CLUSTER® RESOURCES INC.

Prioritization and Fairshare

Why Job Prioritization and Fairshare?

- Maximize system utilization while…
- Giving preference to specific users and projects while…
- Ensuring users' jobs do not sit in the queue too long.

Basically, balancing site goals with fairness

Fairness

- Definition:
  - giving all users equal access to compute resources
  - incorporating historical resource usage, political issues, and job *value*

Moab provides a comprehensive and flexible set of tools allowing the ability to address the many and varied fairness management needs.

http://clusterresources.com/moabdocs/6.0managingfairness.shtml

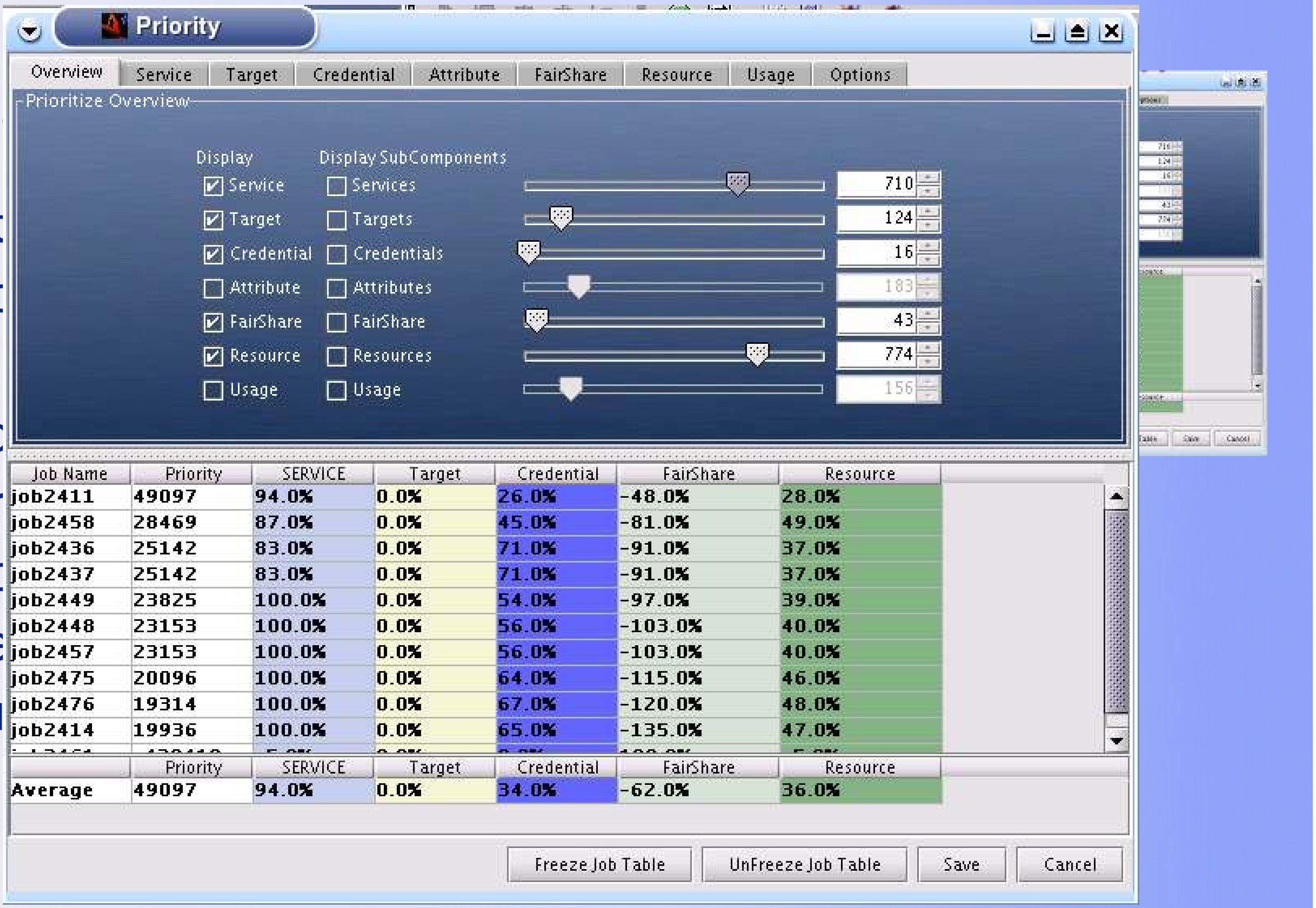
General Fairness Strategies

- Maximize Scheduler Options -- Do Not Overspecify
- Keep It Simple – Do Not Address Hypothetical Issues
- Seek To Adjust User Behaviour,
  Not Limit User Options
- Allow Users to Specify Required Service Level
- Monitor Cluster Performance Regularly
- Tune Policies As Needed

CLUSTER RESOURCES INC.

