Google-Tec Innovation Cell – Brief Project Proposal Format

Proposal title –Capacity Planning of Highly-Available, Highly-Reliable, Highly-Scalable and High-Performance Information Technology Infrastructure Using Self-Similar Models and Principal Component Analysis

Principal investigator (s) – Dr. Raúl V. Ramírez Velarde

Goal - To use advanced theories such as Principal Component Analysis, Self-similarity and Large Deviations to create stochastic models that will derive easy to use equations for the design of large-scale information technology infrastructure.

Problem - When designing highly-available, highly-reliable, highly-scalable and high-performance information technology infrastructure for modern large-scale services, it becomes necessary to characterize individual and global behavior of users and services and scale them to hundreds of thousands of potential users using replication and redundancy, and cluster and grid technologies. In order to do that, traces containing millions of records of thousands of users must be analyzed to obtain information such a first, second, third and fourth cumulants, self-similarity behavior, autocorrelation and covariance. The presence of different types of users and different types of services introduces a multidimensionality difficulty. For example, a video server can have thousands of videos available for users and the designers of the services may not know in advance what will be the users' preferences; or a world accessible service can have thousands of individual applications that present different performance profiles; or a peer.-to-peer server can experience thousands of combinations of user behavior (connection time, number of connections, connection workload, etc,.). For that, Principal Component Analysis can be used to extract the necessary knowledge which involves the decomposition and analysis of very large matrices. Furthermore, almost all traffic, behavior and performance profiles measured so far exhibit self-similarity. Self-similarity means that the behavior of services will not smooth under averaging and the aggregation of data flows will not behave as a weighted average, that is to say, that the systems are not Ergodic. Also, simple to solve models such as Markov and Erlang will not work anymore and even normal queuing theory results are not valid.

Approach – We will first understand how self-similarity affects availability, scalability, reliability and performance of IT services. We will focus on multimedia services such as audio, video and Virtual Reality, and on Script oriented services, such as those found on LMS, CMS, CRM, e-Commerce, etc.

Then, we will understand the effects of multidimensionality of services and users and collect information about those services and users, and develop algorithms that using grid and cluster technology will analyze the data using Principal Component Analysis. Models will be tested using simulation results.

Design parameters will also be tested against Monte Carlo simulations and some architectures will be tested in virtual test beds implemented over the grid.

Status – Currently, research is being conducted on the development of stochastic self-similar model that predict IT data flow behaviors. Also, Principal Component Analysis is being used to reduce dimensionality of the types of services in order to determine appropriate operational parameters. PCA also helps to find clusters in order to create models of behaviors.

Monte Carlo simulations are being considered to relax the strictly analytic focus of our research to one in which simulation establishes operational parameters even if closed formulas cannot be derived. Grid technology is being considered to be able to manipulate matrices of millions of observations for thousands of users. Virtual test beds are being considered to simulate a real implementation of the IT infrastructure.

What we need is **two graduate students**, preferably Ph. D. students, that will improve stochastic models, develop grid algorithms for data analysis and design Monte Carlo simulations. We need **two full-time developer**s that will develop the test beds and collect information that will be input into the analytic models and simulations. We also need Google's experience in cluster and grid implementation. We need to set up an **experimental grid laboratory** that will allow us to test our developments and to establish a model for capacity planning. Some large scale matrix computations will be needed such as PLS_Toolbox for Matlab. We believe this project should be completed in a time frame of **two years**.

Expected Outcomes within 6 months

Within six months of the project we expect to have collected traces of at least two types of information technology services: video on demand and a learning management system. We also expect to have proven the validity of certain stochastic models that are under development through the use of simulation that uses the collected traces. We also expect to have advanced in the principal component analysis of services and establish preliminary results on the validity of using such techniques to attack the multidimensionality problem. Issues that will remain unresolved are benefits of analytic approximation based solutions to the models against Monte Carlo derived solutions, design of Monte Carlo simulations to derive operational parameters, consideration of other dimension reduction techniques such as independent component analysis, etc.