Data Networks
and
Data Network Traffic

Walter Willinger
AT&T Labs – Research
Florham Park, NJ 07932-0971
Email: walter@research.att.com

April 28, 2000

Outline

- General Introduction
  - What is the Internet?
  - On Internet traffic
- Network (Traffic) Modeling: Research Problems
  - Modeling network traffic is easy, understanding it is hard ...
  - What to do about a new breed of network measurements?
  - How to visualize measurements from large-scale networks?

What Is The Internet?
A Short Answer ...

- *An <u>internetwork</u> made up of separate subnetworks*
- *A collection of computers capable of communicating with each other using a "common language" (i.e., a standard set of protocols such as IP, TCP, HTTP).*
- Terminology:
  - "Hosts" are endpoints
  - "Routers" are internal switching stations
  - "Autonomous systems (ASs)" are sub-networks under a single administrative authority
  - "Links" are (physical) connections between hosts/routers

What Is The Internet?

Another Short Answer ...

- A Large-Scale, Complex System That Is
  - constantly changing
  - extremely heterogeneous
- Some Other Characteristics:
  - rapid (exponential) growth
  - highly engineered
  - distributed control

The Internet
A Large-Scale System

- Number Of Internet Hosts (Courtesy of Vern Paxson)
  - Early 1989: about 80,000
  - Early 1990: about 200,000
  - Late 1996: about 10,000,000
  - Early 1999: about 45,000,000
  - Early 2000: around 60,000,000
- Number Of ASs (US)
  - Nov 1997: 3000 nodes, 5150 edges
  - Dec 1998: 4400 nodes, 8250 edges

The Internet
A Highly Heterogeneous System

- Link Speed (Bandwidth)
  - Modem connections: Kbps range
  - Fiber optic links: Mbps/Gbps range
- Latency
  - LAN: negligible
  - WAN: speed of light matters
  - Satellite: can be orders of magnitude larger than in WANs

The Internet
A Highly-Engineered System

- Packet-Oriented Service
  - Information is broken down into "packets"
  - Each packet knows where it goes (router checks header)
  - Excess packets are queued in the buffers at the routers
  - If buffer capacity is exceeded, routers drop packets
- Exploiting Packet Loss
  - Used as feedback to the sender to reduce sending rate
  - TCP/IP comes with end-to-end flow control mechanism
  - Causes intricate interactions among the different connections

Internet Traffic
From Temporal to Spatio-Temporal And Beyond

- The Temporal Dynamics
  - Traditional perspective, with focus on single link/router
  - Traffic rate processes, arrival processes
- The Spatio-Temporal Dynamics
  - Traffic as a function of time and space
  - Open research problem
- The Spatio-Temporal+ Dynamics
  - Traffic as a function of time, space, and network layer ...

Internet Traffic
There Is No Such Thing As "Typical" ...!

- Applications (Courtesy of V. Paxson)
  - Oct.1992: LBNL made 45 www connections
  - Oct.1998: LBNL made 700,000 www connections each day
  - Recently: There are an estimated 9,000,000 napster users
- Sites (Courtesy of V. Paxson)
  - netnews: 15% at LBNL, 60% at USC (Jan. 1991)
  - Mbone: 20% at DEC-WRL, 50% at Bellcore (Jan. 1995)
  - www: 70% on MCI backbone link, 37% at UCB (late 1997)
  - napster: 10% (non-academic site), more than 50% at an university site (recent)

Internet and Internet Traffic
Challenges

- Diversity And Change
  - Over time, across sites
  - How the network is used, by whom
- Taking Complexity Serious
  - A single node/queue is not a network
  - Traffic is not *exogenously* generated but is the result of *endogenous* conditions that constantly change

Internet and Internet Traffic
Unique Opportunities

- Highly Engineered Structure
  - Layered protocol architecture, routing
- Unique Measurement Capabilities
  - The Internet is a collection of computers
  - Computers are good at measuring
- It is difficult to think of any other area in the sciences where the available data provide such detailed information about so many different facets of (user and network) behaviors.

Network Traffic Modeling
The Conventional Approach

- Traditional Methods From "Statistical Inference"
  - "Squeezing a small data set dry"
  - Obsession with finding the "true model"
  - Looking for the "best-fitting" model
- "Black-Box" Modeling Approaches
  - Purely descriptive
  - Provide no insights and/or physical understanding
  - Highly sensitive to underlying networking conditions
- Outcome: **obvious, non-informative, nothing new**.

Toward A "New Kind" Of Traffic Modeling

Requirements: Measurement-Based and Physical

- Focus On "Scientific Inference Methods"
  - Not being concerned with finding "best-fitting" model
  - Avoid tinkering with fine details of the data
  - Focus on robust properties ("invariants") in the data
  - Fully exploit the availability of many large data sets
- Focus On "Physical-Based" Approaches
  - Incorporate network-host interactions (e.g., protocols)
  - Exploit hierarchical networking structure
  - Robust under constantly changing networking conditions
- By-Product: **Parsimonious models of network traffic**

An Example
The "Fractal" Nature Of Network Traffic

- **Scientific Inference:**
  - The discovery of fractal traffic is the result of an extensive analysis of a large number of large data sets
  - It is easy to distinguish between "fractal" or "self-similar" traces and traditionally assumed network traffic
- **Importance Of Independent Follow-Up Studies**
- **Physical-Based Construction:**
  - Provides an answer to "Why is network traffic self-similar?"
  - Results in a detailed physical understanding of self-similarity in the network context (i.e., "heavy tails")

An Example (Cont.)
The "Fractal" Nature Of Network Traffic

- New Insights Into Network Traffic Dynamics
  - Due to user/session characteristics
  - Has little to do with the network
  - Is likely here to stay – because "heavy tails" are here to stay (e.g., file sizes, properties of web documents)
  - Mathematical foundation for effectiveness of IP switching
  - Mathematical basis for novel Web-caching strategies
- An Answer That leads To New Questions
  - Where is the network (e.g., protocol effects, competition)?
  - There must be more to network traffic than self-similarity!

Beyond Network Traffic Modeling

New Developments

- New Breed Of Network Measurement Infrastructures
  - Network-wide
  - Data collection from potentially thousands of probes distributed across the global Internet
  - Example: NIMI
- New Breed Of Network Measurements
  - Huge volumes, high semantic context
  - Yields information about spatio-temporal+ dynamics

Beyond Network Traffic Modeling

New Challenges

- New Measurements Require New Analysis Methodologies
  - How to fully exploit the available data
  - What to look for, how to correlate data across probes
- Urgent Need For Adequate Visualization Techniques/Tools
  - How to handle large-scale networks?
  - How to deal with spatio-temporal processes?
  - How to display local information without losing the global perspective?