The Design and Implementation of a Virtual Cluster Management System

Hidemoto Nakada [1], Takeshi Yokoi [1], Tadashi Ebara [1, 2]
Yusuke Tanimura [1], Hirotaka Ogawa [1], Satoshi Sekiguchi [1]
1.National Institute of Advanced Industrial Science and Technology
2.SUURI Giken

Grid Technology Research Center AIST

AIST

Background

- **Computer Virtualization**
  - Virtual computers contribute reduction of management cost
- **Virtual Computer → Virtual Cluster**
  - For further reduction of management cost

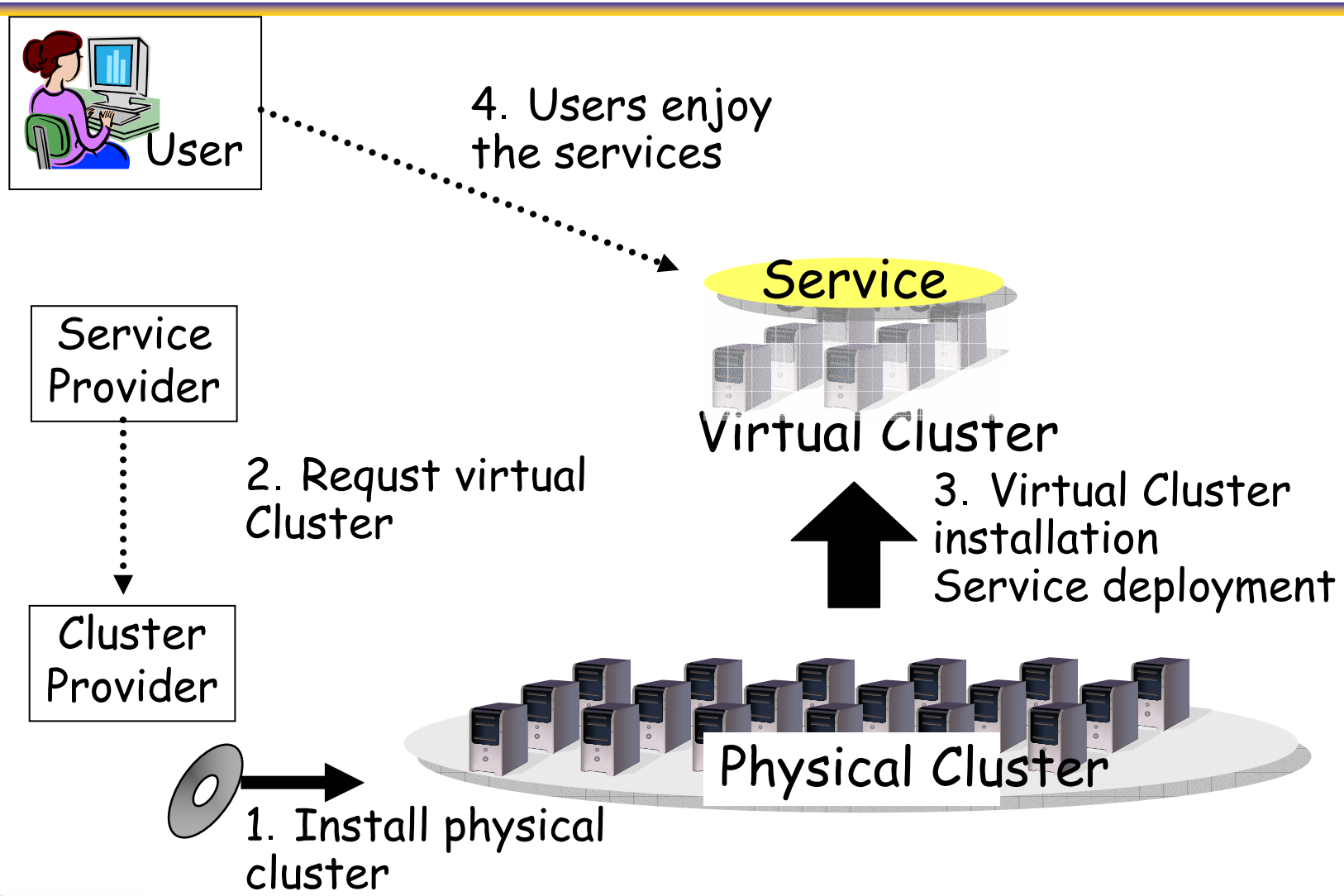
- **What is Virtual Cluster ?**
  - Not mere a group of virtual computers
    - Software configuration, management tools
    - Ex. User namespaces management
  - Computer virtualization is not enough
    - Storage
    - Network

Goal

- **Virtual Cluster**
  - For specific time period, a virtual cluster, with specified software installed, is provided.
  - Users have total control over the cluster
    - Modifications of configuration are allowed
  - Assumed time period: few days – few months.

