



Delivering IaaS for the Greek  
Academic and Research Community



Vangelis Koukis  
vkoukis@cslab.ece.ntua.gr  
Technical Coordinator, ~okeanos Project

# Outline

- ◆ ~oceanos IaaS
- ◆ Compute
- ◆ Synnefo architecture
- ◆ Network
- ◆ Storage
- ◆ Upcoming goals



# Outline

- ◆ ~okeanos IaaS
- ◆ Compute
- ◆ Synnefo architecture
- ◆ Network
- ◆ Storage
- ◆ Upcoming goals



# Motivation

- ◆ Deliver IaaS to GRNET's customers
  - ➔ direct: IT depts of connected institutions
  - ➔ indirect: university students, researchers in academia
  
- ◆ Other IaaS efforts
  - ➔ Amazon EC2 not an end-user service
  - ➔ Need to develop custom UI, AAI layers
  - ➔ Vendor lock-in
  - ➔ Unsuitable for IT depts
    - persistent, long-term servers, custom networking requirements
  - ➔ Gain know-how, build on own IaaS → new services



## ~okeanos IaaS

### ◆ Infrastructure...

- ➔ Compute (Virtual Machines)
- ➔ Network (Virtual Networks)
- ➔ Storage (Virtual Disks)

### ◆ ... as a Service

### ◆ Users manage resources over

- ➔ a simple, elegant UI, or
- ➔ a REST API, for full programmatic control



## ~okeanos Project

- ◆ Goal: Production quality IaaS
  - working Alpha in coming month
- ◆ A jigsaw puzzle of many pieces
- ◆ Presentation focuses on software infrastructure
- ◆ Synnefo
  - custom cloud management software to power ~okeanos
  - Google Ganeti backend
- ◆ Current & future goals for Compute, Network, Storage





# Outline

- ◆ ~okeanos IaaS
- ◆ **Compute**
- ◆ Synnefo architecture
- ◆ Network
- ◆ Storage
- ◆ Upcoming goals



# IaaS – Compute (1)

## ◆ Virtual Machines

➔ powered by KVM

- Linux and Windows guests, on Debian hosts

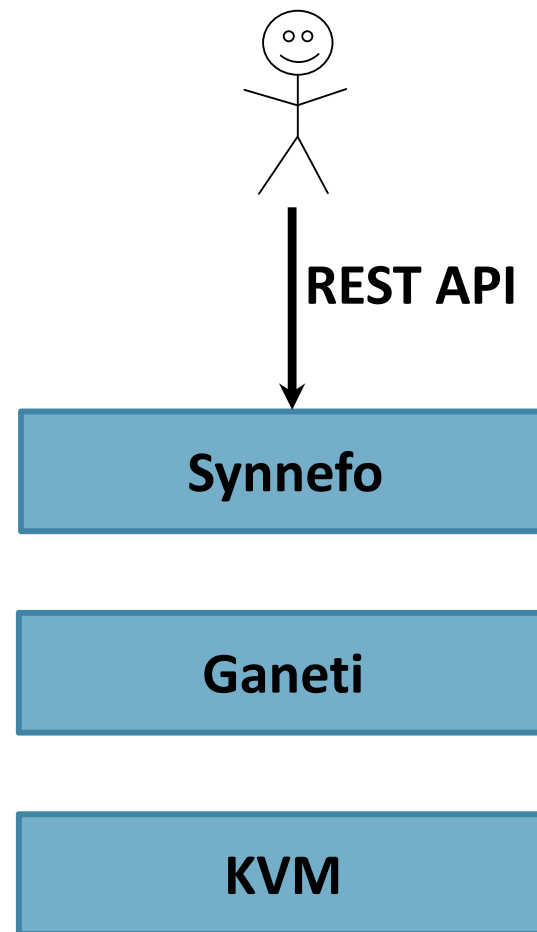
➔ Google Ganeti for VM cluster management

➔ accessible by the end-user over the Web or programmatically (OpenStack Compute v1.1)

◆ Initial target is longer-term, persistent VMs (students, University IT depts)

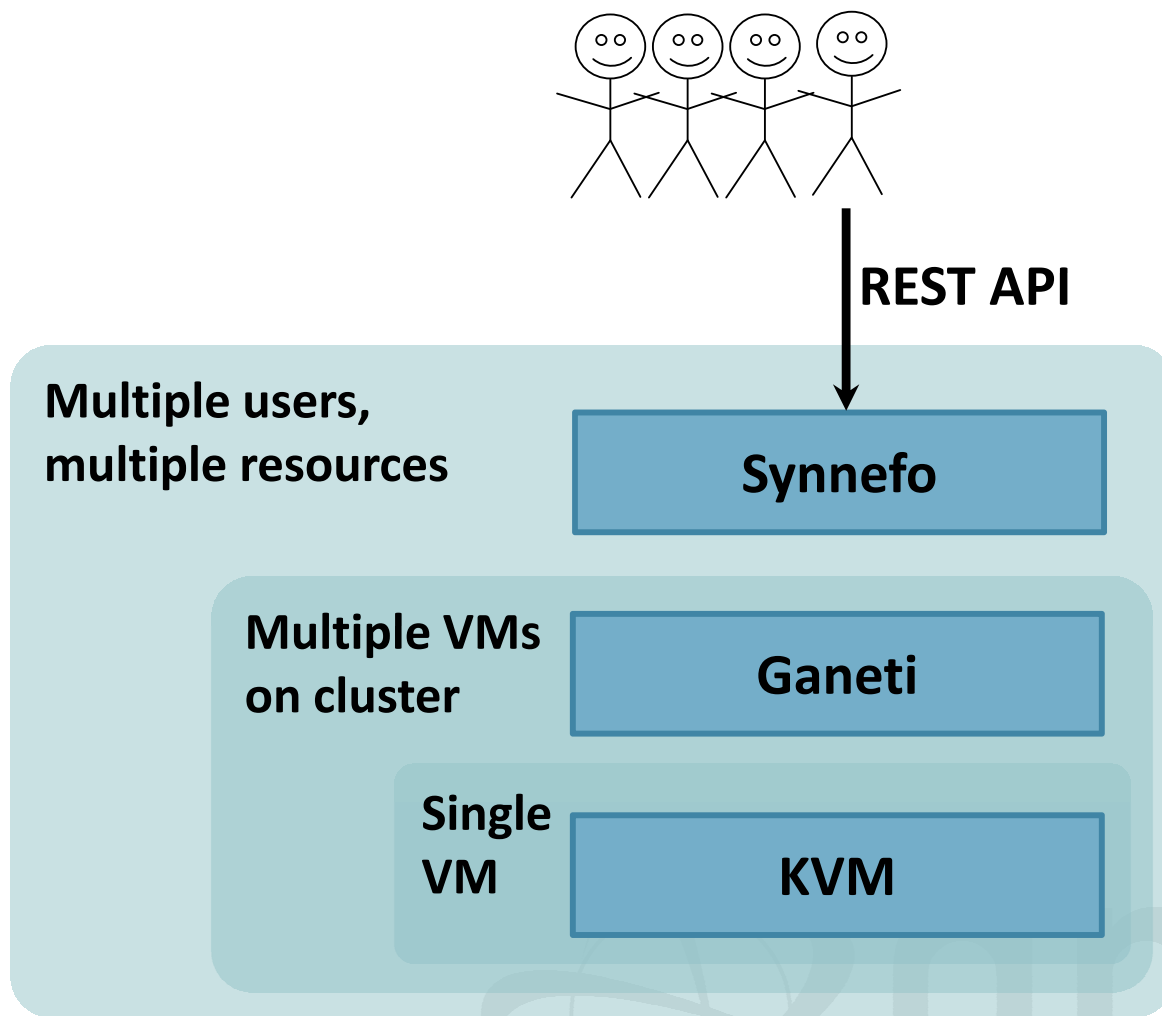


# Software Stack



grnet

# Software Stack



glnet

## IaaS – Compute (2)

### ◆ User has full control over own VMs

#### ➔ Create

- Select # CPUs, RAM, System Disk
- OS selection from pre-defined Images
- popular Linux distros (Fedora, Debian, Ubuntu)
- Windows Server 2008 R2