Pri

- 2-t
- Inc
  su
- Cc
  tar
  job
- Ne
- Tu



Display	Display SubComponents	
☑ Service	□ Services	710
☑ Target	□ Targets	124
☑ Credential	□ Credentials	16
□ Attribute	□ Attributes	183
☑ FairShare	□ FairShare	43
☑ Resource	□ Resources	774
□ Usage	□ Usage	156

Job Name	Priority	SERVICE	Target	Credential	FairShare	Resource
job2411	49097	94.0%	0.0%	26.0%	-48.0%	28.0%
job2458	28469	87.0%	0.0%	45.0%	-81.0%	49.0%
job2436	25142	83.0%	0.0%	71.0%	-91.0%	37.0%
job2437	25142	83.0%	0.0%	71.0%	-91.0%	37.0%
job2449	23825	100.0%	0.0%	54.0%	-97.0%	39.0%
job2448	23153	100.0%	0.0%	56.0%	-103.0%	40.0%
job2457	23153	100.0%	0.0%	56.0%	-103.0%	40.0%
job2475	20096	100.0%	0.0%	64.0%	-115.0%	46.0%
job2476	19314	100.0%	0.0%	67.0%	-120.0%	48.0%
job2414	19936	100.0%	0.0%	65.0%	-135.0%	47.0%

	Priority	SERVICE	Target	Credential	FairShare	Resource
Average	49097	94.0%	0.0%	34.0%	-62.0%	36.0%

How Moab Calculates Priority

<COMPONENT WEIGHT>

*

<SUBCOMPONENT WEIGHT>

*

<PRIORITY SUBCOMPONENT VALUE>

- Component default weight = 1
- Subcomponent default weight = 0
  - QUEUETIME = 1

Job Prioritization – Component Overview

- Service
  - Level of service delivered or anticipated
  - Includes queue time, xfactor, bypass, policy violation, startcount, deadline, and user priority
- Target
  - Desired service level - scheduler does ‘all in its power’ to meet scheduling targets
  - Provides exponential factor growth
  - Includes target queue time, target xfactor

Job Prioritization – Component Overview

- Credential
  - Based on credential priorities
  - Includes user, group, account, QoS, and class

```
# moab.cfg

CREDWEIGHT      1
USERWEIGHT      1
GROUPWEIGHT      10

USERCFG[john]   PRIORITY=2000
USERCFG[paul]   PRIORITY=1000
GROUPCFG[staff] PRIORITY=10000
```

Job Prioritization – Component Overview

- Resource
  - Based on requested resources
  - Includes nodes, processors, memory, swap, disk, walltime, proc-seconds and proc-equivalents
- Resource Scenarios:
  - Favor large resource jobs
  - Level the response time distribution across large and small jobs
  - Improve system utilization
- Golf ball and sand analogy

Job Prioritization – Component Overview

- Usage
  - Based on utilized resources
  - Includes resources utilized, resources remaining, percent walltime consumed, and execution time
  - Useful in preemption based scheduling
- Fairshare
  - Includes user, group, account, QoS, and class fairshare
  - Includes current based on historical resource consumption
  - usage metric of jobs per user, procs per user, and ps per user
  - May allow prioritization with ‘cap’ fairshare target
  - Steer workload toward a particular usage mix across credentials