- **Proposes a Virtual Cluster Management System**
  - Using Rocks, user specified applications and management tools are automatically installed and configured
  - Virtualization of computer, storage and network
    - Computer - VMware Server
    - Storage – iSCSI + LVM
    - Network - VLAN

Scenario

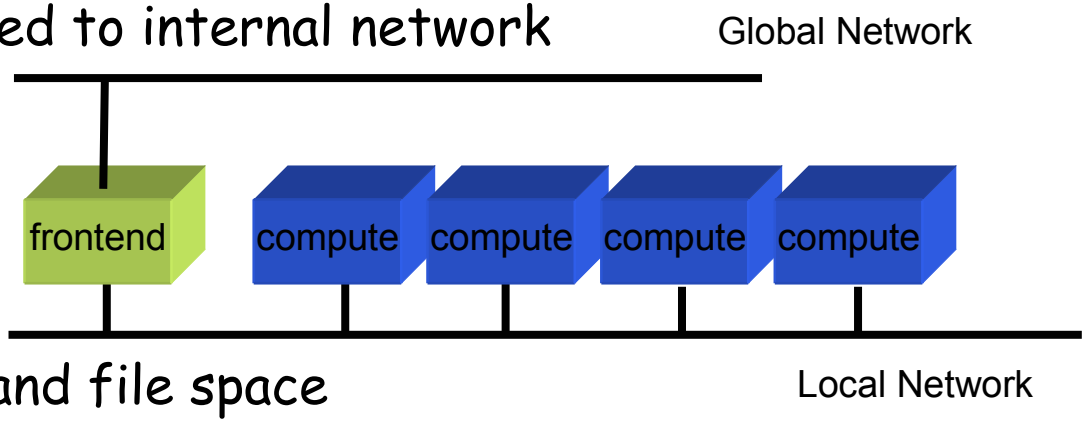

User
4. Users enjoy the services
Service
Virtual Cluster
Service Provider
2. Requst virtual Cluster
3. Virtual Cluster installation
Service deployment
Cluster Provider
Physical Cluster
1. Install physical cluster

Other examples of usage

- **At Class Room**
  - Allocate virtual clusters for each group of students
  - Students can try configuration and installation
    - Can restore to the original state
  - Wakes up same time weekly
- **On demand computer farm expansion**
  - Temporally expand computer farm to meet deadline☺
  - Transparent for users, with grid technology
  - Database and applications are automatically deployed

Requirements for Virtual Clusters

- **For Service Providers, looks same as the physical clusters**
- **Nodes and Networks**
  - One front-end node and worker nodes
  - The front-end acts as router for external network
  - Worker nodes are attached to internal network
    - Internal network is safe



- **Configuration**
  - Shared user name space and file space
  - Operation utilities are installed
    - Monitoring systems
    - Batch queuing systems
- **Storage**
  - Shared storage
  - Scratch file system on each node

Requirements for Virtual Cluster Management System

- **Automatic deploy and configuration of applications**
  - Complicated configuration over several nodes
  - Routing, etc.
- **Computer Virtualization**
  - Single physical nodes may host plural virtual nodes
- **Storage Virtualization**
  - Flexible storage management
    - Independent of physical disk configuration
  - Centralized management to decrease management cost
- **Network Virtualization**
  - With commonly used bridged connection, virtual nodes shares network with real nodes
    - Inappropriate for virtual cluster: separation is needed

Proposed System (1)

