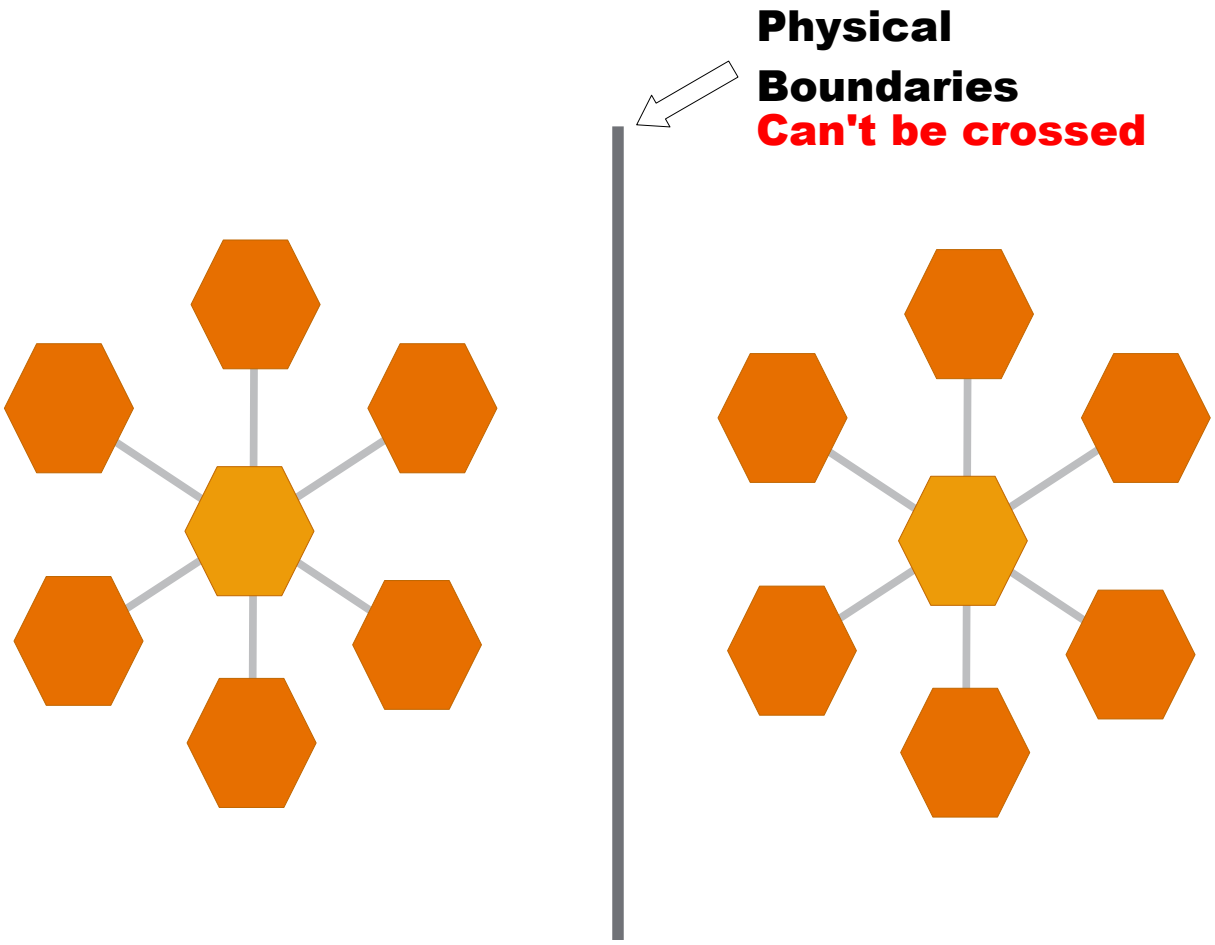
Service Domain Manager Basic and Concepts

Richard Hierlmeier
Sun Microsystems, Inc.