http://www.clusterresources.com/moabdocs/5.1.2priorityfactors.shtml#attr

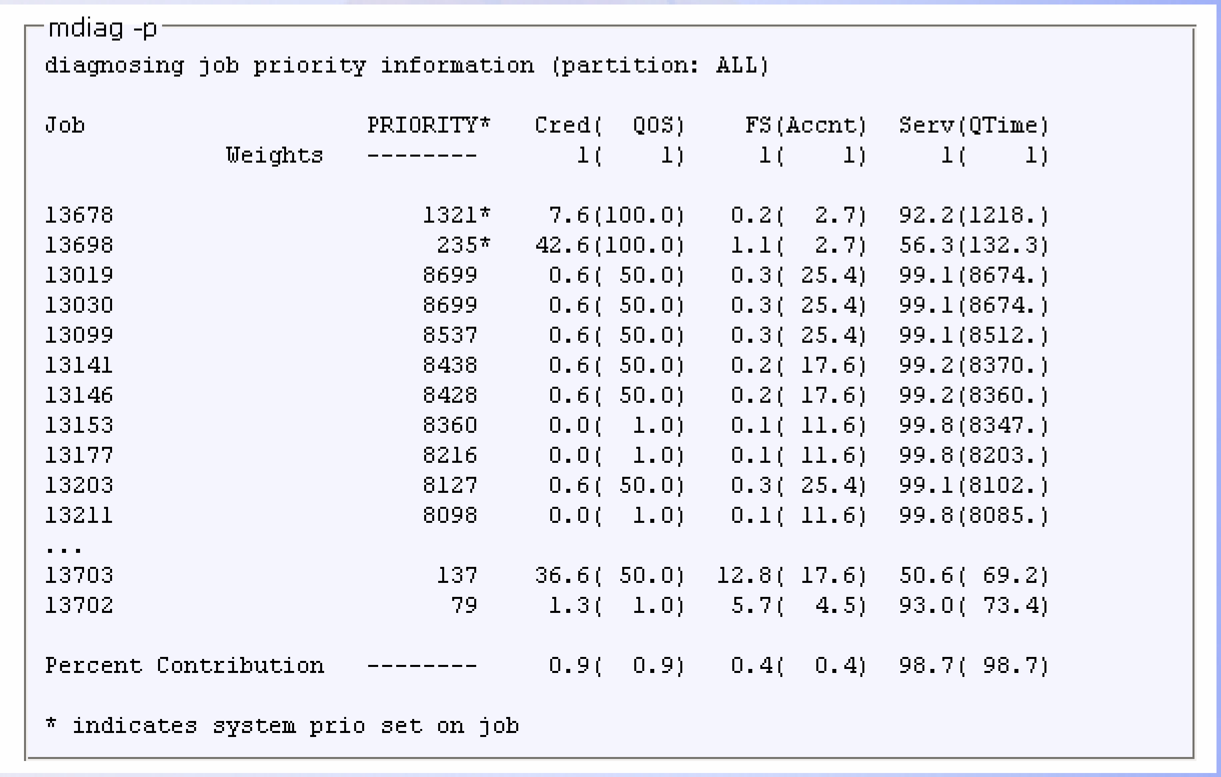
http://www.clusterresources.com/moabdocs/5.1.2priorityfactors.shtml#usage

Job Prioritization – Component Overview

- Job Attribute
  - Allows prioritization based on:
    - current job state – (ie. favor suspended jobs)
    - job's requested node features
    - job attributes (ie, preemptible or interactive)
    - requested licenses, network consumption, or generic resource requirements
  - Useful in preemption based scheduling

```
# moab.cfg

ATTRWEIGHT 100
ATTRATTRWEIGHT 1
ATTRSTATEWEIGHT 1
ATTRGRESWEIGHT 5

# favor suspended jobs
# disfavor preemptible jobs
# favor jobs requesting 'matlab'

JOBPRIOF   STATE[Running]=100   STATE[Suspended]=1000   ATTR[PREEMPTEE]=-200   ATTR[gpfs]=3   GRES[matlab]=400

# map node features to job features

 NODETOJOBATTRMAP   gpfs,pvfs
```

mdiag -p



mdiag -p

```
diagnosing job priority information (partition: ALL)

Job                    PRIORITY*   Cred(  QOS)    FS(Accnt)  Serv(QTime)
             Weights   --------       1(    1)     1(    1)     1(    1)

13678                      1321*    7.6(100.0)   0.2(  2.7)  92.2(1218.)
13698                       235*   42.6(100.0)   1.1(  2.7)  56.3(132.3)
13019                      8699     0.6( 50.0)   0.3( 25.4)  99.1(8674.)
13030                      8699     0.6( 50.0)   0.3( 25.4)  99.1(8674.)
13099                      8537     0.6( 50.0)   0.3( 25.4)  99.1(8512.)
13141                      8438     0.6( 50.0)   0.2( 17.6)  99.2(8370.)
13146                      8428     0.6( 50.0)   0.2( 17.6)  99.2(8360.)
13153                      8360     0.0(  1.0)   0.1( 11.6)  99.8(8347.)
13177                      8216     0.0(  1.0)   0.1( 11.6)  99.8(8203.)
13203                      8127     0.6( 50.0)   0.3( 25.4)  99.1(8102.)
13211                      8098     0.0(  1.0)   0.1( 11.6)  99.8(8085.)
...
13703                       137    36.6( 50.0)  12.8( 17.6)  50.6( 69.2)
13702                        79     1.3(  1.0)   5.7(  4.5)  93.0( 73.4)

Percent Contribution   --------     0.9(  0.9)   0.4(  0.4)  98.7( 98.7)

* indicates system prio set on job
```