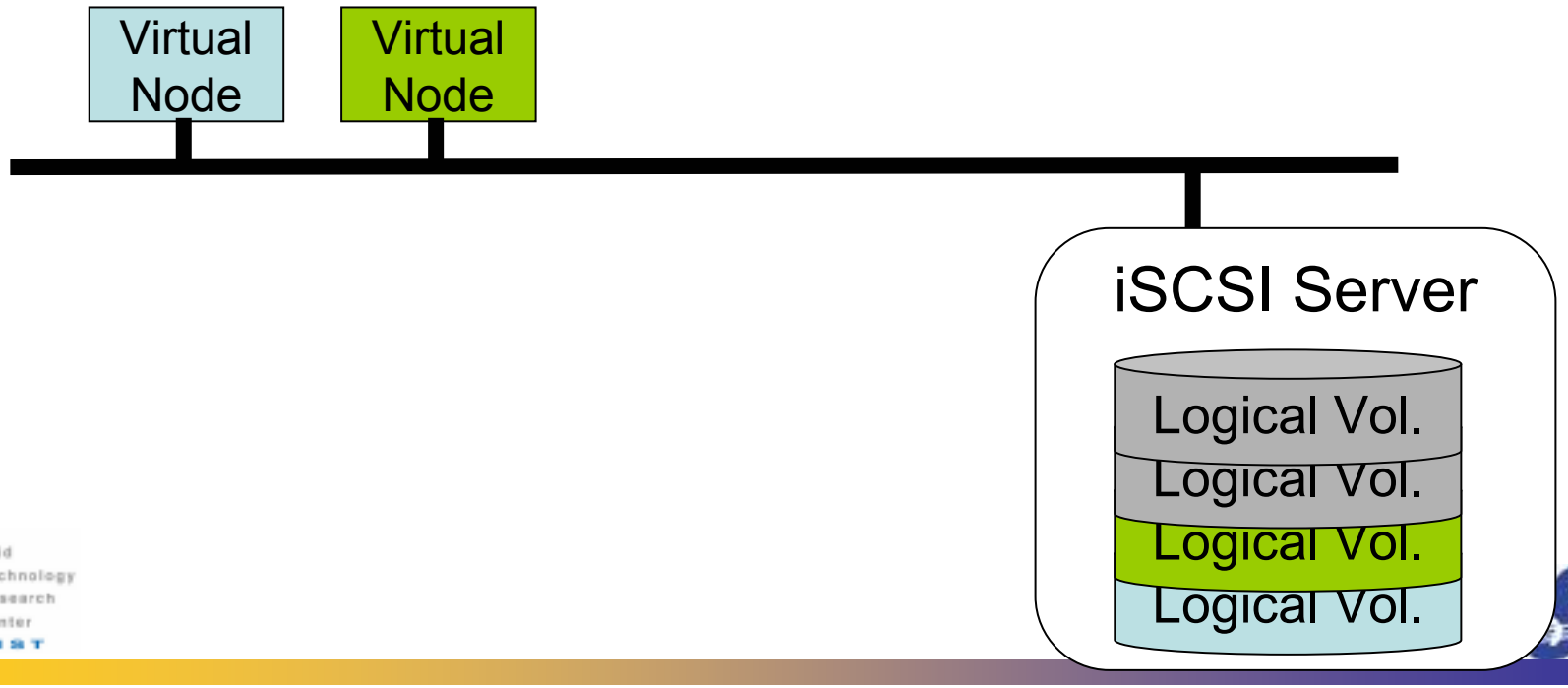
- **Automated application installation and node configuration.**
  - Leverage Rocks, Cluster installation tool.
    - Developed by UCSD as a part of NPACI project
    - Widely use with for cluster management
    - Plenty amount of Rolls(meta packages) are there
      - Covers most scientific computing applications and middlewares
      - No need to re-package them

Proposed System (2)

- **Computer Virtualization**
  - VMware Server
    - Freely available VMM with full virtualization
- **Storage Virtualization**
  - iSCSI + LVM (Logical Volume Manager)
    - iSCSI for location transparency
    - LVM for easy storage management
- **Network Virtualization**
  - Tagged VLAN
    - Logically separate networks of virtual clusters on a physical cluster

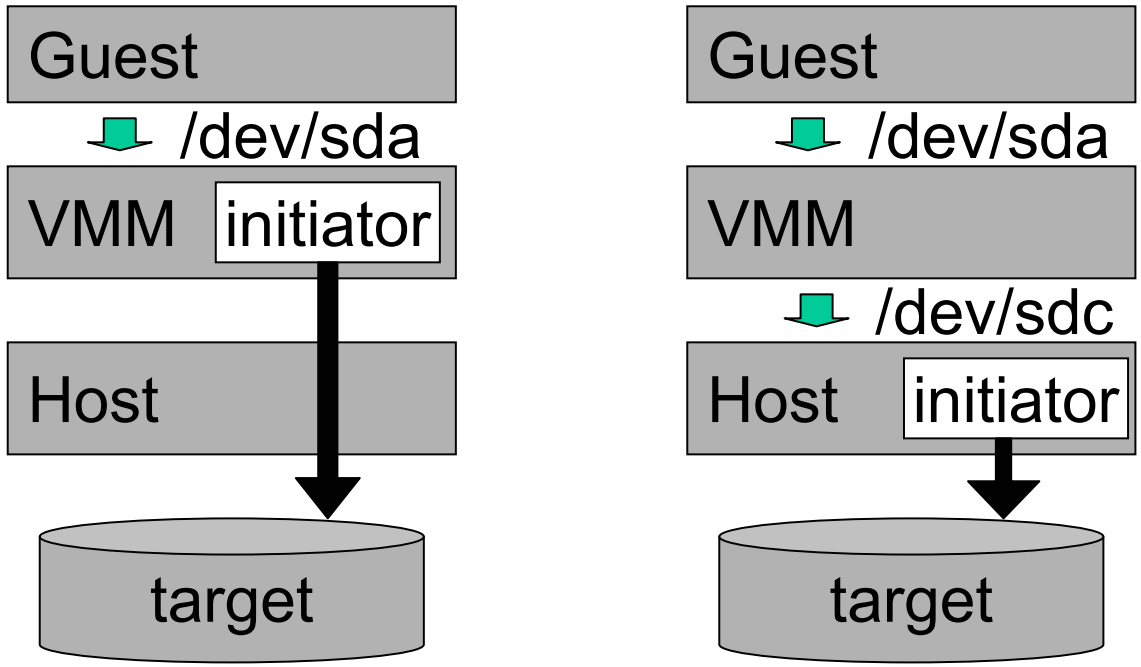
Storage Virtualization

- **Virtualize away storage from physical substance (i.e. disks), to reduce management cost**
  - iSCSI for location transparency
    - Enables centralized management.
  - LVM to enable arbitrary storage configuration, independent of physical disk configuration



iSCSI and VMM

- **Problem: VMware Server does not support iSCSI**
  - Work around: Host OS attaches the iSCSI volumes and exposes them to VMM

