A new Self-Adaptive disPatching System(SAPS) for local cluster

Bowen KAN
(Email: kanbw@ihep.ac.cn)
IHEP CC

CHEP2015, Okinawa, April 2015

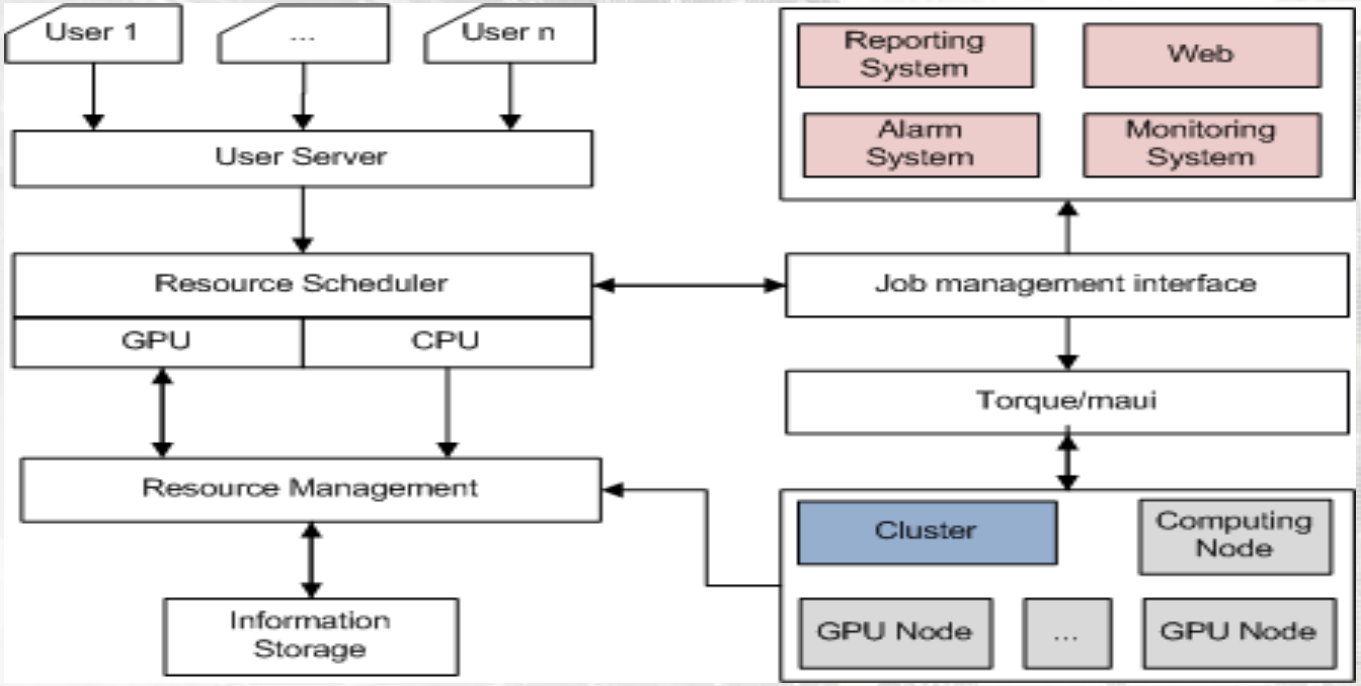
Overview

- Problems to torque/maui of IHEP PC farming
- How to resolve the problems we met
- Experimental results
- More tools provided
  - Monitoring management
  - Reporting system
  - Alarm system
  - Web interface
- Conclusion

Problems to torque/maui of IHEP PC Farm

- Insufficient functionality
  - 256 gpu cards need to be scheduled
  - No GPU scheduling policy provided by maui
- Resource inefficiency
  - Resource utilization is limited by static jobs_max_running
- System fragility
  - Unexpected errors happened to work nodes frequently
  - Communication between torque server and good work nodes was blocked by the error nodes
  - A large number of jobs were scheduled to the error work nodes and failed – black hole

