Grid Engine Administration

Installation Considerations

This module covers

- Pre-install considerations
- Manual installation
- Automated installation
- The new GUI installer
- Spooling
- CSP Installation
- Shadow masters

Pre Install Verification

AKA *‘Things I wish someone had told me … ‘*

Forward and Reverse DNS Resolution

- SGE is obscenely sensitive to name resolution issues
  - Most installation failures tend to be hostname & DNS related
- Reverse DNS resolution is nice
  - Better to not have it than to have it badly configured
- Helpful hint:
  - Always test with the actual binaries SGE uses to query DNS
  - Verify that SGE utilbin binaries return same results as OS tools:

```
[root@dcore-amd sge-6s2u1]# hostname
dcore-amd.sonsorol.net

[root@dcore-amd sge-6s2u1]# /opt/sge-6s2u1/utilbin/lx26-amd64/gethostname
Hostname: dcore-amd.sonsorol.net
Aliases:  dcore-amd
Host Address(es): 66.92.70.152

[root@dcore-amd sge-6s2u1]# /opt/sge-6s2u1/utilbin/lx26-amd64/gethostbyaddr 66.92.70.152
Hostname: dcore-amd.sonsorol.net
Aliases:  dcore-amd
Host Address(es): 66.92.70.152
[root@dcore-amd sge-6s2u1]#
```

Other things to verify before you install

- Consistent UID/GID mapping
  - How you implement does not matter
  - What matters is that everything is globally consistent
  - Verify UID/GID cluster-wide for SGE admin account and others

- Make sure your chosen group ID range is *really* unique
  - SGE asks for a GID range to use internally
  - Used for resource utilization monitoring
  - Range is arbitrary but defines max # of jobs that can run on one host

- Verify shared file system mount options
  - Another thing that can get out of sync on a cluster and cause odd problems
  - Root squash OK
  - SETUID squash not OK

Tips for Apple OS X people …

- Just because Mac OS X lets you put spaces and funky capitalization into your hostname does not mean that this is a good thing to do.

- Your qmaster machine does not need to be called ”j0ez fuNky Xserve”.

- Feel free to do whatever you want with the computer name as it applies to “Bonjour” (multicast DNS) network sharing, but keep the core system hostname something reasonable.

- Grid Engine and other Unix-ish bits under the hood of your OS X system will thank you for doing this.

- Actually, now that I’m on this topic, use the same conservative naming approach for XRAID storage arrays and local disk partitions.

Tip for SGE 6.2u3 Users

- Take a look at the installer GUI
  - Seriously. It’s nice.
- Find it at:
  - $SGE_ROOT/start_gui_installer

Pre Install Decisions

Decisions to make …

- How many cells will there be?
  - Stick with the default cell name of ‘default’ is easiest
- v6.2 and later
  - Pick a “name” for your cluster
  - Or else be prepared to be confused by “sge.6444” startup scripts
- Allocate roles among your hosts:
  - Master host
  - Shadow master host
  - Admin host
  - Submit host
  - Execution host
- Layout and location of SGE root
  - Shared vs local
- Administrative user account name

Decisions to make ...

- Enable JMX?
- Install/enable SDM?
- Using CSP-secured mode?
- Service and port definitions
  - /etc/services vs. NIS vs. environment variables
  - IANA recently assigned port numbers:

```
sge_qmaster         6444/tcp   Grid Engine Qmaster Service
sge_qmaster         6444/udp   Grid Engine Qmaster Service
sge_execd           6445/tcp   Grid Engine Execution Service
sge_execd           6445/udp   Grid Engine Execution Service
```

- Classic vs. Berkeley spooling?
- Combined execd spooling or local execd spooling?
  - By default exec hosts will log into the shared SGE root
  - For performance reasons, local non-shared directory can be specified
  - Typically a performance vs. convenience decision
- Decide on a first pass queue structure
- Think about the first pass policy configuration

Installation user account

- If not 'root'
  - Only that user can use grid engine
  - Qrsh, qtcsh, qmake and tight PE jobs prohibited
- Installation as root generally required
- Running as root not required
  - SGE can run as an unprivileged user