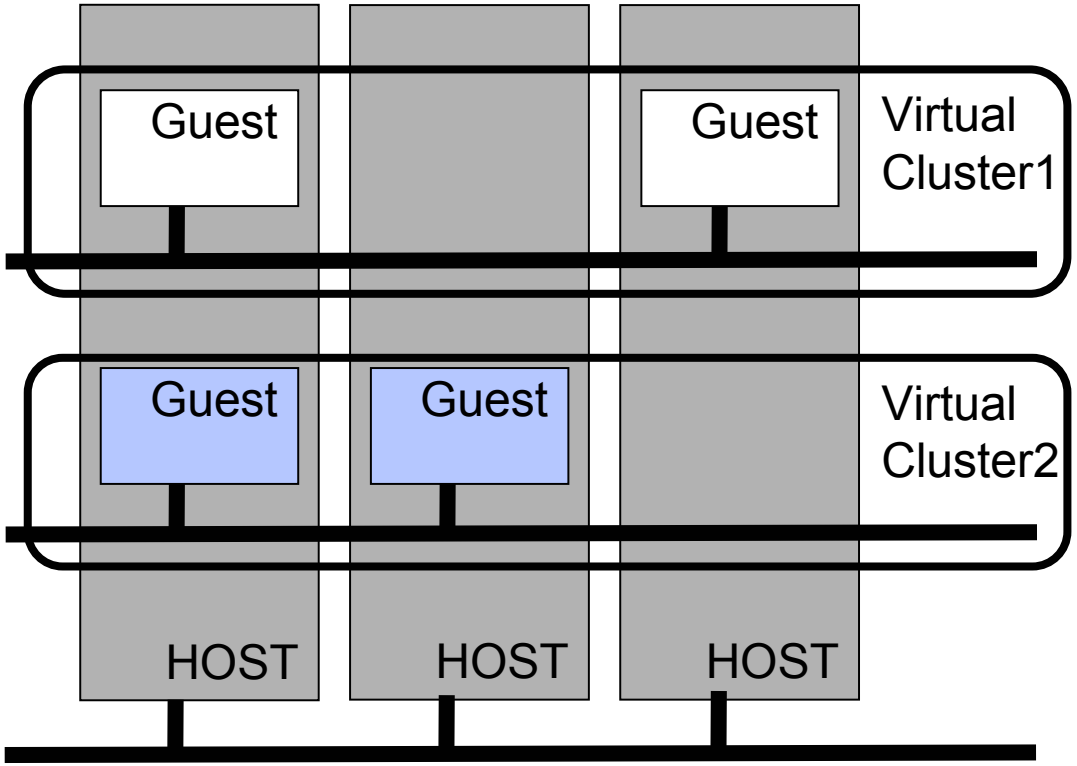


Ideal: VMM does support iSCSI

Real: VMM does not suppot iSCSI

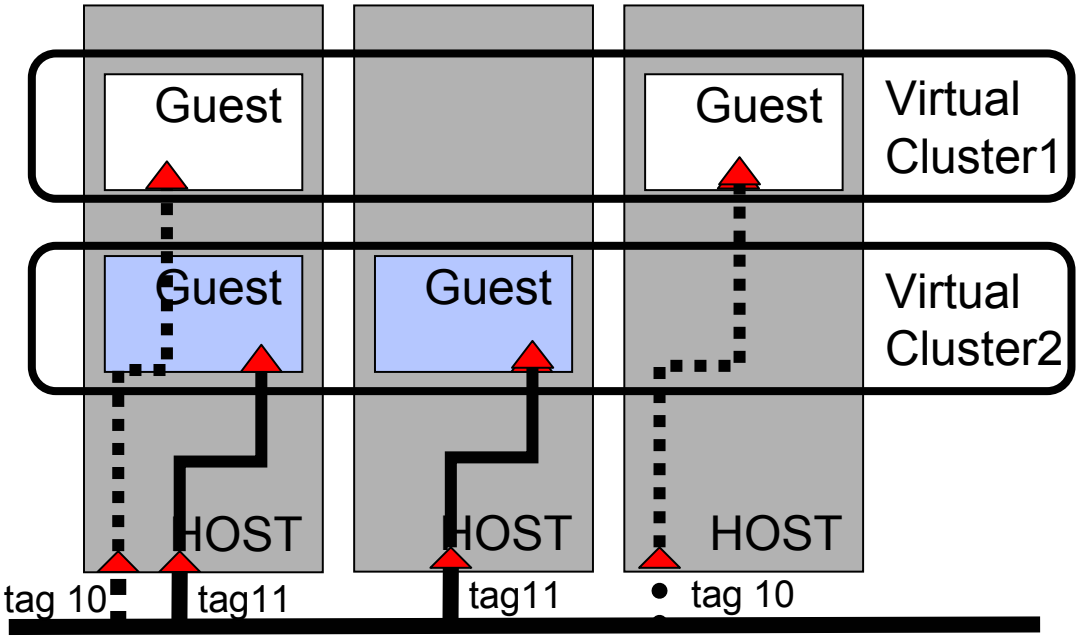
VLAN for separation of virtual clusters

- Each virtual cluster have its own dedicated internal network
- A node in a virtual cluster cannot peek in the network of other virtual clusters.



Separation fo Virtual Cluster with tagged VLAN

- **Host node maps a tagged VLAN with a virtual cluster instance**
  - Host node manages several tagged network interfaces
  - Host node maps one of them to the guest network interface
- **No configuration required within the virtual node**
  - Configuration in virtual nodes could be changed by the user.

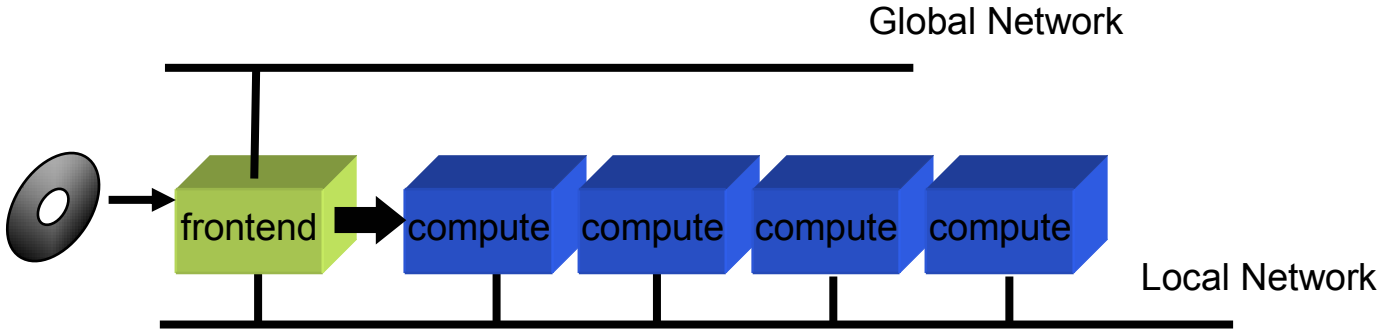