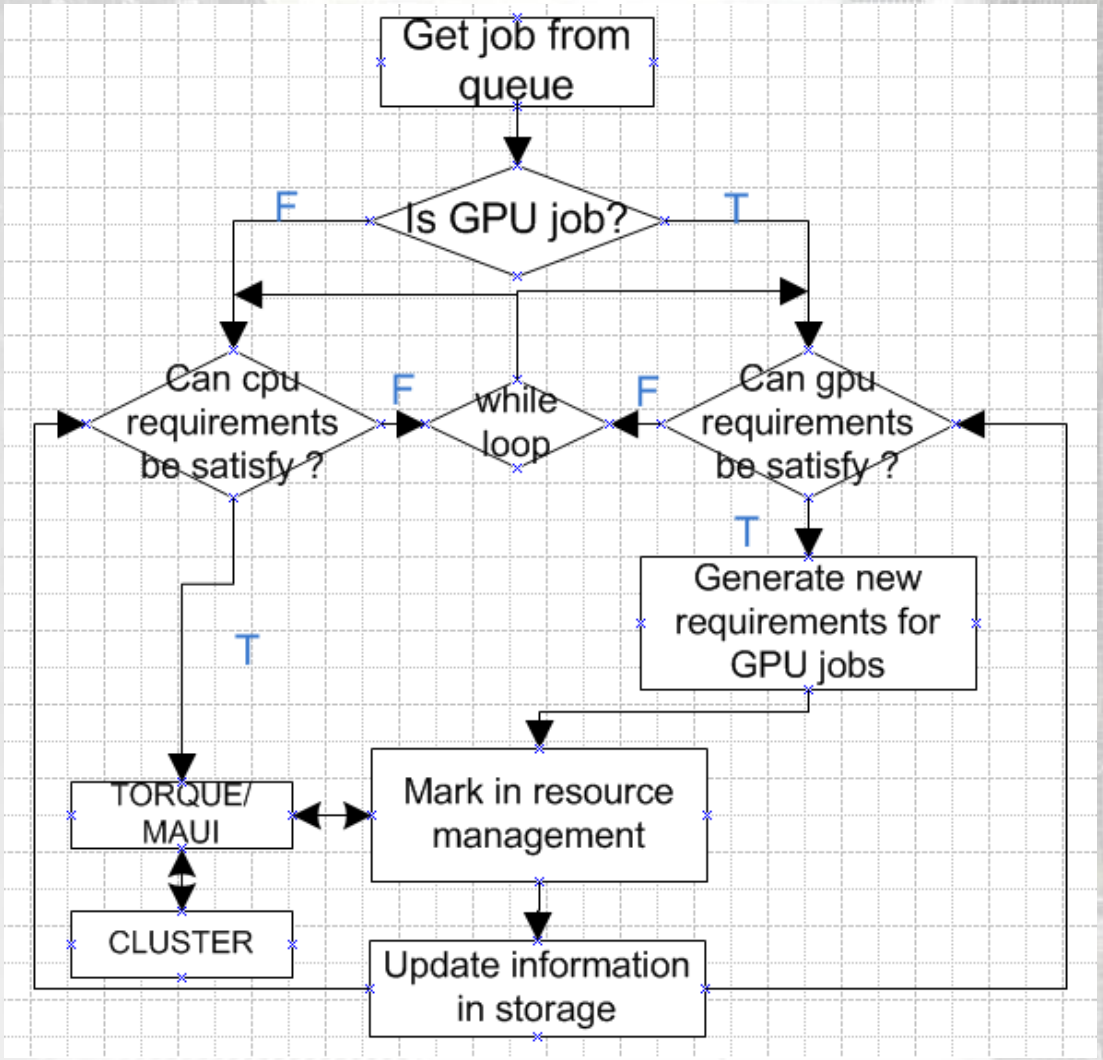
Solutions ---- Architecture(SAPS)



- GPU scheduling function added
- Jobs Max-running tuned dynamically depending on the number of active users and free resources
- Error nodes be detected and recovered in real-time automatically

GPU scheduling

- GPU job->
- Satisfy job's requirement accounting to information storage->
- Generate new requirements for GPU job ->
- Mark and re-map between resource and jobs->
- Update information of storage