#### ➔ Start, Shutdown, Reboot, Destroy

#### ➔ Out-of-Band console over VNC for troubleshooting



## IaaS – Compute (3)

- ◆ REST API for VM management
  - ➔ OpenStack Compute v1.1 compatible
  - ➔ 3rd party tools and client libraries
  - ➔ custom extensions for yet-unsupported functionality
  - ➔ Python & Django implementation
- ◆ Full-featured UI in JS/jQuery
  - ➔ UI is just another API client
  - ➔ All UI operations happen over the API



# Why Ganeti?

- ◆ No need to reinvent the wheel
- ◆ Scalable, proven software infrastructure
  - ➔ Built with reliability and redundancy in mind
  - ➔ Combines open components (KVM, LVM, DRBD)
  - ➔ Well-maintained, readable code
- ◆ VM cluster management in production is serious business
  - ➔ reliable VM control, VM migrations, resource allocation
  - ➔ handling node downtime, software upgrades



## Why Ganeti?

- ◆ GRNET already has long experience with Ganeti
  - ➔ provides ~280 VMs to NOCs through ViMa service
  - ➔ involved in development, contributing patches upstream
  
- ◆ Build on existing know-how for ~okeanos
  - ➔ Common backend, common fixes
  - ➔ reuse of experience and operational procedures
  - ➔ simplified, less error-prone deployment