Overview of Rocks

- **Cluster installation system developed by UCSD, as a part of NPACI effort.**
- **Supports Cluster Installation and Cluster Management.**
  - "Roll" defines 'Macro-package' for each application
    - Ex. HPC Roll, Grid Roll
  - "Appliance" defines roles of nodes
    - Ex. Compute Node, Database Node
  - Cluster monitoring by Ganglia
  - User management by 411

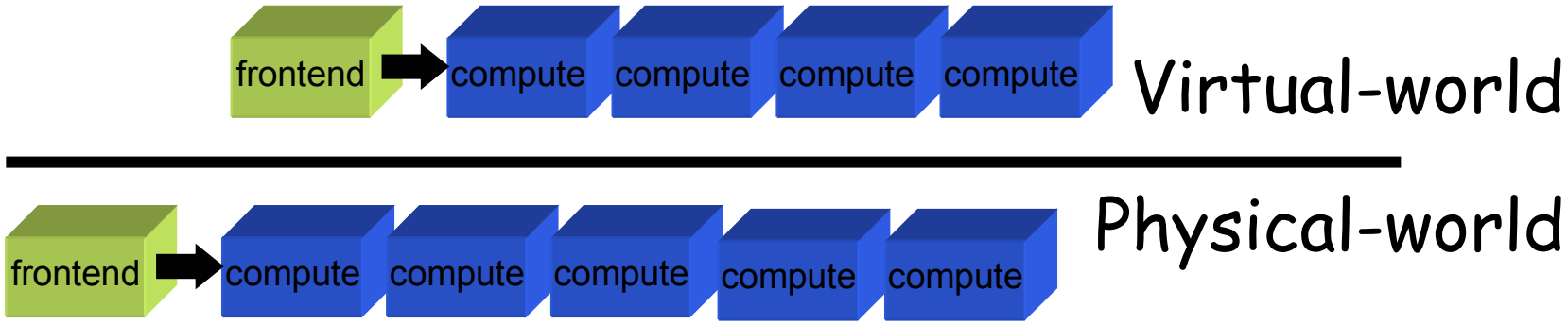
Cluster installation with Rocks

- **Install a front-end from CD (or from central server on network)**
- Power on compute nodes one by one
  - **Each node automatically gets packages from the front-end and installed.**
  - **Node numbers are implicitly determined by the order of power-on**