Service Level Priority Example

- A site wants to do the following:
  - Favor jobs in the low, medium, and high QOS's so they will run in QOS order
  - balance job expansion factor
  - use job queue time to prevent jobs from starving

```
# moab.cfg

QOSWEIGHT                      1
XFACTORWEIGHT                  1
QUEUETIMEWEIGHT               10
TARGETQUEUETIMEWEIGHT    1

QOSCFG[low]           PRIORITY=1000
QOSCFG[medium]       PRIORITY=10000
QOSCFG[high]          PRIORITY=100000
QOSCFG[DEFAULT]    QTTARGET=4:00:00
```

Credential and Service Priority Example

```
# moab.cfg

# Service Priority Factors
SERVWEIGHT 1
XFACTORWEIGHT 10
QUEUETIMEWEIGHT 1000

# Credential Priority Factors
CREDWEIGHT 1
USERWEIGHT 1
CLASSWEIGHT 2
USERCFG[john]                 PRIORITY=200
CLASSCFG[batch]               PRIORITY=15
CLASSCFG[debug]               PRIORITY=100       XFWEIGHT=100
ACCOUNTCFG[bottomfeeder]      PRIORITY=-5000     QTWEIGHT=1     XFWEIGHT=0
```

Priority Caps

Limit the priority contribution due to a particular priority factor

```
#moab.cfg

XFACTORWEIGHT          1
XFACTORCAP             1000

QUEUETIMEWEIGHT 10
QUEUETIMECAP           1000

QOSWEIGHT          1
QOSCAP                 10000
```

Manual Job Priority Adjustment

Sometimes you need to….

- Run an admin test job as soon as possible
- Pacify a disserviced user

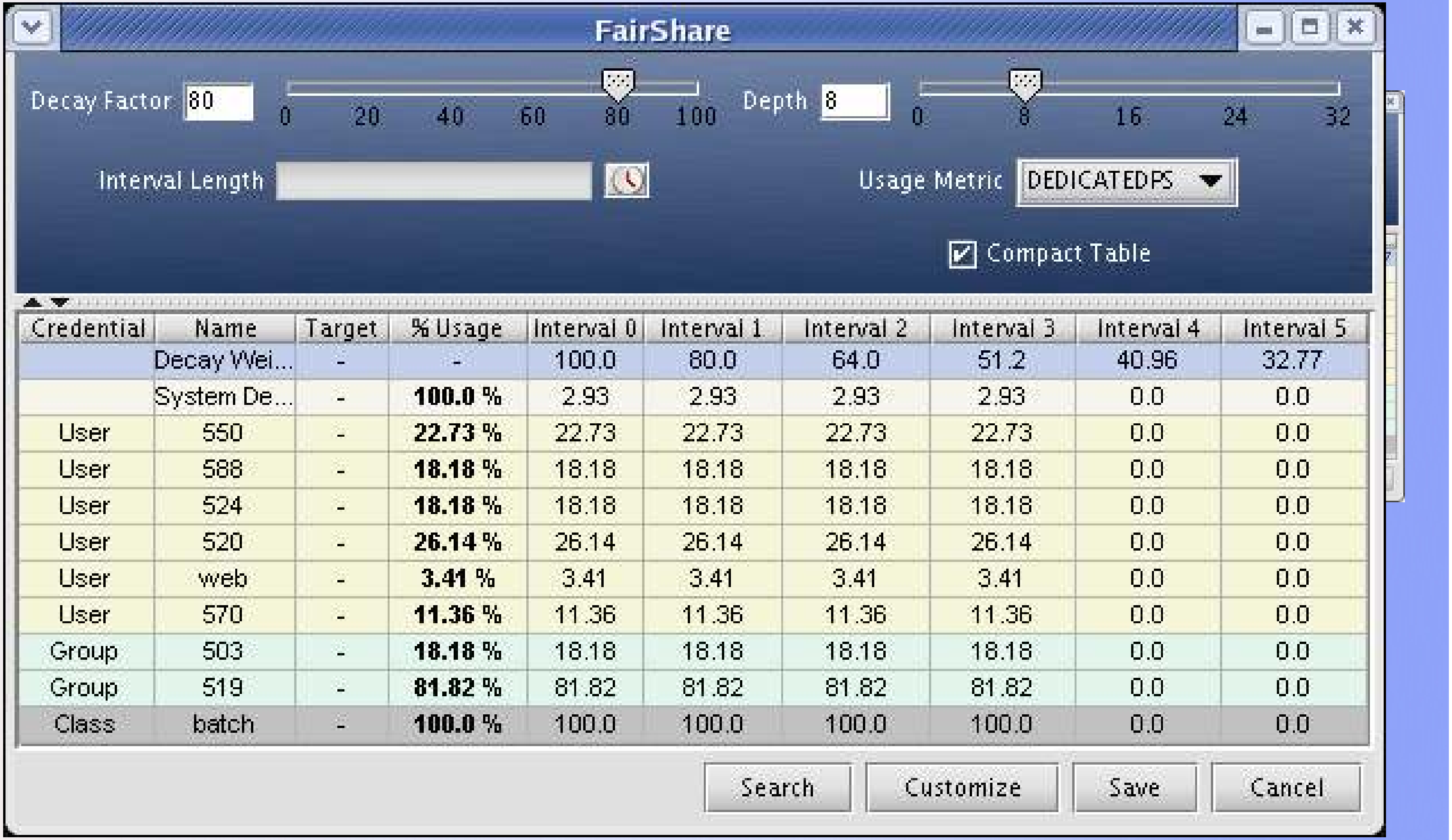
Use the Setspri command:

- setspri [-r] *priority jobid*

Example: setspri 1 cluster.25

User Selectable Priority with QOS

- Enable Access to multiple QOS with own charging rate, priority and target service levels
- Based on job importance, users can select the desired QOS
- Allows users to jump ahead of other users if they are willing to pay the associated costs



Decay Factor 80

Depth 8

Interval Length

Usage Metric DEDICATEDPS

☑ Compact Table

Credential	Name	Target	% Usage	Interval 0	Interval 1	Interval 2	Interval 3	Interval 4	Interval 5
	Decay Wei...	-	-	100.0	80.0	64.0	51.2	40.96	32.77
	System De...	-	**100.0 %**	2.93	2.93	2.93	2.93	0.0	0.0
User	550	-	**22.73 %**	22.73	22.73	22.73	22.73	0.0	0.0
User	588	-	**18.18 %**	18.18	18.18	18.18	18.18	0.0	0.0
User	524	-	**18.18 %**	18.18	18.18	18.18	18.18	0.0	0.0
User	520	-	**26.14 %**	26.14	26.14	26.14	26.14	0.0	0.0
User	web	-	**3.41 %**	3.41	3.41	3.41	3.41	0.0	0.0
User	570	-	**11.36 %**	11.36	11.36	11.36	11.36	0.0	0.0
Group	503	-	**18.18 %**	18.18	18.18	18.18	18.18	0.0	0.0
Group	519	-	**81.82 %**	81.82	81.82	81.82	81.82	0.0	0.0
Class	batch	-	**100.0 %**	100.0	100.0	100.0	100.0	0.0	0.0

Fairshare Parameters

- FSINTERVAL - duration of each fairshare window
- FSDEPTH - number of fairshare windows factored into current fairshare utilization
- FSDECAY - decay factor applied to weighting the contribution of each fairshare window
- FSPOLICY - metric to use when tracking fairshare usage