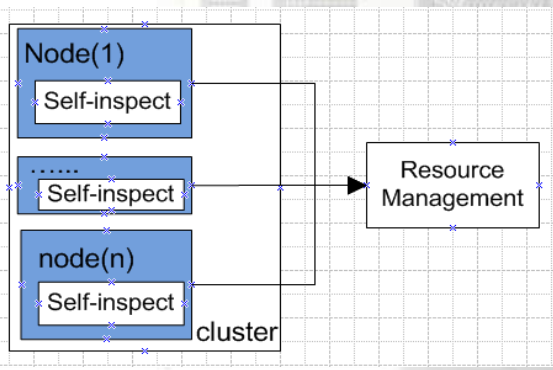
How to resolve the Resource inefficiency

- Reason
  - The amount of active users is uncertain
  - Fixed value of jobs Max-running number blocks the user to get more free resources
- Solution
  - Jobs Max-running number of queue is tuned dynamically according to the active users, running jobs, queuing jobs and idle resources.
  - The amount of active users: whose queuing jobs are more than (1/20) *(the number of idle resource) in one queue.
  - Max-running number (MRN)of the queue equal to :

$$mrn = \frac{(idle\ resources) * (active\ users)}{(queuing\ jobs)}$$

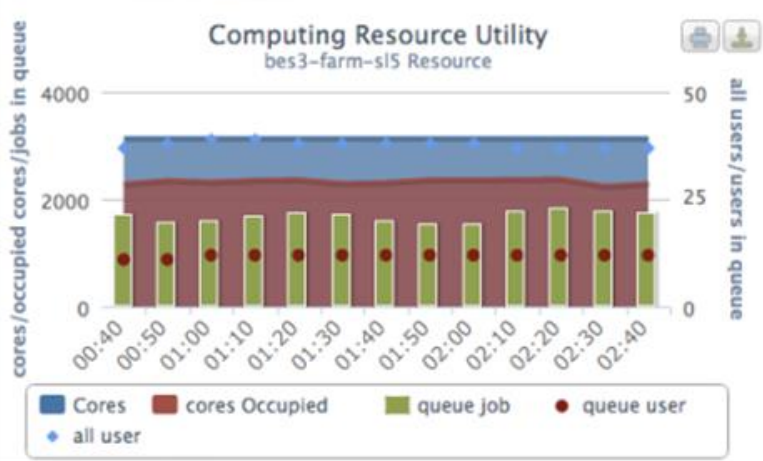
How to increase the system stability

- Reason
  - Unexpected Error happened to work nodes frequently
    - Communication is blocked between the work nodes and scheduling server.
    - A large number of jobs were scheduled to error nodes and kept failing
- Solution
  - Computing nodes self-inspection
    - Mem, cpu, zombie process, large file jobs and so on
  - Excluded the error nodes from the cluster in real-time
  - the error fixed nodes would be included automatically