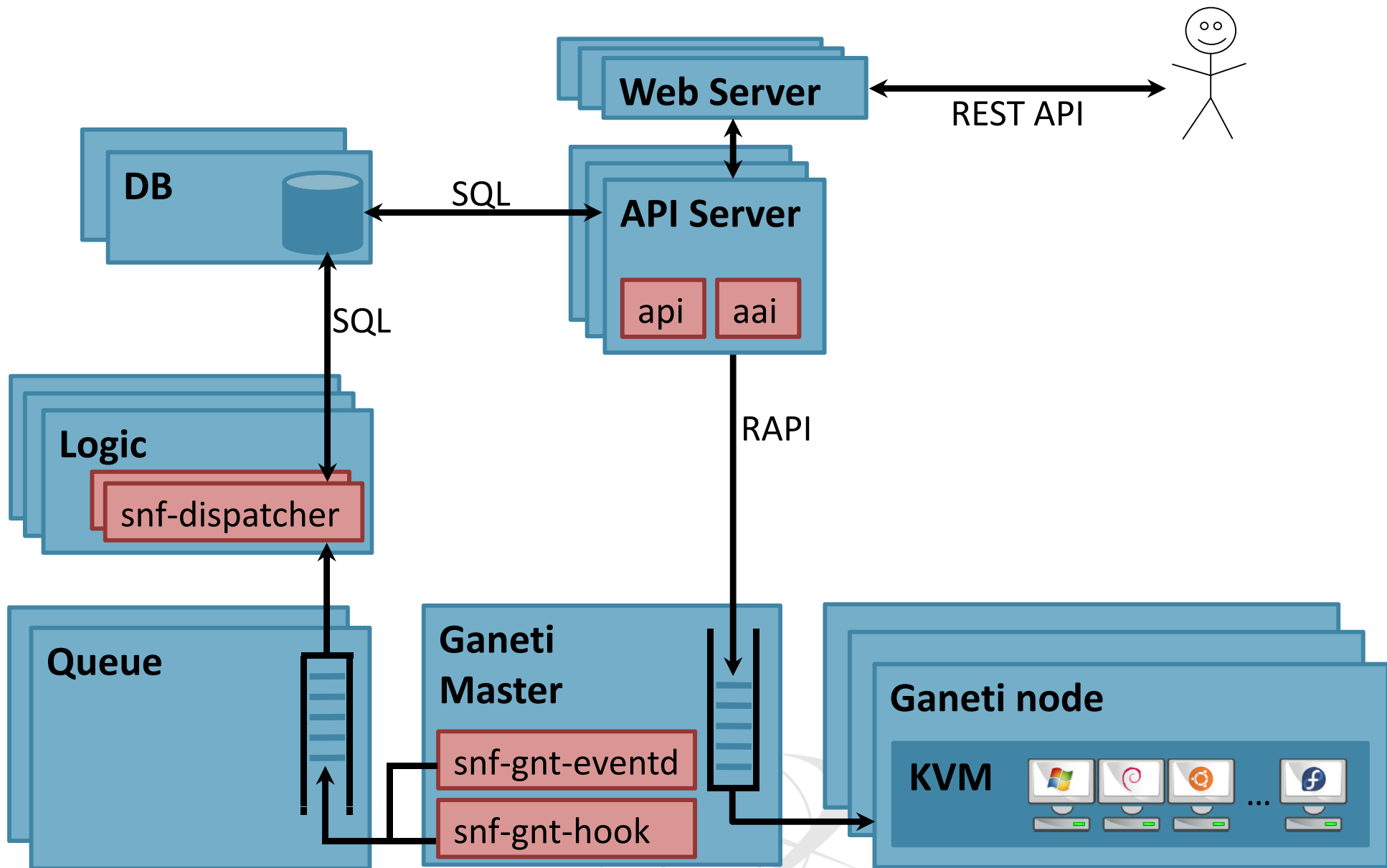


# Outline

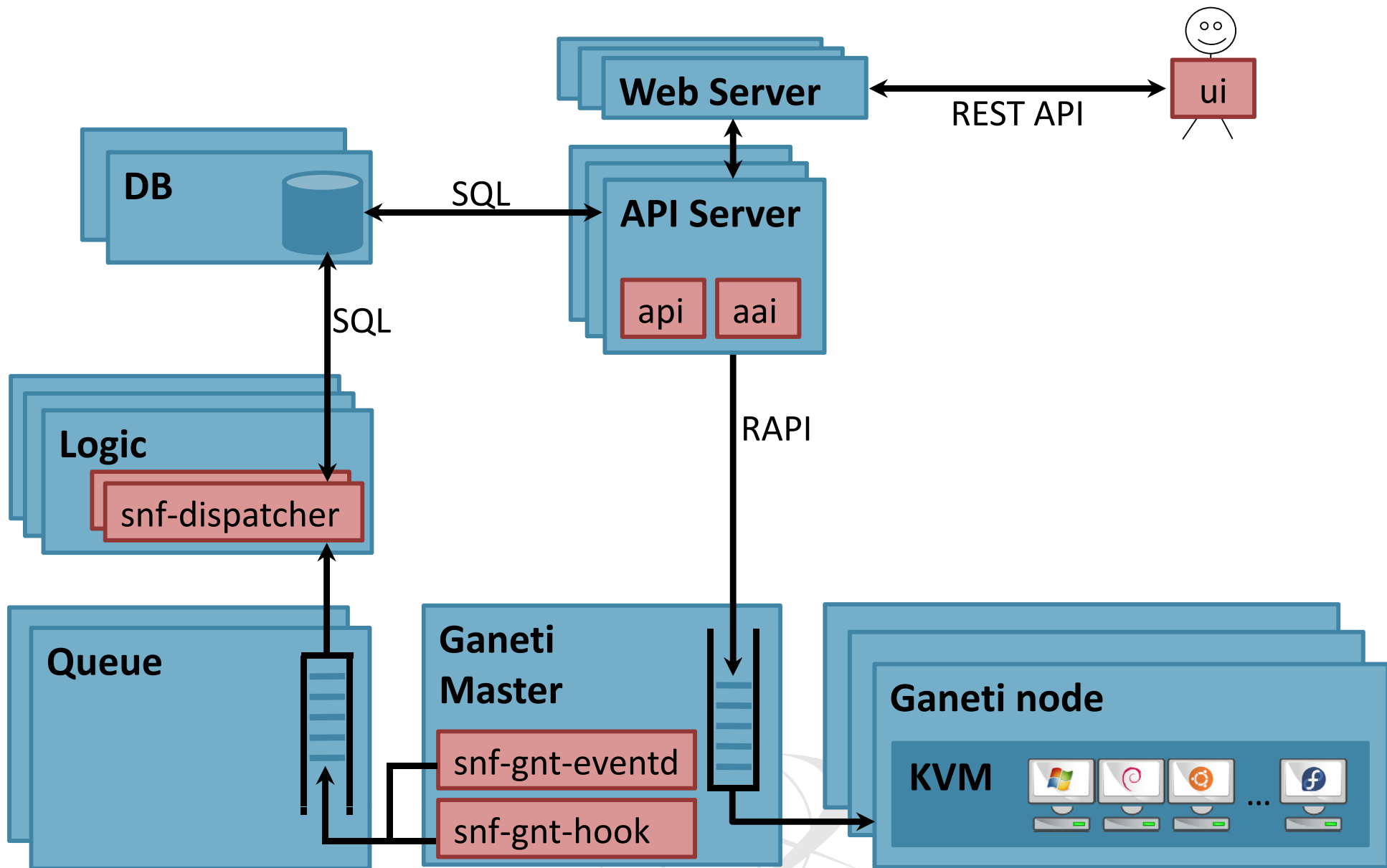
- ◆ ~okeanos IaaS
- ◆ Compute
- ◆ Synnefo architecture
- ◆ Network
- ◆ Storage
- ◆ Upcoming goals



