Measurement questions and the infrastructure to store and analyze the data needed to answer them

Steven Bauer
MIT

2/10/2016

Outline

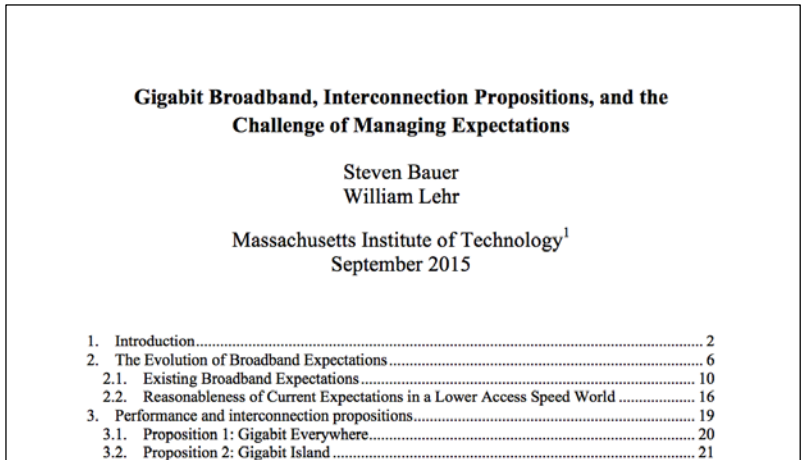
- Measuring the gigabit Internet
- Our efforts to build infrastructure for storing and analyzing large amounts of data from multiple measurement efforts
  - CAIDA/MIT congestion project
  - FCC MBA analysis
  - MBM measurements (future)

Why performance expectations of regulators, users, edge providers matter for coming gigabit Internet

1. Current performance expectations are not appropriate during the transition to a gigabit broadband world
2. Current measurements tools / techniques are not adequate in a gigabit broadband world
3. **Potential to delay or disrupt deployment of very high-speed broadband**

What are reasonable expectations for gigabit broadband?

- Prop #1: Gbps everywhere
- Prop #2: Gbps island
- Prop #3: Gbps in aggregate only
- Prop #4: Gbps somewhere
- Prop #5: Growing toward Gbps paths



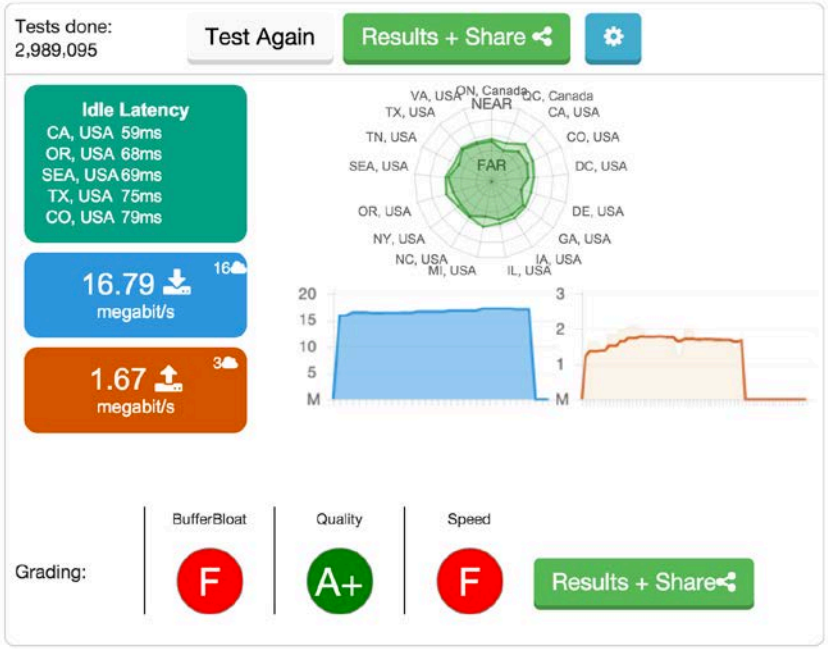
Gigabit Broadband, Interconnection Propositions, and the Challenge of Managing Expectations

