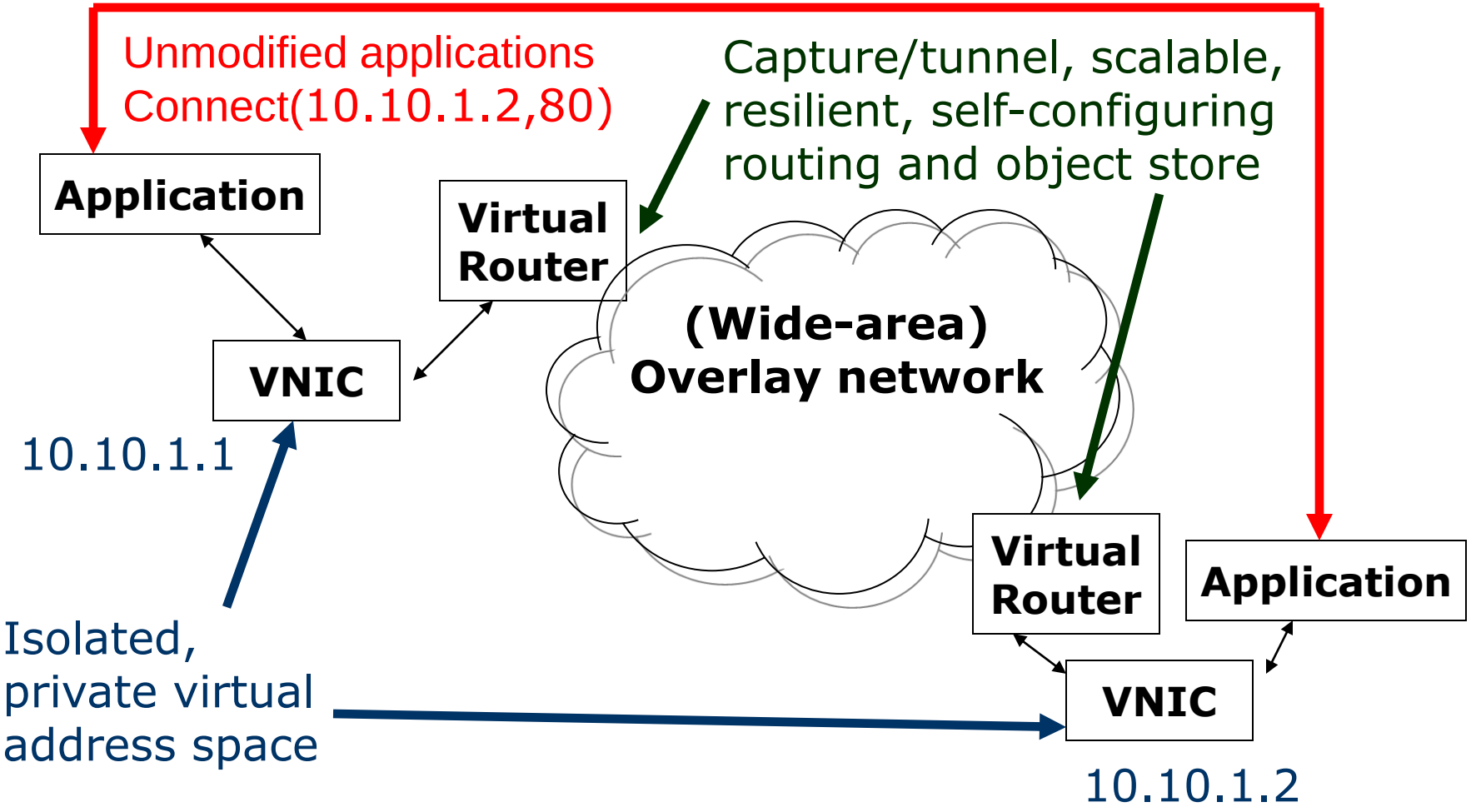


Virtual cluster appliances

- Virtual appliance + virtual network



Virtual network architecture



Demonstration

- A virtual appliance cluster

Q & A

Background

- Virtual appliances
  - Encapsulate software environment in image
    - Virtual disk file(s) and virtual hardware configuration
- The Grid appliance
  - Encapsulates *cluster* software environments
    - Current examples: Condor, MPI, Hadoop
  - Homogeneous images at each node
  - *Virtual LAN* connecting nodes to form a cluster
  - Deploy within or across domains