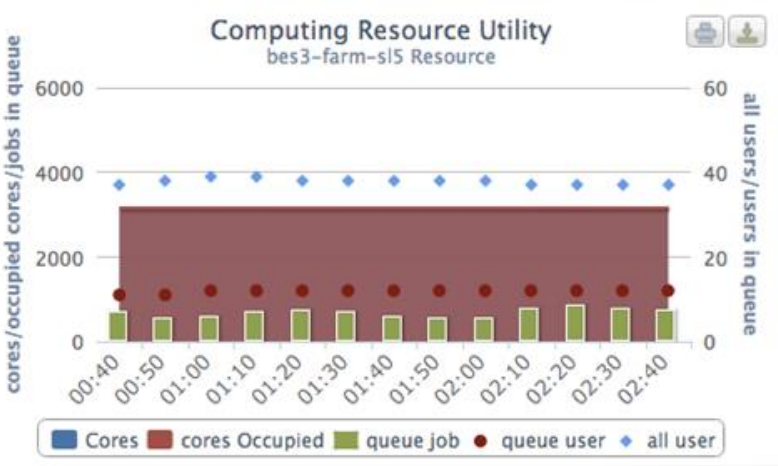


Experimental Results

Before



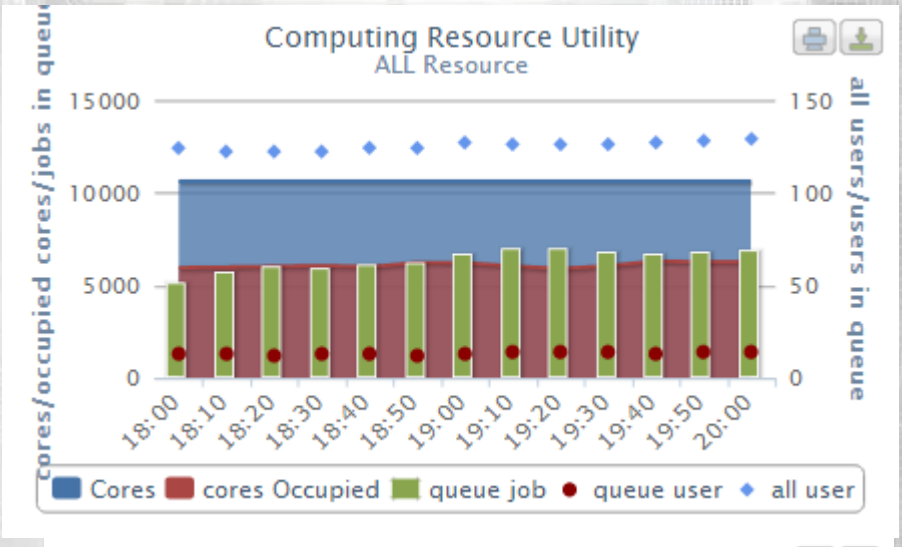
Now

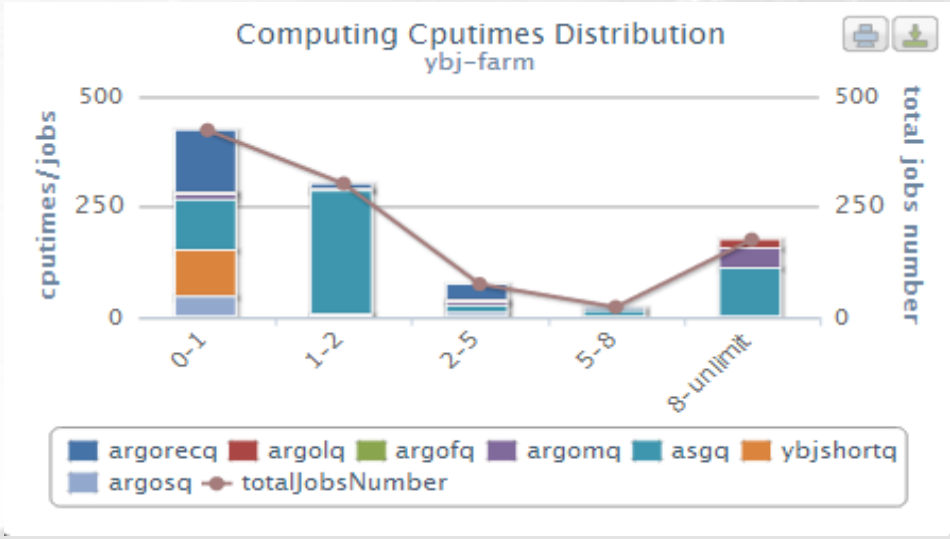


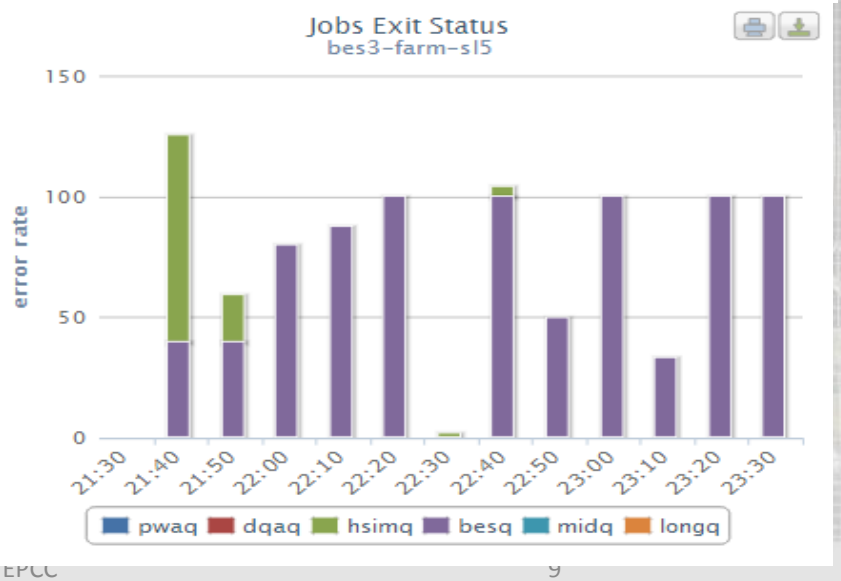
- Queuing jobs reduced sharply
  - Before: almost 2000
  - Now: 500
- Utilization of resources increased significantly
  - From 74% to 100%