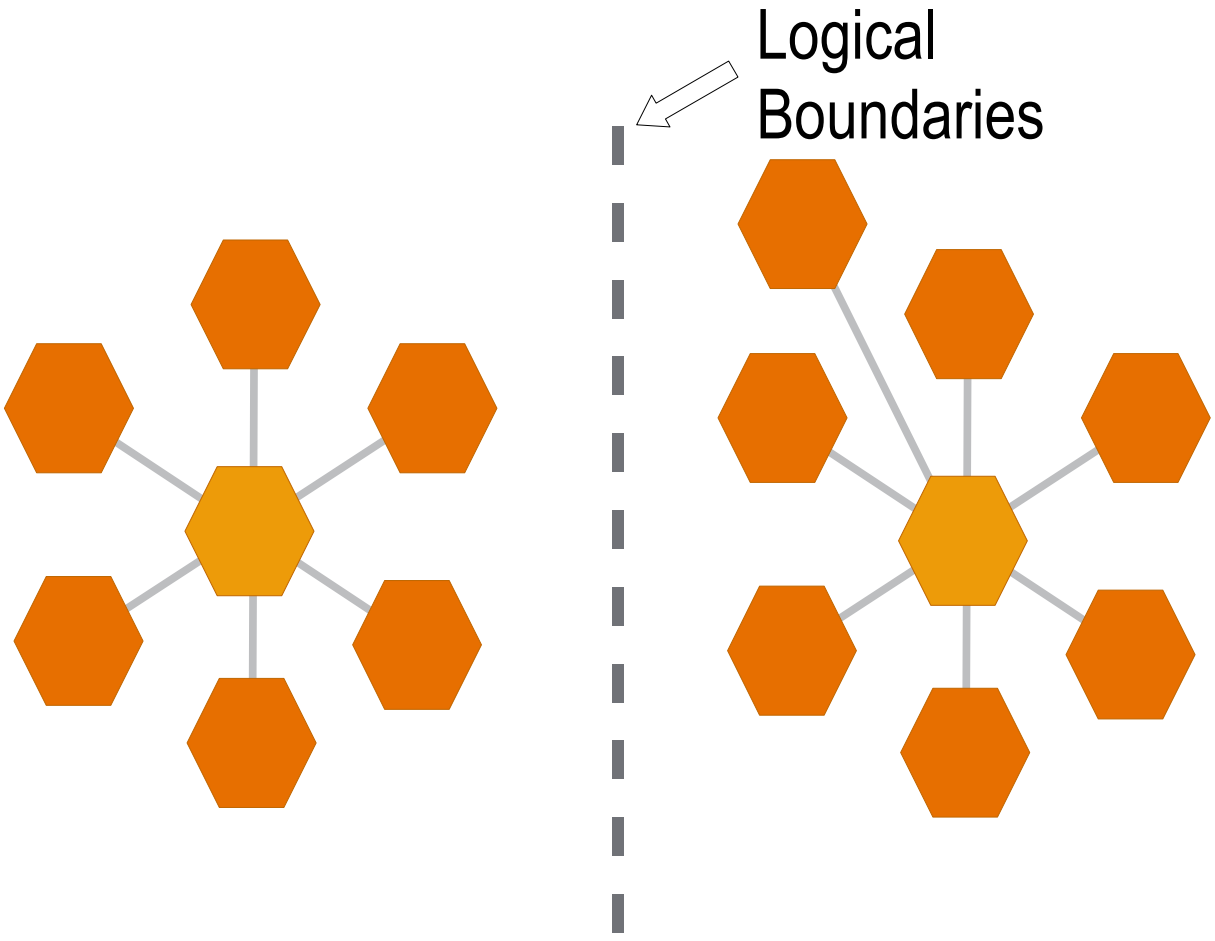
Agenda

- Overview
- Use Cases
- Architecture
- Manage Resources
- SLOs
- Monitoring
- Manage Grid Engine Clusters
- Future Plans
- Q & A