File Permissions & File systems

- Unprivileged SGE user must have consistent read/write access to the SGE root directory on all hosts
- NFS root-squash is OK
- setuid squash not OK
  - SGE will perform setuid operations to “become” the user who submitted a task

About “GID Range”

- Each job gets an additional job ID
  - Attached to job and all child processes
  - SGE uses this to track wayward tasks
- execd_param   ENABLE_ADDGRP_KILL=true
  - Additional group ID used for killing jobs
- gid_range defines the values for these supplementary ids
  - Is also a limit on “max jobs per host”
  - Cant have more jobs than range in gid_range

Spooling

- Very important decision
  - Unlike almost every other SGE option, spooling method can’t be changed without reinstallation
- Two choices
  - Binary Spooling (via berkeley-db)
  - Classic Spooling (plaintext files)

Binary Spooling

- Currently the default option
- Advantages
    - SGE developers don’t reinvent the wheel
        - Let database pros handle the database
        - Leverage features and future work of bdb4 developers
            - Replication, failover, etc.
    - Performance
        - If you need to perform 150 qsubs per second …

Binary Spooling

- Disadvantages
  - Critical state files now in binary form
  - Grid Engine H/A features are compromised if NFSv3 used
    - NFSv4 required for berkelyDB files on NFS
  - Qmaster can spool to a remote Berkeley DB server via RPC
    - This allows use of shadow masters*
      - *Pending job scripts are not spooled to the BDB RPC server
    - RPC has no real security model

Classic Spooling

- Advantages
  - Plaintext flat files
  - Easy to backup, rsync, edit, etc.
  - Easy H/A options
    - Master and all shadow masters simply share a common NFS mount & all spool files

Classic Spooling

- Disadvantages
  - Performance
  - Need to beware of OS level open filehandle limits in some cases
  - Performance hit possible on any system with extremely high task throughput
    - Many tens of thousands of jobs per day …

High Availability Approaches

- Classic
  1. Make the NFS fileserver fast and H/A
     - Use standard shadow master failover techniques
- Binary
  1. Use NFSv4
  2. Find a parallel/cluster filesystem for master and shadow hosts that does not break berkeley-db usage
  3. Build a clustered-for-HA RPC database host*
     - *Otherwise RPC database host is a single point of failure
     - RPC spool over secure network to the H/A database host
     - Test what happens to pending jobs when failover occurs

My $.02 on spooling

- Disadvantages of binary spooling may outweigh the benefits
- Most sites should start with classic spooling
  - Small systems or low-throughput sites will not notice any performance difference
  - For sites that do encounter performance hits
    - Not that hard to capture current SGE config and simply reinstall SGE w/ binary spooling enabled
- If H/A is a requirement
  - Far safer and conservative to stick with classic spooling

Submit & Admin Hosts

- There may be more than you think!
- Some pre-install considerations
  - All nodes should be submit hosts when ..
    - Have a workflow involving active tasks that may submit new work or alter other tasks
  - All nodes should be administrative hosts ..
    - So that nodes can auto-provision themselves
      - Not a SGE thing but a commonly encountered site practice

Shadow Masters

Shadow Masters

- Primary failover system for SGE
- Specific implementation may depend on how $SGE_ROOT is shared and spooling method used
- Primary requirements
    - All shadow masters have $SGE_ROOT access
    - All running `sge_shadowd` daemon
    - All shadow masters listed in shadow_masters file
        - `$SGE_ROOT/$SGE_CELL/common/shadow_masters`

Shadow Masters

- Parameters to care about
  - `SGE_CHECK_INTERVAL`
    - How often sge_shadowd checks heartbeat file
      - `$SGE_ROOT/$SGE_CELL/spool/qmaster/heartbeat`
  - `SGE_GET_ACTIVE_INTERVAL`
    - How long the heartbeat file needs to be unchanged before a shadow takeover is initiated
  - `SGE_DELAY_TIME`
    - Controls length of shadowd pause when takeover fails on systems with multiple shadow masters

