something

Paper #0, 4 Pages

Abstract

FCC reclassifying broadband and broadband speeds
HD content, video traffic, CDNs
How should ISPs progress?
Want to know the current utilization during peak periods
Better way to offer broadband

1 Introduction

The FCC has been charged with controlling broadband in this country Their control on ISPs shows that it is an important institute for the future of broadband deployment and use in US Its responsibility is to increase deployment to rural areas, increase US ranking, get everyone to have broadband connectivity.

The core of the Internet lies here, service lies here. FCC is the most important commission for the future of Internet access of eyeball American users right now. Recent Internet trends have changes so much with videos and end-user data.

Who pays for this? How do we solve issues between content servers, multiple transport providers, and the end-user? FCC is the one setting down the laws. Economies of play change with FCC

Thus anything the FCC comes up with must be checked and sanitized. From next steps, to valid metrics to evaluate broadband usage, as well as the definition of broadband. Currently FCC bases its decisions on aggregates throughout this huge country, and divides regions based on rural and urban. It also compares it self to other developed nations. The most recent report as of 2015 Jan [1] redefines BB speeds as 25/3 and mentions that rural deployment is not keeping up based on this definition of broadband.

Section 706 of the Telecommunications Act of 1996, as amended (1996 Act), requires the Commission to determine and report annually on "whether advanced telecommunications capability is being deployed to all Americans in a reasonable and timely fashion."

1.1 Issues for Inquiry

In the FCC report No. 14-113, on Aug 5, 2014 [**?**], the FCC asks some relevant questions about broadband usage, and requests comments from the community to improve its decision making process. We summarize their comments and hypothesis as follows:

Hypothesis 12: *Household Bandwidth Scenarios* (Table 2, [**?**]). The typical bandwidth a household may need today varies between 4 to 10 Mbps for low to high usage households during peak period. Is this still valid with continuous introduction of new services and connected devices?

Hypothesis 13: *Peak Usage Time.* Should bandwidth requirements for a typical household be assessed during peak Internet usage periods, from 7 pm to 11 pm on weeknights? Is the "peak usage time" an efficient metric, or should the average usage of a household over a day be considered instead? Does establishing a reasonable household usage scenario during peak periods assist the Commission to identify a benchmark?

Hypothesis 14: *Broadband Speed Benchmarks.* SG: make this another subsection with following points 14-...? What is the right benchmark to represent moderate use for a mid-range needed by a 3 user household? Even though the commission has recently decided to set the benchmark to 25/3 [], anticipating future usage, the growth in Internet usage with Netflix super HD [] and our analysis 4 show that setting such limits is not cool. 15: How should the Commission forecast future household broadband uses to justify such a benchmark?

Hypothesis 16: emphAdoption Based Benchmarks. Similarly for uplink, the benchmark is based on 70% adoption rate - does it even make sense? 17. Symmetrical usage like video calls - does it impact aggregate usage at all?

Hypothesis 19: Does it make sense to base the benchmark on the fastest speed tier for which a substantial portion of the consumers subscribe. How should the Commission define "substantial portion" and how should we interpret such demand?

Hypothesis 22. *Other Speed Benchmarks* Broadband requirements are not uniform throughout the nation. Some users will have significantly greater needs. Should FCC opt for multiple benchmarks depending on user scenario, usage, occupation, and different benchmarks for schools, libraries, etc.? TODO: characterize differing usage even in 100 Mbps/250 tier – a user taxonomy, include Sandvine report taxonomy and show of variance in aggregate users – we will show extreme variance in the same band of users and motivate a need of new benchmarks instead of speed. Note: does this end up motivating a case for non-net neutrality based on low usage vs high usage? did FCC take the wrong decision – if we could show our data set uses completely different set of sites etc...

Hypothesis 28. *Data Usage.*

SG: price of tier increases but comcast usage is same here?

What we don't do: latency, application usage, mobile speed benchmarks.

But (as we see in policy papers) this may not be the right way to go about stuff. Aggregates do not show the right representation of users. And aggregates will not give the right policy model for offering broadband. Now that ISPs are regulated it is important that broadband be offered based on the users demands instead of extracting money out of users who don't need to, or letting users buy highly expensive plans of the ISP when the trouble is that the service they are connecting to is not fast enough.

Also section 706 [Section 706 Advanced Services Inquiry] is shit (based on policy papers) and may not be the right way to go. Although this deals with deployment, hard limits don't make sense. We need to look at sliced measurements from home users, how much data do they use even when given the ability to max out their limits.

Thus we do Comcast experiment to get data directly from urban city home user gateways on uploads and downloads. We want to comment on high tier users to find out if FCC limits even make sense with usage patterns at all??

We get data from Comcast to comment on some of these questions in urban single city high tier controlled set. Our analysis sets the baseline and framework for sanitizing the FCC with real measurements, easily available at ISPs, so that future Internet deployment and categorization is not a fuck up.

(A roadmap of the paper as follows: ...)

2 Data Source

Our dataset consists of network usage byte counters reported by Comcast gateways every 15 minutes from October 1, 2014 to December 29, 2014. There are two sets of broadband tiers that were used to collect this data: controlset, consisting of homes and businesses with a 105 Mbps access link, and the testset, consisting of homes and businesses that were paying for a 105 Mbps access link, yet were receiving 250 Mbps instead. Users in the test set were selected randomly and were not told that their access bandwidth has been increased. There were more than 15000 gateway devices in the control set, with varying usage over the three months, and about 2200 gateway devices in the test set. TODO: confirm - these were reported by Comcast gateways right?