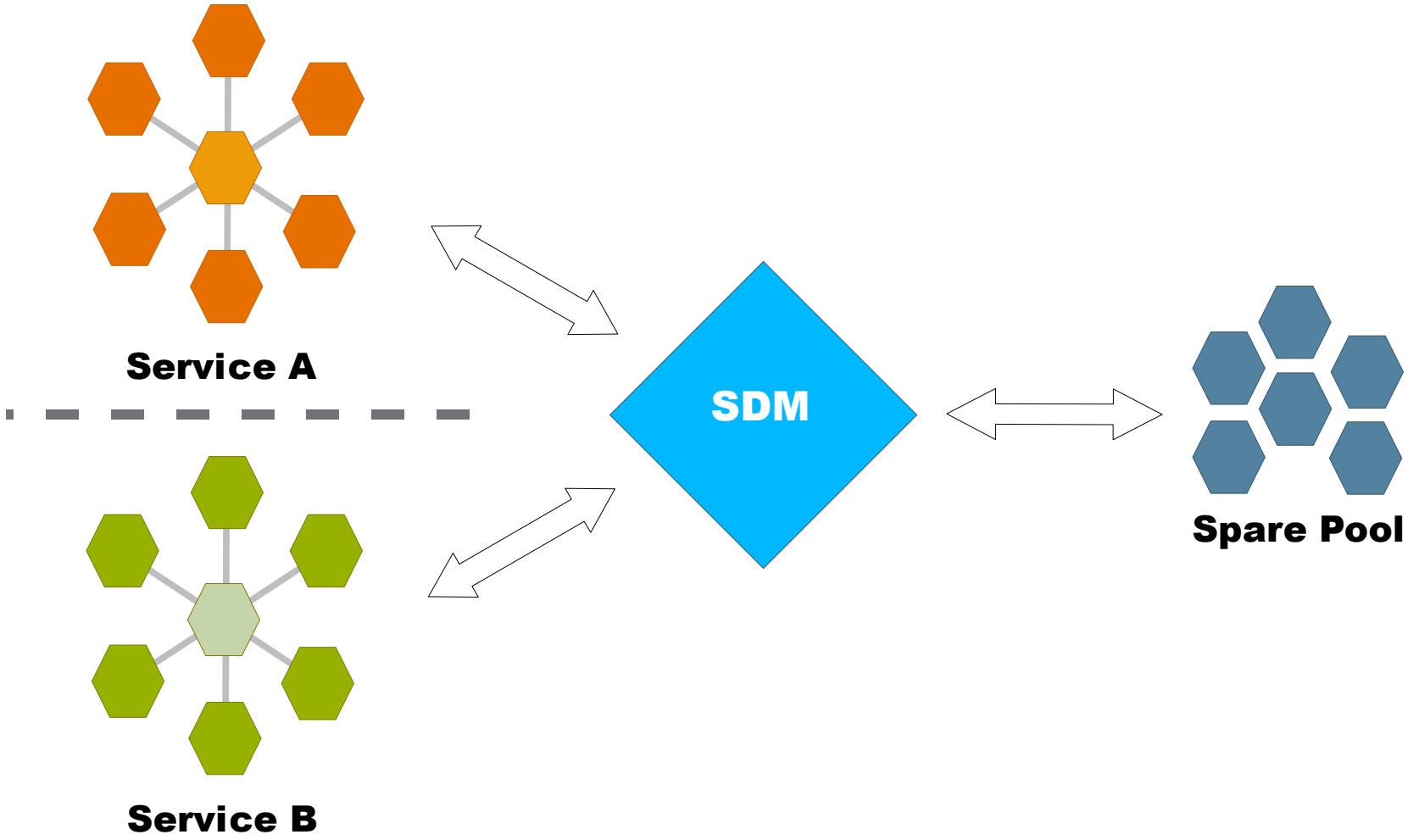
In the Beginning



Along Came Sun Grid Engine

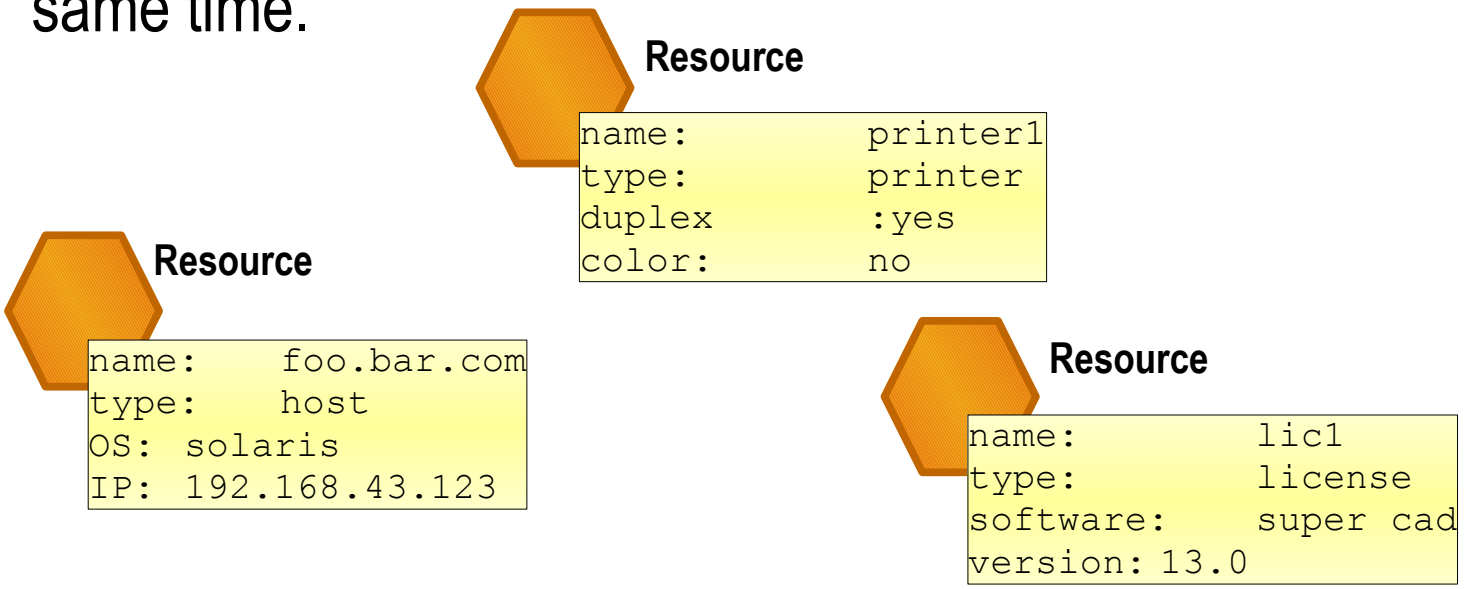


Service Domain Manager



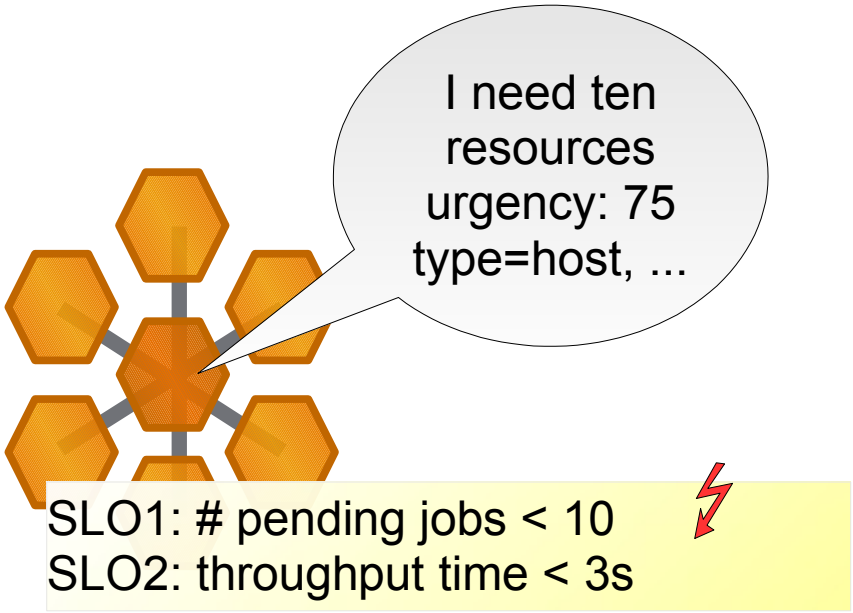
Services

- A Service is a piece of software. It can be a database, an application server or any other software. The only constrain is that the software has to provide a service management interface.
- A Resource is something a Service uses to provide the Service. If you give a Service more Resources, it can do more work in the same time.



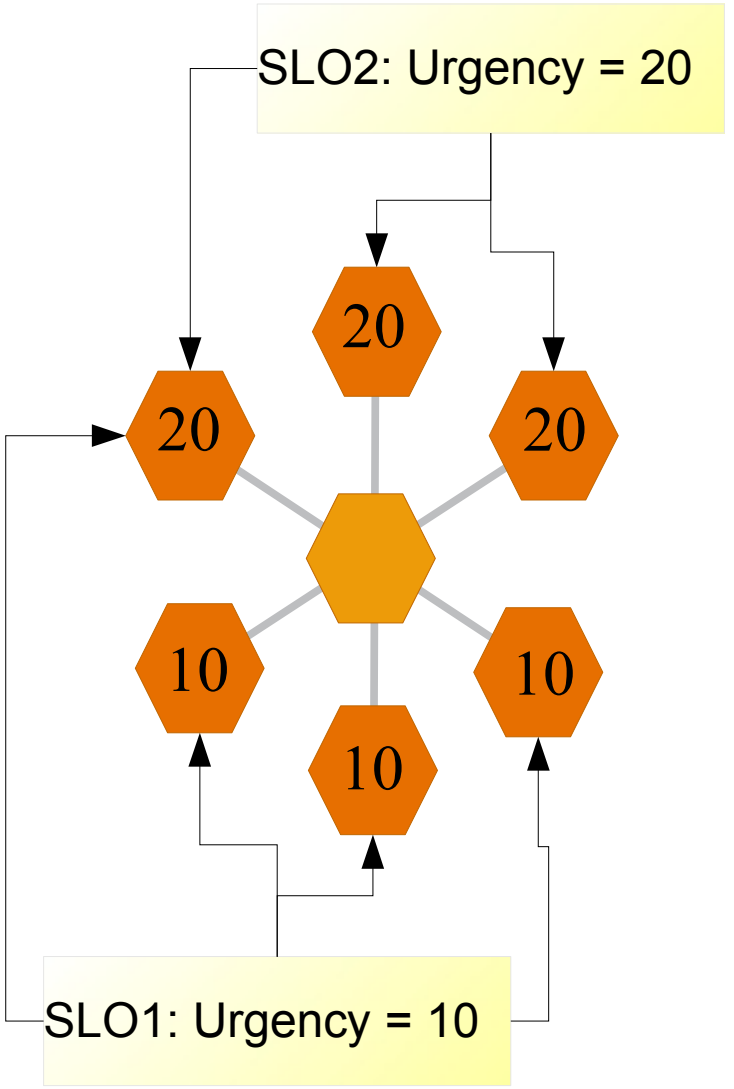
KPIs ad SLOs

- Based on Key Performance Indicators (KPI) a set of Service Level Objectives (SLO) are defined SLOs for each Service
- If a SLO is not met it produces a Need.
  - > Urgency
  - > Quantity
  - > What type of Resource is needed