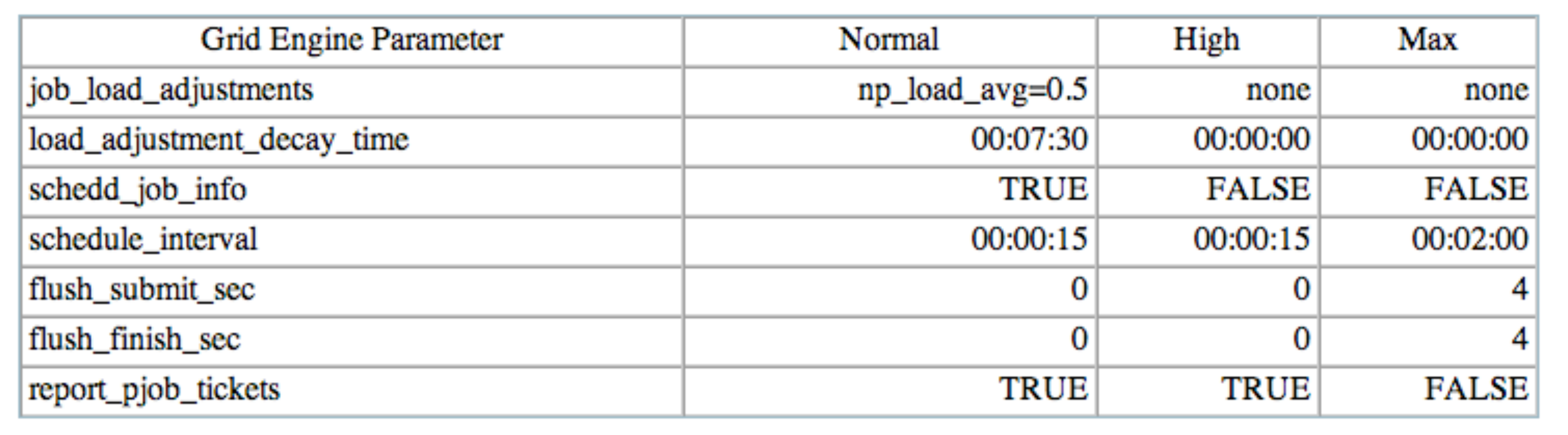
Shadow Masters

- How it works
  1. Qmaster updates heartbeat file every 30 seconds
  2. Shadow checks heartbeat according to SGE_CHECK_INTERVAL
  3. If shadow discovers no heartbeat change, pause for one more SGE_CHECK_INTERVAL
  4. If still no change, start waiting on SGE_GET_ACTIVE_INTERVAL
  5. If still no change, start takeover

Other

Predefined Tuning Profiles

- Some tuning options offered during install
  - Normal, High & Max
  - Not a big deal during install, whatever is chosen can trivially be changed later
  - What actually changes:



Grid Engine Parameter	Normal	High	Max
job_load_adjustments	np_load_avg=0.5	none	none
load_adjustment_decay_time	00:07:30	00:00:00	00:00:00
schedd_job_info	TRUE	FALSE	FALSE
schedule_interval	00:00:15	00:00:15	00:02:00
flush_submit_sec	0	0	4
flush_finish_sec	0	0	4
report_pjob_tickets	TRUE	TRUE	FALSE

Grid Engine “host_aliases” file

- Deal with multi-homed hosts
  - Very common problem:
    - ./act_qmaster is a FQDN not reachable by cluster compute nodes
    - host_aliases is the solution

- Default location
  - $SGE_ROOT/$SGE_CELL/common/host_aliases
- Simple format:
  - <name>   <alias to use>

```
chrisdag-aliased              10.10.10.99
chrisdag.colo.bioteam.net     10.10.10.99
chrisdag.local                10.10.10.99
dhcp-034-192.gfdl.noaa.gov    10.10.10.99
```

