Scheduling & SLA @oVirt

8/11/2012

Doron Fediuck
Supervisor
Red Hat

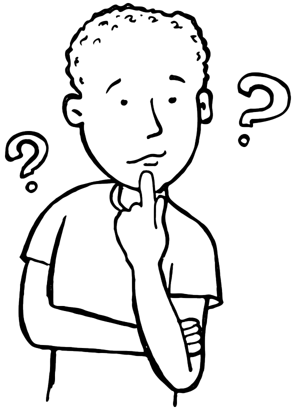
oVirt

Overview

SLA
Scheduling

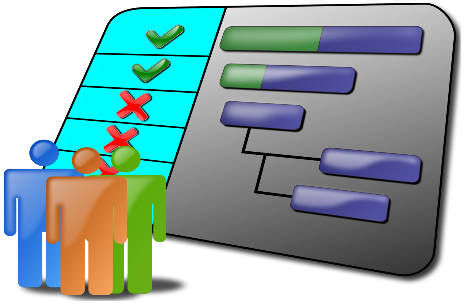
Overview: SLA

- SLA: Service Level Agreement
  - Ensures Quality of Service (QoS) based on parameters and a schema.
  - ISP
    - Schema would be Internet access.
    - Parameters: Up/Down bandwidth, MTTR (Mean Time To Recover), etc.
- In cloud computing this is becoming crucial, as we're providing IaaS



Overview: Scheduling

- Placing a VM on a host
- Schedule various host tasks



Machine re-assignment problem[1]

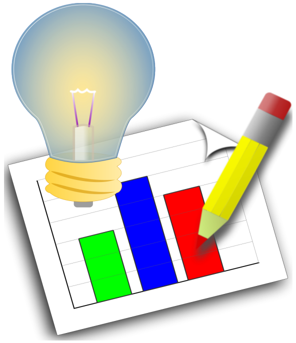
- Defined by Google; assign each process to a machine. All processes already have an original (unoptimized) assignment. Each process requires an amount of each resource (such as CPU, RAM, ...)
- A solution to this problem is a new process-machine assignment which satisfies all hard constraints and minimizes a given objective cost

Found to be mathematically NP-Complete (**can't be solved**)

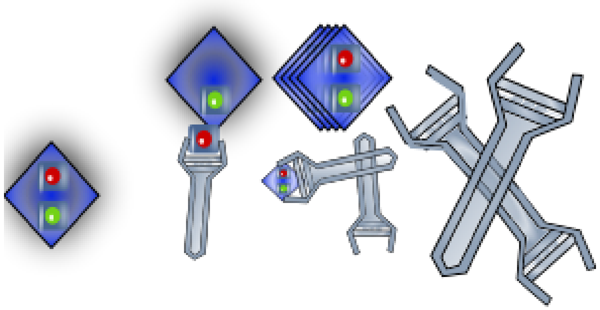
[1] http://challenge.roadef.org/2012/en/

Overview: Scheduling & SLA

So what CAN we do?



- Optimize scheduling scenarios
  - Scheduling improvements
  - Integration with external systems
- Gradually introduce SLA elements into oVirt
  - Add various features which will function as a toolbox
  - Prepare the infrastructure for advanced SLA concepts