Both the testand controlsets were collected from users in Salt Lake City, Utah, to avoid any biases in behavior based on location. Although this dataset corresponds to just one ISP, we believe that it is broadly representative of urban users in the US in the same, or higher broadband bandwidth tier (100 Mbps). Thus, we use this data to draw general conclusions about behavioral change with link capacity TODO: (add more here...)

SG: Supplement the data with bismark?

SG: My Speed Test Usage Patterns data - Any chance?

2.1 Data Description

Comcast splits the controlset into 8 separate pools on different date ranges and gateways TODO: confirm if there is repeated device IDs in control1-8. . Each dataset contains the following relevant fields: Device ID, sample period time, service class, service direction, IP address, and the bytes transferred in the 15 minute sample slot, as described in table 1. TODO: find out more about service class name, and IP addresses being the same across all sets

Field	Description
Device_number	Arbitrarily assigned CM device identifier
end_time	Fifteen minute sample period end time
date_service_created	Service start (not used in our analysis)
service_class_name	Used to differentiate data application
cmts_inet	Cmts identifier (derived from ip address)
service_direction	1-downstream, 2-upstream
port_name	Cmts port descriptor
octets_passed	Byte count
device_key	not used in our analysis
service_identifier	Service id (not used in our analysis)

Table 1: *Field Descriptions for Comcast Dataset by Comcast*

2.2 Data Processing

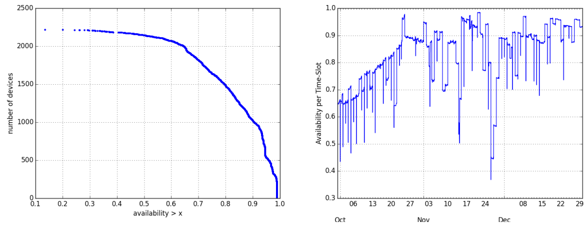
To process the large amount of data (more than 15000 unique devices, 96 time slots per day, 3 months, multiple service classes per time-slot), we first split the data by direction into uplink and downlink. The nature of the questions we ask in section 1 encourages us to concentrate on the downstream data, although we present similar results for upstream data in section 4. We also do not use the service class name identifier in our analysis, which is internal to Comcast. TODO: Our analysis showed that ignoring the service class identifier does not change our conclusions on usage patterns and statistics. Thus we only considered overall usage for each gateway device in each time slot, in uplink and downlink direction.

Note that 15 minute time slots were synchronized to the same time stamp. They were off only by a few seconds (under 30, as claimed by Comcast). As our analysis deals with aggregated patterns on a granularity of 15 minutes, we believe this time synchronization to be irrelevant.

2.3 Data Sanitization

Our initial analysis of data transferred per time slot showed that certain gateway devices were responsive only for brief periods. We also noticed that certain time slots had a very low response rate throughout the dataset. TODO: Why? asked comcast .

We evaluate the fraction of responsiveness of a gateway throughout the dataset, as well as the fraction of responsiveness per time slot, and call this the **availability**. Figure 1 shows how the number of devices decreases for a higher availability requirement. Based on the common trend of this plot throughout the testand controldatasets, we decided to



(a) *Availability by device* **(b)** *Availability by date*

Figure 1: *Availability, based on gateway device responsiveness.*

only choose gateway devices with an availability of at least 0.8.

On exploring control4, we noticed that the dataset spanned the October-December period, but for the first week there were 15000 unique devices reporting their usage statistics, while after the first week, the number dropped to 5000. Furthermore, control5 and control6 reported usage for 5000 devices in the month of November, but only a 100 devices in December. We did not want stray devices impacting our measure of availability, therefore we sliced the controldatasets to monthly date ranges with a minimum of 4000, or at least half the total unique devices present.

Finally, we sliced the sanitized testset based on the date range of each individual controlset for comparison. We compared each of these tests individually to ensure that there are no outliers. We refer to the testand controlsets in this case simply as datasets $1-8$. To analyze data by each month, we also sliced and combined the sanitized data to give us controland testdata for October, November, and December, referred to as *oct*, *nov*, *dec*. Finally, we combine all controlsets to form a large concatenated dataset over the same date range as the complete testdataset, and we refer to this simply as $full$. A description of these sanitized sets is provided in table 2

TODO: Figure 1 Make common **EPS** plot of availability – 8 control sets (half filtered) + test sets vs availability.

SG: Figure 1a should show why we chose 0.5 as threshold, Figure 1b should show why we expected 3k-4k sanitized devices per set

TODO: Table 2 recheck all numbers and rewrite para accordingly, it seems control4 did not end at just one month - continued till December

TODO: Table 2 turn it around

3 Methodology

4 Results

Q/A
Plots

5 Discussion

results that are contradictory
results that data suggests that is wrong
- suggest that this may be due to representation

a better way to measure and offer broadband based on utilization?

6 Related Work

fcc reports + sandvine on usage patterns
papers studying broadband vs utilization
Note: this can go in the intro if lack of space

Dataset	set$_1$	set$_2$	set$_3$	set$_4$	set$_5$	set$_6$	set$_7$	set$_8$	set$_{oct}$
Start Time	2014-09-30	2014-10-01	2014-10-01	2014-11-01	2014-11-01	2014-11-01	2014-11-01	2014-12-01	2014-10-01
End Time	2014-10-31	2014-10-31	2014-10-312	2014-12-29	2014-12-29	2014-12-29	2014-12-29	2014-12-29	2014-10-30
Devices$_{control}$	3627	4033	3969	1266	3632	3852	3644	4277	11629
Devices$_{test}$	1481	1481	1481	1481	1481	1481	1481	1481	1481

Table 2: *Sanitized Dataset Description: Most of our analysis will be based on set$_{full}$ unless otherwise stated*

References

[1] Federal Communications Commission. International Broadband Data Report (Fourth), February 2015. (Cited on page 1.)