SAPS monitoring management

- Get data from resource management.
- Data is converted to json files.
- graphical display page based on json file provided.





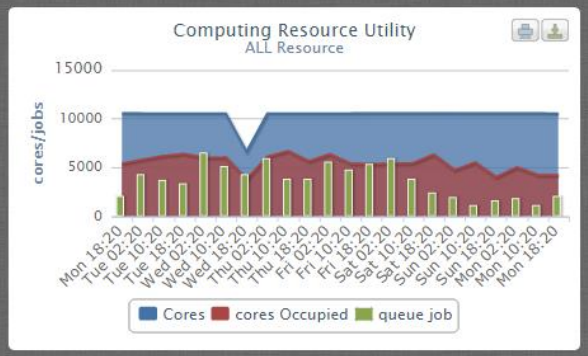


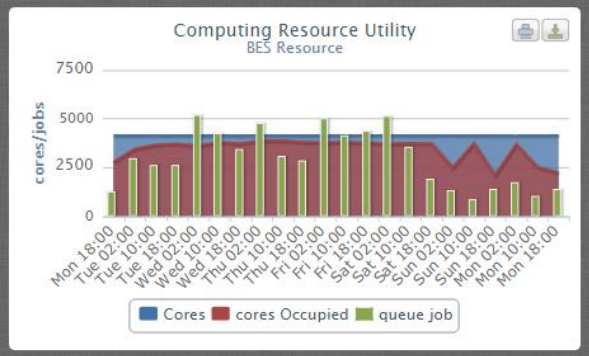
Local Cluster Monitor

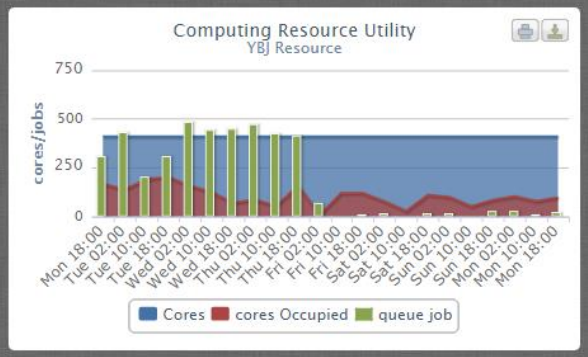
2015-3-30 19:37:22

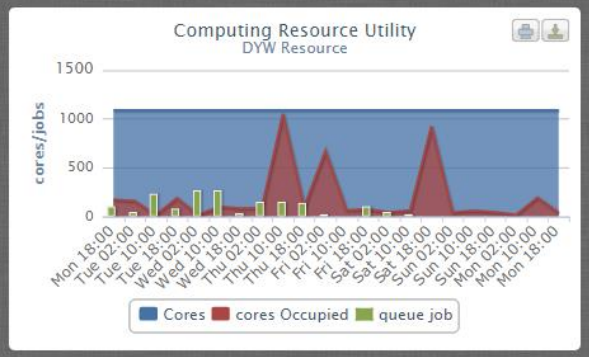
Status | Application-cores | Resource-cores | Application-time | Utilization-time | Resource-time | Cputime | ExitStatus | ExitReason