Grid appliance in a nutshell

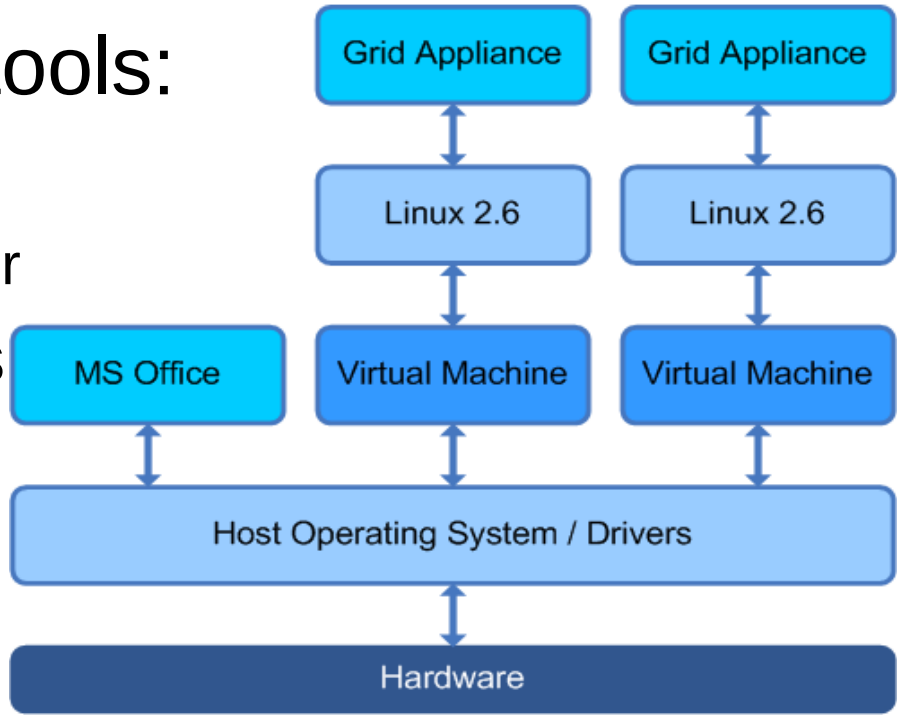
- Plug-and-play clusters with a pre-configured software environment
  - Linux + (Hadoop, Condor, MPI, …)
  - Scripts for zero-configuration
  - “Virtual machine” appliance; open-source software runs on Linux, Windows, Mac
- Hands-on examples, bootstrap infrastructure, and zero-configuration software – *you’re off to a quick start*

Grid appliance in a nutshell

- Creating an equivalent Grid on your own resources, or on cloud providers, is also easy
- Deploy image on FutureGrid, Amazon EC2
- Copy the same appliance to clusters, PC labs
- Simple deployment and management of ad-hoc clusters
  - Opportunistic computing
  - Testing, evaluation
  - Education, training

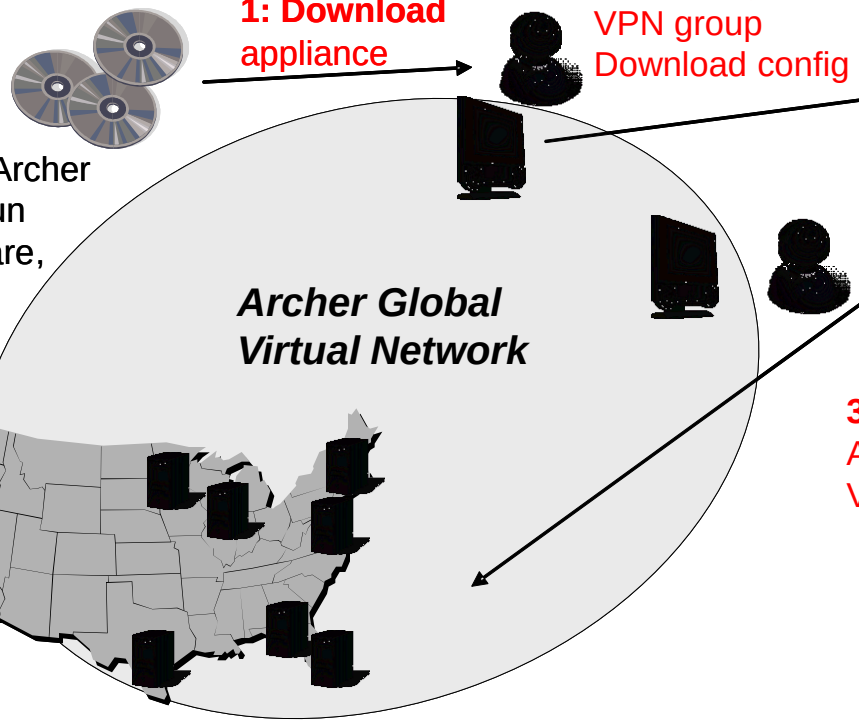
Example: Desktop Grids

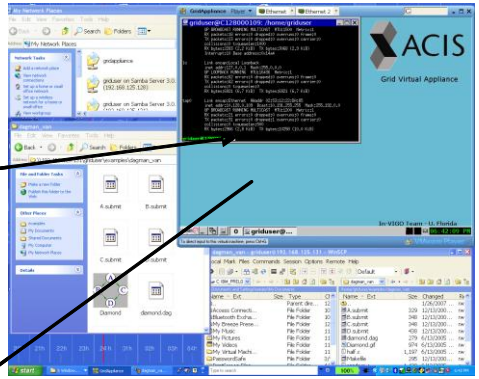
- Reuse wealth of O/S tools:
  - VM image = files
    - Copy, compress, transfer
  - VM instance = process