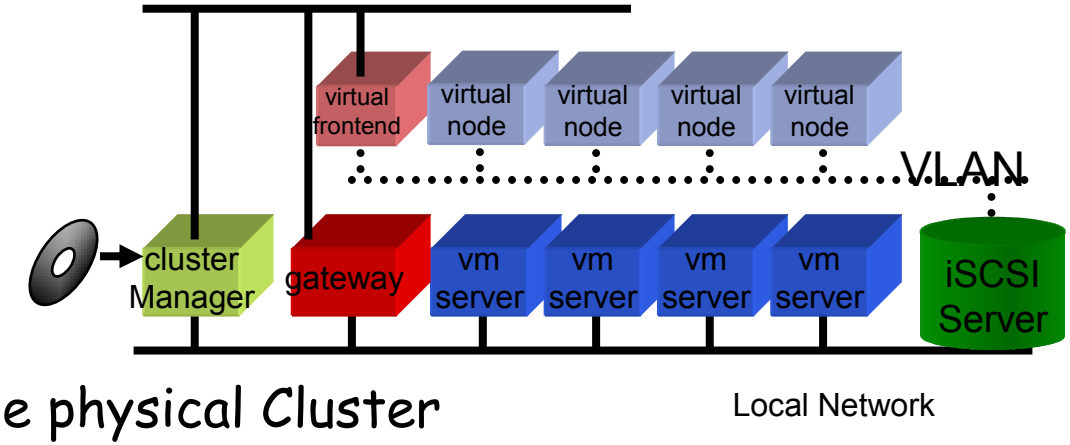
Virtual Cluster and Rocks

- **Install 'virtual front-end' as a virtual node**
  - From the virtual front-end other nodes are installed



- **The physical cluster, including the virtual cluster management system, is also managed by Rocks**
  - Physical cluster management is also easy

Configuration of the proposed virtual cluster



- **Four types of nodes**
  - Cluster Manager
    - Just One for the whole physical Cluster
  - Gateway Nodes
    - Host virtual frontend nodes
    - Have access to the external network
  - VM Server Nodes
    - Hosts virtual compute nodes
  - Storage Nodes
    - Manages disks and provides iSCSI access

