Throughput scheduler

Stephan Grell

Software Engineer

Sun microsystems

We make the net work.

Overview

- Scope
- Profiling
- Areas for improvements
- Finished enhancements
  - Impact of the finished enhancements
- Upcoming enhancements
- Scheduler configuration
  - Scheduler configuration interface
- What to expect

Scheduler - scope

- Support high throughput scenarios
- Max utilization of the compute resources
- Making the scheduler efficient
- Understanding the scheduler
  - > Profiling

Profiling - scheduler

- Schedd_param: profile = true

```
PROF: SGEEE job ticket calculation: init: xx s, pass 0: xx s, pass 1:
        xx s, pass2: xx s, calc: xx s
PROF: SGEEE job sorting took xx s
PROF: SGEEE update orders: job orders: xx s, update orders: xx s
PROF: SGEEE job ticket calculation took xx s
PROF: scheduled in xx (u xx + s xx = xx): detailed information
PROF: send orders took: xx s
PROF: schedd run took: xx s (init: xx s, copying lists: xx s)
```

-> format might be changed at any time
-> Stored in the scheduler message file

included in:

- V 5.3 P5
- V 6.0

areas for scheduler improvements

- Duplicating data for a scheduler run
- Computing functional tickets
- Hard/ soft requests
- Sending tickets to qmaster

Scheduler - finished improvements

- Hard/ soft requests matching
  - New implementation

included in:
- V 6.0

- Computing functional tickets
  - New implementation

included in:
- V 5.3 P5
- V 6.0

- Sending ticket to qmaster

included in:
- V 6.0

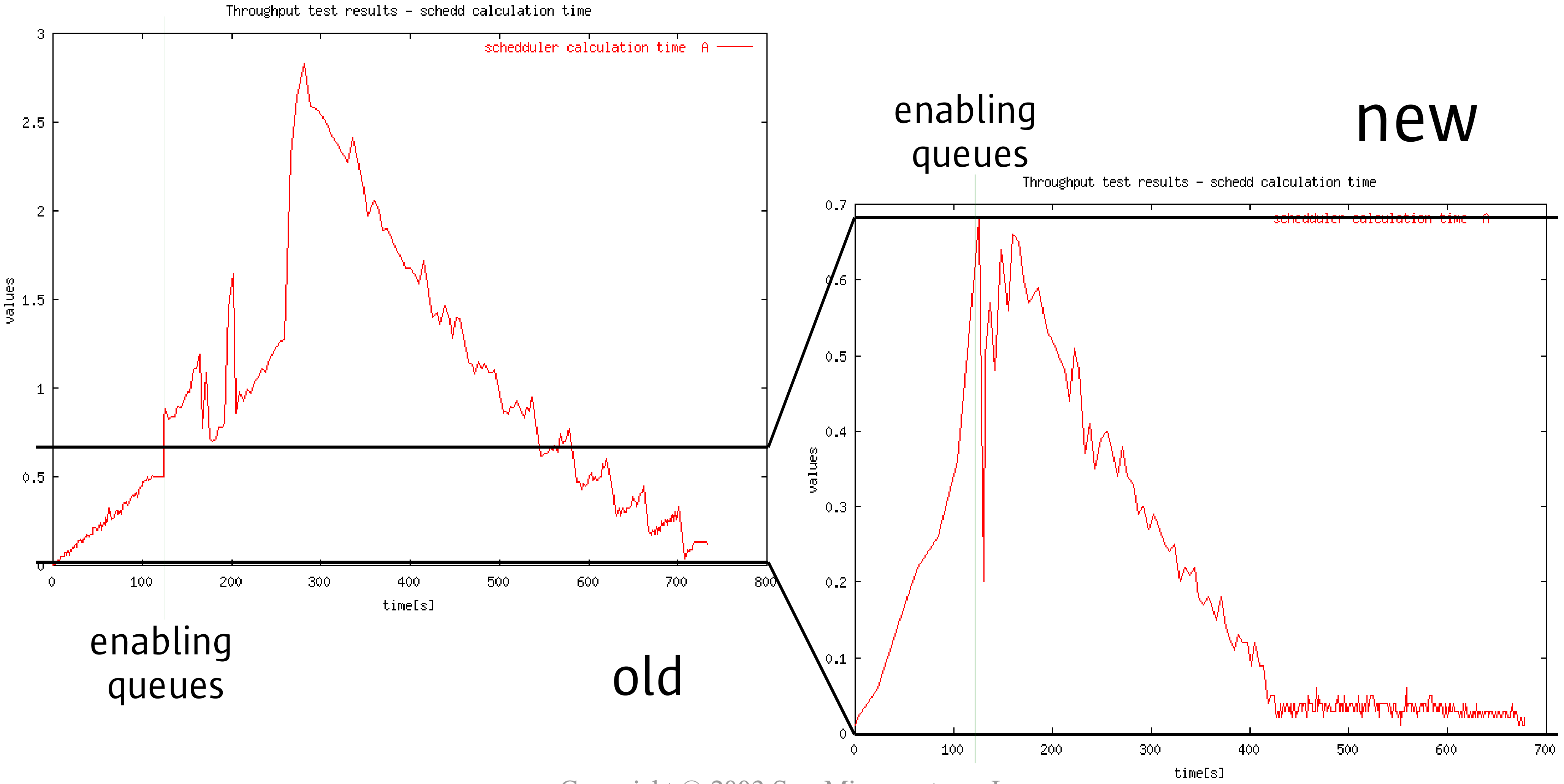
  - New schedd_param: -> no pending job tickets
    - REPORT_PJOB_TICKETS = true/ false
  - sgeee_scheduling_interval = 0:0:0 -> no running job tickets