Steven Bauer
William Lehr

Massachusetts Institute of Technology[1]
September 2015

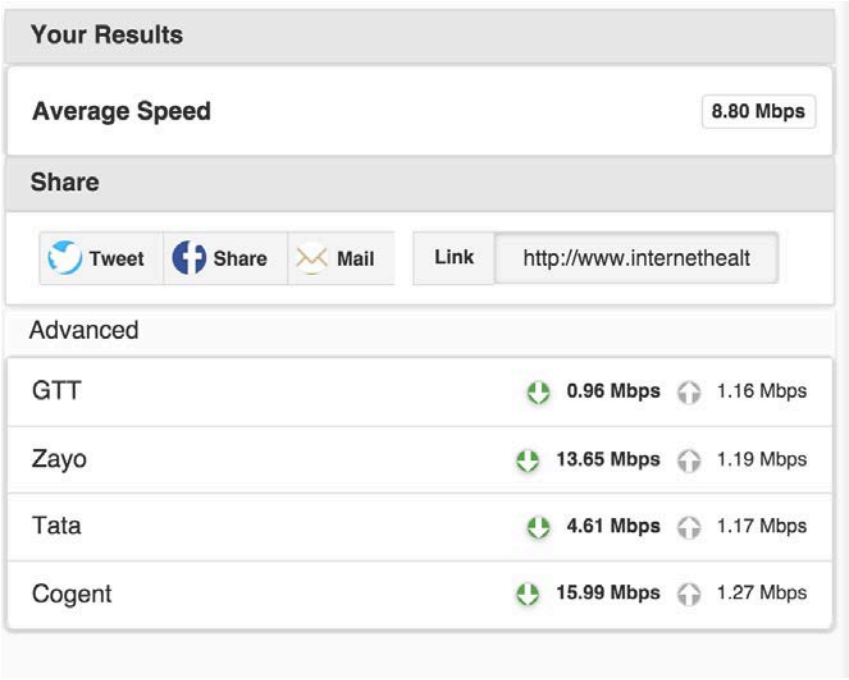
DSLReports Speed Test

- Multiple simultaneous TCP tests to different destinations
- Multiple streams to each destination
- Destination servers are not necessarily nearby
- Result is aggregate throughput achieved



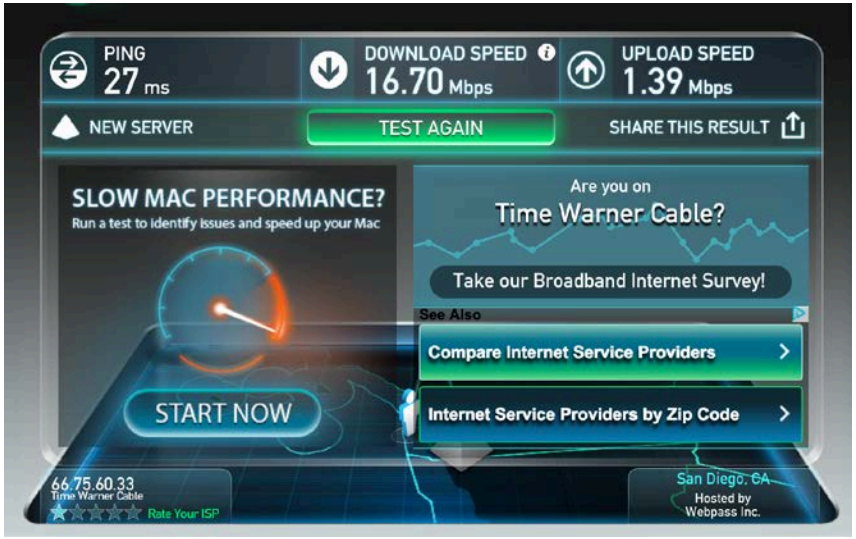
Internet Health Test

- Sequential single stream NDT tests
- Servers are selected to cross interconnects
- Result is average of all sequential tests