Grid Engine “sge_aliases” file

- Alias file system paths
  - $SGE_ROOT/$SGE_CELL/common/sge_aliases
  - Format
    - `<src path>    <submit host>   <exec host>   <replacement path>`
- Very useful when
  - SGE uses a path that exists on the qmaster but nowhere else
- Example
  - Small cluster, qmaster node mounts SAN volume for NFS export:

```
/Volumes/XSAN/VOL1/Users   *    * /Users
```

Template driven autoinstallation

SGE Auto installation tools can be flaky*

- Fail silently when problems are encountered
- Syntax of the install templates is pretty picky and sensitive to typos, spaces & mistakes
- Assume passwordless RSH/SSH remote command execution already exists
- Very often I find:
  - Manual installation on smaller clusters (30 nodes or less) is easier
    - Far faster than test/debug/fix/test cycle with the SGE autoinstall tools
- If you have a large cluster (and passwordless SSH)
  - Often a better practice to roll your own scripts to automate SGE setup/teardown on compute nodes
- If you want to stay with the SGE auto install tools
  - Start with a “known good” template from a friend or the mailing list
  - Test after each minor modification

- * My biased opinion, of course!

SGE Auto Installation (Remote)

- Start with a copy of template:

```
# cd $SGE_ROOT/$SGE_CELL/util/install_modules/
# cp ./inst_template.conf $SGE_ROOT/config.txt
```

- Install qmaster + execd on master host:

```
# cd $SGE_ROOT
# ./inst_sge -m -x -auto ./config.txt
```

SGE Autoinstallation (local)

- Kickstart or system imaging friendly
- Scripted SSH into node, or %post script:

```
cd /usr/local/sge;
./inst_sge -x -auto -noremote ./template.conf
```

When auto install fails

- Check /tmp/ for installation log messages
- Edit the “inst_sge” script
  - Trigger verbose output
    - Edit first line:
      - “#!/bin/sh -x”
- Rinse, repeat …