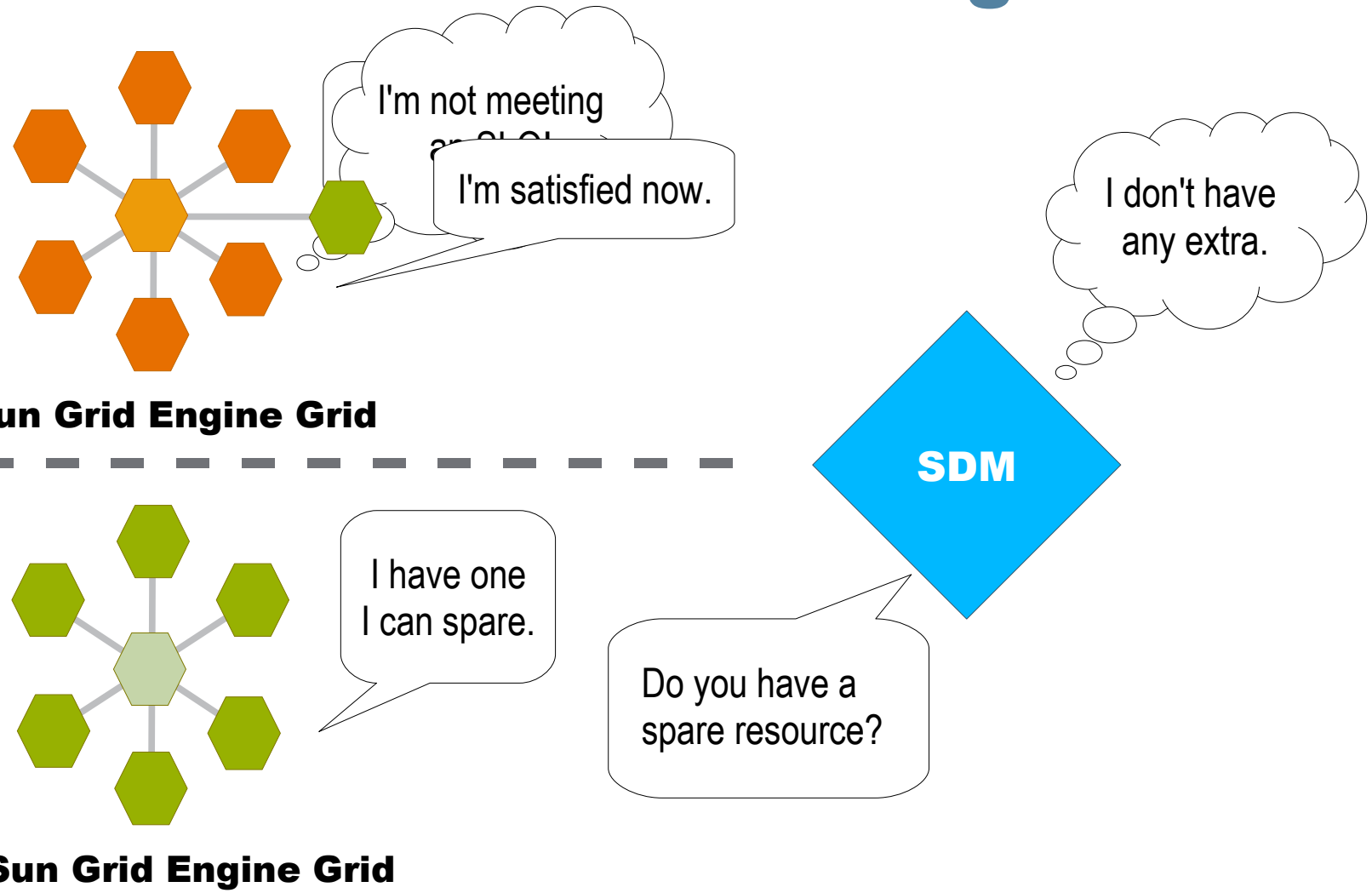


Urgency and Usage

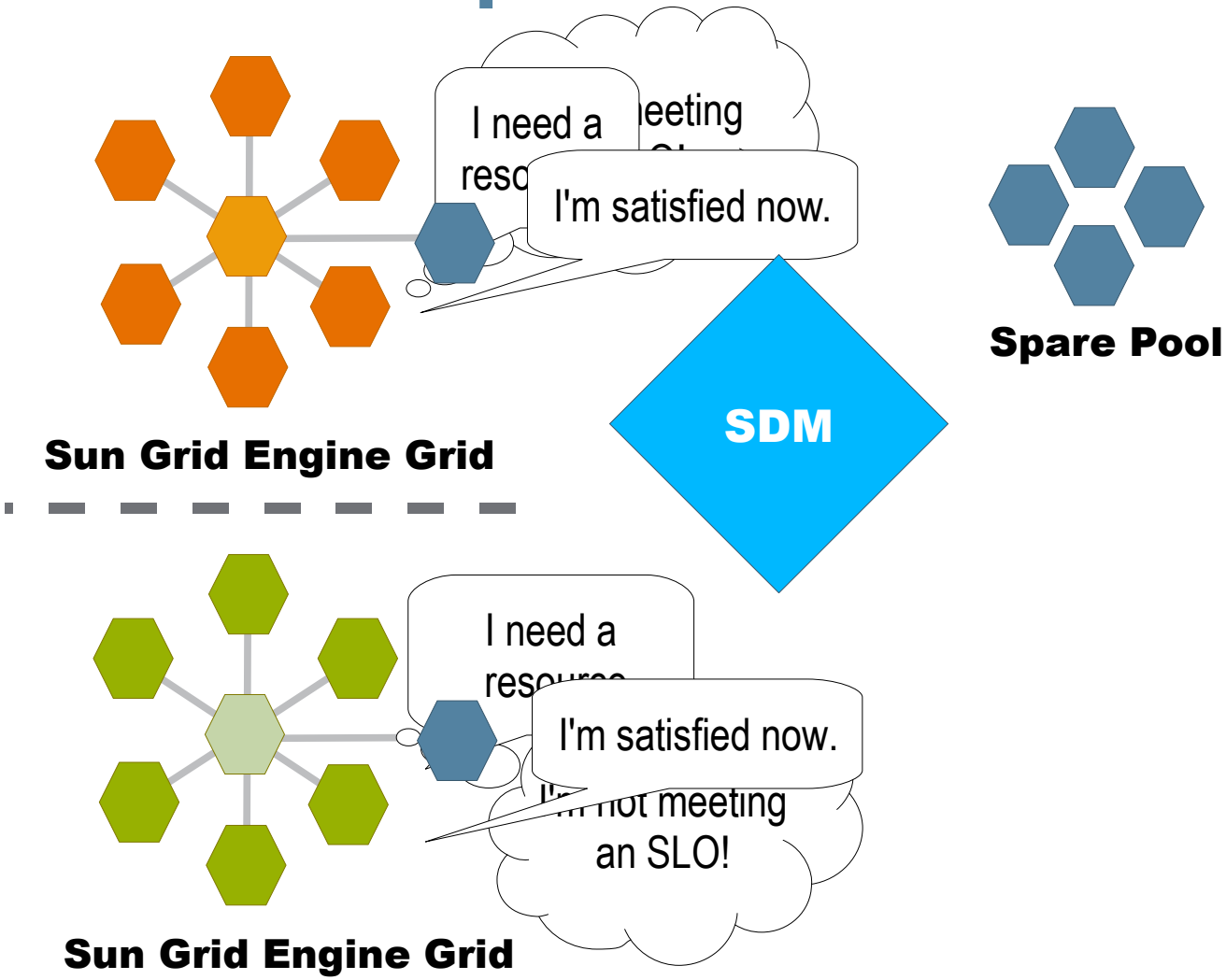
- SLO defines with the Urgency how important the produced Resource Request is
- Whenever Resource is needed to met a SLO the Resource gets as Usage the Urgency of the SLO
- When two or more SLOs need a Resource the Usage will be the maximum Urgency of the SLO