oVirt

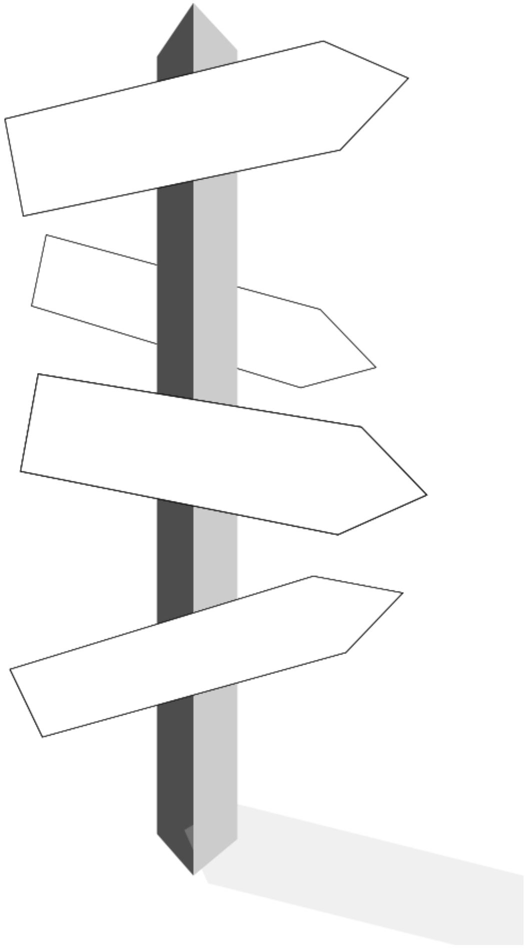
Scenarios

What is it good for, anyway?

Scenarios

SLA Based

- *Multi Tenancy / cloud models: capping, quotas*
- *VM HA*



Scheduling based

- *Memory over commitment*
- *Power saving policies*
- *KSM performance: negative affinity*
- *Advanced scheduling*
  - *Time based: turn on/off at a given time*
  - Various algorithms implementations
  - Statistic-based scheduling

Scenarios: SLA

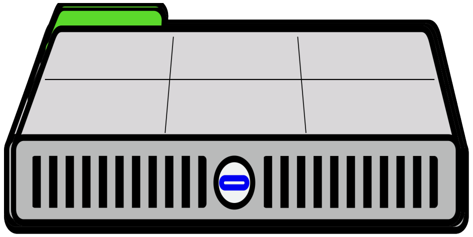
Private-cloud / multi-tenancy models

- Limitations / Capping (CPU, RAM, TBD...)
  - Allow limiting a VM's resource consumption
  - Provide better control on VM behavior and prevent a VM from going wild.
- *Quota*
  - *Management level limitations*

Scenarios: SLA

VM High Availability

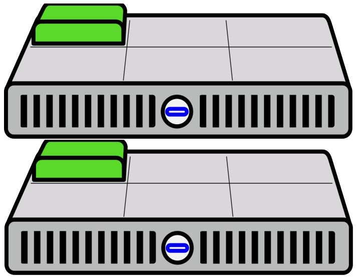
- Host level: Tagged hosts should be used when scheduling HA-VMs.
- VM level: allow auto-reset when guest fails (blue screen, etc.)
- Application level: monitor specific application(s) and act accordingly (reset, migrate, etc) when it stops responding



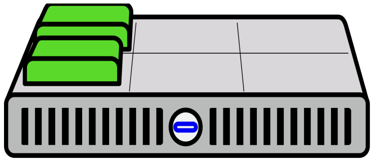
Scenarios

VM affinity (co-location, Positive / Negative)

- Negative affinity
  - One VM 'repels' the other
  - HA via separate host VM placements

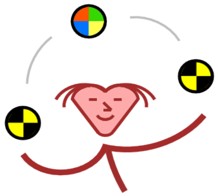


- Positive affinity
  - One VM 'attracts' the other VM
  - Grouping all VMs with the same OS will get best KSM results.
  - Licensing pricing model in some OSs
  - Simple maintenance and power saving
  - Traffic monitoring for specific VMs



Scenarios: Scheduling

HW utilization: Memory Over Commitment



- Allow running more VMs than available physical memory

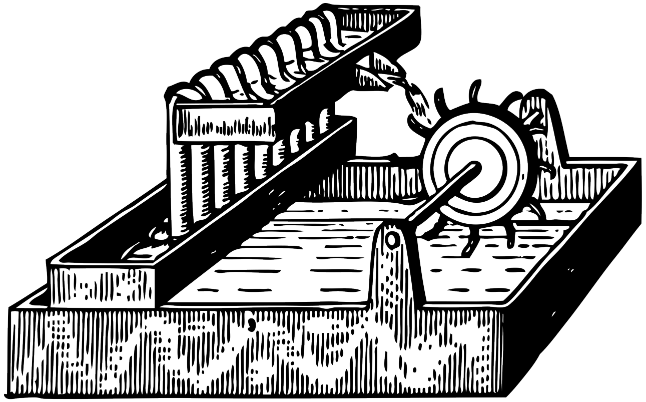
Power saving policies

- Shutdown idle VMs
- Gather all VMs to several hosts (load balancing, already exists) and shut down / suspend unused hosts.