# Synnefo deployment



# Synnefo deployment

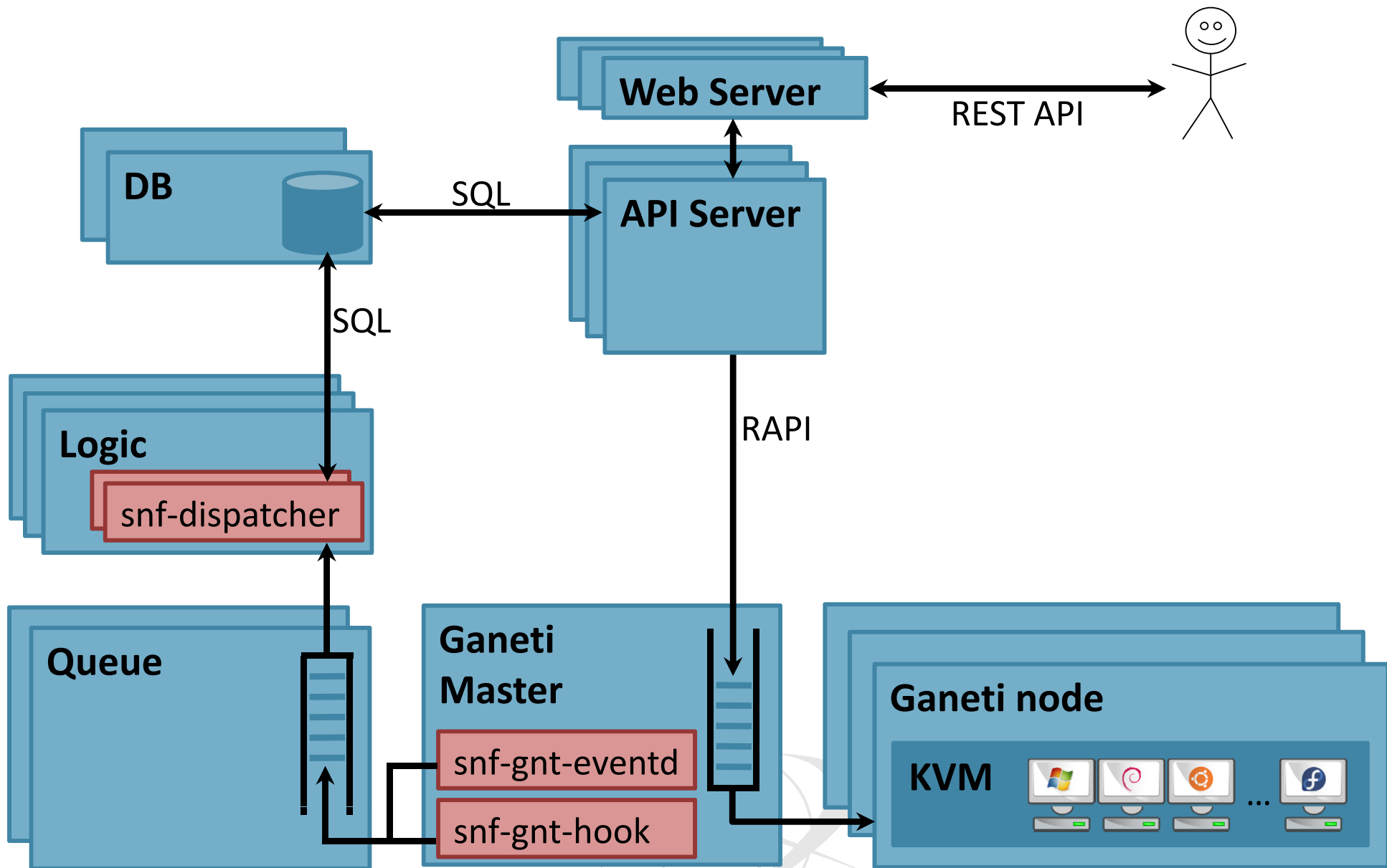


# Asynchronous design

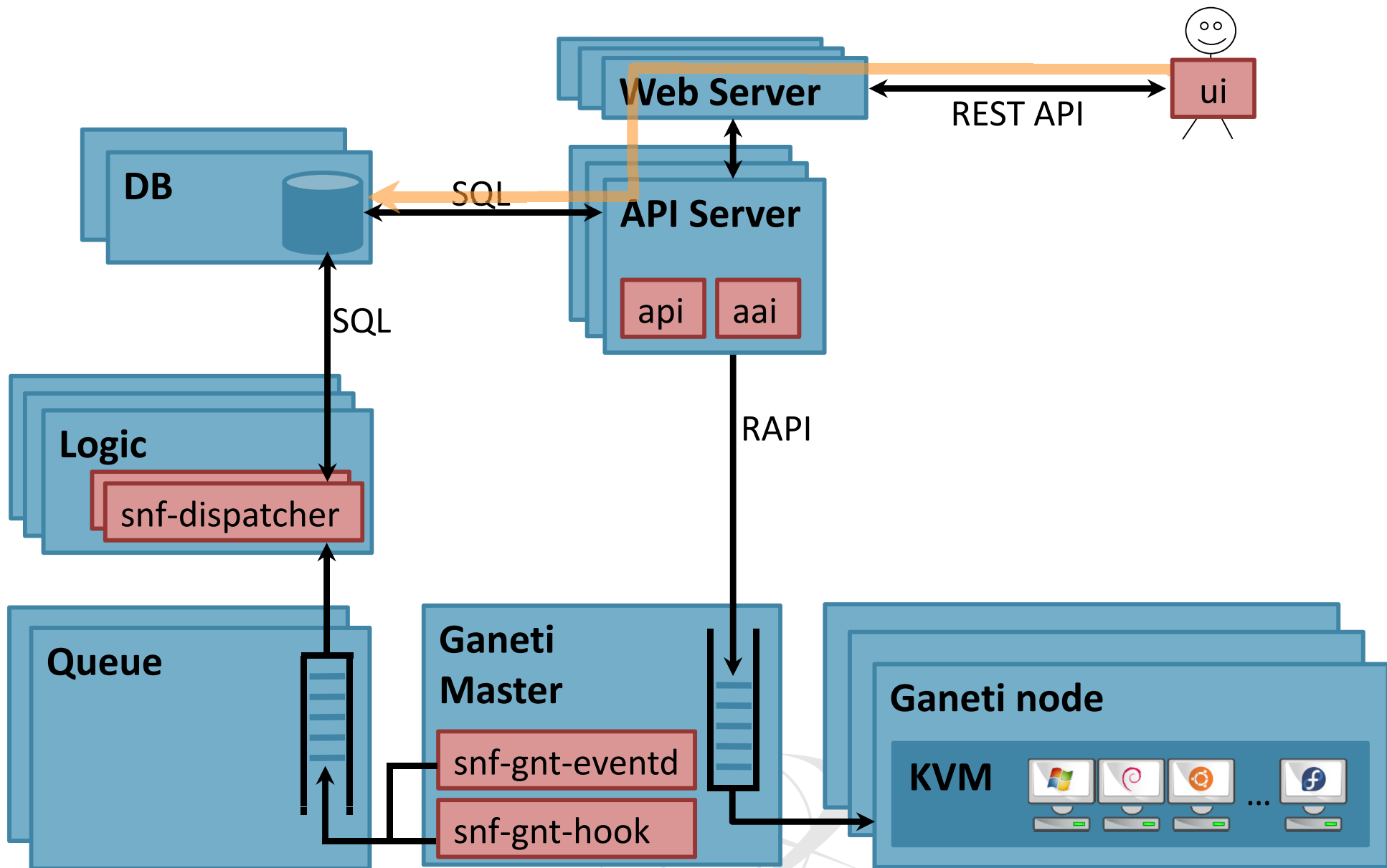
- ◆ DB contains All state needed to handle API queries
  - ➔ no need to reach the backend
  - ➔ Ganeti GetInstanceInfo() is a proper job, too slow
- ◆ Two distinct paths, *effect* and *update*
- ◆ *Effect* changes to VMs
  - ➔ when servicing API requests to modify VM state
  - ➔ issue commands to Ganeti backend, over RAPI
  - ➔ ACK reception of request to user
- ◆ *Update* DB, when interesting things happen
  - ➔ user or *admin* initiated
  - ➔ Queue notifications to Message Queue, over AMQP



# Synnefo deployment



# Synnefo deployment

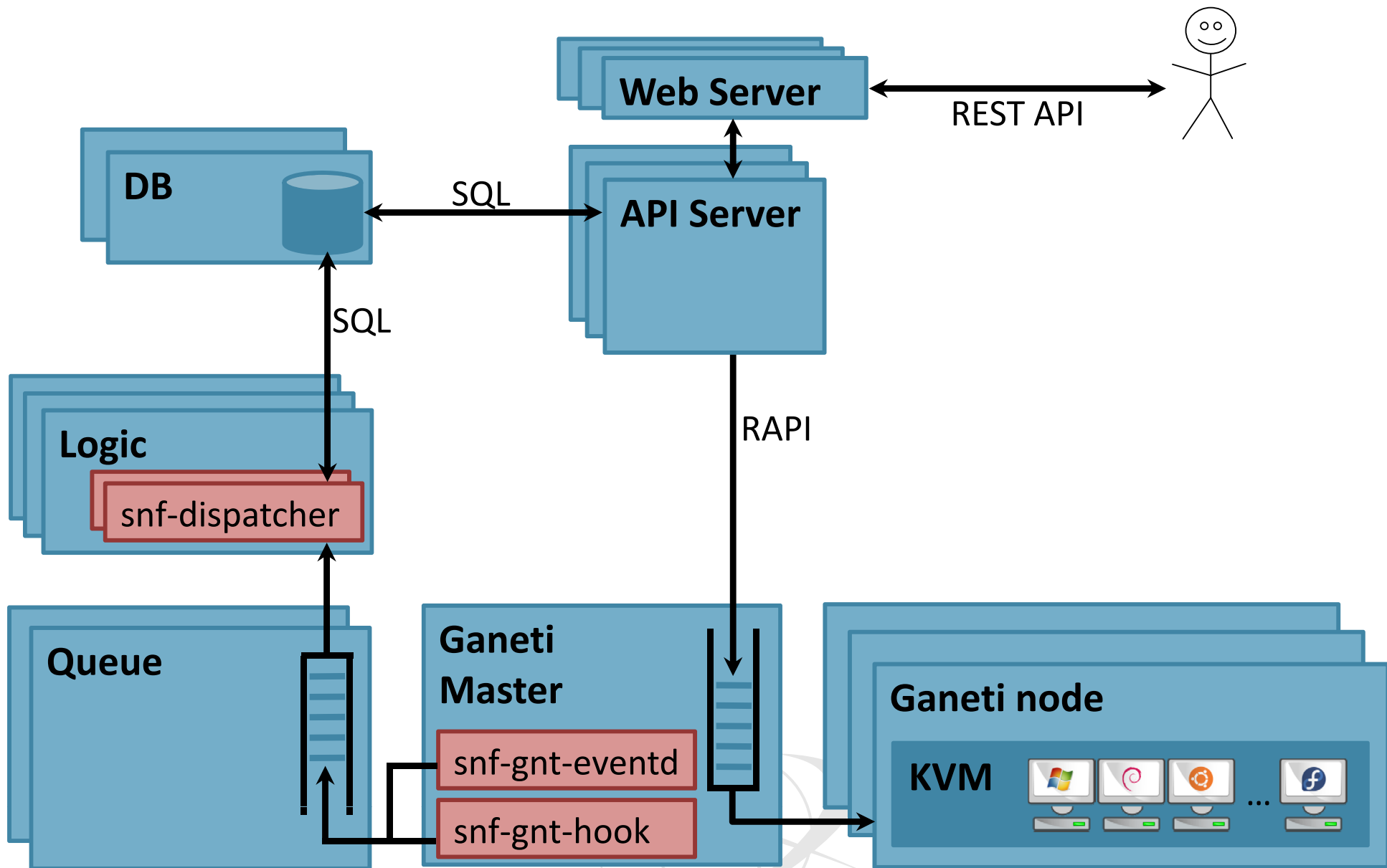


## The “effect” Path

- ◆ Reception of API request to modify VM state (e.g.,  
PUT /servers over HTTP)
- ◆ API enforces access rights and policy
  - ➔ Ganeti knows no cloud users or access rights
- ◆ Need to translate from Openstack Compute to backend ops (e.g., CreateInstance())
- ◆ Asynchronous request processing
  - ➔ Return HTTP 202 Accepted
  - ➔ it's up to the API client to poll for completion