- Easy install on typical systems
  - KVM, VirtualBox: open-source
  - VMware Player/Server/Workstation

Appliance/GroupVPN Example





Free pre-packaged Archer Virtual appliances - run on free VMMs (VMware, VirtualBox, KVM)

CMS, Wiki, YouTube: Community-contributed content: applications, datasets, tutorials

3. Boot appliances
Automatic connection to group VPN – self-configuring DHCP

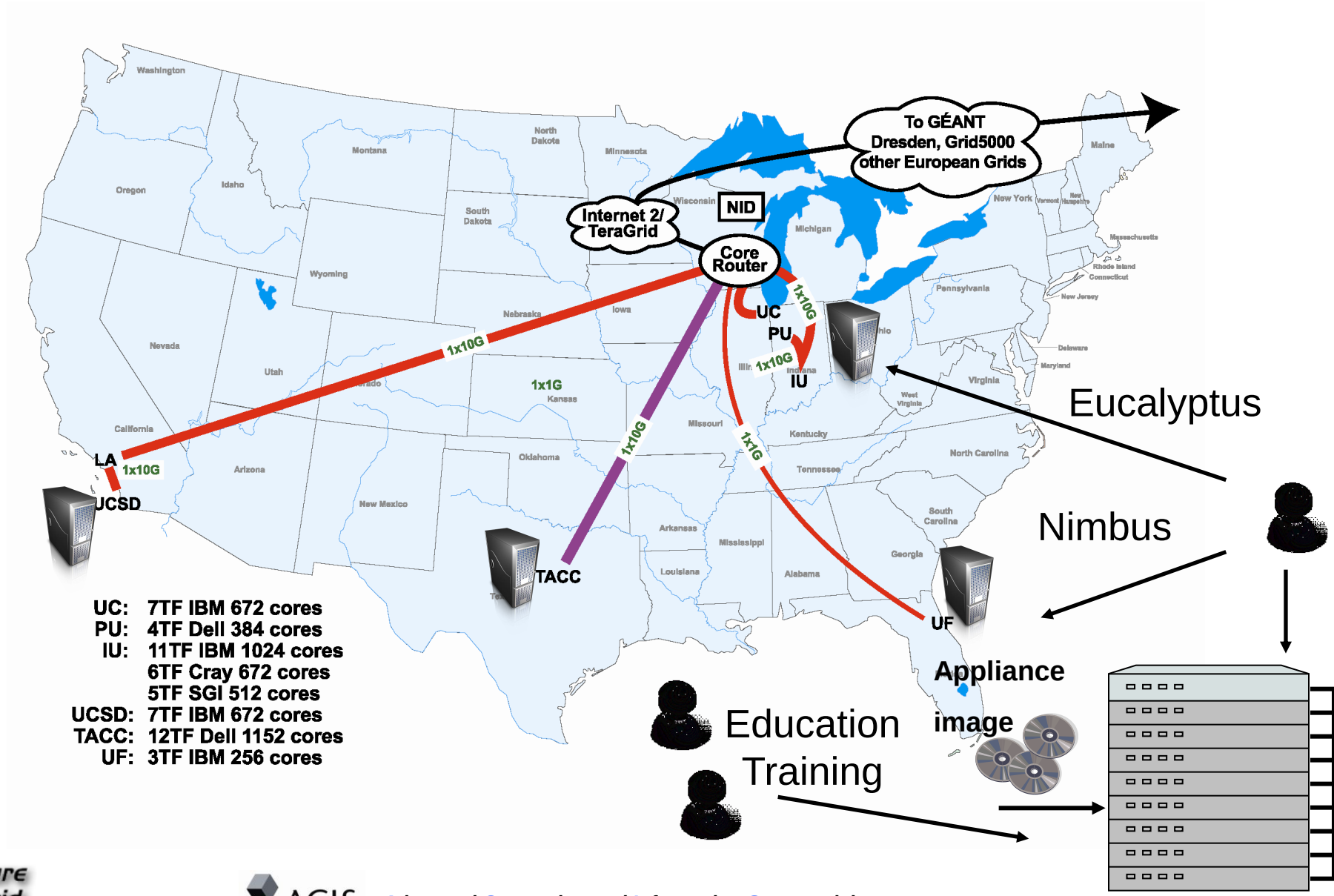
Middleware:
Condor scheduler
NFS file systems

Archer seed resources
450 cores, 5 sites

Cloud deployment

- Cloud meaning Infrastructure-as-a-Service
  - Pay as needed
    - Elasticity – you typically only need cycles near conference deadlines
      - 100 nodes for two weeks vs 4 nodes for a year?
    - Management, cooling, power costs are not an issue
  - Amazon EC2 pricing today makes it a viable option
    - On-demand: $0.085/hour (1 core, 1.7GB), $0.34/hour for large (2 cores, 7.5GB)
      - $2856 for 100 small nodes for 2 weeks
    - Reserved: $228 fee, then $0.03/hour
    - Research credits available through grants
  - Research infrastructures
    - FutureGrid; Science Clouds
  - Private clouds