FSINTERVAL and FSDEPTH

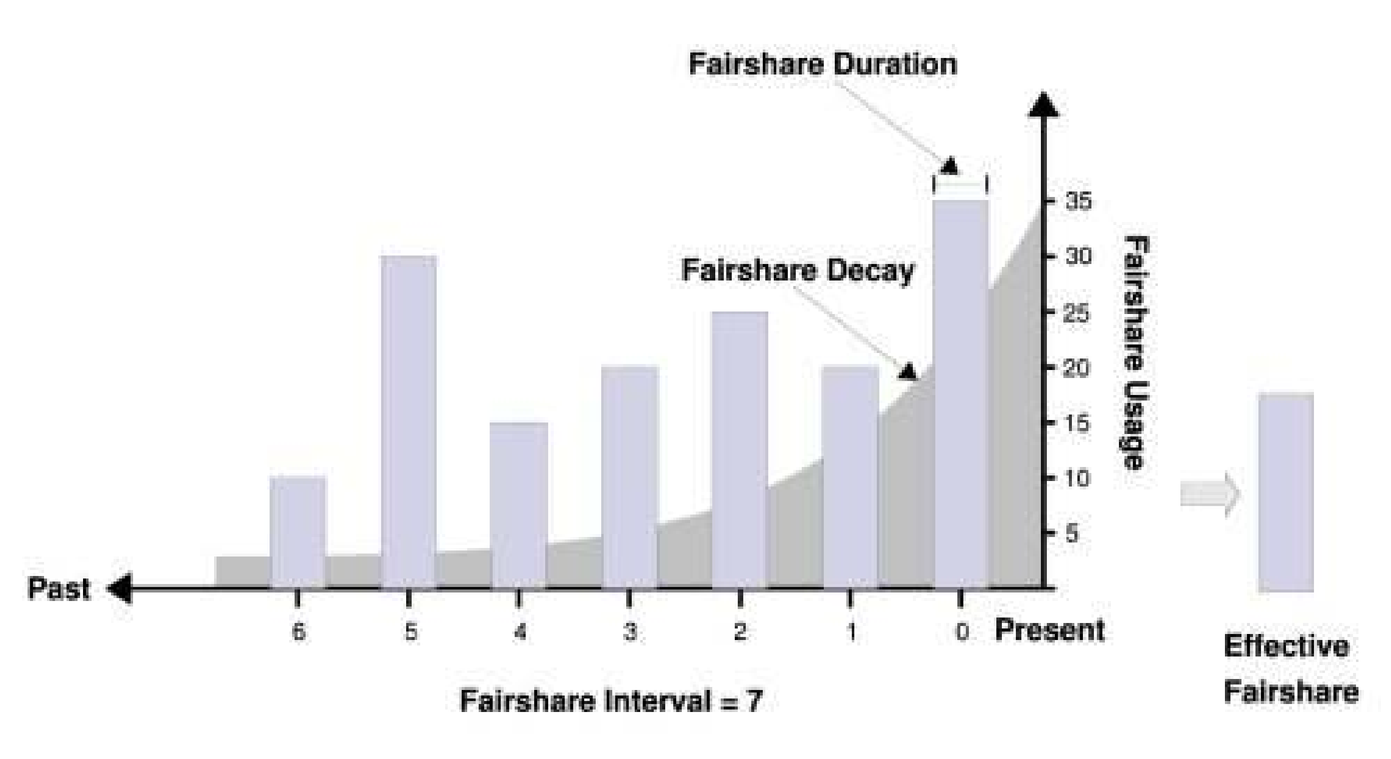
FSINTERVAL * FSDEPTH = Total time evaluated by fairshare

FSDECAY

- Value between 0 and 1
- Smaller the value, the more rapid decay
- More windows will causes decay factor to degrade more quickly

FSPOLICY

- DEDICATEDPES - processor-equivalent seconds
- DEDICATEDPS - processor seconds
- PDEDICATEDPS - processor seconds dedicated to each job with per node usage scaled by the node processor speed attribute.
- SDEDICATEDPES - processor-equivalent seconds dedicated to each job with per node usage scaled by the node speed attribute.
- UTILIZEDPS - processor seconds utilized by each job


Fairshare Duration
Fairshare Decay
35
30
25
20
15
10
5
Fairshare Usage
Past
6
5
4
3
2
1
0
Present
Fairshare Interval = 7
Effective
Fairshare

Fairshare Targets

Affects Job Priority

Target Type	Description
Ceiling (-)	Only adjusts job priority down when usage exceeds target
Floor (+)	Only adjusts job priority up when usage falls below target
Targets	Adjusts job priority up or down to meet target

Fairshare Target Example

```
# moab.cfg

FSWEIGHT    1
FSUSERWEIGHT 100

USERCFG[john] FSTARGET=16.5+
USERCFG[DEFAULT] FSTARGET=10
```