Scenarios: Scheduling

Advanced VM scheduling

- Time based: turn on/off at a given time
- Various algorithms implementations
- Statistic-based scheduling

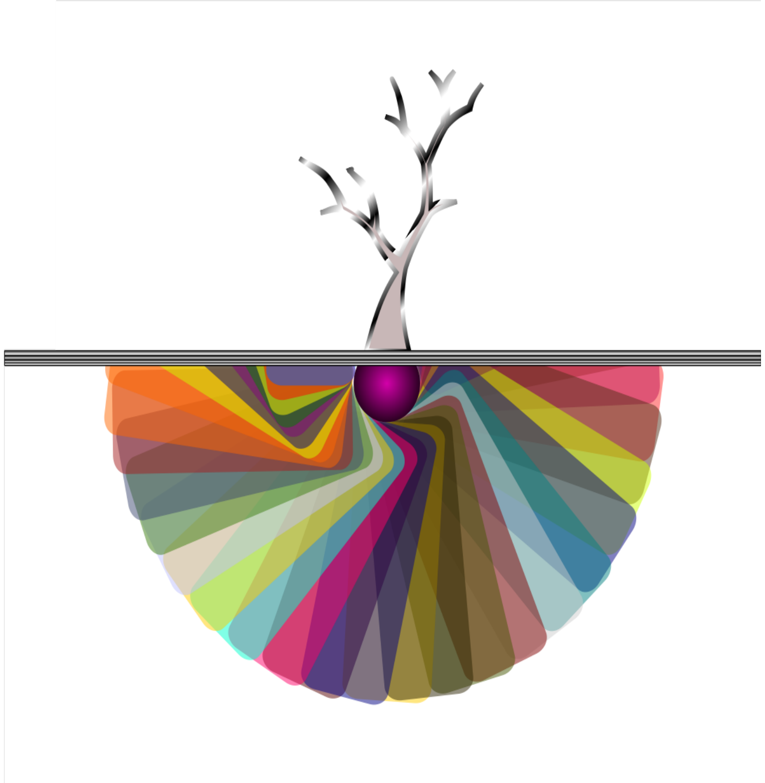


Scheduling considerations

Consider while scheduling...

Each VM and host has meta-data crucial for scheduling

- Resources
  - Connection to network RED
  - Storage usage (DB in a guest)
  - HA reservations
- Topologies
  - CPU pinning
  - NUMA



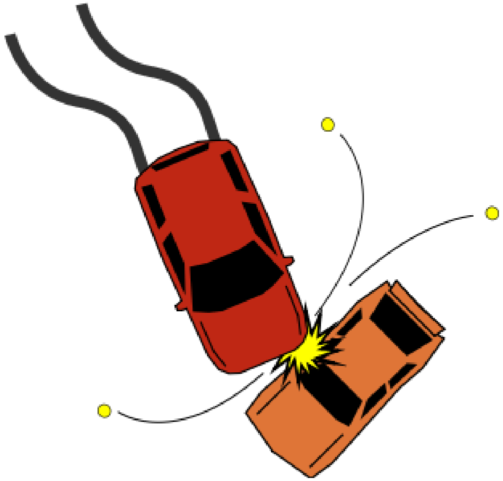
Consider while scheduling...

Resource mapping should be preserved after migration

- What happens when destination host will not support it?

Avoid collisions

- Host-Pinning / HA vs Power savings
- CPU-pinning vs NUMA / KSM
- Optional vs Mandatory VM network



Naive rule: specific settings will override the general policy

- Host-Pinning overrides Power savings

Scheduling & SLA Today

What do we have so far?