Example – FutureGrid



Grid appliance: under the hood

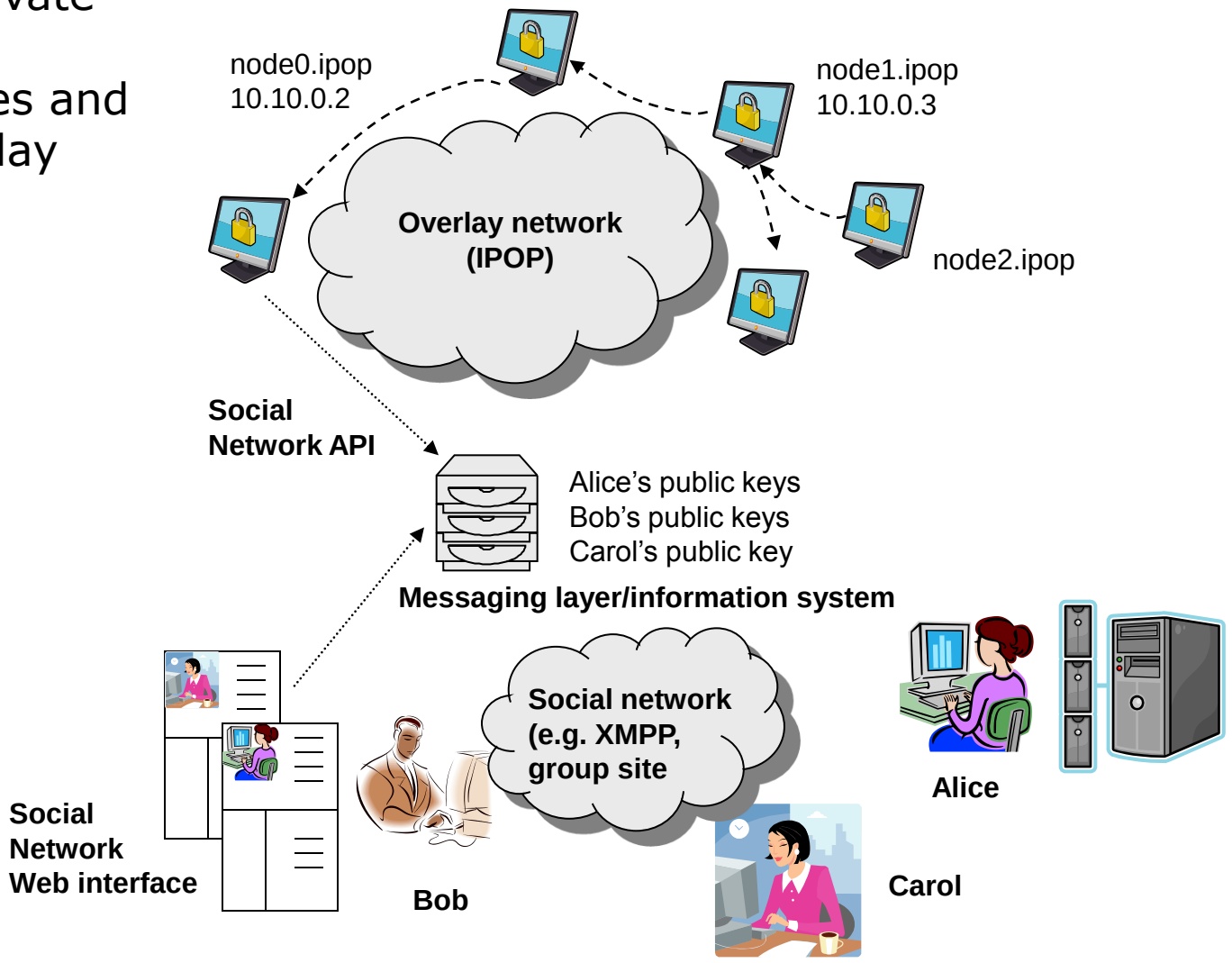
- VM instances + GroupVPN + Grid/cloud middleware
  - VM instances (Xen, Vmware, KVM, …) provide:
    - Sandboxing; software packaging; decoupling
    - Can be provisioned ad-hoc or through Cloud middleware
  - Virtual network (UF's GroupVPN) provides:
    - Virtual private LAN over WAN; self-configuring and capable of firewall/NAT traversal
  - Grid/cloud middleware (Condor, Hadoop, MPI):
    - Scheduling, data transfers, …
    - *unmodified*

Virtual network: GroupVPN

- Key technique: IP-over-P2P (IPOP) tunneling
  - Interconnect VM appliances
  - *VMs perceive a virtual LAN environment*
- *Self-configuring*
  - Avoid administrative overhead of typical VPNs
  - NAT and firewall traversal
- *Scalable and robust*
  - P2P routing deals with node joins and leaves
- *Networks are isolated*
  - One or more private IP address spaces
  - Decentralized DHCP serves addresses for each space