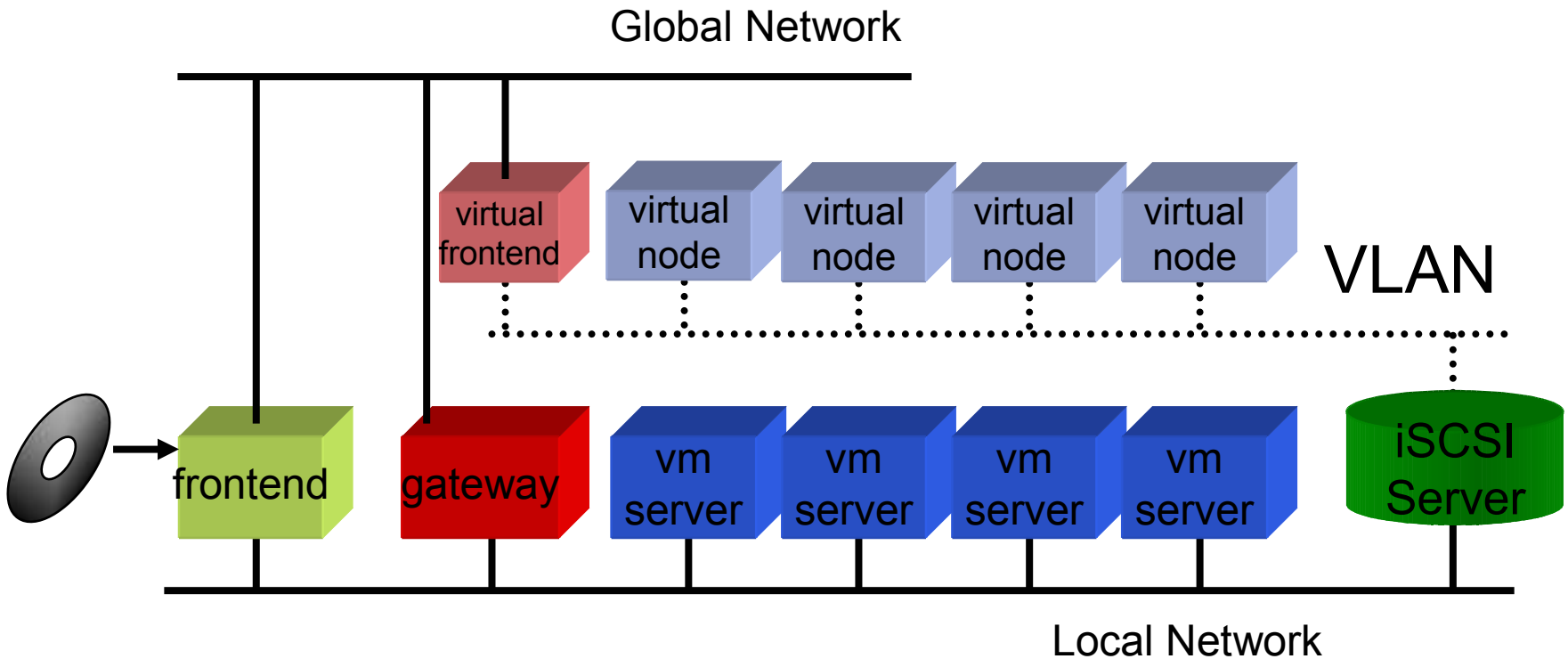
Operation steps

1. **Service Provider makes reservation for a virtual cluster via web based interface**
   - Start time, end time, amount of memory, amount of storage
   - Roll, Appliance
   - ssh public key to access the virtual front-end
2. **On the start-up time**
   - A Virtual cluster will be set up.
     - Storage and VLAN tag are allocated
   - A Rocks Cluster is installed in the virtual world
     - Virtual front-end is installed
     - Virtual-nodes are installed from the virtual front-end

Operation steps (2)

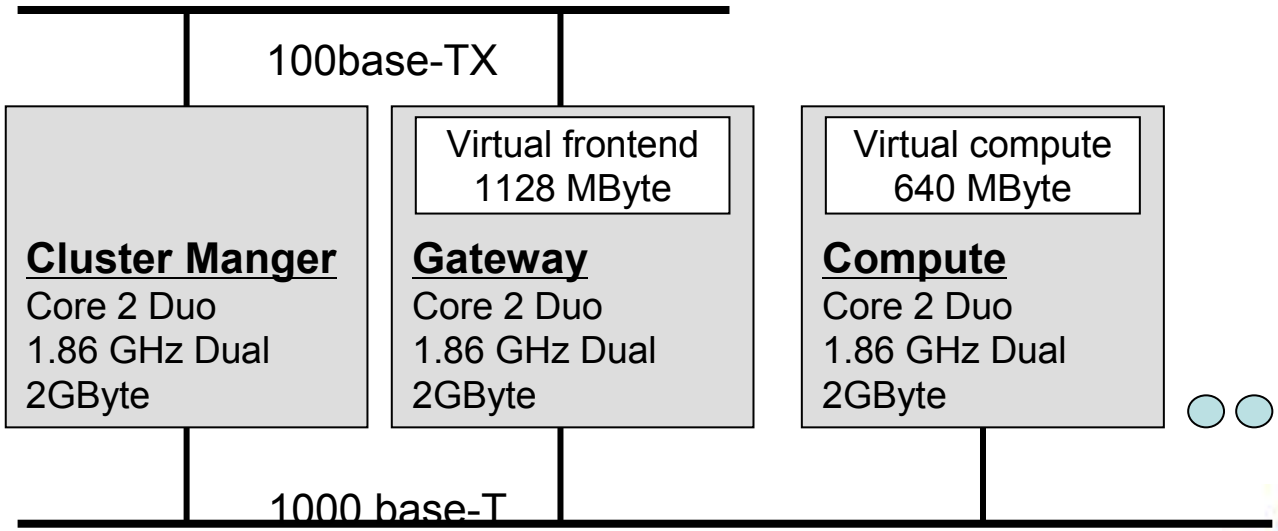
3. **When all the installation finishes,**
   - Pass the control over the virtual cluster to the service provider.
   - The service provider now can log in using the ssh key, and do anything they want.
4. **On reservation end time**
   - Release allocated resources, i.e. storage and virtual computers, and VLAN tag
   - Virtual computers are just shut off