Scheduling & SLA Today

Existing Algorithms

- Even distribution
- Power saving

Current scheduling

- Running a VM
  - Basic validations
  - HasMemoryToRunVM
  - Use the relevant selection algorithm to find the best host

Scheduling & SLA Today

Current scheduling

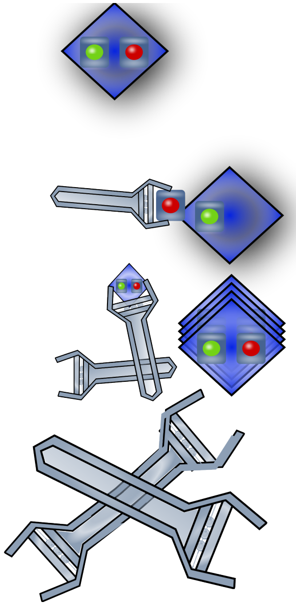
- Migrating a VM
  - Same validations as with running a VM
  - Avoid selecting current host
  - HasCpuToRunVM
  - Use the relevant selection algorithm to find the best host

- Load balancing (cluster policy)
  - Time based polling, using one of the current selection algorithms to migrate VMs as needed.

Scheduling & SLA Today

New features 3.1 introduced

- Enabling memory balloon by default[1]
  - Deflated, may be used externally
- CPU pinning[2]
  - Specific and range pinning topology
  - Migration allowed
    - No validation on destination host.



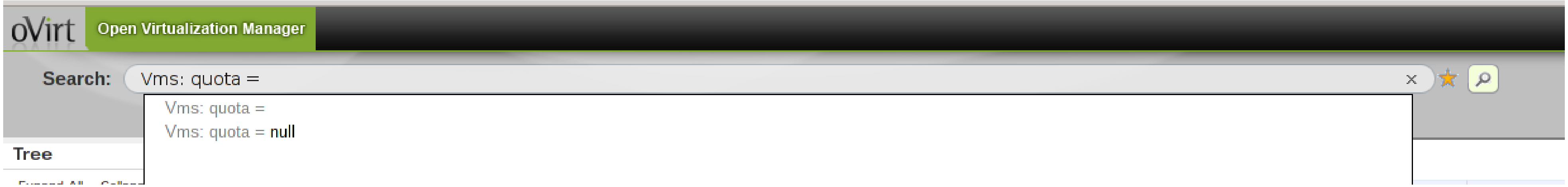
[1] http://wiki.ovirt.org/wiki/Features/Design/memory-balloon

[2] http://wiki.ovirt.org/wiki/Features/Design/cpu-pinning

Scheduling & SLA Today

Quota[1]

- Control resource allocation
- Storage quota
- Cluster (Memory+CPU) quota
- Disabled (default), audit and enforcing modes
- Search-queries (VMs, templates and disks)