Scheduler - performance test

- 1 dedicated server (qmaster, scheduler)
- 3 exec hosts
- 10 queues per exec host
- 5 slots per queue => 150 slots in the system
- 3000 jobs with 5 sec. run time
- enabling queues after 1500 jobs were submitted

Scheduler - impact

scheduler calculation time



enabling queues

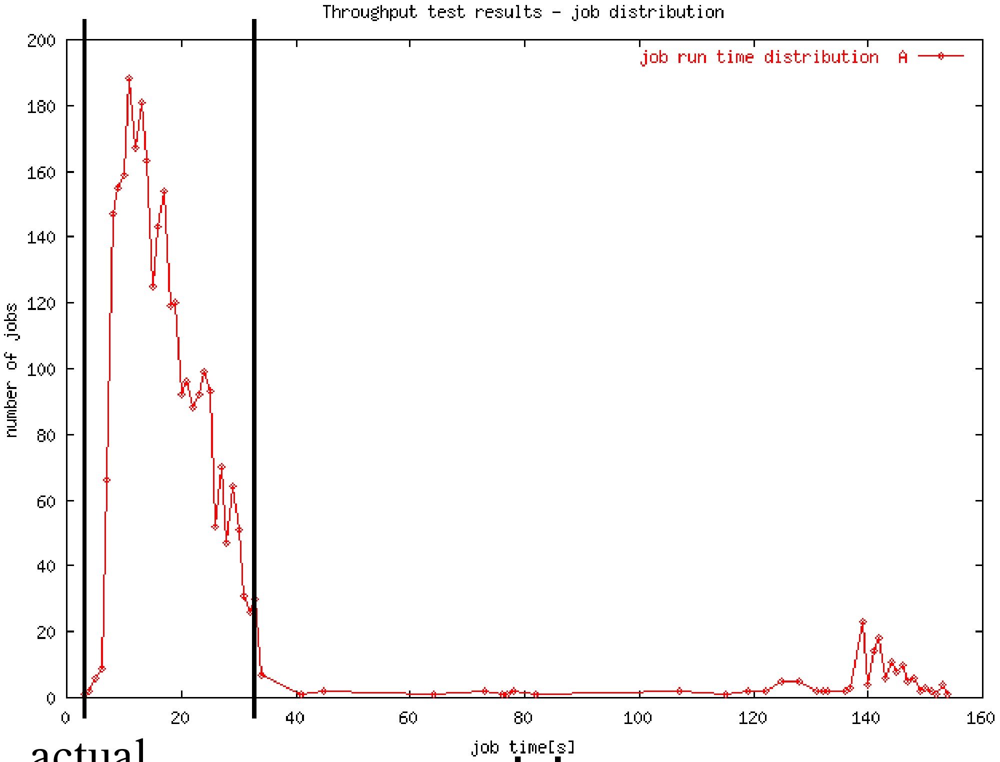
old

enabling queues

new

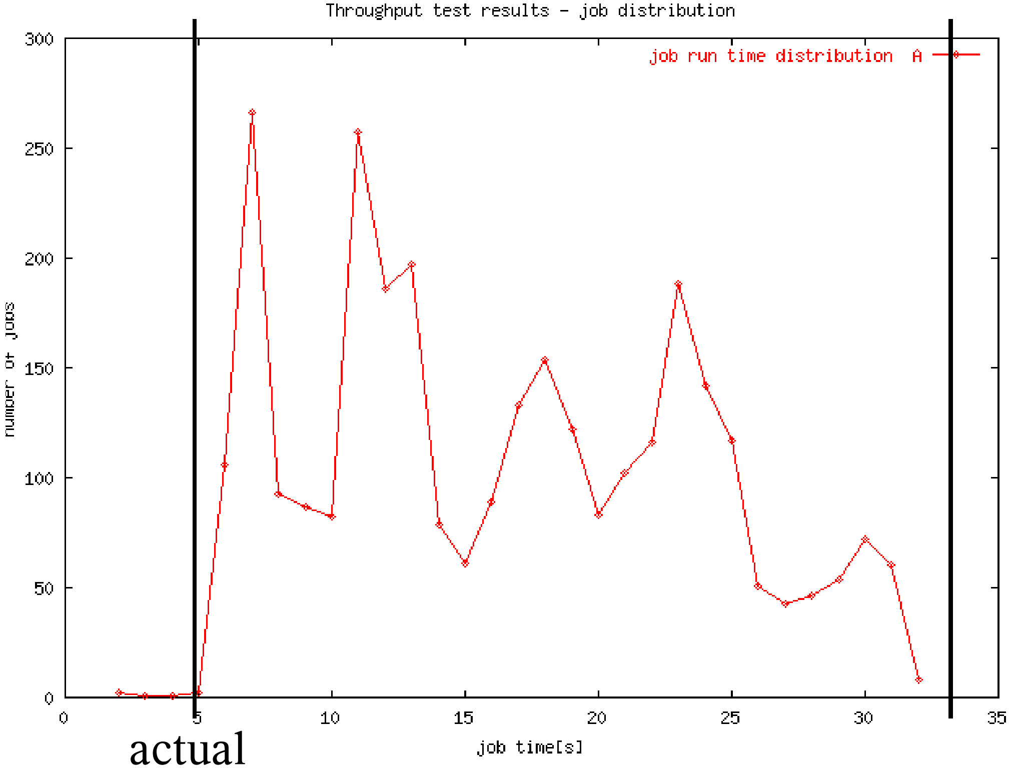
Scheduler - impact

perceived job run time



actual
job run
time = 5s

old



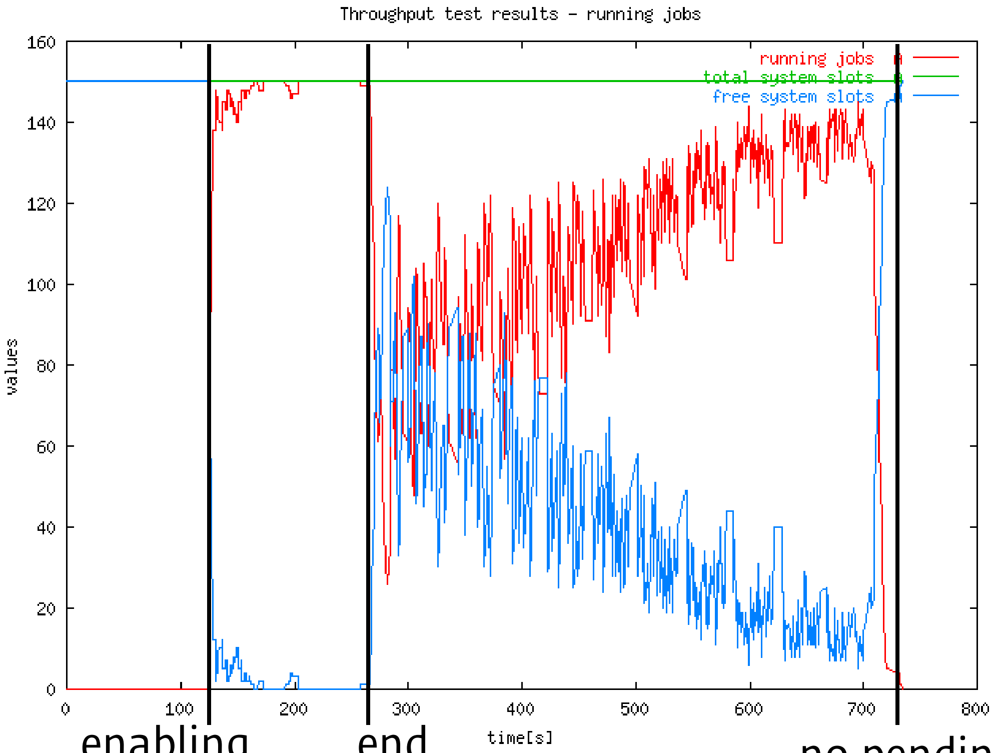
actual
job run
time = 5s

new

Scheduler - impact

slots utilization (150 slots)