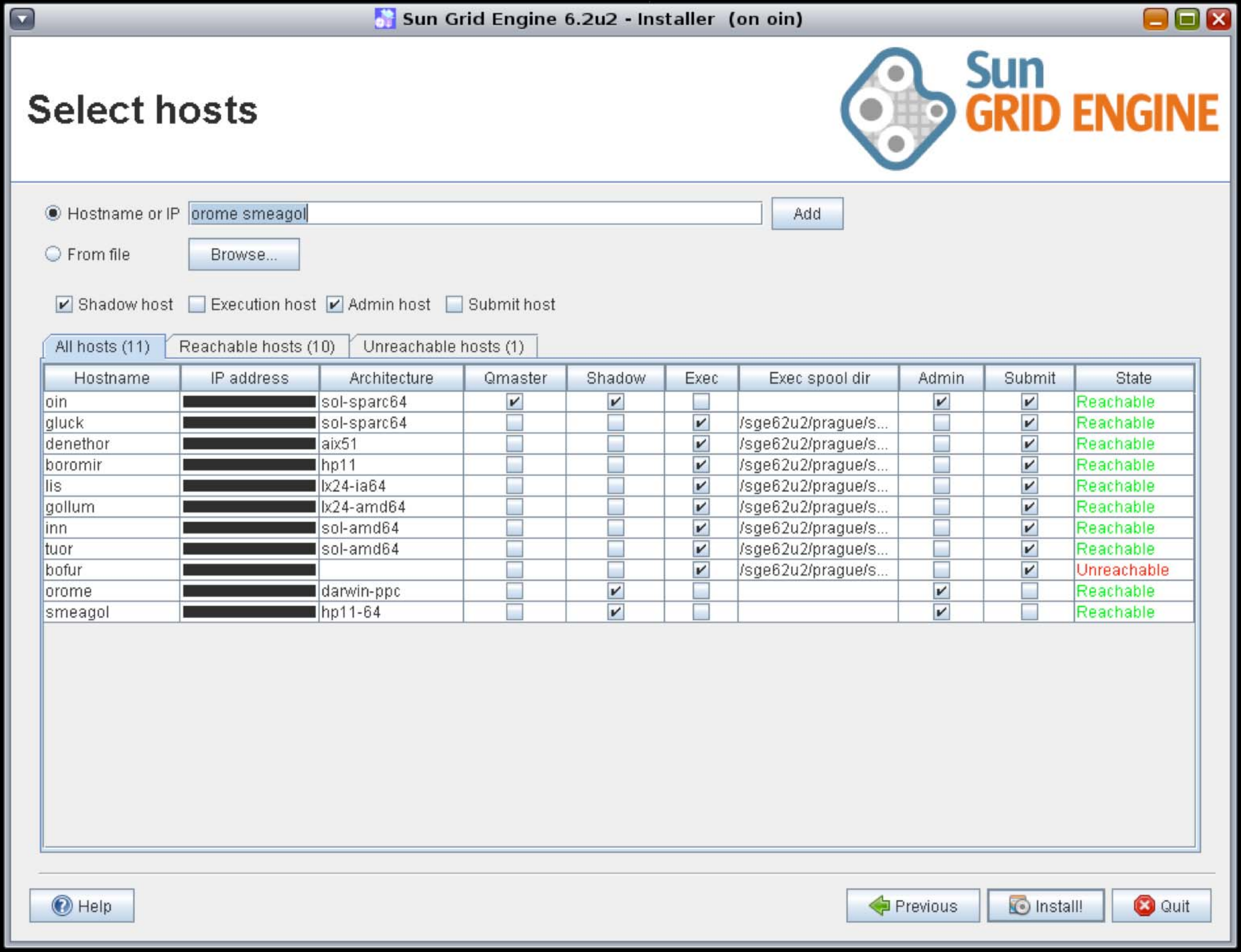
CSP ‘Secure’ Mode

- Certificate Security Protocol
  - Based on OpenSSL
- Only security features provided:
  - Access control
    - Users, hosts all need certificates to communicate with the SGE qmaster
  - Encryption
    - Communication traffic encrypted

Installing in CSP Mode

- Set up the certificate authority (CA)
  - `# ./install_qmaster -csp`
  - … once CA is setup, standard qmaster install continues
- Create user list
  - Automated script then creates user keys
- User keys installed:
  - `$SGE_ROOT/$CELL/util/sgeCA/sge_ca -copy`

New GUI Installer (since 6.2u2)



Sun Grid Engine 6.2u2 - Installer (on oin)

Select hosts

● Hostname or IP: orome smeagol

○ From file

☑ Shadow host □ Execution host ☑ Admin host □ Submit host

All hosts (11) | Reachable hosts (10) | Unreachable hosts (1)

Hostname	IP address	Architecture	Qmaster	Shadow	Exec	Exec spool dir	Admin	Submit	State
oin	[redacted]	sol-sparc64	☑	☑	□		☑	☑	Reachable
gluck	[redacted]	sol-sparc64	□	□	☑	/sge62u2/prague/s...	□	☑	Reachable
denethor	[redacted]	aix51	□	□	☑	/sge62u2/prague/s...	□	☑	Reachable
boromir	[redacted]	hp11	□	□	☑	/sge62u2/prague/s...	□	☑	Reachable
lis	[redacted]	lx24-ia64	□	□	☑	/sge62u2/prague/s...	□	☑	Reachable
gollum	[redacted]	lx24-amd64	□	□	☑	/sge62u2/prague/s...	□	☑	Reachable
inn	[redacted]	sol-amd64	□	□	☑	/sge62u2/prague/s...	□	☑	Reachable
tuor	[redacted]	sol-amd64	□	□	☑	/sge62u2/prague/s...	□	☑	Reachable
bofur	[redacted]		□	□	☑	/sge62u2/prague/s...	□	☑	Unreachable
orome	[redacted]	darwin-ppc	□	☑	□		☑	□	Reachable
smeagol	[redacted]	hp11-64	□	☑	□		☑	□	Reachable

SGE training, consulting and special projects - BioTeam Inc. - http://www.bioteam.net

