Enabling Grids for E-sciencE

ASC: an Adaptive Scheduling Controller

SCOPE
Università degli studi di Napoli Federico II

Authors
Dr. BARONE, Giovanni Battista (University of Naples Federico II), Dr. BOTTALICO, Davide (University of Naples Federico II), Dr. BOCCIA, Vania (University of Naples Federico II), Dr. CARRACCIUOLO, Luisa (CNR), Dr. DORIA, Alessandra (INFN), Dr. DI MARTINO, Catello (University of Naples Federico II), Dr. D'ANGELO, Giovanni (University of Naples Federico II)

Co-Authors
Dr. ESPOSITO, Christian (University of Naples Federico II), Dr. PETRILLO, Mauro (CEINGE), Dr. BUSIELLO, Gianluca (CEINGE), Dr. CANTELE, Giovanni (CNR), Dr. VITAGLIANO, Giuseppe (University of Naples Federico II), Dr. POLLIO, Maurizio (University of Naples Federico II), Dr. PARDI, Silvio (INFN)

1: The problem

Resources

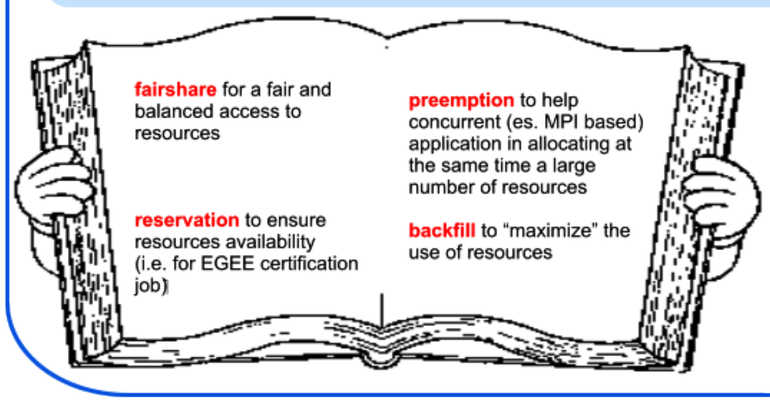
The deployment, management and TCO of large computing environments always involve huge investments. These systems, once in production, have to meet the needs of users belonging to large and heterogeneous communities: only an efficient and effective use of these systems can repay the investment made

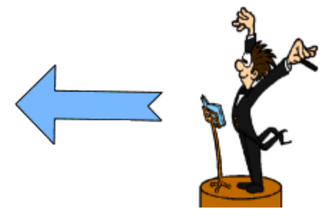
Users

Resources Manager & Scheduler

2: The scheduling system

ASC relies on top of a Maui-Torque scheduling system configured to use a combination of fairshare, reservation, preemption and backfill mechanisms, in order to achieve an efficient and effective use of the system





MAUI-Torque Scheduling system

3: Some remarks

Such kind of system manages heterogeneous communities of users but is not able to manage the changes in the user type

→

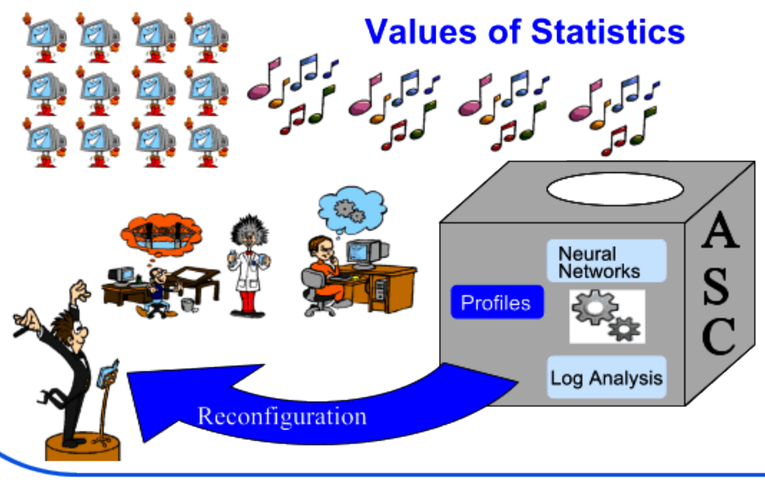
We have to define a mechanism which is able to "recombine/reconfigure" scheduling system on the basis of a change in the user type

→

ASC: Adaptive Scheduling Controller

Analyzing system behavior by some key-statistics (queue waiting time, jobs throughput, resource usage, and so on) a control loop dynamically defines a new set of Maui key-parameters values.