GroupVPN Overview

Bootstrapping private links through Web 2.0 interfaces and IP-over-P2P overlay tunneling



Creating your own GroupVPN

- Setting up and managing typical VPNs can be daunting
  - VPN server(s), key distribution, NAT traversal
- GroupVPN makes it simple for *users to create and manage virtual cluster VPNs*
- Key insights:
  - Web 2.0 interface: create/manage user groups
  - All the complexity of setting up and managing VPN links is automated

GroupVPN Web interface

- You can request to join or create your own VPN group
  - Determines who is allowed to connect to virtual network
- You can request to join or create your own appliance group
  - Determines priorities of users on resources owned by their groups

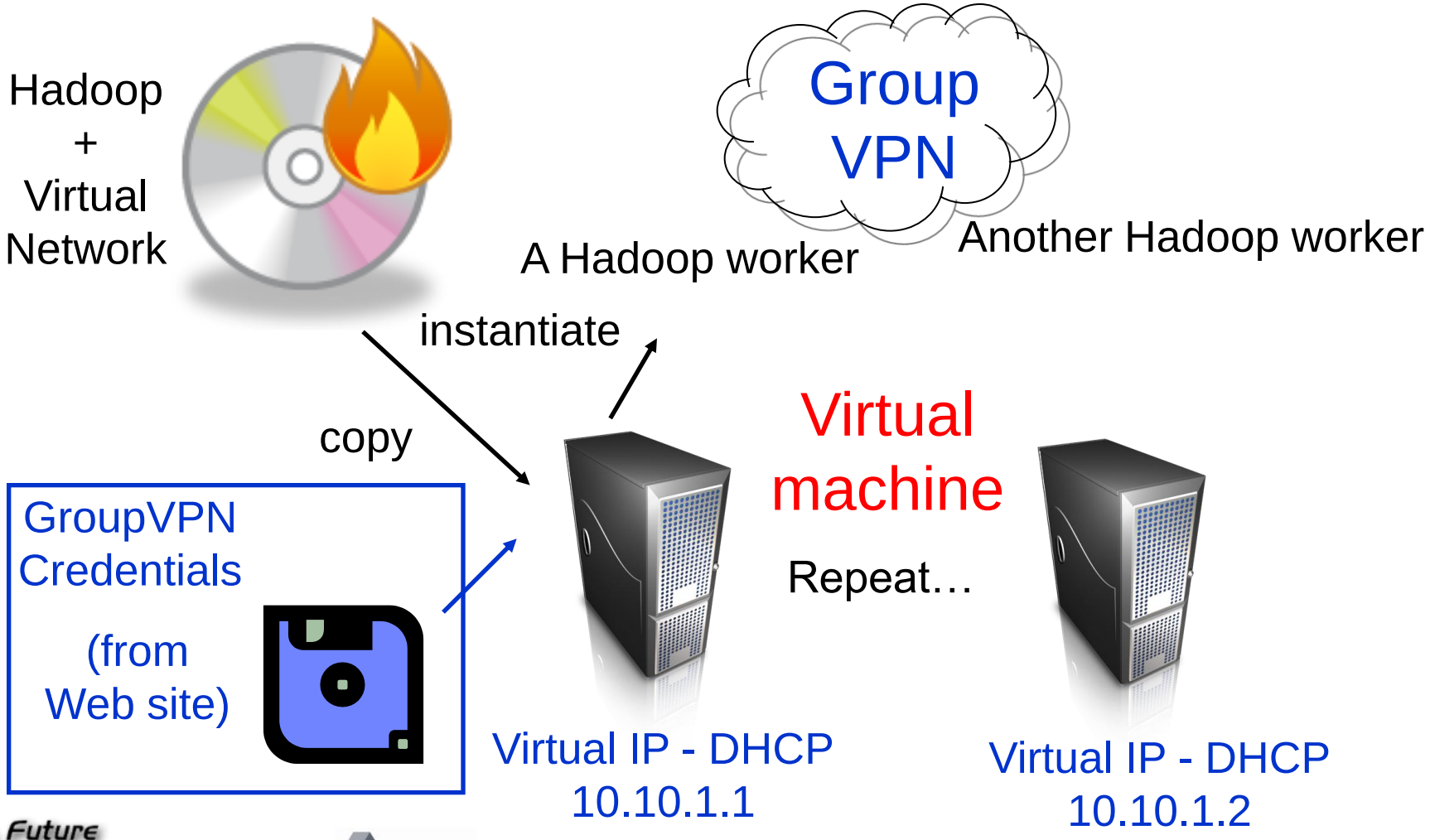
Demonstration

- GroupVPN user interface

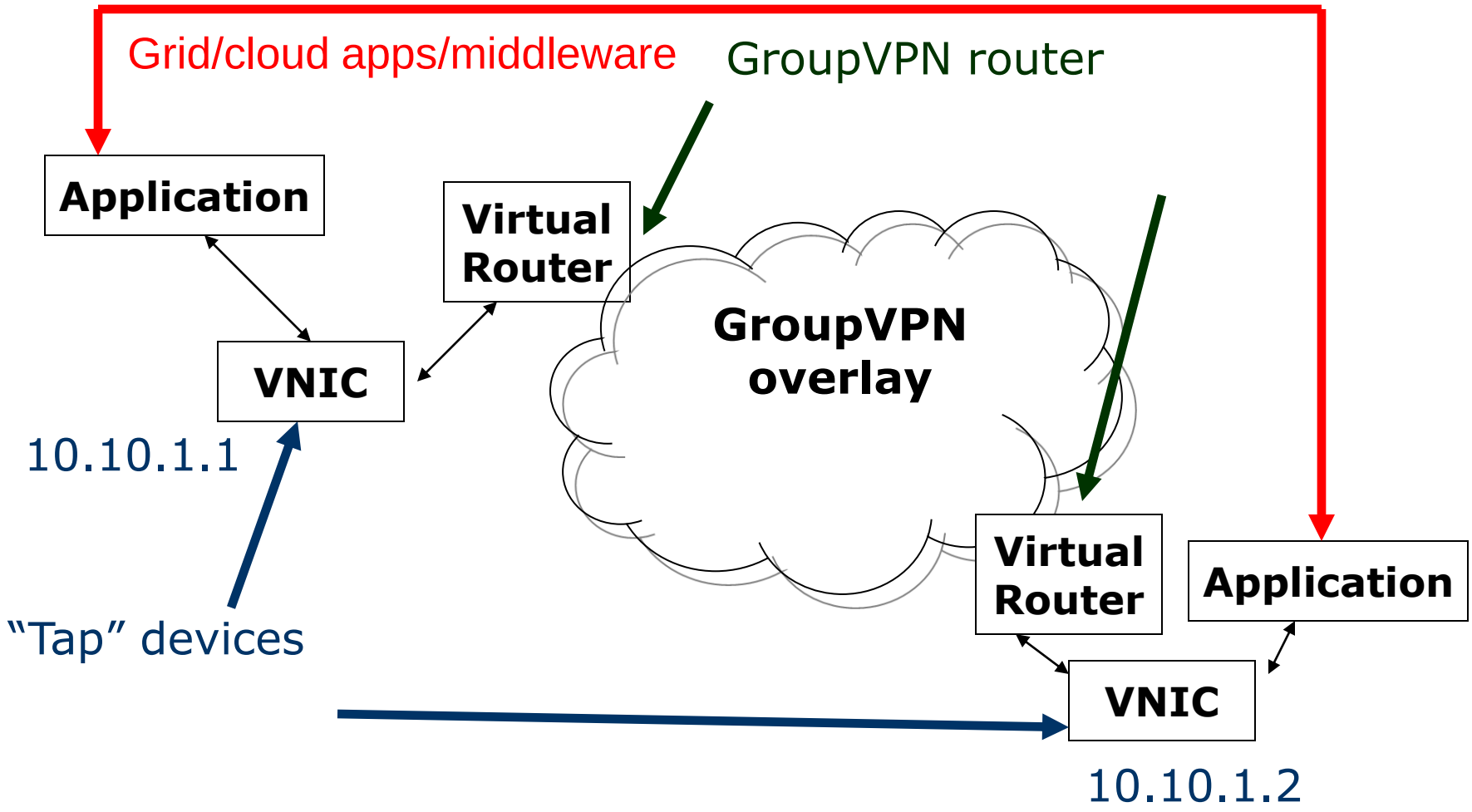
Q & A

Deploying virtual clusters

- Same image, different VPNs

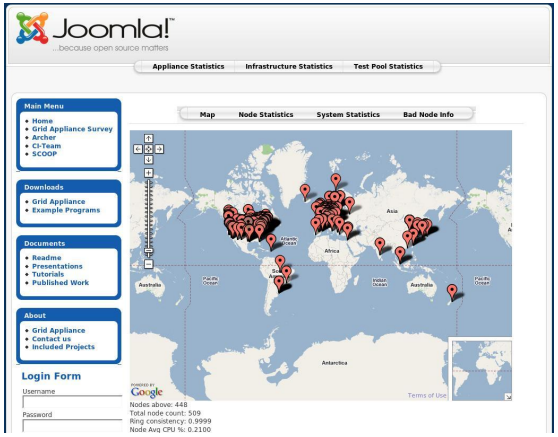


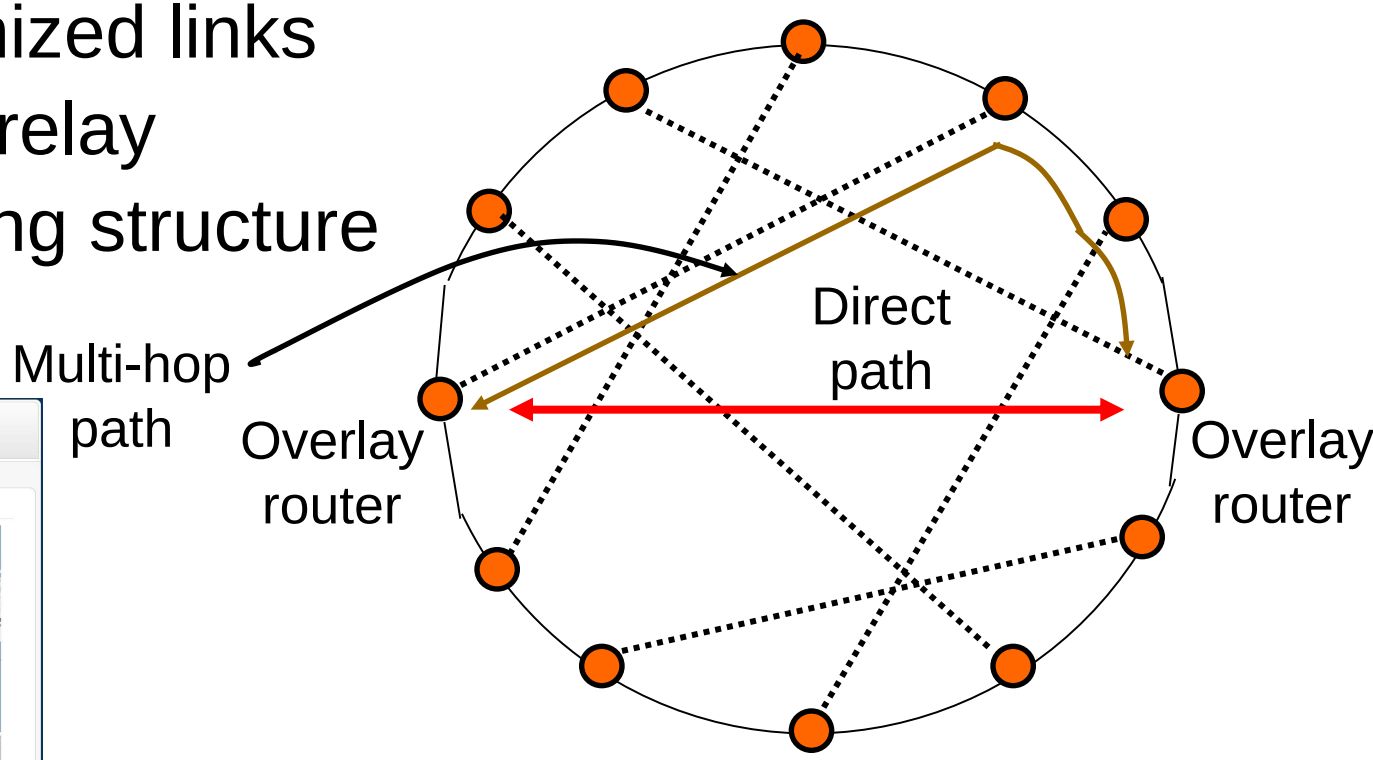
GroupVPN architecture



Under the hood: overlay architecture

- Bi-directional structured overlay (Brunet library)
- Self-configured NAT traversal
- Self-optimized links
  - Direct, relay
- Self-healing structure





Cloud deployment approach

- Generate virtual floppies
  - Through GroupVPN and GroupAppliance Web interface
- Deploy appliances image(s)
  - FutureGrid (Nimbus/Eucalyptus), EC2
  - GUI or command line tools
  - Use APIs to copy virtual floppy to image
- Submit jobs; terminate VMs when done

FutureGrid example - Nimbus

- Example using Nimb

workspace.sh --deploy
/tmp/floppy-worker.zip
https://f1r.idp.ufl.futuregrid.org:8443/wsrf
/services/Work --
file /tmp/outp
/tmp/grid-appliance.xml --deploy-mem
1000 --deploy-du --trash-at-
shutdown Trash
displayname
/home/renato/.ssh/id_dsa.pub

GroupVPN floppy

Nimbus service endpoint

Metadata – points to image on Nimbus server

SSH public key to log in to instance

FutureGrid example - Eucalyptus

- Example using Eucalyptus (or ec2-run-instances on Amazon E

GroupVPN floppy image

Image ID on Eucalyptus server

euca-run-instances ami-fd4aa494 -f floppy.zip --instance-type m1.large -k keypair