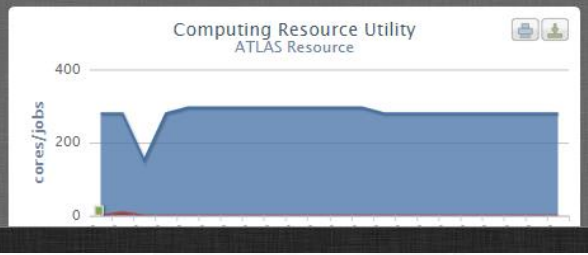
Hour Day Week Month Year

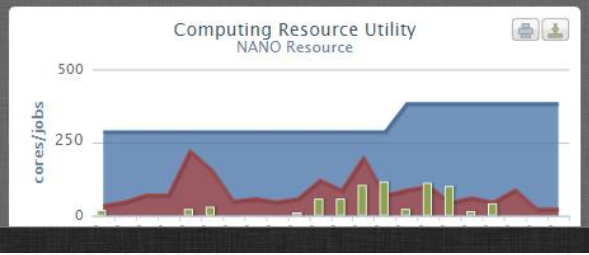












SAPS alarm and reporting system

- Alarm system
  - Alarm to failed job
  - Alarm to job lost
- Reporting system
  - According to the results of the reservation
  - Report will be sent to user periodically

Dear vindy,

Your Failed Jobs fininshed during last hour are listed as follows:

ExitStatus	JobID	Queue	JobName	WallDurationSeconds	CpuDurationSeconds	MemoryReal	MemoryVirtual
271	17710821.pbssrv.ihep.ac.cn	dp2q	job_001.sh	61797	3306	227556	934012	bws0354.ihep.ac.cn/7
271	17710823.pbssrv.ihep.ac.cn	dp2q	job_002.sh	61797	2813	229872	935064	bws0347.ihep.ac.cn/2
271	17710824.pbssrv.ihep.ac.cn	dp2q	job_003.sh	61797	2763	230852	934924	bws0346.ihep.ac.cn/0
271	17710826.pbssrv.ihep.ac.cn	dp2q	job_004.sh	61797	2874	231500	935460	bws0347.ihep.ac.cn/7
271	17710831.pbssrv.ihep.ac.cn	dp2q	job_008.sh	61732	2988	231584	938296	bws0217.ihep.ac.cn/1
271	17710835.pbssrv.ihep.ac.cn	dp2q	job_010.sh	61733	2892	223404	932932	bws0544.ihep.ac.cn/3
271	17710839.pbssrv.ihep.ac.cn	dp2q	job_013.sh	61734	2969	218836	928268	bws0583.ihep.ac.cn/10
271	17710842.pbssrv.ihep.ac.cn	dp2q	job_015.sh	61617	2786	231408	940780	bws0530.ihep.ac.cn/2
271	17710843.pbssrv.ihep.ac.cn	dp2q	job_016.sh	61617	3155	229972	939220	bws0598.ihep.ac.cn/1
271	17710848.pbssrv.ihep.ac.cn	dp2q	job_019.sh	61500	3015	234168	942536	bws0504.ihep.ac.cn/0
271	17710850.pbssrv.ihep.ac.cn	dp2q	job_020.sh	61498	3021	235216	940112	bws0504.ihep.ac.cn/6
271	17710851.pbssrv.ihep.ac.cn	dp2q	job_021.sh	61497	2998	231836	937208	bws0520.ihep.ac.cn/5
271	17710852.pbssrv.ihep.ac.cn	dp2q	job_022.sh	61497	3116	228844	938384	bws0566.ihep.ac.cn/0
271	17710854.pbssrv.ihep.ac.cn	dp2q	job_023.sh	61497	2960	224824	934364	bws0580.ihep.ac.cn/7
271	17710856.pbssrv.ihep.ac.cn	dp2q	job_024.sh	61498	3013	225128	934360	bws0555.ihep.ac.cn/2
271	17710859.pbssrv.ihep.ac.cn	dp2q	job_027.sh	61383	416	218516	927640	bws0577.ihep.ac.cn/10

Please consult http://cuc.ihep.ac.cn/Monitor/# 'ExitReason' or following reasons to check the reason of failed jobs.

ExitStatus	Reason
-2	error duing job execution.
1	1. can not find input files; 2. empty input files; 3. permission denied on job output directory; 4. wrong data file(s).
2	job submission error, with the wrong job submission script.
6	unknown reason, please contact to admin.
11	exceeds the boundary of job array.
126	can not excute the command in the job.
127	1. with successful result, but job turns into a zombie process, needs to be killed by pbs 2.with failed result, can not execute the command in the job.
134	memory leak
137	unknown reason, please contact to admin.
139	unknown reason, please contact to admin.
143	job deleted by user.
250	unknown reason, please contact to admin.
265	job failed.
267	job failed: array out of boundary inside the job script.
271	job exceeds its walltime or cputime.