4: The ASC architecture



The default profile of the ASC control loop, based on automated log analysis and neural network techniques, can be chosen among a set of available profiles, each one identifies a target class of applications/users (e.g., parallel jobs, multi-thread jobs, concurrent jobs, and so on).

5: Example of Key statistics or metrics

System effectiveness ratio E

$$E = \frac{\sum_{i=1}^{n} p_i \cdot t_i}{P \cdot T}$$

System Makespan Mk

$$Mk = \max_{i=1,\ldots,n} (t_i + q_i)$$

Queue waiting time average Q

$$Q = \frac{\sum_{i=1}^{n} q_i}{n}$$

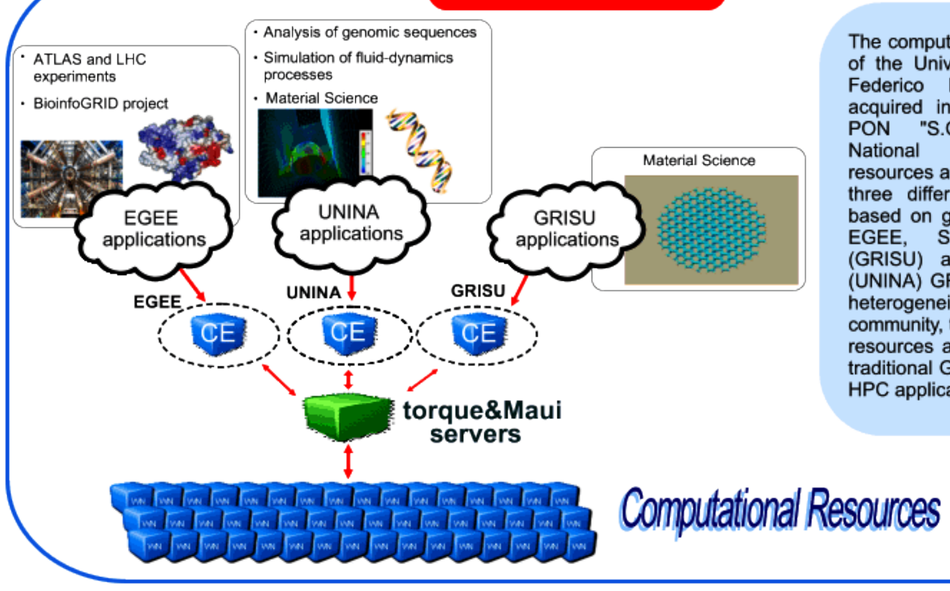
"In reality, the metrics attempt to formalize the real goal of a scheduler:
- *Satisfy the users*
- *Maximize the profit"*

(Feitelson at al.)

- *E* can express measure about "the effective system performance" in a real-world operational environment where different type of applications (parallel and sequential, short and very long, and so on) are executed
- *Mk* and *Q* can be useful to describe users satisfaction
- obviously other metrics, or combination of those described here, can be considered.

Where:
- n Number of jobs running
- p_i Number of processors utilized by job
- t_i Wall clock run time required by job
- q_i Queuing time required by job
- P Total number of processors in the system
- T Wall clock run time for all jobs completion

6: The testbed



The computational resources of the University of Naples Federico II, have been acquired in the context of PON "S.Co.P.E." Italian National project. The resources are shared among three different contexts all based on gLite middleware: EGEE, Southern Italian (GRISU) and metropolitan (UNINA) GRIDs. Due to the heterogeneity of the user community, the computational resources are used both for traditional GRID jobs and for HPC applications

7: Some performace results and Conclusions

The benchmark used for performance evaluation employs a suite of codes that reflects the current distribution and type of jobs running on computational resources of the testbed.
The suite is composed of application related to:
1) Material Science (based on quantum espresso MPI based software)
2) Analysis of genomic sequences
3) Physics Analysis for ATLAS experiments
The jobs of the suite are organized in three blocks (short, medium and long) on the basis of different duration and are submitted to the scheduling system in an order given by a pseudo-random number generator. The duration of the jobs varies from about 15 minutes to 36 hours. The benchmark has been executed for three different values of m (the number of tasks) on the same set of P cores.