Fairshare Caps

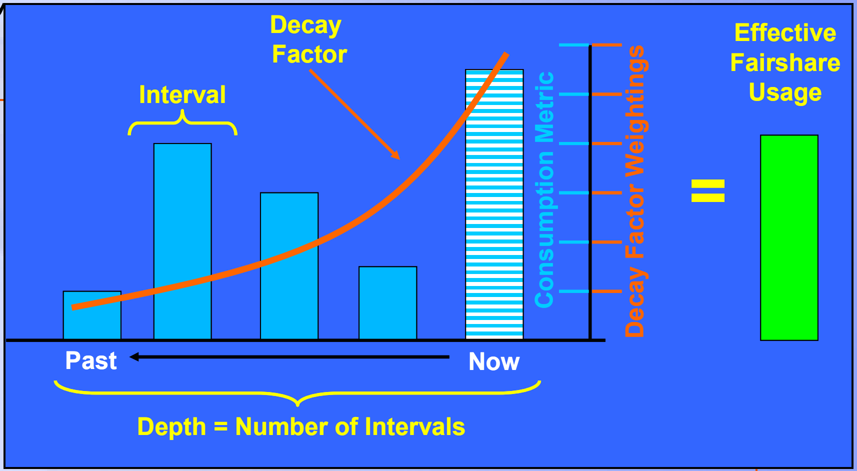
Affects Job Eligibility

```
# moab.cfg

FSPOLICY   DEDICATEDPS
FSINTERVAL  12:00:00
FSDEPTH 14

ACCOUNTCFG[marketing]    FSCAP=16500
ACCOUNTCFG[DEFAULT]    FSCAP=10%
```

CLUSTER RESOURCES INC.

Fairshare Example

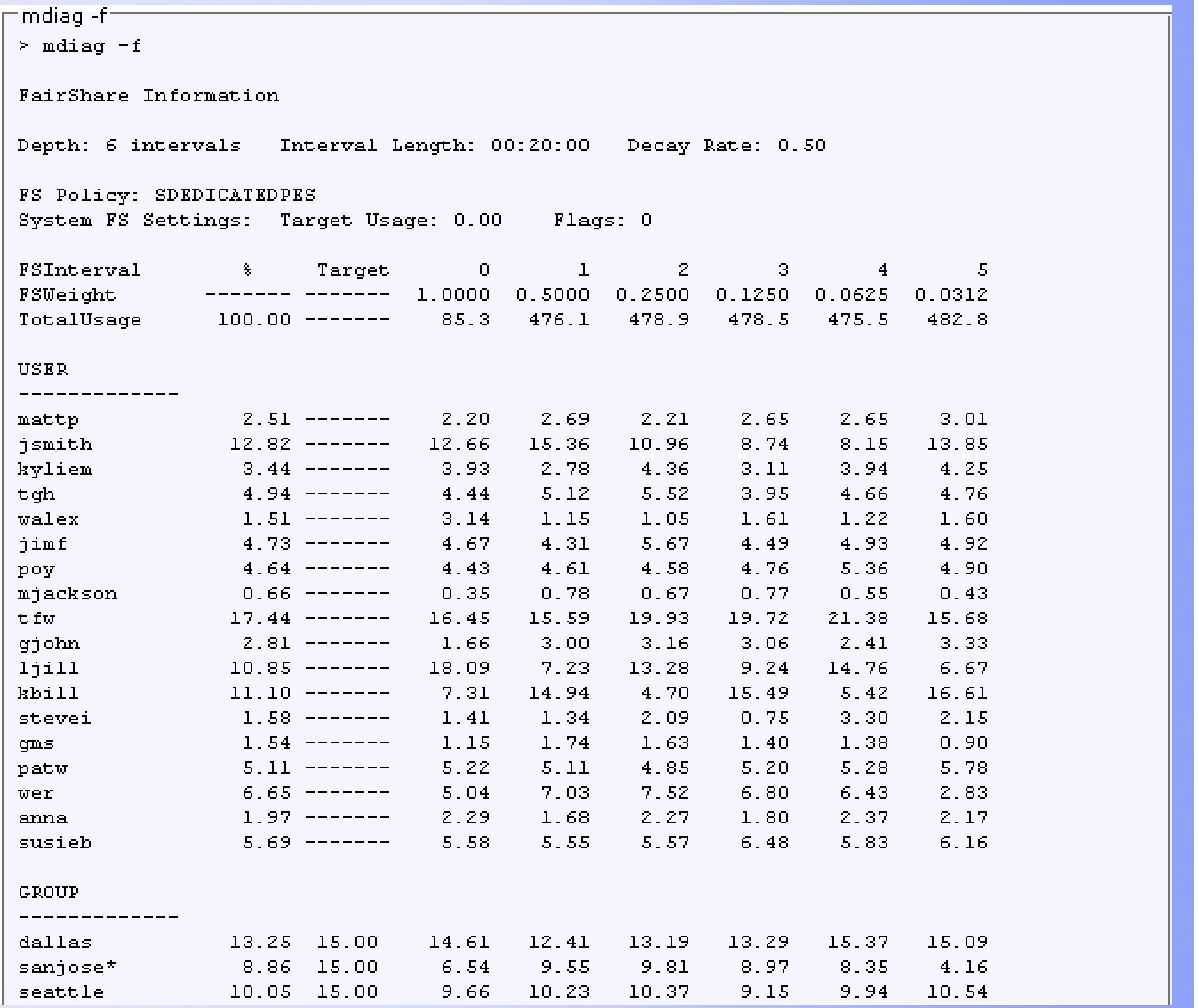
```
# moab.cfg
FSINTERVAL                 12:00:00
FSDEPTH                    4
FSDECAY                    0.5
FSPOLICY                   DEDICATEDPS

# all users should have a fs target of 10%
USERCFG[DEFAULT] FSTARGET=10.0

# user john gets extra cycles
USERCFG[john] FSTARGET=20.0

# reduce staff priority if group usage exceed 15%
GROUPCFG[staff] FSTARGET=15.0-

# give group orion additional priority if usage drops below 25.7%
GROUPCFG[orion] FSTARGET=25.7+

FSUSERWEIGHT     10
FSGROUPWEIGHT 100
```



