ABSTRACT

This study focuses on a computational grid which enables global sharing of the extensive computing power of heterogeneous resources for problem-solving in a large class of engineering, scientific and demand-driven commercial domains. This study develops grid resource management with failure handling services and it is one of the most challenging issues in a highly dynamic, heterogeneous and failure-prone grid environment. The primary function of Grid Resource Management System (GRMS) is to provide suitable resources for handling failures before and after scheduling the user job to complete successfully within the deadline. The present grid environment helps grid users to solve the computational and data-intensive applications without giving a guaranteed service. This work gives an assurance to the Quality of Service (QoS) requirements such as Deadline, Cost, Reliability, and Availability. By using Service Level Agreement (SLA), QoS can be made use of grid computing attractive for commercial applications.

This study tries to develop an Integrated Proactive and Reactive Failure Handling Framework to deliver the guaranteed service to the grid users which may minimize the runtime failures and job migrations. It allocates the best suitable resources to run a job and manage the failures like resource, process, match-making, and time-out. It returns the job successfully to the grid user within the stipulated budget and deadline in the end. This framework discusses group scheduling strategy for failure management which enables to minimize the communication time and finally it proposes adaptive failure prevention methods to minimize the number of reschedulings and job migrations. This work further evaluates different components of the integrated framework using different strategies and evaluated using the performance parameters such as resource utilization, job success rate and job completion time. This framework is also experimented using grid simulator GridSim toolkit 5.2 and verified with the existing Min-Min scheduling strategy. The proposed failure handling framework comprises five failure handling strategies namely Resource Failure Handling (R-FH) strategy, Process Failure Handling (P-FH) strategy, QoS Failure Handling (Q-FH) strategy, Group Scheduling Strategy for Failure Management (GSS-FM) and Adaptive Failure Prevention (A-FP) strategy. The R-FH strategy is introduced to reduce the runtime failures in the provision of reliable job execution. It selects the best suitable resources for considering QoS requirements such as
Deadline, Cost, Reliability, and Availability. After the selection, to provide the reliable service to the grid user, the strategy handles the failure of resources with minimum check-pointing time and recovery time. The P-FH strategy is developed to handle the process failure and resource failure for the better reliable job execution. During the execution of the user job, the P-FH strategy can recover the failed job using monitoring-service with minimum completion time whenever the process failure occurs due to CPU and memory overloads. The performance of P-FH is further improved by assuring the availability of checkpoints for the successful recovery of the failed job within the deadline. Simulations are performed to show the impact of reliability, availability, and speed of resources. Finally, it is proved that the proposed P-FH strategy outperforms the widely used Min-Min strategy.

With the necessity to handle match-making and time-out failures in the provision of guaranteed job execution along with the resource and process failures have resulted in the development of Q-FH model. Inclusion of Match-making Failure Handling (M-FH) strategy in Q-FH model allocates the suitable resources that fit the user-specified requirements to complete the user job within the specified time. In this strategy, the time deadline is relaxed using SLA renegotiation to match more user requests based on the capability of the resources to increase the job success rate. Inclusion of Time-out Failure Handling (T-FH) strategy in Q-FH model determines the required reliable resources in advance for job submission to avoid runtime failures. It also handles time-out failure by transparently redirecting the reserved job to the suitable resources without abnormally terminating the currently executing job to reduce the number of abnormal job terminations and hence improves the job success rate further. The requirement of combining jobs as a group at the meta-scheduler before deployment to reduce the communication overhead resulted in the introduction of Group Scheduling Strategy for Failure Management (GSS-FM). It combines the user jobs into job groups according to the size of jobs and computation power of reliable and available resources. It allocates them to the selected resources before the deployment to minimize the communication time and maximize the resource utilization. However, there is a need to provide suitable failure prevention methods using resource load prediction before scheduling the user jobs to reduce the number of reschedulings. An Integrated Computational Grid Framework for Proactive, Reactive Failure Handling Services with an Adaptive Failure Prevention Strategy (IF-PRAFS) is recommended for the further improved resource management. It selects the
suitable prevention actions prior to the resource allocation itself to reduce the runtime failures and job migrations. Results of the P-FH, Q-FH, GSS-FM, and IF-PRAFS strategies have shown the significant improvements in job success rate, resource utilization and reduction in job completion time for the submitted jobs compared to Min-Min strategy and IF-PRAFS strategy performs well compared to GridWay metascheduler.

IF-PRAFS is developed as an integrated computational grid framework for handling failures proactively and reactively to improve the service of resource selection, allocation, scheduling and monitoring in the high failure prone grid environment to achieve 99.2% resource utilization, 93.2% job success rate, and 45% reduction in job completion time. The resource, process, match-making and time-out failures are handled to provide higher reliable services and the resources are utilized to its maximum level to benefit the service provider. Implementation of SLA guarantees the service delivery for the clients and made the grid technology applicable in the commercial arena. The failure handling strategies which are presented for grid environments can be extended to fit into the resource allocation of cloud computing environments. The different components of IF-PRAFS can be converted into appropriate services for cloud resource selection, allocation, scheduling and monitoring to solve the issues of reliability, availability and service agreements. Thus, the development of integrated computational grid framework for proactive and reactive failure handling services is provided for the reliable improvement and guaranteed service to commercial grid applications.