**Service   
“Hypervisor”**

Technical description

**Supplier:**

EDS Systems OÜ

Reg.nr.: 11384133

Sadama 4, 10111 Tallinn

Telephone: +372 6409600

Fax: +372 6409606

[www.edss.ee](http://www.edss.ee/)

e-mail: [info@edss.ee](mailto:info@edss.ee)

**Document revisions**

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Version | Initiatior | Changes |
| 29.07.2013 | 0.1 | Eugene Istomin | Initilal |
| 12.09.2013 | 0.2 | Eugene Istomin | Changes in common information |

1. **Common information**

**Service goal:**

provide a secure, proven, scalable,network-agnostic and near-native performance virtualization platform with flexible configuration.

Service concept:

* Shared-nothing architecture
* PV (Xen), HVM(Qemu), PVonHVM(PV GPL/VirtIO) modes, PVH mode
* NUMA-optimised VM placement
* flow-based network configuration/management/MAC/ACL
* fully-reserved on-fly VM disks allocation
* zero-overhead VM snapshots
* zero-overhead fast VM creation
* weight-based CPU accounting
* on-fly resource allocation (mostly linux)
* hardware firmware updating using common infrastructure
* IPMI-ready configuration
* autogenerated configuration for network layer such LACP
* {V/VX}LAN-agnostic configuration & isolation using openflow model

Arguable solutions:

* Shared-nothing architecture → no updates/reboots without whole server interrupt
* Openflow controller → SPOF, additional latencies
* Kernel locking → unable to use kernel > 3.4 (OVS, OCFS2,Xen netback)

Service main functionality:

* create disk images
* create VM configurations
* run VMs
* on-fly resource allocating
* update HW firmware
* update IPMI configuration
* NUMA reassign
* create snapshot (by request or by schedule)
* make a network layer

Software,protocols, technologies:

* Xen (4.3) – hypervisor
* Qemu (gte 1.5) – virtualization platform
* OpenvSwitch (gte 1.10) – SDN solution for complex network flows management
* OCFS2 – as /storage file system

Current limitation:

* Unable to use monitoring and logging templates. TBD: implement.
* managing OpenFlow rules for more than one server is not supported. TBD: centralized OpenFlow controller
* MS Windows VM hosting needs detailed testing. TBD: testing.
* 3.10 kernels not supported. TBD: ovs patching
* non-kthread model, netback proccess pinning, “options netbk copy\_skb=1 bind=1” in modprobe. TBD: waiting for efficient xen netback model (xen 4.4).
* autogenerated configuration for network layer based on igb Ethernet cards. TBD: make a field with network card driver.
* hardware firmware update is not flexible. TBD: make a fields with HW specification.
* IPMI/RAID monitoring is out of scope. TBD: implement.
* Template storage resource location is not flexible. TBD: make a fields with DD of template storages.
* XSM-Flask is not used. TBD: research.
* GRUB1-support only. TBD: research.
* VM image resizing is not supported. TBD: research
* HW parameters is not filled automatically. TBD: implement.
* Sudoers in not templated. TBD: implement.
* Management user creation is not templated. TBD: implement.
* Remus replication is out of scope. TBD: research and implement.
* IPoIB/RoCE is out of scope. TBD: research and implement.
* Non-working dependencies Systemd timers services . TBD: testing.
* Snapshot service leads to CPU stall. TBD: testing, ocfs2 maillists.
* Snapshots can't be backuped to another host. TBD: bacula integration.
* Snapshot needs some management. TBD: implement.
* All key resources in Dom0 must be limited by cgroup. TBD: research.
* HVM VM w/o PV drivers doesn't belongs to ovs switch. TBD: research.
* No support for connecting multiple OVS instances. TBD: research.
* No port mirroring feature.TBD: research.
* DHCP servers is not allocated in interfaces. TBD: implement.
* Storage driver domains are not supported. TBD: implement.

Non-automated:

* ocfs2 fs creation
  + mkfs.ocfs2 -b 4KB -C 1MB -N 4 -T vmstore -L "storage" --fs-features=local,backup-super,sparse,unwritten,inline-data,metaecc,refcount,xattr,indexed-dirs,discontig-bg /dev/{DEVICE}

**Software/technology considerations:**

1. Xen

* there is a system job for snapshots creation - “dom0\_job”

**Service actions map**

*0-once*

* “hypervisor--main”
  + diskmount
  + reset ovs-switch
  + initial folders creation

*1-conf*

* "hypervisor--boot"
  + update of /boot/grub/menu.lst
* "hypervisor--hw\_raid"
  + downloading a RAID controller firmware to /tmp/firmware
* "hypervisor--dom0\_scripts"
  + updating all custom xen scripts
  + updating /etc/faster
  + reenabling custom systemd services
* "hypervisor—ipmi”
  + updating IPMI settings
* "hypervisor—network"
  + updating management interface names in udev
  + updating OVS interface names in udev
  + configuring OVS network modes (Singleport, LACP\_L3+4)
  + initial interfaces creating
  + including interfaces to OVS bridges
  + apply udev rules without server reboot

*run*

* + getting all VMs allocated to this HV
  + generating xl configuration file (if not exist)
  + cloning/copying system disk (if not exist, double checking)
  + allocating data disks, DRBD metadata disks (if not exists, double checking)
  + autostart VM on creation

1. **Support and maintenance**

Chef pattern:

hypervisor

Chef deploy variants:

* hypervisor-common

Chef code reusage:

* templates
  + xen
  + udev
  + \_base
* files
  + xen
  + network

Data safety information

* all downloaded templates are stored in /storage/templates
* all used templates are stored in /storage/images/!master
* all VM images made by external systems are stored in /storage/images/infrastructure
* all VM images made by hand are stored in /storage/images/manual
* all VM configurations are stored in /storage/xl
* all VM snapshots are stored in /storage/backup
* extended acl must be saved by DS solution

Monitoring templates

* Linux - Common
* Disks – Storage
* Services – Network – OVS
* HW – RAID – {RAID\_CONTROLLER}
* HW – IPMI - {VENDOR}

Logging templates

* Linux - Common
* Xen
* HW – RAID – {RAID\_CONTROLLER}