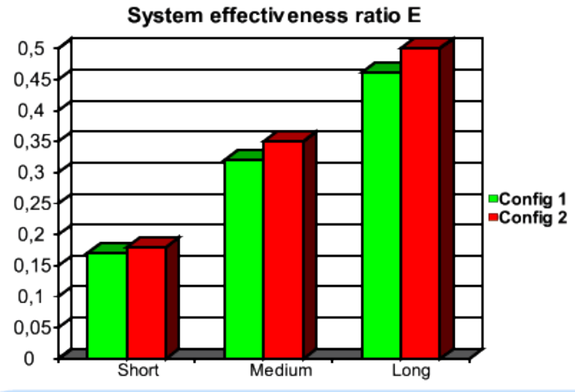
m/P	*E*	*Mk* (secs)	*Q* (secs)	*MPI Long App Q* (secs)
2,50	0,17	138219	1335,00	3896
5,00	0,32	144485	1794,00	12902
7,50	0,46	185203	3598,00	31161

The 1th Configuration: General Purpose Configuration

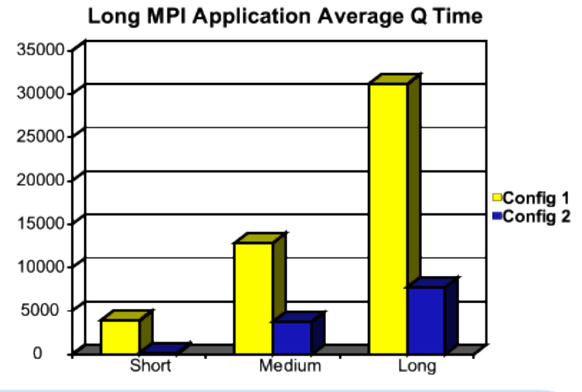
m/P	*E*	*Mk* (secs)	*Q* (secs)	*MPI Long App Q* (secs)
2,50	0,18	138139	7646	216
5,00	0,35	150591	11181	3744
7,50	0,50	183955	15460	7721

The 2nd Configuration: HPC aware configuration

System effectiveness ratio E



Long MPI Application Average Q Time



Conclusions:
Here we describe the work made to devise an adaptive scheduling controller (ASC), which aims to gain a balanced, efficient and effective use of the computing resources by heterogeneous communities.
We also show the importance of the most proper choice for the metrics, that we consider a very significant part of the ASC system, used to evaluate both system performance and user satisfaction for each community type.

8: References

- D.J. Quinn, D. Jackson, Q. Snell and M. Clement,"*Core Algorithms of the Maui Scheduler*", In Job Scheduling Strategies for Parallel Processing (JSSPP 2001), pp. 87-102, Springer-Verlag, 2001. Lecture Notes in Computer Science Vol. 2221.
- D. G. Feitelson, L. Rudolph, U. Schweigelshohn, K. Sevcik, and P. Wong, *"Theory and practice in parallel job scheduling"*. In Job Scheduling Strategies for Parallel Processing(JSSPP 1997), pp. 1-34, Springer-Verlag, 1997. Lecture Notes in Computer Science Vol. 1291.
- A. T. Wong, L. Oilker, W. T. C. Kramer, T. L. Kaltz and D. H. Bailey, *"Evaluating System Effectiveness in High Performance Computing Systems"*, LBNL Technical Report #44542, November 11, 1999.
- L. Merola on behalf of the SCoPE project, *"The S.Co.P.E. Project"*, Final Workshop of Grid Projects "PON RICERCA 2000-2006, AVVISO 1575", February 2009.