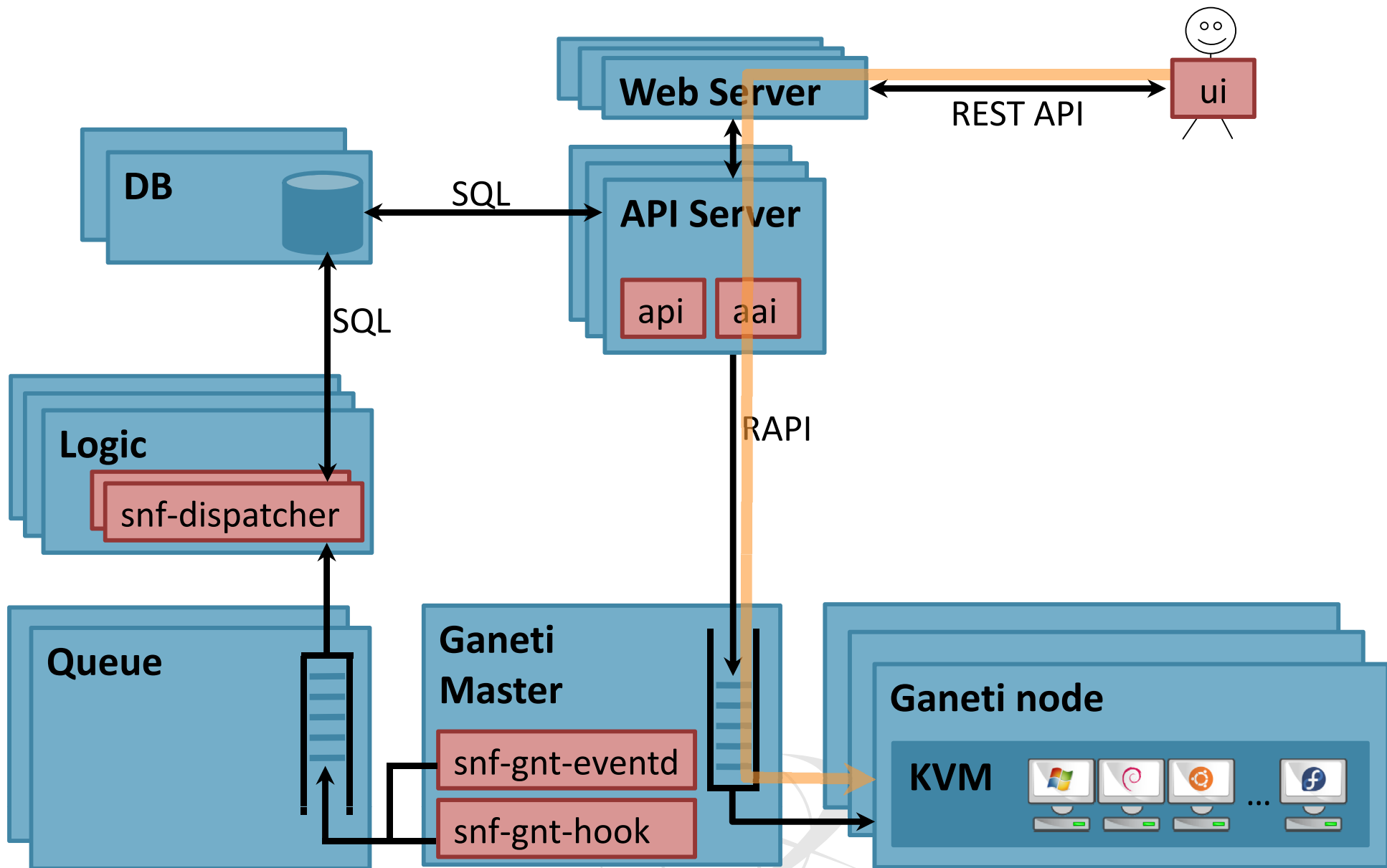


# Synnefo deployment





# Synnefo deployment

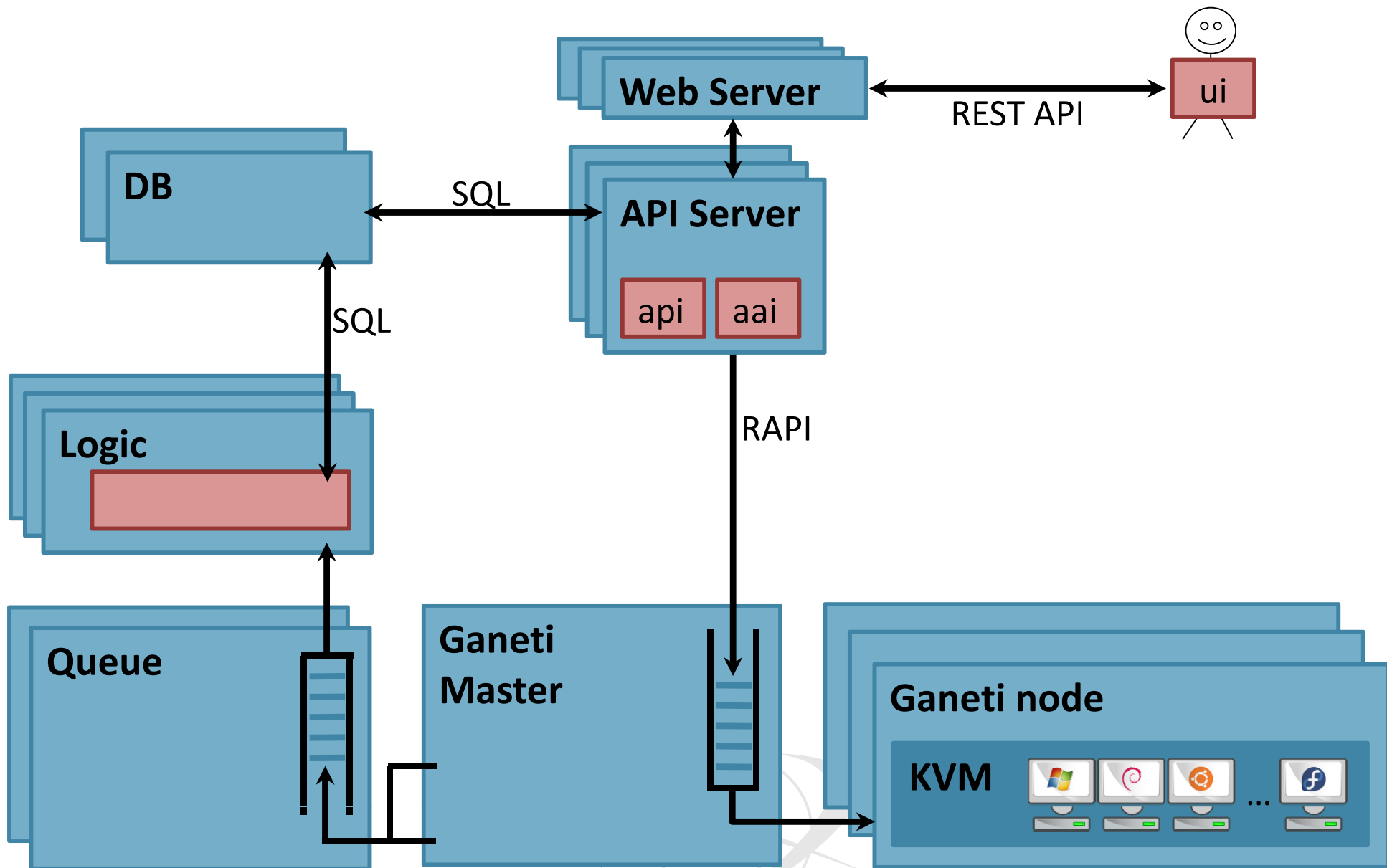


## The “update” path

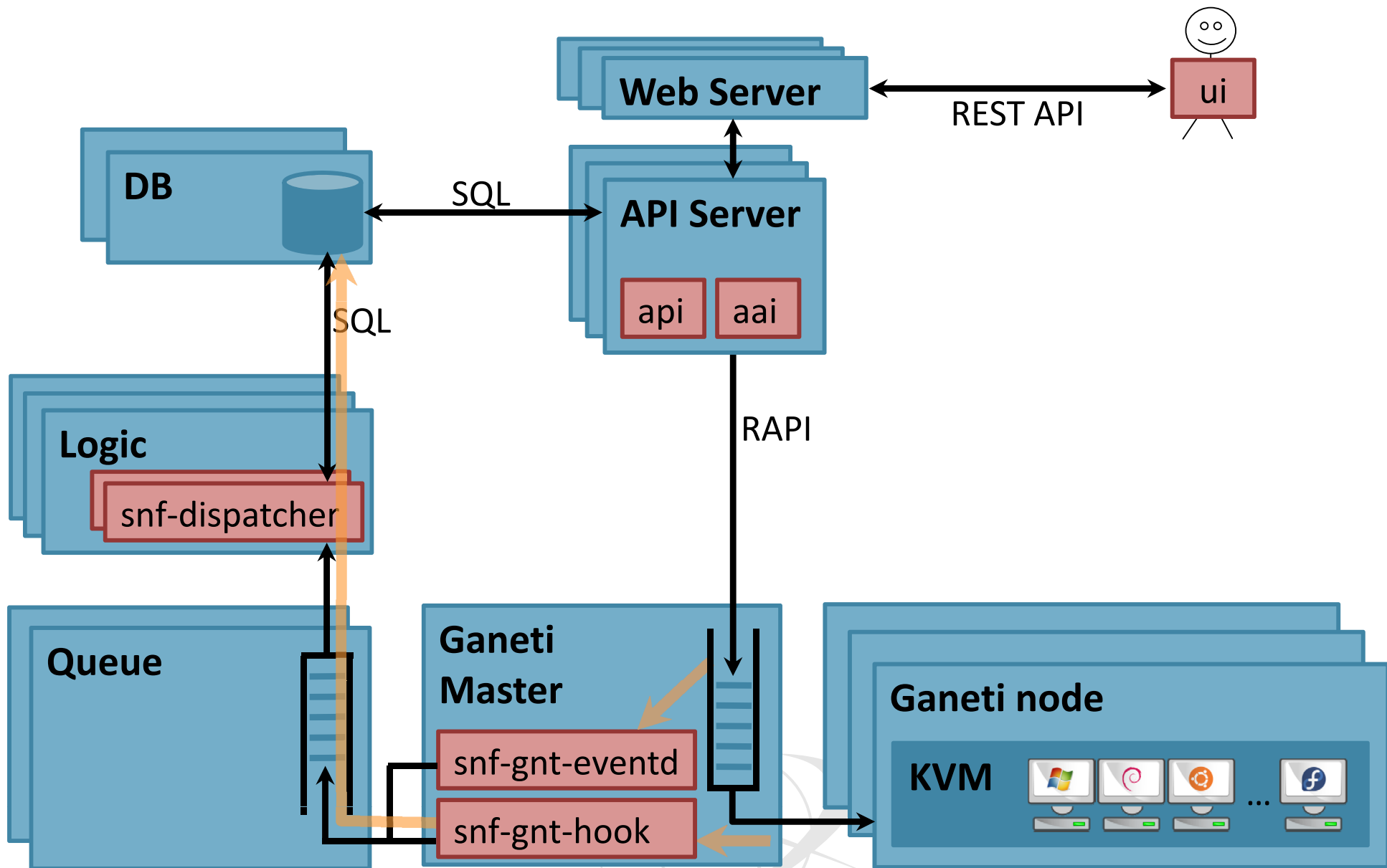
- ◆ May run at any time
- ◆ Completely decoupled from “effect” path
- ◆ Design goal:
  - ➔ Ganeti admins free to bypass frontend
  - ➔ Synnefo adapts
- ◆ Synnefo logic triggered on backend events
  - ➔ Ganeti operation progressing in the queue
  - ➔ Synnefo hook running inside Ganeti
    - Hooks run at various phases in a VM’s lifecycle