old



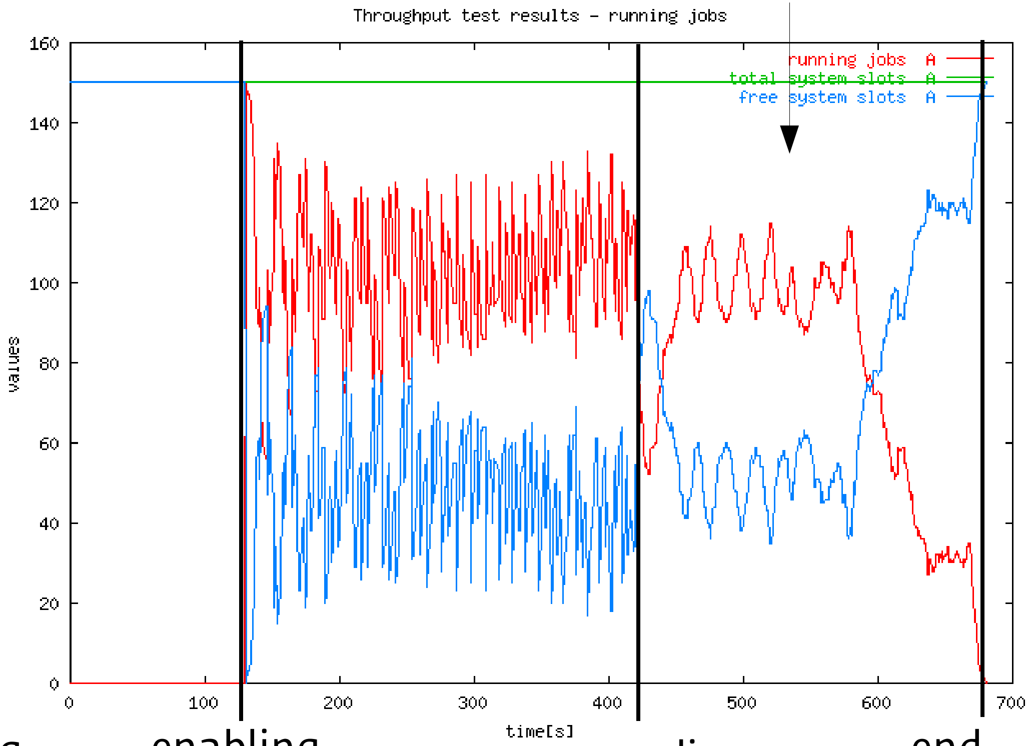
enabling queues 1500 jobs

end submit of 3000 jobs

no pending jobs

new

all incoming jobs are scheduled immediately



enabling queues 1500 jobs

no pending jobs

end submit of 3000 jobs

Scheduler - upcoming improvements

- related to Hard/ soft requests

to come in:

- V 6.0

later:

- Sequential slots in exec daemons
- reduce scheduler internal status creation time
- Optimize sending tickets to qmaster

Scheduler - Tuning guide

Tuning...	effect
• Scheduler monitoring	good
• Finished jobs	okay
• Load and suspend thresholds	good
• Load adjustments	good
new in V 6.0:	
• Report tickets	
– Pending job tickets	good
– Running job tickets	okay

http://gridengine.sunsource.net/project/gridengine/howto/tuning.html

Scheduler - configurations

Tuning...	effect
• Scheduling-on-demand	difficult
• Job verification	do not use

Scheduler - admin interface

- Scheduler performance tuning with one interface
- fast switching between different profiles
- three predefined profiles
  - standard
  - fast
  - aggressive

Expected features

- Profiling

  => Yes!

- Efficient scheduler

  => We will go as far as possible

- Max utilization of the compute resources

  => Depends on the other SGE modules as well

Stephan Grell

Stephan.Grell@sun.com

http://sun.com/grid

Sun microsystems
We make the net work.