Search: Vms: quota =

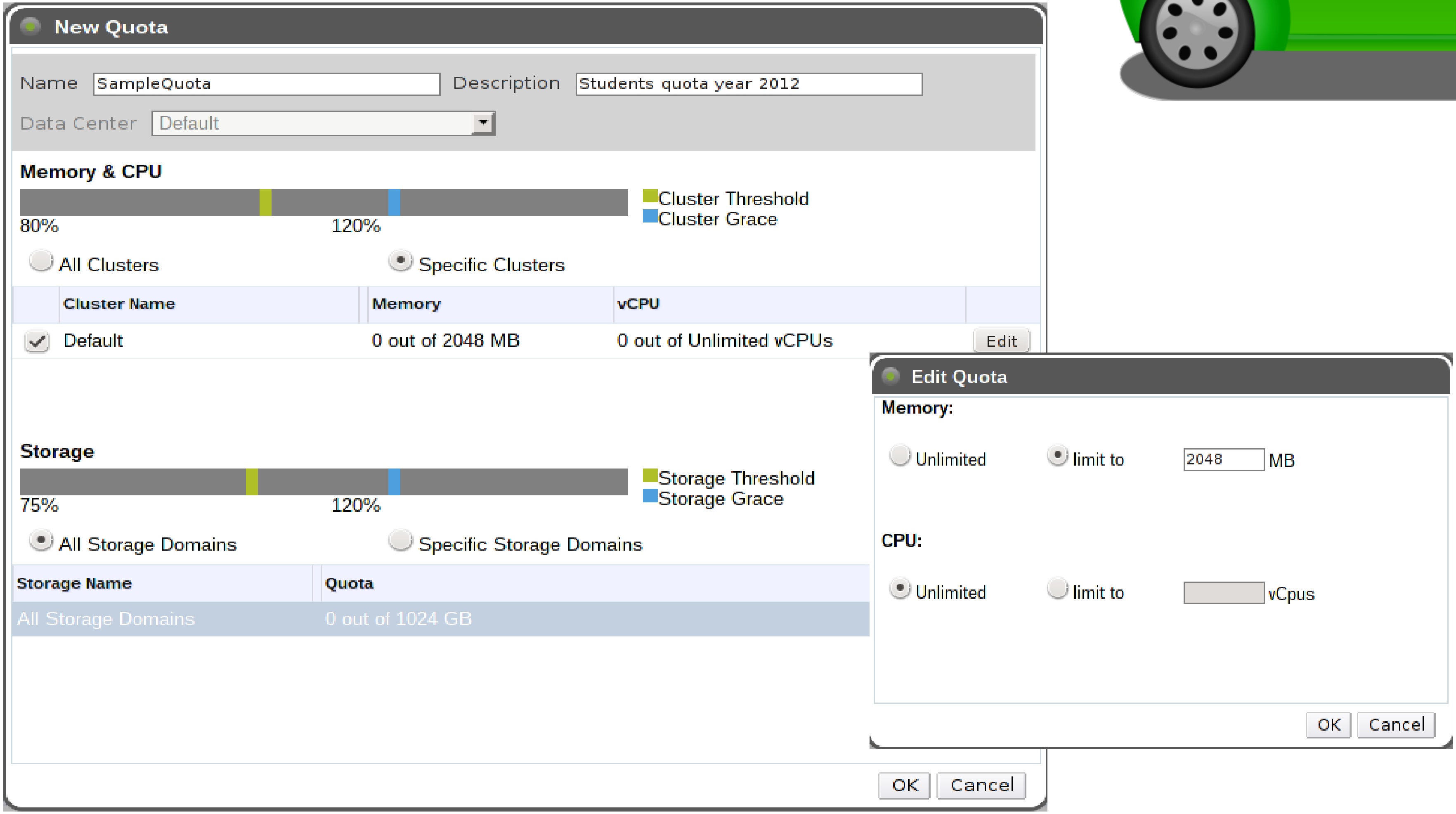
Vms: quota =

Vms: quota = null

[1] http://wiki.ovirt.org/wiki/Features/Design/Quota

Scheduling & SLA Today

Quota sample



New Quota

Name: SampleQuota
Description: Students quota year 2012
Data Center: Default

Memory & CPU

○ All Clusters ● Specific Clusters

	Cluster Name	Memory	vCPU	
✓	Default	0 out of 2048 MB	0 out of Unlimited vCPUs	Edit

Storage

● All Storage Domains ○ Specific Storage Domains

Storage Name	Quota
All Storage Domains	0 out of 1024 GB

Edit Quota

Memory:

○ Unlimited ● limit to 2048 MB

CPU:

● Unlimited ○ limit to vCpus

oVirt

Work in Progress

So what are we doing?

Work in Progress

Pluggable scheduling architecture[1]

- Replace or add to internal scheduler
- Allow users to write their own scheduler
- API based
- Community friendly
- Actually, needed by community...