# Synnefo deployment

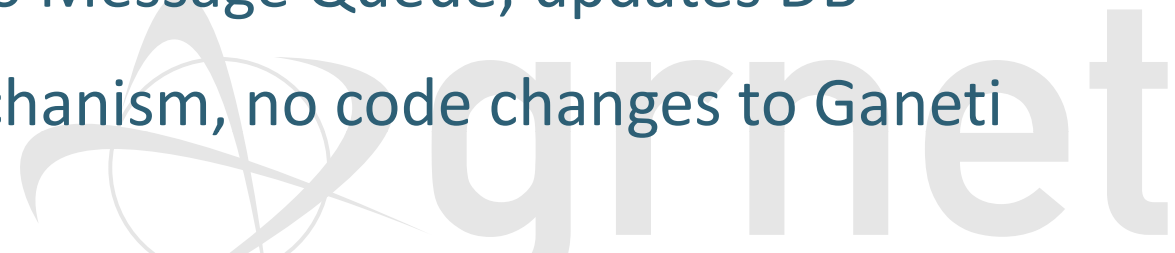


# Synnefo deployment



# The Ganeti event daemon

- ◆ Ganeti master manages job queue
  - ➔ Jobs pass *Queued, Waiting, Running*, end up in *Canceled, Success, Error*.
- ◆ Need a way for Synnefo to monitor job progress
- ◆ Synnefo-specific solution: Ganeti event daemon
  - ➔ Passively monitor the Ganeti job queue
  - ➔ Notifications over AMQP on job progress
  - ➔ Synnefo logic listens to Message Queue, updates DB
  - ➔ `inotify()`-based mechanism, no code changes to Ganeti