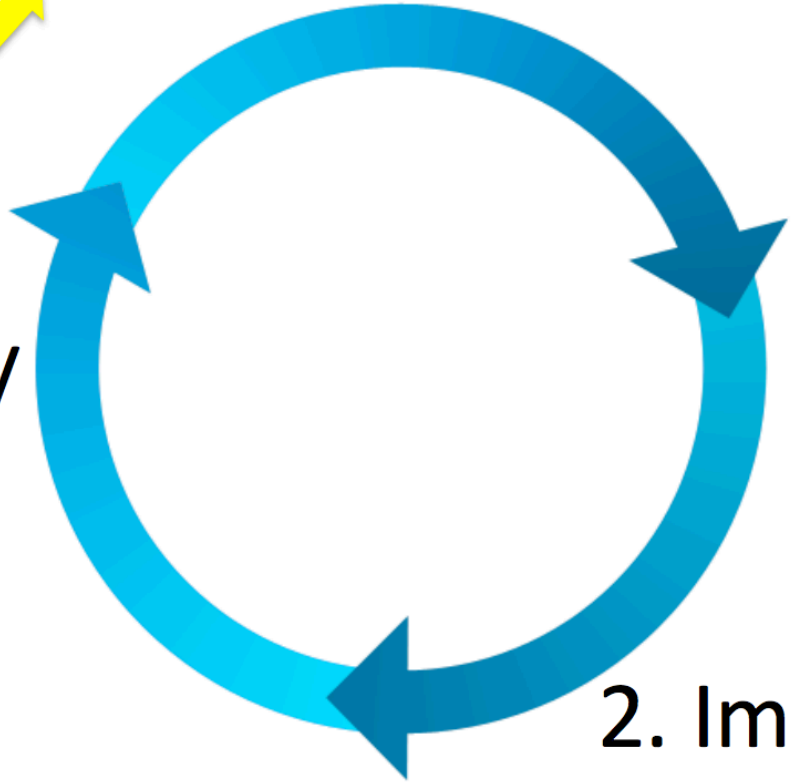
Speedtest.net

- Multiple concurrent TCP streams to single server
- Closets server is selected
- Result is aggregate throughput achieved by all streams



1. Research Ideas / Questions

We talk to each other about these a lot



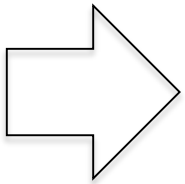
We should talk to each other about the engineering details more. It is fun, educational, therapeutic.

2. Implementation

3. Evaluation / Results

Last year

Analytic code was mostly hand crafted python code
Single machine
Data store in text files



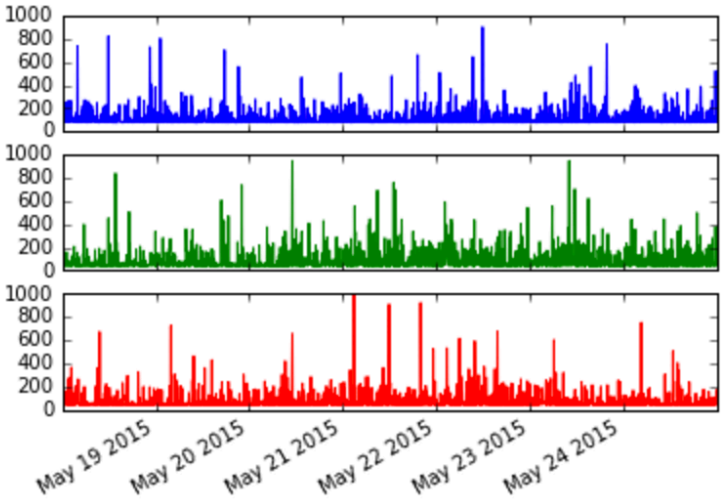
This year

Heavy use Spark, Pandas, etc.
OpenStack cluster
Apache Parquet + Postgres

75%	1	1.432361e+09	77.000000	102.225000	118.350000
max	1	1.432512e+09	990.900000	948.000000	901.700000

In [78]:
```
ptid1[['far_min', 'middle_min', 'near_min']].plot(kind='line', subplots=True, legend=False)
```

Out[78]:
```
array([<matplotlib.axes._subplots.AxesSubplot object at 0x7f56aa087350>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x7f56aa004d50>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x7f56aa035510>], dtype=object)
```



In [67]:
```
from wavelets import *
%matplotlib inline
```

Dataflow / Apache Beam?

“Unified programming model for developing and executing a wide range of data processing patterns including ETL, batch computation, and continuous computation.”

“Dataflow provides programming primitives **such as powerful windowing and correctness controls that can be applied across both batch and stream based data sources.** Dataflow effectively eliminates programming model switching cost between batch and continuous stream processing by enabling developers to express computational requirements regardless of data source.”