GUI Installer

- Very nice!
  - Beta release late in ‘08
  - Flickr tour of beta with comments:
    - http://www.flickr.com/photos/chrisdag/sets/72157611344682697/
  - Production release with SGE 6.2u2
- Java driven
  - Requires Java 5 or later
  - Works fine with X11 over SSH
  - Lubomir’s prep screencast:
    - http://blogs.sun.com/lubos/entry/preparing_the_environment_for_sge

GUI Installer comments

- Historically SGE power users/admins do not use the GUI tools
- Full auto cluster install requires passwordless access anyway
- But …
- What I like about the GUI
  - Wildcard hostnames! IP address ranges
  - Nice to see non-Motif GUI development within SGE
  - Easier than the template-driven autoinstall
  - In particular I like the pre-install testing that it does

Lab Time - SGE Installation

- Task 1
  - Full manual install qmaster & execd
- Task 2
  - Build a template; perform full automatic install
- Task 3
  - Experiment with GUI installer

After installation ...

- Check out your bootstrap file
  - `cat $SGE_ROOT/$SGE_CELL/common/bootstrap`

Final note for Mac users

- SGE's init scripts are not reliable on modern Apple OS X systems
  - Apple has switched from SystemStarter() to launchd() framework
  - SGE seems unreliable now under the old SystemStarter framework
  - BioTeam has published launchd script creator tools:
    - http://blog.bioteam.net/2008/07/15/sge-launchd-script-maker-for-apple-os-x-105-leopard/

Grid Engine Upgrades & Backup

Upgrade Options

- SGE 5.x to 6.0 or 6.1
  - Changes between 5.x and 6.x are so fundamental a clean reinstall is almost always the best option
- SGE 6.0x to 6.1x Upgrade
  - Upgrade scripts for 6.0u2 and later
  - Prior to 6.0u2 a clean reinstall is best
- Point updates (Example: 6.1u3 -> 6.1u4)
  - Upgrade scripts not necessary
    1. Move sge_shepherds if needed
    2. Shutdown SGE
    3. Drop new binaries into place; restart SGE

6.0 to 6.1 Upgrades

- Tutorial by Marco
  - http://gridengine.sunsource.net/servlets/ReadMsg?list=users&msgNo=21820
  - This is also linked on http://gridengine.info and shows up in Google searches
- Summary
  1. Backup existing system*
  2. Shut down existing system
  3. Unpack new distribution
  4. Run "./inst_sge -upd" to upgrade spool
  5. Restart SGE

Performing Point Release Upgrades

- Point Release Upgrade Howto for 6.0
  - http://gridengine.sunsource.net/install60patch.txt

- Point Release Upgrade Howto for 6.1
  - http://gridengine.sunsource.net/install61patch.txt

- Point Release Upgrade Howto for 6.2
  - http://gridengine.sunsource.net/install62patch.txt

Bugfix Lists

- 6.2 Issues Fixed
  - http://gridengine.sunsource.net/project/gridengine/62patches.txt
- 6.1 Issues Fixed
  - http://gridengine.sunsource.net/project/gridengine/61patches.txt
- 6.0 Issues Fixed
  - http://gridengine.sunsource.net/project/gridengine/60patches.txt
- Comments
  - Extremely useful docs
  - If you need more info on a particular issue
    - Go to http://gridengine.sunsource.net/servlets/ProjectIssues
    - Type in the Issue Number and press “Find”
    - http://gridengine.info offers HTML version w/ Issue links embedded into the document

Grid Engine Backups

- Backup

  -
    ```
    cd $SGE_ROOT; ./inst_sge -bup
    ```
- Restore from backup

  -
    ```
    cd $SGE_ROOT; ./inst_sge -rst
    ```
- Backup scripts are nice
  - Can be template-driven (automated)
  - Makes nice datestamped tarballs
- Classic spooling & feeling lazy?
  - rsync is your friend!

    -
      ```
      mkdir sge-backup; rsync -av $SGE_ROOT ./sge-backup/
      ```

Questions?

- Optional diversions we can pursue if there is interest…
  1. Change execd spool location to simulate switch from NFS to local disk spooling
  2. Examine file and directory differences in classic vs. binary spooling installations
  3. Scheduler profiles: activate ‘on demand’ scheduling