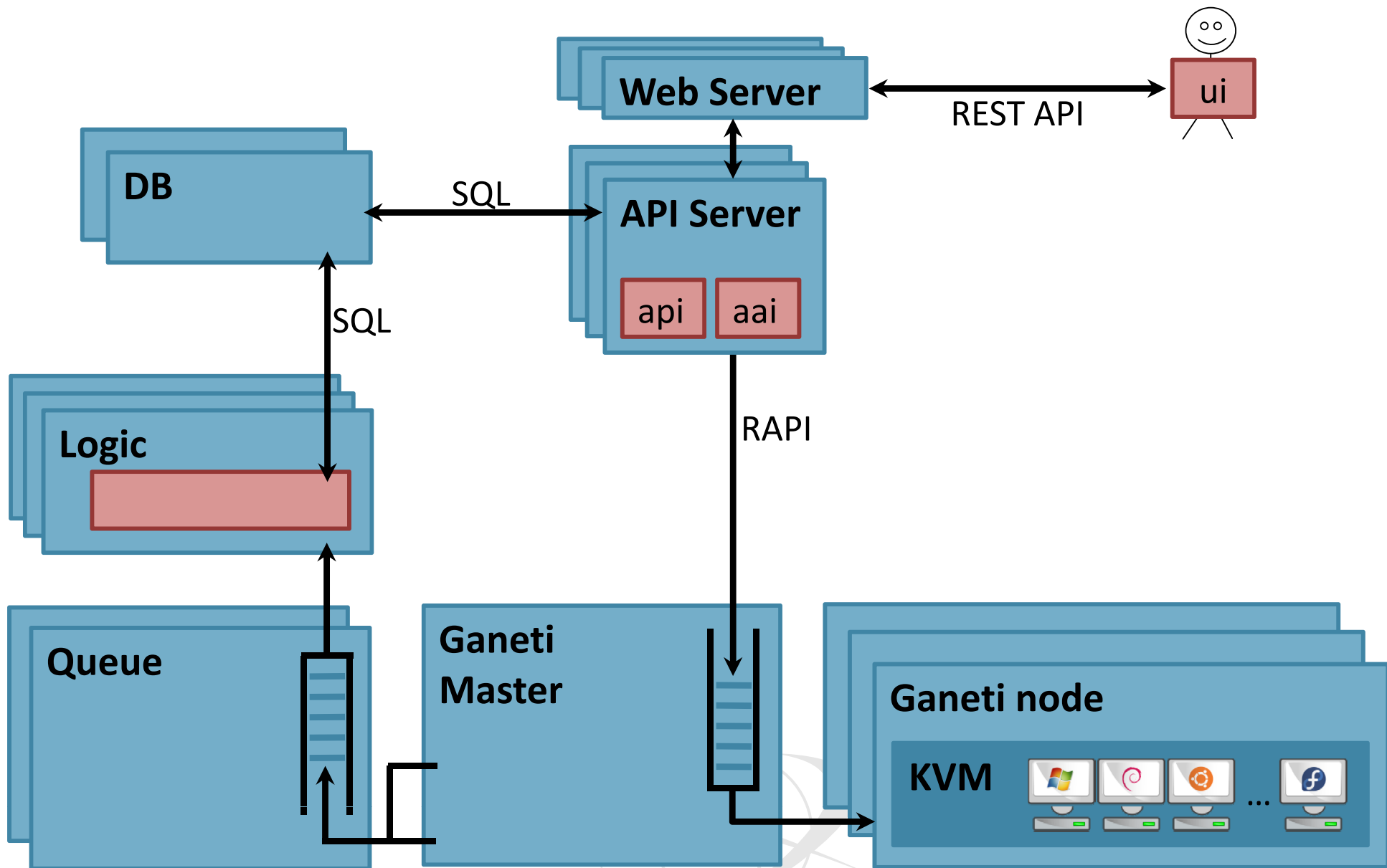


# The Synnefo hook in Ganeti

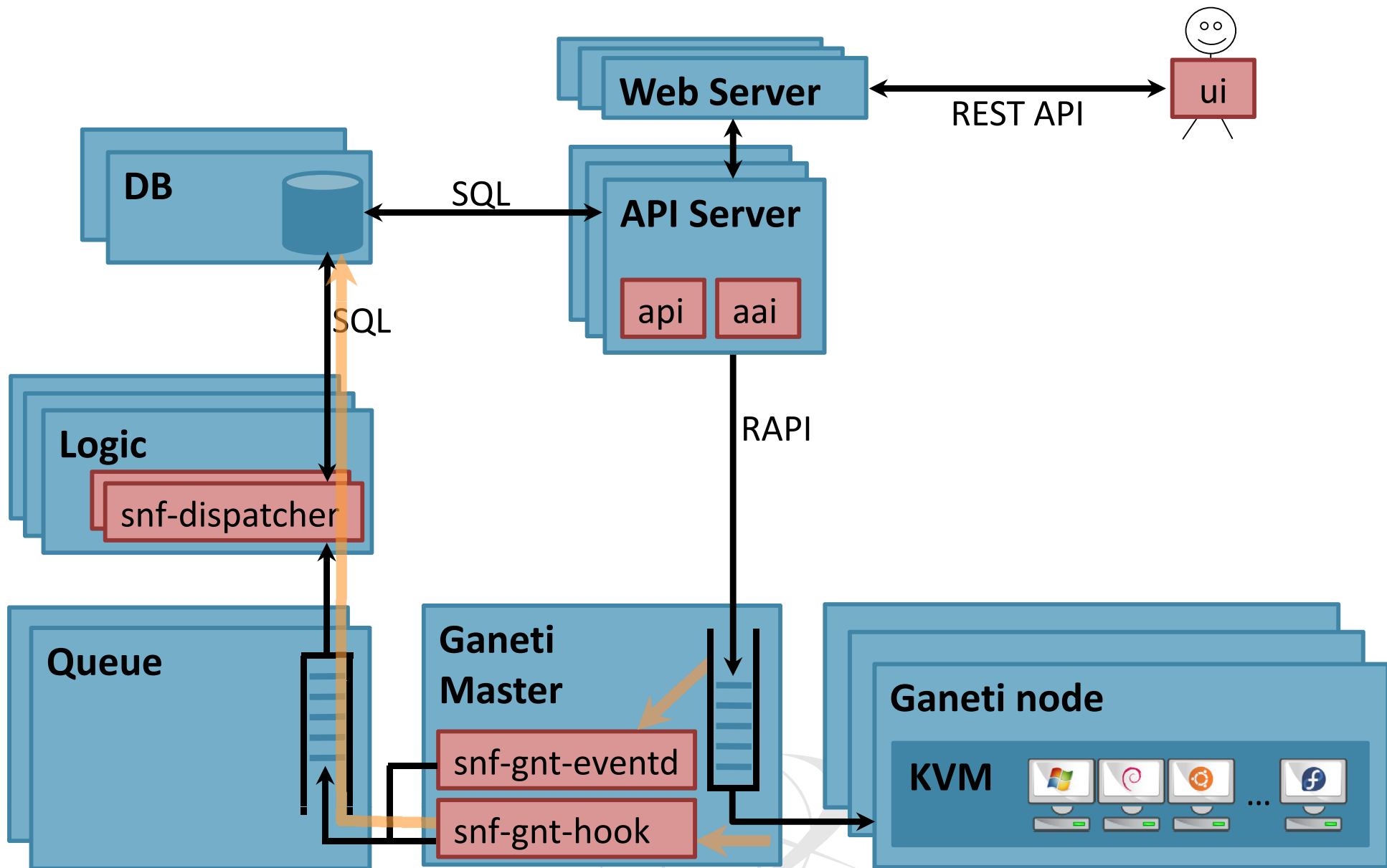
- ◆ Different phases in a VM's lifecycle
  - ➔ {pre, post} – {add, start, stop, reboot, modify}
- ◆ Run Synnefo-specific hook in `post-*`
- ◆ Pushes VM configuration notifications to MQ
  - ➔ e.g., NIC setup



# Synnefo deployment



# Synnefo deployment





## Reconciliation with Ganeti

- ◆ What if the MQ is down, and messages are lost?
  - ➔ Ganeti is the Single Source of Truth for VM state
- ◆ Reconcile DB state asynchronously
  - ➔ On success notification for a Ganeti `GetInstanceInfo()` op
- ◆ Triggered periodically, e.g., using `cron`
  - ➔ or even by the administrator,  
running `gnt-instance info` manually



# Outline

- ◆ ~okeanos IaaS
- ◆ Compute
- ◆ Synnefo architecture
- ◆ **Network**
- ◆ Storage
- ◆ Upcoming goals



## IaaS – Network - Functionality

- ◆ Dual IPv4/IPv6 connectivity for each VM
- ◆ Easy, platform-provided firewalling
  - ➔ Array of pre-configured firewall profiles
  - ➔ Or roll-your-own firewall inside VM
- ◆ Multiple private, virtual L2 networks
- ◆ Construct arbitrary network topologies
  - ➔ e.g., deploy VMs in multi-tier configurations
- ◆ Exported all the way to the API and the UI

grnet

# IaaS – Network - Implementation

- ◆ Custom modifications to Ganeti
  - ➔ IP pool management for the public network
- ◆ Custom-written DHCP server over NFQUEUE
- ◆ Custom interface handling scripts
  - ➔ Enforce VM networking configuration
- ◆ Private Networks
  - ➔ Alpha: pre-provisioned bridges to 802.1Q VLANs
  - ➔ Later on: MAC-prefix based filtering