[1] http://wiki.ovirt.org/wiki/Features/SLA_PluggableArchitecture

Work in Progress

Smart Scheduler
Integrating BMC

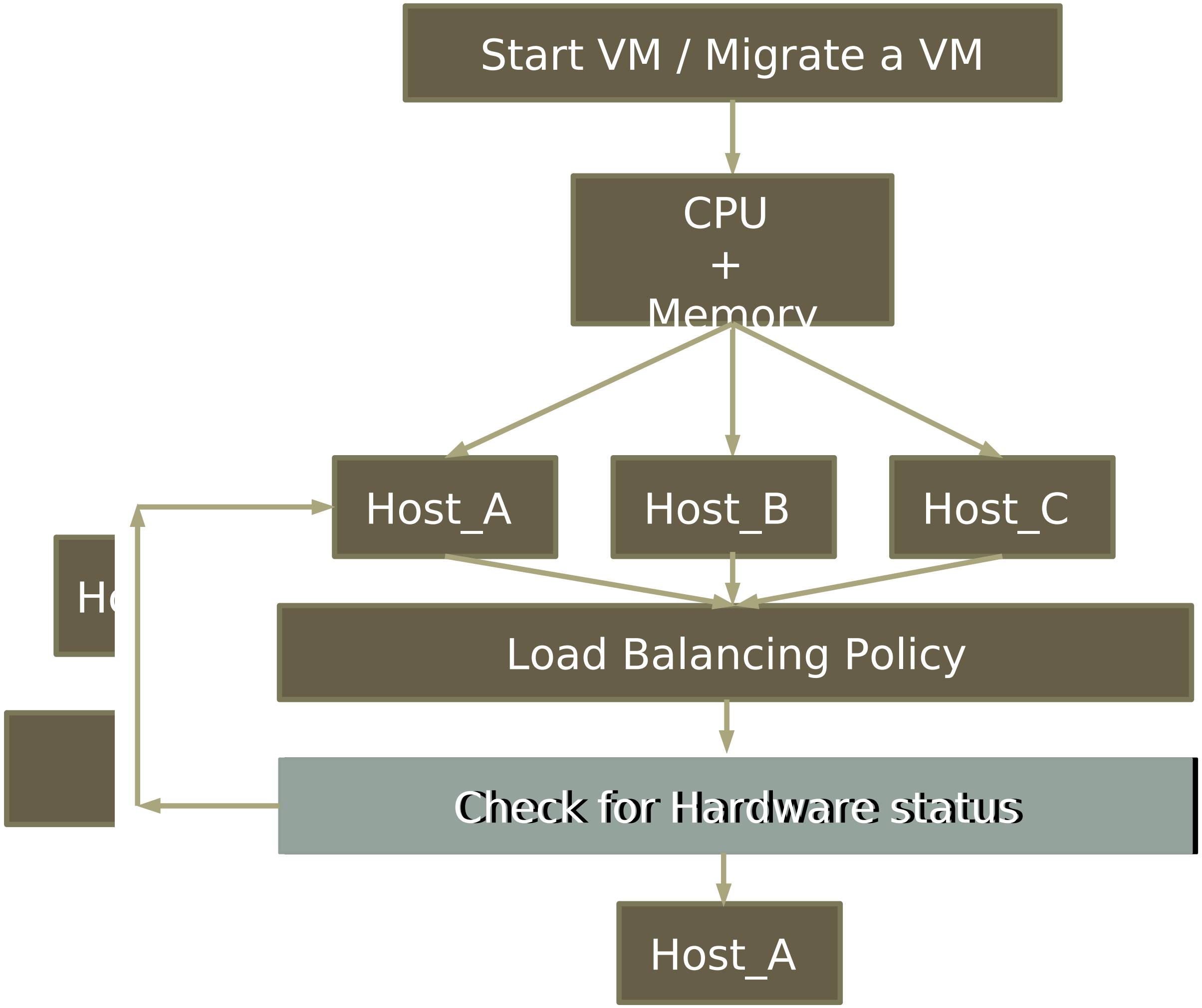
Srinivas Gowda G
Surya Prabhakar
Dell India R&D

Presented

Last month

In Bangalore

oVirt workshop



Work in Progress

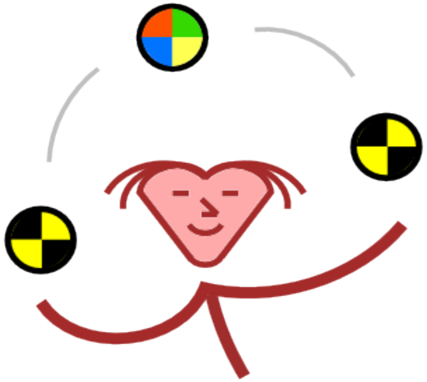
- Rules engine (RBMS) integration
  - Based on pluggable scheduling
  - Currently suggested as a POC
  - Tomorrow in László Hornyák's session on Drools integration.

