

**CHENNAI – PONDICHERRY**

**EVALUATING REPLICATION FOR PARALLEL JOBS: AN EFFICIENT APPROACH**

**ABSTRACT:**

Many modern software applications rely on parallel job processing to exploit large resource pools available in cloud and grid infrastructures. The response time of a parallel job, made of many subtasks, is determined by the last subtask that finishes. Thus, a single laggard subtask or a failure, requiring re-processing, may increase the response time substantially. To overcome these issues, we explore concurrent replication with canceling. This mechanism executes two job replicas concurrently, and retrieves the result of the first replica that completes, immediately canceling the other one. To analyze this mechanism we propose a stochastic model that considers replication at both job-level and task-level. We find that task-level replication achieves a much higher reliability and shorter response times than job-level replication. We also observe that the impact of replication depends on the system utilization, the subtask reliability, and the correlation among replica failures. Based on the model, we propose a resource-provisioning strategy that determines the minimum number of computing nodes needed to achieve a service-level objective (SLO) defined as a response-time percentile. This strategy is evaluated by considering realistic traffic patterns from a parallel cluster, where task-level replication shows the potential to reduce the resource requirements for tight response-time SLOs.