Availability, Performance and Cost Analysis for Large Scale Cloud

Prof. Kishor S. Trivedi
ECE Dept., Duke University, E-mail: kst@ee.duke.edu

We believe that research is needed in scalable methods for the availability, performance, energy consumption, resiliency and cost analysis of cloud systems through high fidelity analytic models and optimization. Cloud based systems are inherently large scale, distributed, almost always virtualized, and operate in automated shared environments. Performance and availability of such systems are affected by a large number of parameters including characteristics of the physical infrastructure (e.g., number of servers, number of cores per server, amount of RAM and local storage per server, configuration of physical servers, network configuration, persistent storage configuration), characteristics of the virtualization infrastructure (e.g., VM placement and VM resource allocation, deployment and runtime overheads), failure characteristics (e.g., failure rates, repair rates, modes of recovery), characteristics of automation tools used to manage the cloud system, and so on. Because of this, any naive modeling approach will quickly run into state explosion and/or intractable solution. Scale of the cloud makes the problems challenging in terms developing practical solutions. Unfortunately, above described challenges are not currently addressed or are addressed in an ad hoc manner. A few measurement based approaches and simple analytic and simulation models that do not capture the details of cloud resources and cloud management have been tried. A comprehensive and high fidelity modeling approach is sorely needed. Initially, in collaboration with IBM T. J. Watson Research Center, we developed an interacting Markov chain based modeling approach for joint analysis of performance and availability of Infrastructure-as-a-Service (IaaS) cloud [2] . Models developed in this approach take into account resource provisioning decision, VM provisioning, run-time execution and failure-repair of servers. To scale our approach for thousands of servers, we further developed a scalable availability model [4] using interacting Markov chainss. Our recent research also spanned beyond the traditional performance or availability analysis and quantified the notion of resiliency in the context of IaaS cloud [1] . **In future, we propose to extend our work through a three phase research plan:** (i) **validation of the developed models** using opportunistic simulations, controlled experimentations and using the data collected from real cloud (such as IBM, Amazon etc.), (ii) **extension of developed models** to capture the details [3] of failure-recovery modes, priority among the jobs, different workload arrival processes, different types of service time distribution, analysis of energy consumption, cost estimation, capacity planning and extension to other cloud service models (e.g., Platform-as-a-Service and Software-as-a-Service) and deployment models (e.g., hybrid cloud), (iii) **building a predictive tool using the developed models** that will allow the cloud service providers to carry out different types of "what-if" analysis (e.g., SLA analysis, understanding of over-all cloud economics, trade-offs, different optimizations) during design, development, testing and operational phases of a cloud service.

References

[1] R. Ghosh, F. Longo, V. K. Naik and **K. S. Trivedi**, "Quantifying Resiliency of IaaS Cloud," In Proc. RACOS workshop, held in conjuction with SRDS 2010.
[2] R. Ghosh, **K. S. Trivedi**, V. K. Naik and D. Kim, **"**End-to-End Performability Analysis for

Infrastructure-as-a-Service Cloud: An Interacting Stochastic Models Approach**,"** In Proc. *PRDC* 2010.

[3] D. Kim, F. Machida and **K. S. Trivedi**, "Availability Modeling and Analysis of a Virtualized System," In Proc. *PRDC* 2009.

[4] F. Longo, R. Ghosh, V. K. Naik and **K. S. Trivedi**, "A Scalable Availability Model for Infrastructure-as-a-Service Cloud", submitted to DSN 2011.