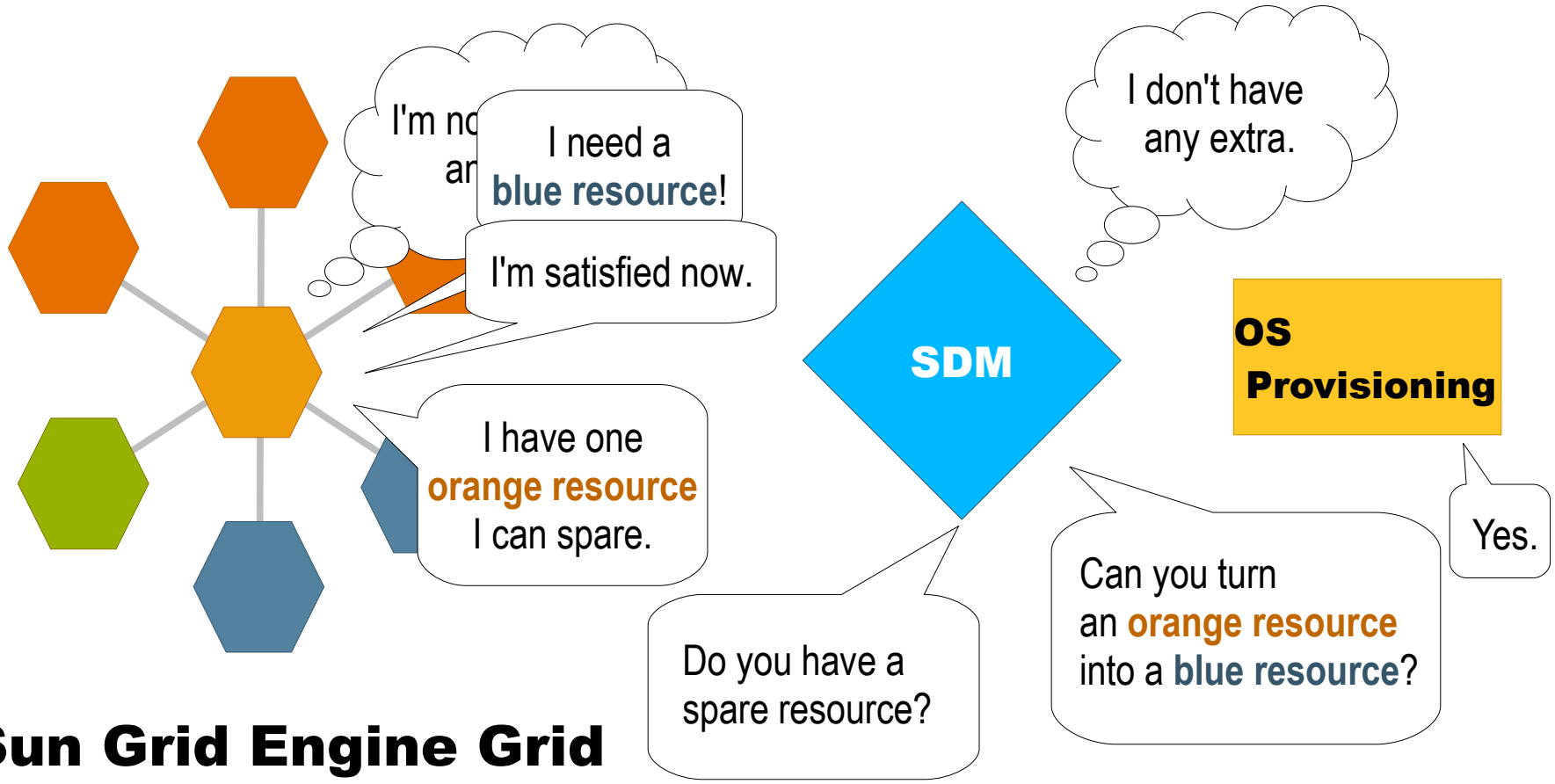
Use Case: Resource Sharing



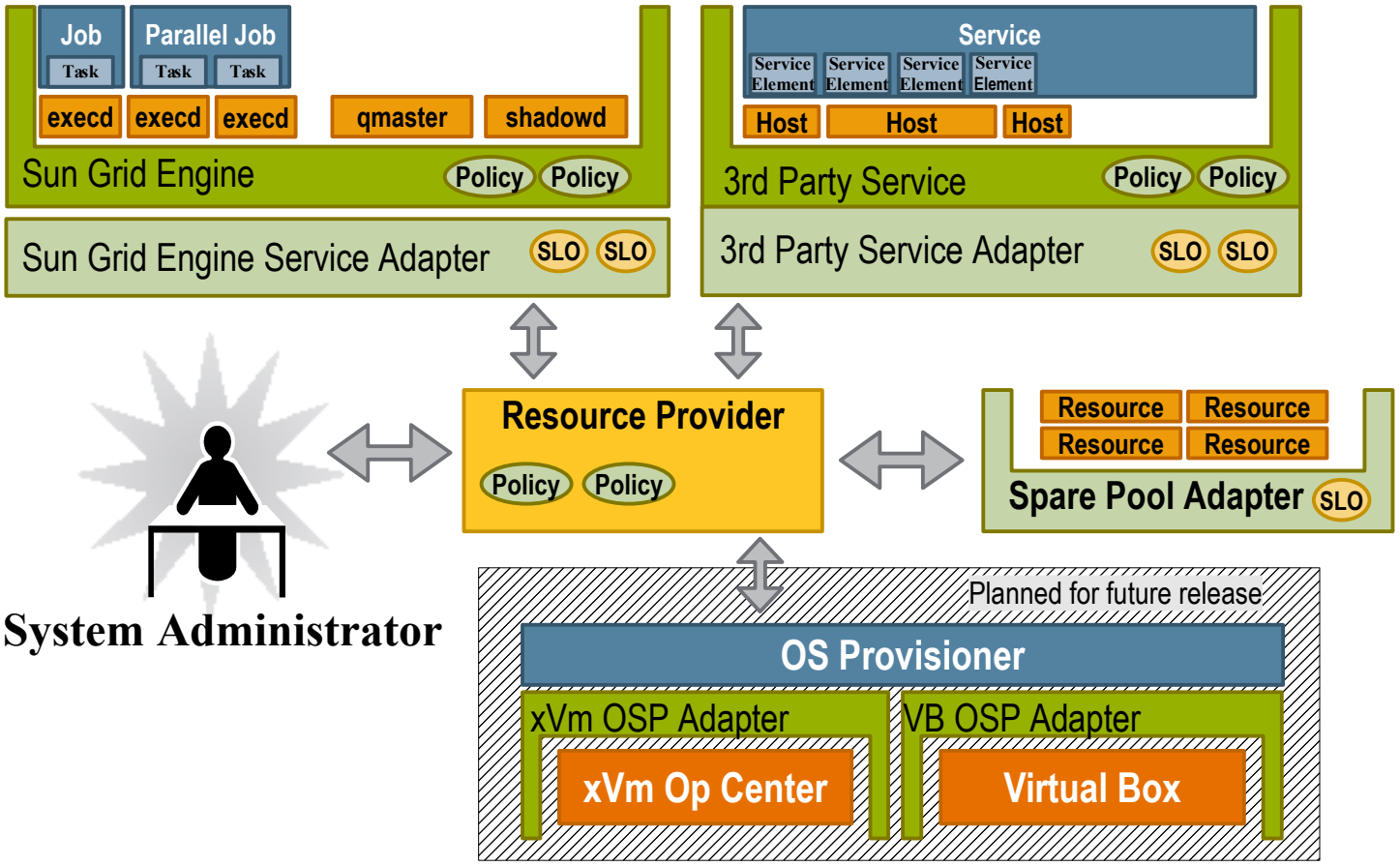
Use Case: Spare Pool



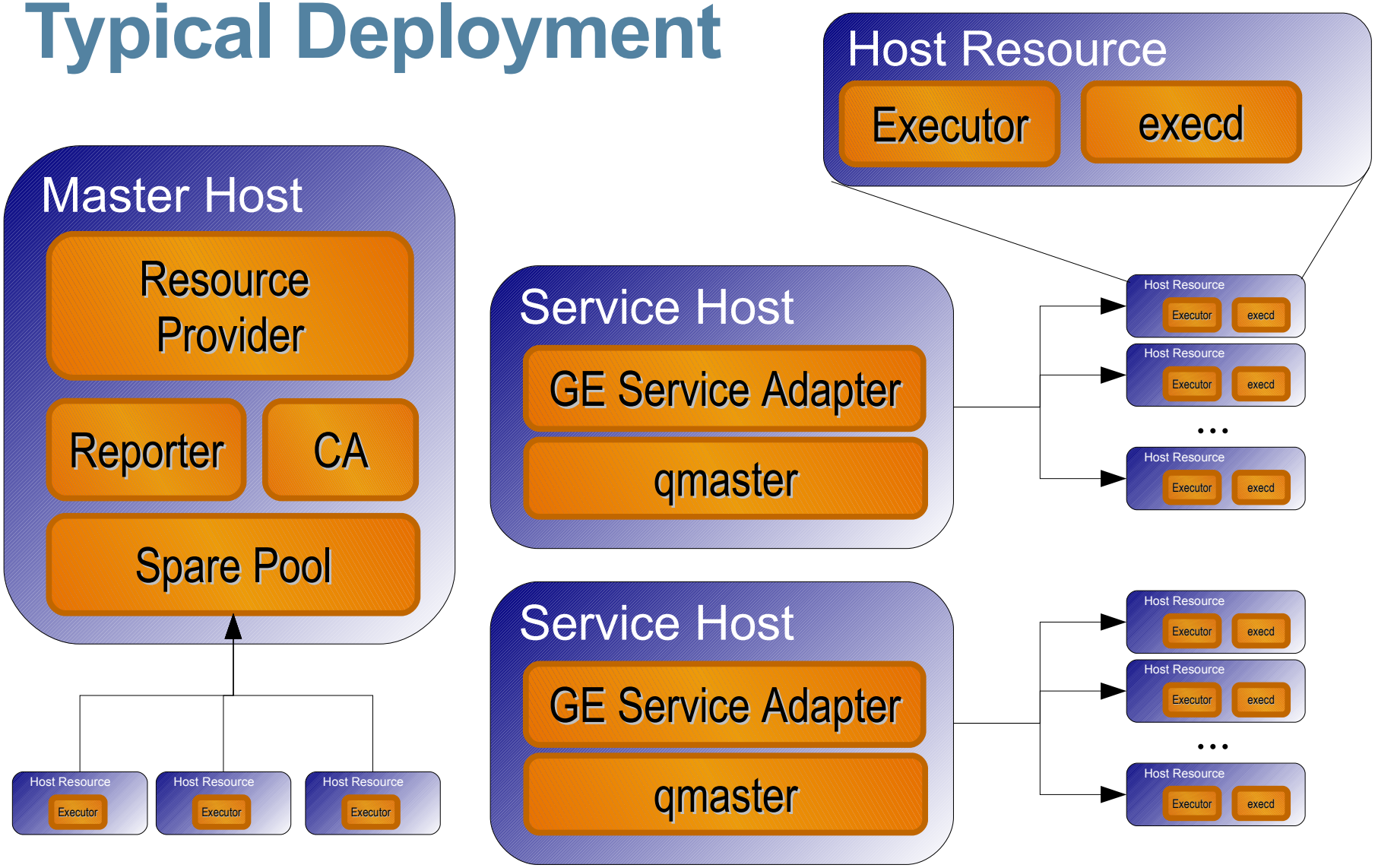
Use Case: OS Provisioning



SDM Architecture



Typical Deployment



User Interface

- SDM provides a Command Line Interface:

```
% sdmadm [global options] <command> [command options]
```

- Provides any functionality to install, administrate, configure and uninstall SDM

```
% sdmadm install_master .....

% sdmadm mod_component_config -c resource_provider


% sdmadm add_resource -r foo.bar

% sdmadm show_slo
```

Manage Resources

- Currently only Host Resources are supported
- A Resource has a state
  - UNASSIGNED The resource is not assigned to any service.
  - ASSIGNING Assignment process is ongoing
  - ASSIGNED The Resource is assigned to a service
  - UNASSIGNING The Resource will be removed from the service
  - INPROCESS Short term state transition inside of ResourceProvider
  - ERROR An action on the resource produces an unrecoverable error. The resource is currently not usable