SSH public key to log in to instance

Demonstration

- Deploying virtual appliance node on FutureGrid
- Configuring Hadoop cluster

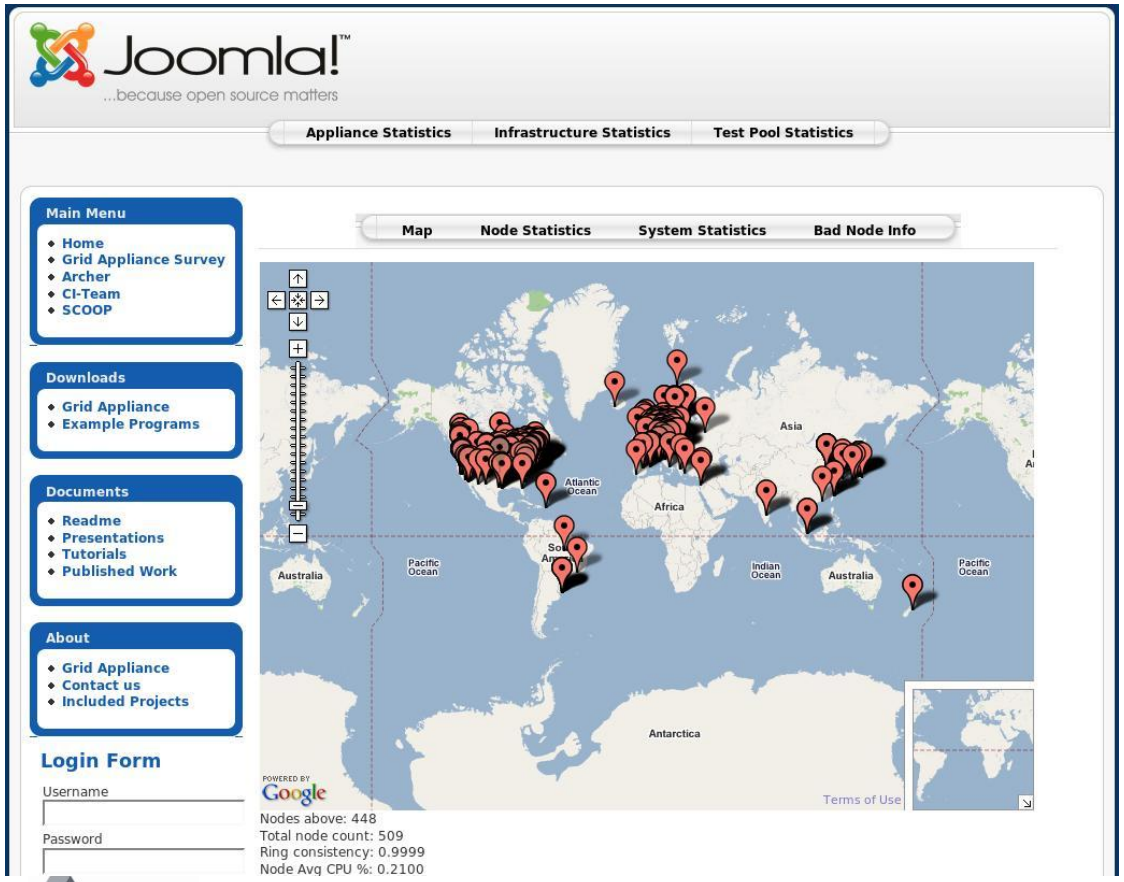
Q & A

Local appliance deployments

- Two possibilities:
  - Share our “bootstrap” infrastructure, but run a separate GroupVPN
    - Simplest to setup
  - Deploy your own “bootstrap” infrastructure
    - More work to setup
      - Especially if across multiple LANs
    - Potential for faster connectivity

PlanetLab bootstrap

- Shared virtual network bootstrap
  - Runs 24/7 on 100s of machines on the public Internet
  - Connect machines across multiple domains, behind NATs



PlanetLab bootstrap: approach

- Create GroupVPN and GroupAppliance on the Grid appliance Web site
- Download configuration floppy
- Point users to the interface; allow users you trust into the group
- Trusted users can download configuration floppies and boot up appliances

Private bootstrap: General approach

- Good choice for single-domain pools
- Create GroupVPN and GroupAppliance on the Grid appliance Web site
- Deploy a small IPOP/GroupVPN bootstrap P2P pool
  - Can be on a physical machine, or appliance
  - Detailed instructions at grid-appliance.org
- The remaining steps are the same as for the shared bootstrap

Connecting external resources

- GroupVPN can run directly on a physical machine, if desired
  - Provides a VPN network interface
  - Useful for example if you already have a local Condor pool
    - Can “flock” to Archer
  - Also allows you to install Archer stack directly on a physical machine if you wish

Demonstration

- Connecting a local appliance to FutureGrid cluster

Where to go from here?

- Tutorials on FutureGrid and Grid appliance Web sites for various middleware stacks
  - Condor, MPI, Hadoop
- A community resource for educational virtual appliances
  - Success hinges on users effectively getting involved
  - If you are happy with the system, let others know!
  - Contribute with your own content – virtual appliance images, tutorials, etc

Questions?

- More information:
  - http://www.futuregrid.org
  - http://grid-appliance.org

- This document was developed with support from the National Science Foundation (NSF) under Grant No. 0910812 to Indiana University for "FutureGrid: An Experimental, High-Performance Grid Test-bed." Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF