The Flexlab Approach To Realistic Evaluation of Networked Systems

Robert Ricci, Jonathon Duerig, Pramod Sanaga, Daniel Gebhardt, Mike Hibler, Kevin Atkinson, Junxing Zhang, Sneha Kasera, and Jay Lepreau

NSDI 2007
April 12, Cambridge, MA
Emulators

• Examples: ModelNet and Emulab
• The Good: Control, repeatability, wide variety of network conditions
• The Bad: Artificial network conditions
Examples: RON and PlanetLab
The Good: Real network conditions, deployment platform
The Bad: Overloaded, few privileged operations, poor repeatability, hard to develop/debug on
Evaluating Networked Systems: Flexlab
Goal: Real Internet within Emulator
The Flexlab Approach

Measure
The Flexlab Approach

Measure

Model
The Flexlab Approach

Measure

Model

Emulate
The Flexlab Approach

Measure → Model → Emulate
The Flexlab Approach

Measure

Model 2

Emulate
Key Points

• Software framework for pluggable network models
• Application behavior can drive measurements & model in real-time
• Application-Centric Internet Modeling
  – High fidelity measurement/emulation technique
  – Includes new techniques for ABW measurement
More in the Paper

- Flexible network measurement system
- Network stationarity results
- Two straightforward network models
- Shared bottleneck analysis
- PlanetLab scheduling delay measurements
Flexlab Architecture
Flexlab: Application

Emulab Host

Application

Application Traffic

Application

Emulab Host
Flexlab: Application Monitor
Flexlab: Network Model

Offered Load Model → Network Model

Emulab Host → Application

App Monitor → Application Traffic

Emulab Host → Application

App Monitor
Flexlab: Measurement Repo.
Flexlab: Path Emulator

- Application Emulab Host
- Offered Load Model
- Network Model
- Path Emulator
- Application Traffic
- Application Monitor
- Measurement Repository

Network Characteristics
ACIM: Application-Centric Internet Modeling
Imagine Ideal Fidelity

PlanetLab Host

Internet

Application Traffic

PlanetLab Host

Emulab Host

Application

Emulab Host

Application
ACIM Architecture

- PlanetLab Sliver
  - Agent
  - Application
  - App Monitor
- Emulab
- Internet
- Measurement Traffic
- Path Emulator
- Application Traffic
ACIM Design Challenges

• Determining when to drop packets
• Finding relationship between throughput and ABW
• Extension to UDP
• CPU starvation on PlanetLab
  – Host artifacts in throughput
  – Packet loss in libpcap
ACIM Path Emulator Parameters

Available bandwidth

Packets enter

Queuing delay

Packets leave

All other delay

Available bandwidth

Available bandwidth

Available bandwidth

Available bandwidth

Available bandwidth
All Other Delay

- Base RTT: Smallest RTT seen recently [Vegas 95]
- Packets saw little or no queueing delay
Packet Loss

- Caused by full queue at bottleneck link
  - Difficult to measure directly
- So measure queue length in time: Max recent RTT - Base RTT
Throughput and ABW
Throughput and ABW

- **Offered load**
- **Available bandwidth**

Bandwidth vs. Time

Delay vs. Time
Throughput and ABW

- Offered load
- Available bandwidth
- Measured throughput
## Throughput and ABW

<table>
<thead>
<tr>
<th>Throughput and ABW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
</tr>
<tr>
<td>Bandwidth</td>
</tr>
<tr>
<td>Offered load</td>
</tr>
<tr>
<td>Available bandwidth</td>
</tr>
<tr>
<td>Measured throughput</td>
</tr>
</tbody>
</table>

- **Throughput and ABW**
- **Time**
- **Bandwidth**
  - Offered load
  - Available bandwidth
  - Measured throughput

**Diagram:**
- Bandwidth axis with three labels:
  - Offered load
  - Available bandwidth
  - Measured throughput
- RTT and Delay axes with one label each:
  - RTT
  - Delay
Throughput and ABW

- Offered load
- Available bandwidth
- Measured throughput

Bandwidth vs. Time

RTT vs. Time
Throughput and ABW

- **Offered load**
- **Available bandwidth**
- **Measured throughput**

**Bandwidth** vs. **Time**

- **RTT**
- **Delay**
Throughput and ABW

- Offered load
- Available bandwidth
- Measured throughput

Bandwidth vs. Time

- RTT
- Delay

Measured throughput and available bandwidth decrease with time as the offered load increases.
Throughput and ABW

- Offered load
- Available bandwidth
- Measured throughput

Time

Bandwidth

RTT

Delay
Throughput and ABW

- Offered load
- Available bandwidth
- Measured throughput

Bandwidth vs. Time

Delay vs. Time

RTT vs. Time
Throughput and ABW

• If (throughput > last ABW measurement), use new value

• Else, look for indications that throughput has reached ABW
  – Socket buffer is filling up AND
  – Recent RTTs have been increasing
  – Using linear regression
ACIM Features

- Precise: assesses only relevant parts of the network
  - Scales in nodes and paths
- Complete: automatically captures all relevant network behavior
  - Simpler to measure e2e effects than find causes
  - Detects rare and transient effects
  - Evokes all reactive network behaviors (except content-based)
  - Rapidly tracks conditions
ACIM Accuracy

- Is ACIM path emulation accurate?
- Is it accurate at fine granularity?
Methodology

- iperf runs in Emulab
- Measurement Agent runs on PlanetLab at UT Austin and AT&T Research
- We added transient TCP cross traffic between these sites
TCP iperf Throughput

Measurement Agent

Flexlab with ACIM
A Real Application

• Does ACIM give accurate results for a real, complicated application?
A Real Application

- Does ACIM give accurate results for a real, complicated application?
- ... does PlanetLab?
A Real Application

• Does ACIM give accurate results for a real, complicated application?

• ... does PlanetLab?

• Can we discover ground truth?
Methodology: BitTorrent

- Two simultaneous instances of reference BitTorrent:
  - One on PlanetLab
  - One in Flexlab
- Eight nodes in US and Europe: One seed, seven clients
- We reduced randomness in BT ... but some still remains
BitTorrent w/ CPU Reservation
BitTorrent w/ CPU Reservation

PlanetLab: 5.2 Mbps average

Flexlab: 5.4 Mbps average
BitTorrent w/o CPU Reservation

PlanetLab: 2.3 Mbps average

Flexlab: 5.8 Mbps average
BitTorrent Bottom Line

• Conclusion: For this experiment, both Flexlab and PlanetLab with CPU reservations give accurate results
  – PlanetLab alone does not

• CPU availability on PlanetLab hurts BitTorrent

• ACIM reduces host resource needs on PlanetLab for this experiment
  – BitTorrent: 36-76% CPU
  – ACIM Agent: 2.6% CPU
  – Factor of 15 - 30 CPU
  – Factor of 4 memory
The Future?

- No need to perfect in PlanetLab:
  - Full resource isolation
  - Total control over hosts
  - Orthogonal control network
- ... use in the emulators that already have them
- Use PlanetLab nodes as NICs
- Conserve resources for deployed services with end users
Conclusion

- New approach to evaluating networked systems
- Separates the network model
- Designed to leverage vibrant measurement and modeling community
- Couples an emulator to an overlay testbed
- ACIM high fidelity emulation technique
- Contact testbed-ops@emulab.net to use
Backup Slides
Why not just add more nodes to every PlanetLab site?

- Remaining problems:
  - Poor repeatability
  - Hard to develop/debug
  - No privileged operations
- Some malicious traffic cannot be tested
- Some Flexlab network models reduce network load
- Emulab node pool stat muxed and shared more efficiently than per-site pools
- Overload can (will?) still happen with PL's pure shared-host model
- Major practical barriers: admin, cost
Flexlab and VINI

- Entirely different kinds of realism and control
- Flexlab: passes "experiment" traffic over shared path
  - Real Internet conditions from other traffic on same path, but app. traffic is not from real users
  - Control: of all software
  - Environment: friendly local dev. environ, dedicated hosts
- VINI: can pass "real traffic" over dedicated link
  - Real routing, real neighbor ISPs, potentially traffic from real users, but network resources are not realistic/representative
  - Dedicated pipes with dedicated bandwidth, that insulate experiment from normal Internet conditions
  - Control: restricted to VINI's APIs (Click, XORP, etc.)
  - Environment: distributed environ; shared host resources
## Change Point Analysis

<table>
<thead>
<tr>
<th>Path</th>
<th>High</th>
<th>Low</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asia to Asia</td>
<td>2</td>
<td>1</td>
<td>0.13%</td>
</tr>
<tr>
<td>Asia to Commercial</td>
<td>2</td>
<td>0</td>
<td>2.9%</td>
</tr>
<tr>
<td>Asia to Europe</td>
<td>4</td>
<td>0</td>
<td>0.5%</td>
</tr>
<tr>
<td>Asia to I2</td>
<td>6</td>
<td>0</td>
<td>0.59%</td>
</tr>
<tr>
<td>Commercial to Commercial</td>
<td>20</td>
<td>2</td>
<td>39%</td>
</tr>
<tr>
<td>Commercial to Europe</td>
<td>4</td>
<td>0</td>
<td>3.4%</td>
</tr>
<tr>
<td>Commercial to I2</td>
<td>13</td>
<td>1</td>
<td>15%</td>
</tr>
<tr>
<td>I2 to I2</td>
<td>4</td>
<td>0</td>
<td>0.02%</td>
</tr>
<tr>
<td>I2 to Europe</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>Europe to Europe</td>
<td>9</td>
<td>1</td>
<td>12%</td>
</tr>
</tbody>
</table>
Simple Static Model

Flexmon All-Sites PlanetLab Measurements

Static Network Model

Datapository

Network Characteristics to Path Emulator
Simple Dynamic Model

Flexmon All-Sites PlanetLab Measurements

Dynamic Network Model

Datapository

Application Network Model from Monitor

Network Characteristics to Path Emulator
**Flexmon Architecture**

- **Manager**
  - Manager Client
  - Manager Client
  - Manager Client
  - Auto-Manager Client

- **Path Prober**
  - Path Prober
  - Path Prober
  - Path Prober

- **Data Collector**
  - Data Collector

- **Datapository**
  - Datapository

- **Flexlab**
  - Flexlab

- **PlanetLab**
  - PlanetLab

- **Emulab**
  - Emulab

- **Shared**
- **Adaptive**
- **Reliable**
- **Controllable**
- **Safe**
- **Accommodates high-performance data retrieval**
CPU Starvation on PlanetLab

- Host Artifacts
  - Long period when agent can't read or write
  - Empty socket buffer or full receive window
  - Solution: Detect and ignore

- Packet loss from libpcap
  - Long period without reading libpcap buffer
  - Many packets are dropped at once
  - Solution: Detect and ignore
Reverse Path Congestion

- Can cause ack compression
- Throughput Measurement
  - Throughput numbers become much noisier
  - We abuse the TCP timestamp option
  - PlanetLab: homogeneous OS environment
  - Extending it would require hacking client
- RTT Measurement
  - Future work
Initial Conditions

• Needed to bootstrap ACIM
  – ACIM uses traffic to generate conditions
  – But conditions must exist for first traffic

• We created a measurement framework
  – All pairs of sites are measured
  – Put data into measurement repository
  – Set initial conditions to latest measurements
Simultaneous TCP iperf

Throughput (Mbps)

Time (seconds)

PlanetLab

Flexlab with ACIM

Throughput (Mbps)

Time (seconds)
Repeatability vs. Fidelity

- Emulab
- Static
- General Internet Model
- Dynamic
- ACIM
- PlanetLab

Higher Network Fidelity

More Repeatable
Throughput and ABW

Agent write()s

Packets On The Wire

Data

Data

Data

Data

ACK

ACK

ACK

ACK

Time

Application
Offered Load

RTT

Throughput

Avail-BW?
Currently available for Beta Testing

http://www.flux.utah.edu/flexlab
UDP Streaming Video

Throughput (Mbps) vs Time (seconds)

- Orange line: Throughput
- Green line: Throughput
Opens Up New Questions

• Further validation

• Accuracy tests at runtime
  – Similar in spirit to Emulab's linktest

• Use to compare models
  – Find which models most appropriate for different classes of applications

• Replay for ACIM

• Study fidelity of different software combinations
  – Different TCP implementation or OS in Emulab