Cross-Disciplinary Approaches

to Empirical Networking Research

Jeffay, Kulkarni, Marron, Smith

OVERVIEW

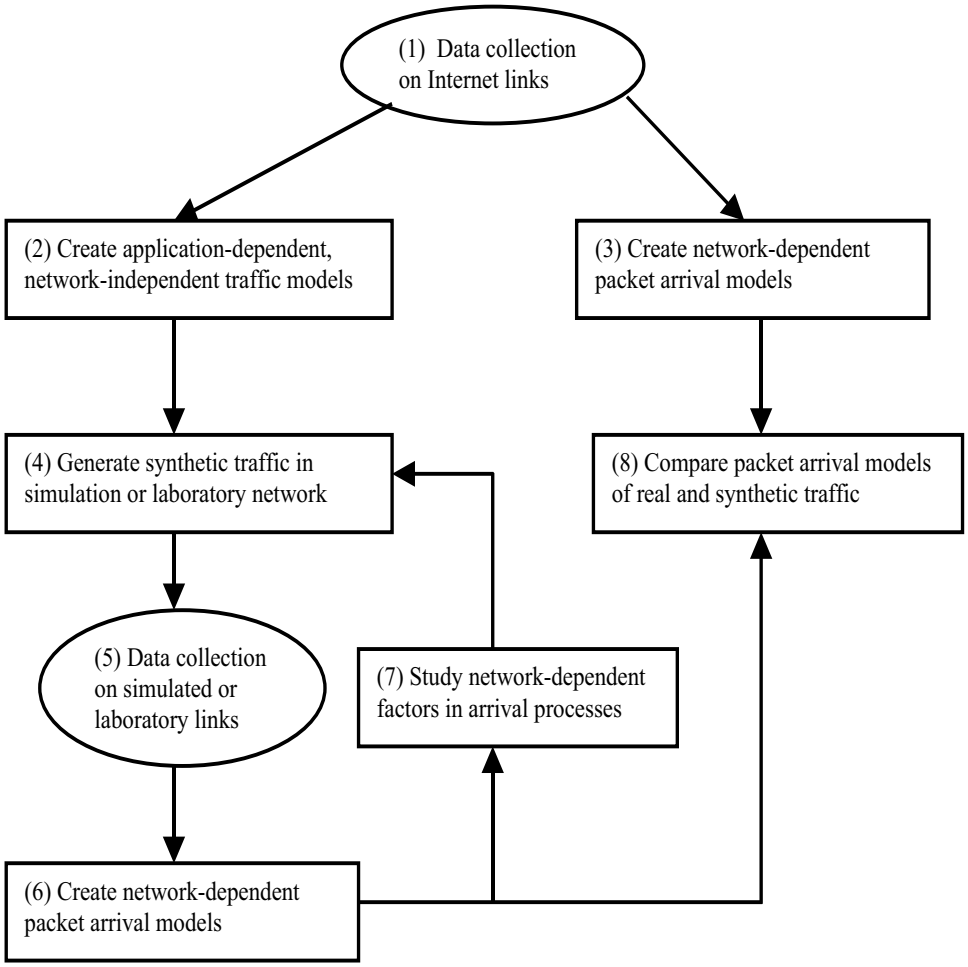

Figure 1: The flowchart showing the proposed plan of our research.

MAJOR GOAL

Develop tools for automatic collection and analysis of Internet traffic data, a continuously updated database of traffic traces, the statistical procedures for estimating the parameters of the models, procedures for synthetic traffic generation, and comparing and evaluating various traffic models.

MeMoSA: a collection of software tools for the **Me**asuring, **Mo**deling and **S**tatistical **A**nalysis of Internet traffic.

SUB-GOALS

1. Develop and continuously updating a terrabyte size database of network traffic traces.
2. Study and develop network-dependent traffic models of link-level and application-level traffic traces.
3. Study and develop network-independent traffic models of user behavior and network work-load generation.
4. Develop a scheme based on statistical clustering procedures to classify the ever-increasing Internet applications into a manageable number of statistically homogeneous traffic classes.
5. Develop the statistical procedures to fit traffic models to traffic traces and evaluate the goodness of fit.
6. Investigate the theoretical and empirical properties of the proposed new cascaded on-off model.

1. DATA BASE

Updated continuously to reflect changing traffic characteristics.

- **Link Traces**
  1. Link Info (speed etc)
  2. Time Info
  3. Time stamp and packet size for each packet in the trace.
- **Connection Traces**
  1. Link Info (speed etc)
  2. Time Info
  3. Origin, destination connection address
  4. Port Number
  5. Time stamp and packet size for each packet in the trace.

2. Network-Dependent Traffic Models

- **Connection Traces**
  1. Renewal Arrivals.
  2. On-Off Model.
  3. Conservative Cascades.
  4. Cascaded On-Off Model.
- **Link Traces.**
  1. Simplest Model: IID.
  2. Markov Modulated Sequence.
  3. Time Series Model.
  4. TES Model.
  5. Self Similar Traffic Models.
  6. Conservative Cascades.
  7. Cascaded On-Off Model.
- **From connection Traces to Link Traces.**

3. Network-Independent Traffic Models

1. Distribution of the number of TCP connections initiated by a user
2. Application specific file size distributions
3. Distribution of the structures and sizes of web pages
4. Distributions of the amount of transmitted data in TCP connections
5. Distributions of think times
6. Time dependent user arrival processes
7. Actual user behavior trajectories

4. Classification of Applications.

A top level hierarchy:

- Each TCP connection carries a single application-level data unit in each direction (a single request-response model). HTTP/1.0 is a member of this class.
- Each TCP connection carries multiple pairs of application-level data units (a multiple request-response model). HTTP/1.1, SMTP, FTP-control, etc., are members of this class. This class should be further refined into different subclasses based on a statistical classification of properties such as the distribution of application-level data unit sizes, the arrival process of data units, etc., so that the members of th class are statistically homogenous.
- Each TCP connection carries a single application-level data unit in only one direction. FTP-data is a member of this class.
- Each TCP connection carries multiple application-level data units in only one direction.

5. Statistical Analysis.

- Statistcal summaries of traces.
- Statisticl estimation of model parameters for selected models
- Simulation from estimated model
- Envelope analysis

6. Proposed Cascaded On-Off Model.

- Let $\{X_i(t) : t \geq 0, i = 1, 2, 3...\}$ be a sequence of independent stationary Continuous Time Markov Chains (CTMC) on the state space $\{0, 1\}$ with rate matrix

$$A_i = \begin{bmatrix} -2^{i-1}\mu & 2^{i-1}\mu \\ 2^{i-1}\lambda & -2^{i-1}\lambda \end{bmatrix}.$$

- Define

$$Z_n(t) = m\left(\frac{\lambda+\mu}{\mu}\right)^n \prod_{i=1}^{n} X_i(t), \quad t \geq 0.$$

- $\{Z_n(t), t \geq 0\}$ is the cascaded on-off model with parameters n, λ and μ. We propose to study its sample path properties, the mean, variance, quantiles and distributions of the on and off times, the autocovariance function of the Z_n process, limiting properties of the Z_n process as $n \to \infty$, queuing analysis with Z_n as the input process, statistical estimation of the parameters λ, μ, and n.

- Study the ramifications of using different scaling instead of 2^i, and of non-Markovian on-off processes at the lowest or the highest level.