

**CHENNAI – PONDICHERRY**

**PERFORMANCE EVALUATION OF CLOUD COMPUTING CENTERS WITH GENERAL ARRIVALS AND SERVICE**

**ABSTRACT**

Cloud providers need to size their systems to determine the right amount of resources to allocate as a function of customer's needs so as to meet their SLAs (Service Level Agreement), while at the same time minimizing their costs and energy use. Queueing theory based tools are a natural choice when dealing with performance aspects of the QoS (Quality of Service) part of the SLA and forecasting resource utilization. The characteristics of a cloud center lead to a queueing system with multiple servers (nodes) in which there is potentially a very large number of servers and both the arrival and service process can exhibit high variability. We propose to use a G/G/c-like model to represent a cloud system and assess expected performance indices. Given the potentially high number of servers in a cloud system, we present an efficient, fast and easy-to-implement approximate solution. We have extensively validated our approximation against discrete-event simulation for several QoS performance metrics such as task response time and blocking probability with excellent results. We apply our approach to examples of system sizing and our examples clearly demonstrate the importance of taking into account the variability of the tasks arrivals and thus expose the risk of under- or over-provisioning if one relies on a model with Poisson assumptions.