# Outline

- ◆ ~okeanos IaaS
- ◆ Compute
- ◆ Synnefo architecture
- ◆ Network
- ◆ **Storage**
- ◆ Upcoming goals



## IaaS – Storage (1)

### ◆ First-phase deployment

- ➔ Ability to customize VM contents based on predefined images of common OSs
- ➔ Redundant storage based on DRBD, VMs survive scheduled node downtime

### ◆ Currently under development:

- ➔ Reliable distributed storage over RADOS
- ➔ Combined with custom software for snapshotting, cloning to provide dynamic virtual storage volumes



## IaaS – Storage (2)

- ◆ Multi-tier storage architecture
  - ➔ Dedicated Storage Nodes (SSD, SAS, and SATA storage)
  - ➔ OSDs for RADOS
- ◆ Custom storage layer
  - ➔ manages snapshots, creates clones over RADOS
  - ➔ OS Images held as snapshots
- ◆ VMs created as clones of snapshots



# Interaction with other GRNET services

## ◆ GRNET AAI Federation

- ➔ Provides the user base for ~oceanos
- ➔ Once authenticated, the user retrieves a Synnefo-specific auth token for programmatic access

## ◆ Pithos storage service

- ➔ Currently being overhauled
- ➔ Aim is to provide the Image service for ~oceanos
- ➔ Sharing a common storage backend





# Outline

- ◆ ~okeanos IaaS
- ◆ Compute
- ◆ Synnefo architecture
- ◆ Network
- ◆ Storage
- ◆ **Upcoming goals**



## Upcoming goals

- ◆ Credit-based resource allocation
- ◆ Abstract away the Ganeti backend, replace with backend connector behind the MQ
  - ➔ Release to community as reference implementation of OpenStack Compute v1.1
- ◆ Support live modification of VMs in Ganeti
- ◆ Snapshots, clones in storage layer
  - ➔ Dramatic decrease in VM initialization time
  - ➔ Support workloads with 100s of ephemeral VMs
    - e.g. for scientific computation, MPI jobs

machines



Create New +



Name: debian-1  
IP: 192.168.32.7

Show: disks | networks | more info

Running



Name: win1  
IP: 192.168.32.13

Show: disks | networks | more info

Running




Name: ubuntu-srv1  
IP: 192.168.32.21

Show: disks | networks | more info

Running



### Manage Tags

Create, edit and delete Tags for machine:  debian-1

Role	Web server	<input checked="" type="checkbox"/>	<input type="checkbox"/>
OS	debian		
Location	office		

## Create New



Name:

IP:

Show:

CPU: 1  
RAM: 4096 (MB)  
System Disk: 40 (GB)

Image: Debian base  
Image Size: 2000 (MB)



Name:

IP: 192.168.32.13

Show: disks | networks | more info



Name: ubuntu-srv1

IP: 192.168.32.21

Show: disks | networks | more info

Running



machines



Create New +



Name: debian-1  
IP: 192.168.32.7

Running  
■ ■ ■ ■

Reboot  
Shutdown  
Console  
Destroy

Confirm

Cancel

Show: disks | networks | more info

CPU: 1  
RAM: 4096 (MB)  
System Disk: 40 (GB)



Labels (4)  
debian  
my\_servers  
critical

Metadata: OS (2)  
Location

Image: Debian base  
Image Size: 2000 (GB)

details

Manage Tags



Name: win1  
IP: 192.168.32.13

Running  
■ ■ ■ ■

Show: disks | networks | more info

CPU: 1  
RAM: 2048 (MB)  
System Disk: 20 (GB)



Labels (4)  
debian  
my\_servers  
critical

Metadata: OS (1)

Image: Windows 2008R2  
Image Size: 11000 (GB)

details

Manage Tags



Name: ubuntu-srv1  
IP: 192.168.32.21

Running  
■ ■ ■ ■

Reboot  
Shutdown  
Console  
Destroy

Confirm

Cancel

Show: disks | networks | more info

Your actions will affect 2 machines

Cancel All

Confirm All

## machines



Create New +


 Name: debian-1  
 IP: 192.168.32.7

Rebooting



Show: disks | networks | more info


 CPUs: 1  
 RAM: 4096 (MB)  
 System Disk: 40 (GB)

 Labels (4)  
 debian  
 my\_servers  
 critical

 Metadata: OS (2)  
 Location

 Image: Debian base  
 Image Size: 2000 (GB)

[details](#)
[Manage Tags](#)

 Name: win1  
 IP: 192.168.32.13

Running



Show: disks | networks | more info

 CPUs: 1  
 RAM: 2048 (MB)  
 System Disk: 20 (GB)

 Labels (4)  
 debian  
 my\_servers  
 critical

Metadata: OS (1)

 Image: Windows 2008R2  
 Image Size: 11000 (GB)

[details](#)
[Manage Tags](#)

 Name: ubuntu-srv1  
 IP: 192.168.32.21

Shutting down



Show: disks | networks | more info



networks



Create New +



Internet

Show: machines (3)

Public Network





Create New +

### Internet Public Network

Show: machines (3)

	debian-1	IPv4: 192.168.32.7
	IPs   Firewall (Off)	
	win1	IPv4: 192.168.32.13
	IPs   Firewall (Off)	
	<ul style="list-style-type: none"><li><input type="radio"/> Unprotected mode (Firewall off)</li><li><input type="radio"/> Fully protected mode (Firewall on)</li><li><input type="radio"/> Basically protected mode (Firewall on)</li><li><input type="radio"/> Custom protection <a href="#">settings</a> (Advanced users only)</li></ul>	<input type="button" value="Apply"/>
	ubuntu-srv1	IPv4: 192.168.32.21
	IPs   Firewall (Off)	

Details



network

Create New



Inter

Show

Name your network

Name:  (\* Required field)

Cancel

Create Network

network

Create New



Internet

Show



prv-net




Show



### Add machine



Select machines to add to: prv-net1

-  debian-1
-  win1
-  ubuntu-srv1

Cancel

Add

## networks



Create New +



Internet

Show: machines (3)

Public Network



prv-net1

Show: machines (2)

Private Network



debian-1



[Connect](#) to manage private IPs



win1



[Connect](#) to manage private IPs



1 Image

2 Flavor

3 Name

Select an OS

system images

custom images

-  **Ubuntu 11.04 Desktop**  
Ubuntu 11.04 Desktop, Vanilla Installation 4000 MB
-  **Fedora 14 Desktop**  
Fedora 14 Desktop, full installation 3300 MB
-  **Fedora 15 Desktop**  
Fedora 15 Desktop Edition 2284 MB
-  **Windows 2008R2 Standard**  
Windows Server 2008 R2 11000 MB

Cancel

Next

1

image

2

Flavor

3

Name

Select CPUs, RAM and Disk Size

small  medium  large  custom

CPUs  1 cores

RAM  2048 MB

Size  20 GB

Your wallet: 10,000 Credits | This setup will cost you: 20 C/hour

Back

Next

1 Image

2 Flavor

3 Name

Confirm your settings

Name:

Image: Debian base

CPUs: 1 cores

RAM: 2048MB

System Disk: 20GB

Cost per Hour: 40 credits

Credits in Wallet: 10.000

Back

Create VM

machines



Success!



Your new machine is now building... (this might take a few minutes)

Password:

wTpguki2

**Write down your password now!**

You will need this later to connect to your machine.  
After closing this window you will NOT be able to retrieve it again.

Thank You!

Questions?