http://clusterresources.com/moabdocs/6.3fairshare.shtml

Fairshare

- Provide cred
  usage distrib
- mdiag –f
- Maintained fo
- Stored in sta
- Shows detail
  metric



```
> mdiag -f

FairShare Information

Depth: 6 intervals   Interval Length: 00:20:00   Decay Rate: 0.50

FS Policy: SDEDICATEDPES
System FS Settings:  Target Usage: 0.00    Flags: 0

FSInterval        %     Target       0       1       2       3       4       5
FSWeight       ------- -------  1.0000  0.5000  0.2500  0.1250  0.0625  0.0312
TotalUsage      100.00 -------    85.3   476.1   478.9   478.5   475.5   482.8

USER
-------------
mattp             2.51 -------    2.20    2.69    2.21    2.65    2.65    3.01
jsmith           12.82 -------   12.66   15.36   10.96    8.74    8.15   13.85
kyliem            3.44 -------    3.93    2.78    4.36    3.11    3.94    4.25
tgh               4.94 -------    4.44    5.12    5.52    3.95    4.66    4.76
walex             1.51 -------    3.14    1.15    1.05    1.61    1.22    1.60
jimf              4.73 -------    4.67    4.31    5.67    4.49    4.93    4.92
poy               4.64 -------    4.43    4.61    4.58    4.76    5.36    4.90
mjackson          0.66 -------    0.35    0.78    0.67    0.77    0.55    0.43
tfw              17.44 -------   16.45   15.59   19.93   19.72   21.38   15.68
gjohn             2.81 -------    1.66    3.00    3.16    3.06    2.41    3.33
ljill            10.85 -------   18.09    7.23   13.28    9.24   14.76    6.67
kbill            11.10 -------    7.31   14.94    4.70   15.49    5.42   16.61
stevei            1.58 -------    1.41    1.34    2.09    0.75    3.30    2.15
gms               1.54 -------    1.15    1.74    1.63    1.40    1.38    0.90
patw              5.11 -------    5.22    5.11    4.85    5.20    5.28    5.78
wer               6.65 -------    5.04    7.03    7.52    6.80    6.43    2.83
anna              1.97 -------    2.29    1.68    2.27    1.80    2.37    2.17
susieb            5.69 -------    5.58    5.55    5.57    6.48    5.83    6.16

GROUP
-------------
dallas           13.25   15.00   14.61   12.41   13.19   13.29   15.37   15.09
sanjose*          8.86   15.00    6.54    9.55    9.81    8.97    8.35    4.16
seattle          10.05   15.00    9.66   10.23   10.37    9.15    9.94   10.54
```