- Internal Quota improvements[1]
  - Filling-in UI gaps
  - Making sure Quota is not skipped by new commands

[1] http://wiki.ovirt.org/wiki/Features/Design/Quota-3.2

Work in Progress

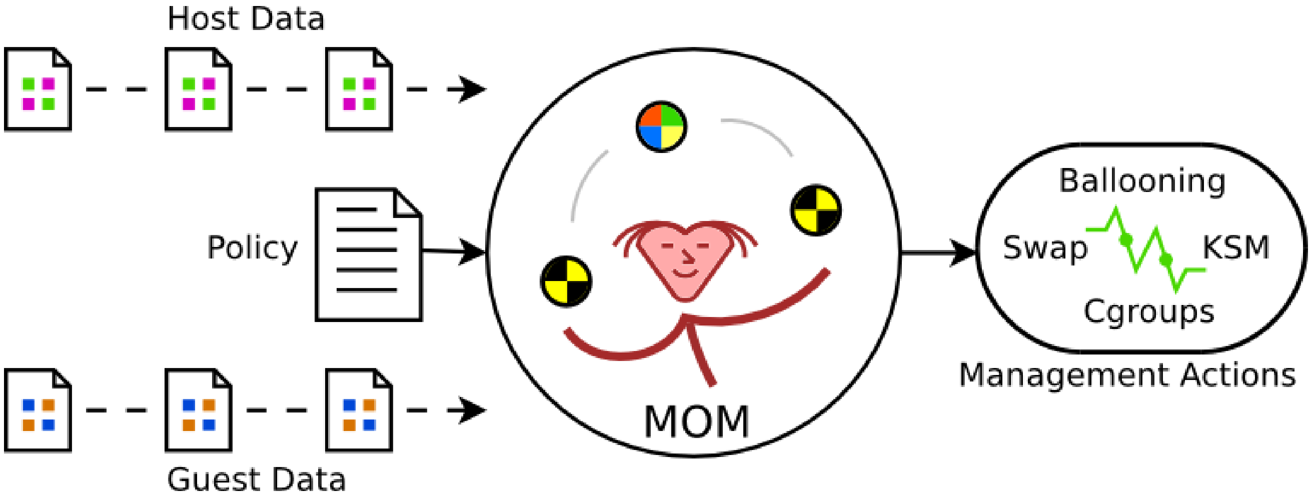
- Integrating VDSM-MoM
  - Written and maintained by Adam Litke
  - Joined oVirt as an incubation project last year
  - Monitors and handles ksm and ballooning
  - Trying to prevent interaction mistakes
    - Ballooning VS KSM