Manage Resources

- A Resource has an Identifier
  - For Host Resource the Identifier is the hostname
  - Before comparing two Host Resource Identifiers a hostname resolving is done.
  - Other Resource types will define different mechanism for building Resource Identifiers

```
% sdmadm sr -r foo.bar
service id     state    type flags usage annotation
-----------------------------------------------------
p31004  foo    ASSIGNED host S     50
```

Manage Resources

- A Resource has properties
  - predefined properties:

```
% sdmadm show_resource_types
name              property                        flags type
-------------------------------------------------------------
host              hardwareCpuArchitecture                String
                  hardwareCpuCount                       Integer
                  operatingSystemName                    String
                  resourceHostname                M      Hostname
                  ...
                  resourceIPAddress                      String
                  static                          M      Boolean
```

  - Any other resource properties are treated as strings
  - Flag M signalizes that the property is mandatory

Manage Resources

- Display Resource properties

```
% sdmadm show_resource -r foo -all
service id     state    type flags usage annotation
-----------------------------------------------------
p31004  foo    ASSIGNED host S     50
    hardwareCpuArchitecture=amd64
    hardwareCpuCount=4
    operatingSystemName=Solaris
    operatingSystemRelease=10.4
    resourceHostname=master
```

Manage Resources

- Modify a Resource

```
% sdmadm mod_resource -r master -all
----- editor -----------------------------------
#
# Resource properties of resource master
#
resourceHostname = master
static = true
hardwareCpuArchitecture = amd64
hardwareCpuCount = 4
operatingSystemName = Solaris
operatingSystemRelease = 10.4
# hardwareCpuFrequency = <String, optional>
# operatingSystemPatchlevel = <String, optional>
# operatingSystemVendor = <String, optional>
# resourceIPAddress = <String, optional>
----- editor -----------------------------------
```

Manage Resources

- Create a new Resource (interactive)

```
% sdmadm add_resource|ar -s spare_pool
------------ editor -------------------------
resourceHostname = <Hostname, mandatory>
static = false
# hardwareCpuArchitecture = <String, optional>
# hardwareCpuCount = <Integer, optional>
...
```

- Create a new Resource (non interactive)

```
% sdmadm ar -r <hostname> -s spare_pool
```

```
% echo "resourceHostname=foo\\n===resourceHostname=foo1" \
    | sdmadm ar -f - -s spare_pool
```

Manage Resources

- Assign a Resource to a Service manually

```
% sdmadm move_resource|mvr -r <resource id> -s <service name>
```

- Depending on the SLOs of the Services it can happen that the Resource will be immediately removed from the service
  - > It can happen that the service reject the resource => resource will be on the black list
  - > A Grid Engine Service will reject a host resource if it can not execute the execd installation on the host

Manage Resources

- Resources can go into ERROR state, e.g.
  - > installation of execd on host resource failed
  - > execd died
- SDM does not touch Resources which are in ERROR state
- Administrator must have a look onto the resource
- Once the problem is solved administrator can reset the resource

```
% sdmadm reset_resource|rr -r <resource id>
```

- GEService will tries to reinstall the execd

SLOs

- A Need specifies
  - How many Resources are needed
  - How important the Resource Request is
  - What Resources are needed

```
quantity: 10
urgency:  50
resourceFilter:

    type = "host" & hardwareCpuCount > 1 &
    (operatingSystemName = "Solaris"  |
     operatingSystemName = "Linux")
```

SLOs

- Supported SLOs
  - FixedUsageSLO
    - gives any proper Resource a fixed Usage
    - produces no Needs

```
% sdmadm mc -c p31004
<common:componentConfig xsi:type="ge_adapter:GEServiceConfig"
                        mapping="default">
    <common:slos>
        <common:slo xsi:type="common:FixedUsageSLOConfig"
                    urgency="50"
                    name="fixed_usage"/>
    </common:slos>
</common:componentConfig>
```

SLOs

- Supported SLOs
  - MinResourceSLO
    - gives any proper Resource a fixed Usage
    - produces Needs if the service has not enough Resources

```
% sdmadm mod_component -c service1
<common:componentConfig xsi:type="ge_adapter:GEServiceConfig"
                        mapping="default">
    <common:slos>
        <common:slo xsi:type="common:MinResourceSLOConfig"
                    urgency="50"
                    min="10"
                    name="min_res"/>
    </common:slos>
</common:componentConfig>
```

SLOs

- Supported SLOs
  - PermanentRequestSLO
    - gives any proper Resource a fixed Usage
    - Has always a Need

```
% sdmadm mc -c spare_pool
<common:componentConfig xsi:type="spare_pool:SparePoolConfig"
                        mapping="default">
    <common:slos>
        <common:slo xsi:type="common:PermanentRequestSLOConfig"
                    urgency="1"
                    quantity="10"
                    name="PermanentRequestSLO">
        </common:slo>
    </common:slos>
</common:componentConfig>
```

SLOs

- Display SLO states

```
% sdmadm show_slo|sslo
service    slo                  quantity urgency request
-------------------------------------------------------------------
sge62      fixed_usage          0        0       SLO has no needs
spare_pool PermanentRequestSLO 10       1       type = "host"
```

- Display usage

```
% sdmadm sslo -u
service    slo                  resource usage
-----------------------------------------
p31004     fixed_usage          master   50
spare_pool PermanentRequestSLO foo      0
```

SLOs

- Requesting specific Resources
  - Any SLO allows the definition of a request filter

```
<common:componentConfig xsi:type="spare_pool:SparePoolConfig"
                        mapping="default">
    <common:slos>
        <common:slo ...>
           <common:request>
               type = “host” & hardwareCpuCount > 1 &
               (operatingSystemName = “Solaris” |
                operatingSystemName = “Linux”)
           </common:request>
        </common:slo>
    </common:slos>
</common:componentConfig>
```

SLOs

- Resource Filtering
  - Any SLO allows the definition of a resource filter. This filter limits the resources which will be considered by the SLO.

```
<common:componentConfig xsi:type="spare_pool:SparePoolConfig"
                        mapping="default">
    <common:slos>
        <common:slo ...>
           <common:resourceFilter>
               type = “host” & hardwareCpuCount > 1 &
               (operatingSystemName = “Solaris” |
                operatingSystemName = “Linux”)
           </common:resourceFilter>
        </common:slo>
    </common:slos>
</common:componentConfig>
```

SLOs

- Supported SLOs
  - MaxPendingJobsSLO
    - Can only be used for a Grid Engine Service
    - If the number of pending jobs exceeds a limit a Need is produced
    - It is possible to define a filter for matching jobs
    - Any Host Resource which has running jobs matching the job filter will get as Usage the Urgency of the MaxPendingJobsSLO

```
<common:slo xsi:type="ge_adapter:MaxPendingJobsSLOConfig"
              max="10"
              urgency="60"
              name="max_pending">
     <ge_adapter:jobFilter>
       arch matches "lx.*" & num_proc = 2
     </ge_adapter:jobFilter>
  </ge_adapter>
</common:slo>
```