Virtual Cluster Installation

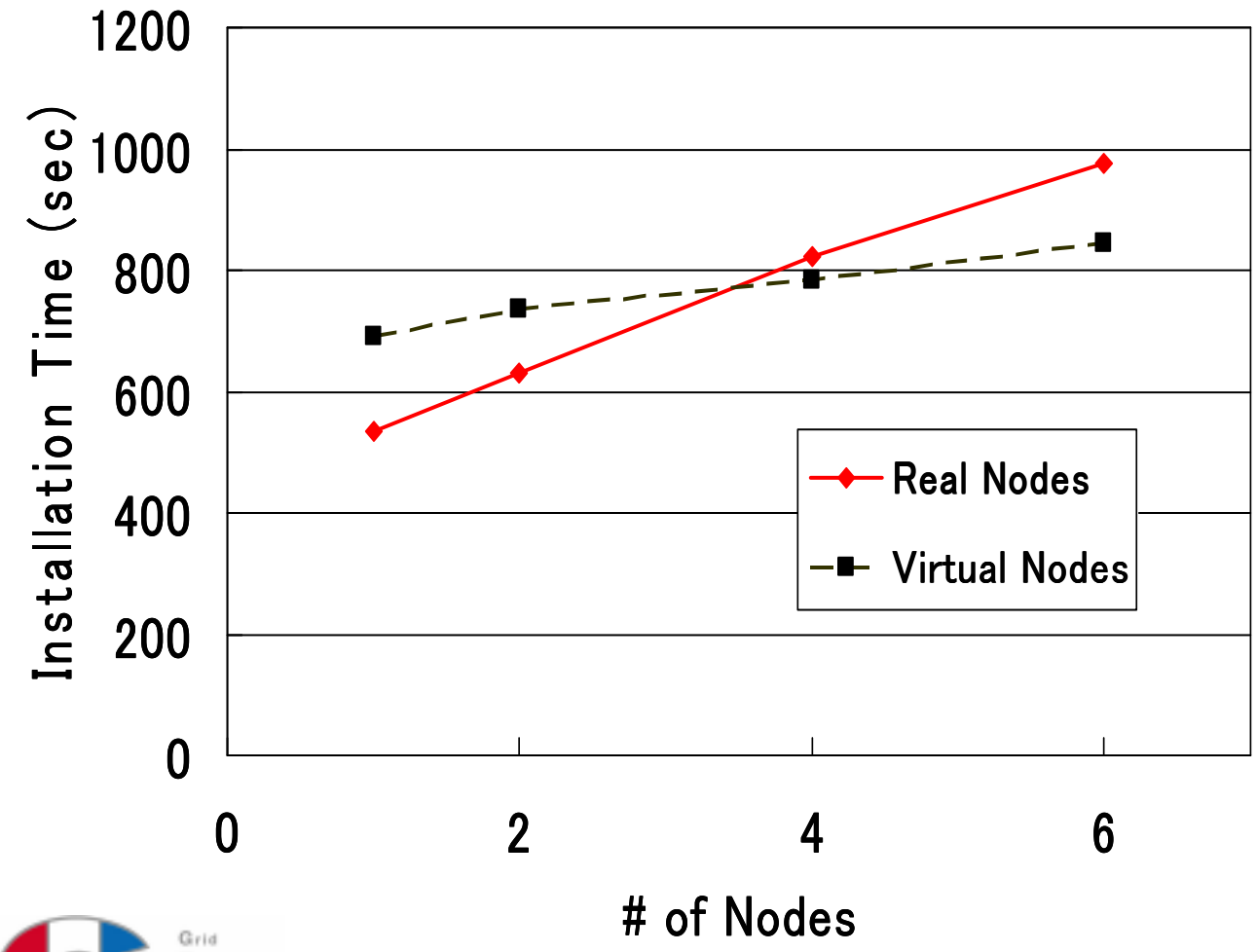


Measurement

- **Measured installation time for clusters**
  - Physical cluster installation
  - Virtual cluster installation
  - For several # of nodes.



Measurement



- Installation time required for virtual cluster is equivalent with physical cluster
- Note: the installed packages are not completely the same

Grid Technology Research Center AIST

AIST

Related work

- **ORE Grid [Nishimura '07]**
  - Leverages Lucie, a cluster installation tool
  - hi speed cluster installation

- **Virtual workspace [Keahey '06]**
  - A part of Globus project
  - Provides Web Service based interface to create a virtualized environment, where users can submit their jobs.
  - Create one virtual node for one job

Related work (2)

- **Xen Cluster with OSCAR [Vallee '06]**
  - OSCAR
    - Cluster deployment tool like Rocks

- **Cisco vFrame**
  - Virtualizes storage and network using Infiniband network , SAN and dedicated swtich.
  - Computers are not virtualized
  - Super expensive.

Summary

- **Proposed a Virtual Cluster Management System**
  - Automatic Virtual cluster deployment and configuration by NPACI Rocks
  - Virtualized computer, storage and network VMware Server
    - iSCSI + LVM
    - VLAN
- **Measured Installation time**
  - Confirmed that the speed is comparable with the real clusters.

Future Work

- **Hide installation cost from service providers**
  - Install virtual nodes in advance
- **Adopt Xen**
  - Rocks4, based on CentOS4 is not compatible with Xen
  - We are waiting for Rocks5, based on CentOS 5
- **Advanced Virtual Storage management**
  - Cluster file system such as Lustre or PVFS for high performance storage
  - No idea how it would work with iSCSI, though
- **Other Operating System / Distributions as Guest**
  - Windows CCS?
- **Implement external interface for cluster reservation**
  - WSRF based ?
  - Waiting for 'standard'...

Future work (2)

- **One virtual cluster over several physical clusters**
  - Provides large virtual clusters with Single System Image
  - Using VPN
  - A demo will be shown at SC'07, Reno

Physical Cluster

Physical Cluster

Acknowledgement

- **We'd like to thank SDSC Rocks team including**
  - Mason Kats
  - Greg Bruno
  - Anoop Rajendra