Work in Progress: Introducing MoM

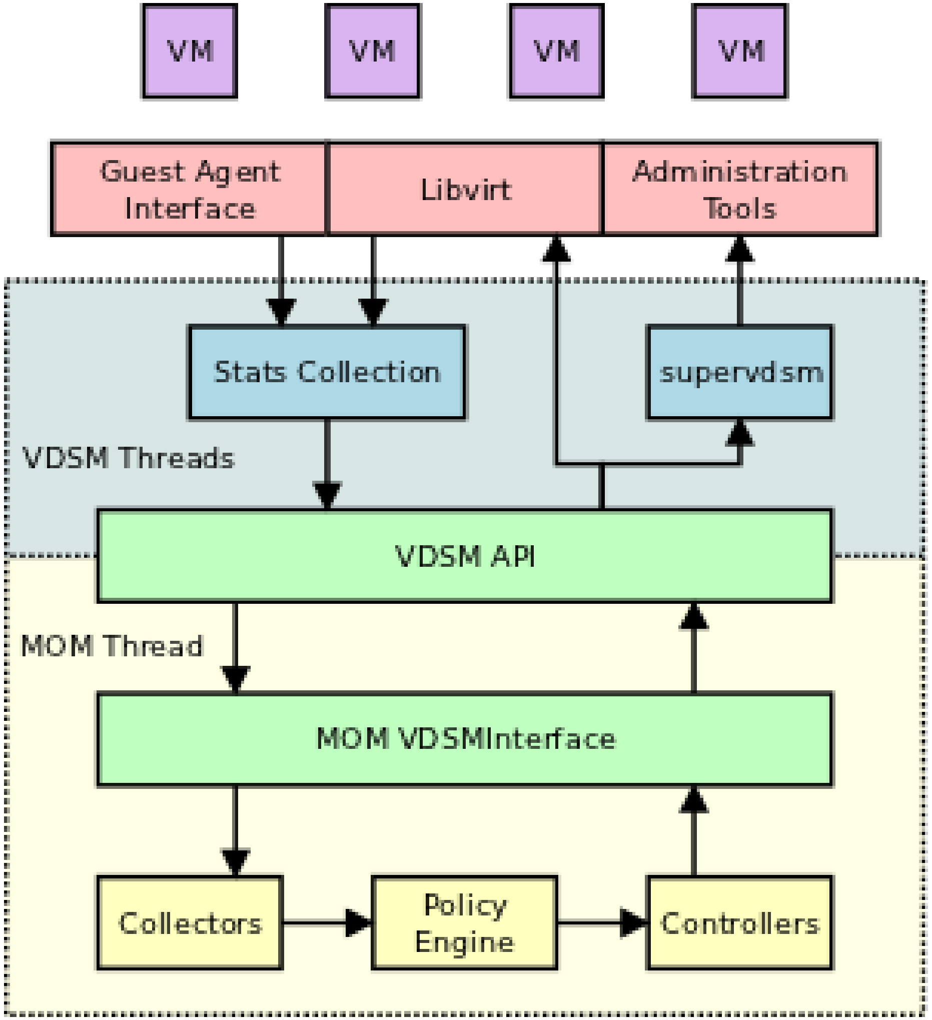
- Guest tracking
- Stats collection
- Fully extensible

- Dynamic policy engine
- Support for ksm and ballooning



Work in Progress: mom integration[1]

- MOM threads run within vdsmd
- Stats collected via the vdsm API
- Ksm / ballooning operations via vdsm API
- VDSM installs a default MOM policy

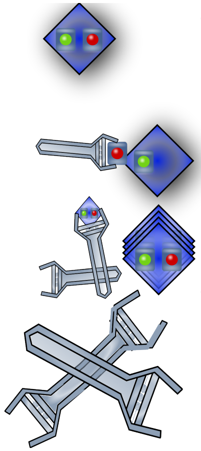


[1] http://wiki.ovirt.org/wiki/SLA-mom

Work in Progress

MoM integration[1]

- MoM is becoming the enforcement agent
- VDSM integration done by Adam Litke and his colleagues (Mark Wu, Royce Lv)
  - Still gaps on engine side.



- Initial phase for basic integration while maintaining ksm functionalities, adding API support for memory balloon
  - Packaging and maintaining (added to Bugzilla)
- Now adding capping (limitations) API support to VDSM
  - CPU & Memory (guaranteed, hard and soft limits)

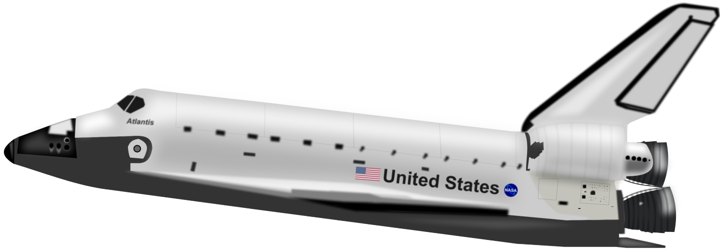
[1] http://wiki.ovirt.org/wiki/SLA-mom

oVirt

Road-map

to Infinity (affinity?) and Beyond!

Scheduling & SLA Road-map



- SLA features
  - VM Watchdog (VM HA)
  - HEAT integration (Application HA)
  - NUMA (numad, auto-numa)
- Scheduling: additional improvments
- Extend MoM capabilities
  - Handle specific VMs
  - Policy resolution (allow policy parts)
  - Limitations for network & storage

oVirt

and now is a good time for....

Questions?

THANK YOU !

http://wiki.ovirt.org/wiki/Category:SLA
engine-devel@ovirt.org
vdsm-devel@lists.fedorahosted.org

#ovirt irc.oftc.net