Contact email: ihep_computing_service@ihep.ac.cn.

IHEP CLUSTER 记帐信息

Application: BES Month: 2014-06

NOTE: Efficiency = CpuTime / Walltime ErrRate=ErrJobs / JobSum Utility = 1- (DeleteJobsWalltime / WallTime)

User	Queue	Group	Job Sum	WallTime(hr)	CpuTime(hr)	QueueTime(hr)	AveQueueTime(hr)	Efficiency	ErrJobs	ErrRate	ErrWallTime(hr)	ErrCpuTime(hr)	DelJobs	DelWalltime(hr)	DelCputime(hr)	Utility
aixc	besq	physics	1788	56901.921	32929.509	4594.560	2.570	0.579	514	0.287	11428.785	9395.255	0	0.000	0.000	1.000
chenjc	besq	physics	5694	54611.207	51880.326	44220.788	7.766	0.950	3	0.001	200.904	200.017	18	1109.386	1100.787	0.980
zhzhang	besq	physics	3491	44886.166	2321.635	33319.971	9.545	0.052	15	0.004	21.336	21.213	444	39085.859	10.164	1.000
kangxl	besq	physics	5553	36065.686	34222.336	48791.377	8.786	0.949	120	0.022	638.663	614.058	1	14.472	13.559	1.000
ripka	besq	physics	4699	33927.294	29786.536	54186.445	11.531	0.878	489	0.104	1474.113	1342.750	1	3.131	2.839	1.000
like	besq	physics	7737	33244.842	18912.539	45401.564	5.868	0.569	26	0.003	216.186	85.300	374	31.608	26.371	0.999
luy	besq	physics	3367	32822.995	27350.491	20550.364	6.103	0.833	1	0.000	0.003	0.001	191	1244.152	1187.342	0.964
liujie	besq	physics	8942	31930.743	24920.000	37417.151	4.184	0.780	34	0.004	100.357	72.013	23	14.958	14.163	1.000
luhn	besq	physics	9983	31467.008	23026.499	66610.050	6.672	0.732	82	0.008	132.339	128.104	390	403.485	141.675	0.995
vindy	besq	physics	3626	26225.833	23495.174	30662.392	8.456	0.896	415	0.114	13649.558	13626.868	2	67.067	66.969	0.997
xiaod	besq	physics	8645	23299.221	13797.914	21776.123	2.519	0.592	3	0.000	22.326	21.445	0	0.000	0.000	1.000
hajime	besq	physics	14419	22808.671	20047.791	63372.184	4.395	0.879	1221	0.085	6167.766	5637.386	507	709.755	498.256	0.978
mahl	besq	physics	3143	21749.344	15864.772	6995.387	2.226	0.729	0	0.000	0.000	0.000	305	512.393	444.801	0.980
liyt	besq	physics	8036	21584.426	20967.496	60020.497	7.469	0.971	74	0.009	4500.665	4493.610	4	205.707	205.526	0.990
zhangjielei	besq	physics	8984	21181.627	18146.131	45477.224	5.062	0.857	467	0.052	1006.149	871.618	4	6.809	4.432	1.000
chuxk	besq	physics	8112	20816.243	11966.454	37939.860	4.677	0.575	581	0.072	899.194	259.841	0	0.000	0.000	1.000
yinjh	besq	physics	6098	20291.268	13282.706	34704.083	5.691	0.655	24	0.004	378.553	356.603	0	0.000	0.000	1.000
liukai	besq	physics	4041	19992.174	18344.883	5763.446	1.426	0.918	15	0.004	10.616	10.453	67	88.502	84.935	0.996

Conclusion

- The SAPS implements the GPU scheduling with multi-core
- Utilization of resources significantly increased
- System stability has been improved
- Monitoring system, alarm system and reporting system are provided by SAPS
- Next plan
  - Jobs Max-running number just focus on queue
  - A finer granular needed: focus on users
  - The element of Jobs Max-running numbers are generated more automatically

- Thank you
- Question ?