Manage a Grid Engine cluster

- Requirements on qmaster side
  - JMX Agent must be enabled
  - SSL encryption for JMX Agent must be enabled
  - GE Service needs keystore for authentication
  - JMX agent must allow authentication with keystores

```
# install_master ...  -jmx ....
Grid Engine JMX MBean server
----------------------------
Please give some basic parameters for JMX MBean server
...
Using the following JMX MBean server settings.
   libjvm_path               >.../jre/lib/amd64/server/libjvm.so<
   Additional JVM arguments  ><
   JMX port                  >54322<
   JMX ssl                   >true<
   JMX client ssl            >true<
...
```

Manage a Grid Engine cluster

```
% sdmadm [global_options] add_ge_service|ags [-start]
         -h host_name -j jvm_name -s service_name [-f file_name]
```

- host_name should be the qmaster host (needs access to SGE_ROOT)
- jvm_name is normally rp_vm

```
<common:componentConfig xsi:type="ge_adapter:GEServiceConfig"
                        mapping="default">
    ...
    <ge_adapter:connection keystore="/var/spool/sgeCA/..."
                           password=""
                           username="sge_admin"
                           jmxPort="54322"
                           execdPort="31005"
                           masterPort="31004"
                           cell="default"
                           root="/opt/sge"
                           clusterName="p31004"/>
     ...
</common:componentConfig>
```

Manage a Grid Engine Cluster

- GE service component observes qmaster
  - If qmaster goes down the service state goes into UNKNOWN state
  - Restarting qmaster brings automatically service state into RUNNING state.
- GE service discovers resources owned by qmaster
  - execds running on a managed host will be shown as “normal” resources
  - execd on qmaster host will be a static resource. Can not be removed from the service

GEService

- How is the execd installation/uninstall performed
  - GEService executes a shell script on the executor of the host resource
  - This shell script gets as first parameter the path to a configuration file
  - shell script and config file is generated out of templates
  - In this templates placeholders will be replaced
  - The path to the templates can be defined in the `<execd>` element
    of the GEService configuration.
    default path to the templates is

```
<sdm dist>/util/templates/ge-adapter/install_execd.sh
<sdm dist>/util/templates/ge-adapter/install_execd.conf
<sdm dist>/util/templates/ge-adapter/uninstall_execd.sh
<sdm dist>/util/templates/ge-adapter/uninstall_execd.conf
```

GEService

- Exit codes of the exec installation script
  - 0 => has been successfully executed (resource will go into ASSINGED once qmaster reports the new execd)
  - 2 => execd installation could not been executed because it is not possible (e.g local spool dir does not exist)
    GEService will reject the resource, RP can assign it to different services
  - In all other cases GEService assumes that the execd installation failed and something on the host has been modified
    Admin has must have a look on it
    Reset of the Resource is necessary

GEService

- Complex to Resource Property Mapping
  - GEService automatically updates the resource properties of the assigned resource
  - Once and EXECD_MOD event occurs the reported complexes value are mapped into resource properties
  - The mapping can be configured
  - So also auto discovered host resources have resource properties

```
% sdmadm sr -r foo -all
service id     state    type flags usage annotation
-----------------------------------------------
p31004  foo    ASSIGNED host       50
    hardwareCpuArchitecture=amd64
    hardwareCpuCount=4
    operatingSystemName=Solaris
    resourceHostname=foo
```

Monitoring

- SDM records all actions on the resource with the Reporter component
- It is configurable how long the history is kept:

```
% sdmadm mc -c reporter
-------- editor output ---------------------------------
<reporter:reporter ...
    fileCount="4"
    fileSize="5242880"/>
```

- Report write the history in `<local spool dir>/spool/reporter` on the master host
- `sdmadm show_history` prints out the content of the history

Monitoring

```
% sdmadm shist -help
Usage: sdmadm [global_options] show_history|shist
    [-ed end_date] [-f advanced_filter] [-r resource] [-s service] [-sd
start_date] [-t type] [-hlp]

Show the data stored by reporter component.

Options:
-ed end_date                        End date.
-f advanced_filter                  The filter that could be compound with one
                                    or more filters
-r resource                         Resource name.
-s service                          Service name.
-sd start_date                      Start date.
-t type                             Event type.

% sdmadm show_history|shist -r foo
22/04/2008 08:00:19.940 RESOURCE_REJECTED          p31004       foo  Cannot execute install script on executor...
22/04/2008 08:00:20.443 RESOURCE_ADD               spare_pool   foo  Resource [foo] is going to be added to sp...
22/04/2008 08:00:20.445 RESOURCE_ADDED             spare_pool   foo  Resource [foo] has been added to spare po...
22/04/2008 08:11:19.685 RESOURCE_REMOVE            spare_pool   foo  Resource [foo] is going to be removed fro...
22/04/2008 08:11:19.691 RESOURCE_REMOVED           spare_pool   foo  Resource [foo] has been removed from spar...
22/04/2008 08:11:19.799 RESOURCE_ADD               p31004       foo  Got add resource request, assigning resou...
22/04/2008 08:11:20.122 RESOURCE_REJECTED          p31004       foo  Cannot execute install script on executor...
22/04/2008 08:11:20.715 RESOURCE_ADD               spare_pool   foo  Resource [foo] is going to be added to sp...
22/04/2008 08:11:20.722 RESOURCE_ADDED             spare_pool   foo  Resource [foo] has been added to spare po...
22/04/2008 08:12:30.030 RESOURCE_REMOVE            spare_pool   foo  Resource [foo] is going to be removed fro...
22/04/2008 08:12:30.031 RESOURCE_REMOVED           spare_pool   foo  Resource [foo] has been removed from spar...
...
```

Future Plans

- Enhance SLOs
- Manage Virtual Resources (like Licenses)
- ECO Computing
  - SparePool switches power off
- OS Provisioning and Virtualisation
  - xVm?
  - VirtualBox?
- More Service Adapters

Resources

- http://wikis.sun.com/display/GridEngine/Grid+Engine
- Open Source Project
  - http://hedeby.sunsource.net
  - users@hedeby.sunsource.net
  - dev@hedeby.sunsource.net
- Beta Packages
  - http://gridengine.sunsource.net/project/gridengine/downloads/62/download.html

Q & A

Service Domain Manager Basic and Concepts

Richard Hierlmeier
richard.hierlmeier@sun.com