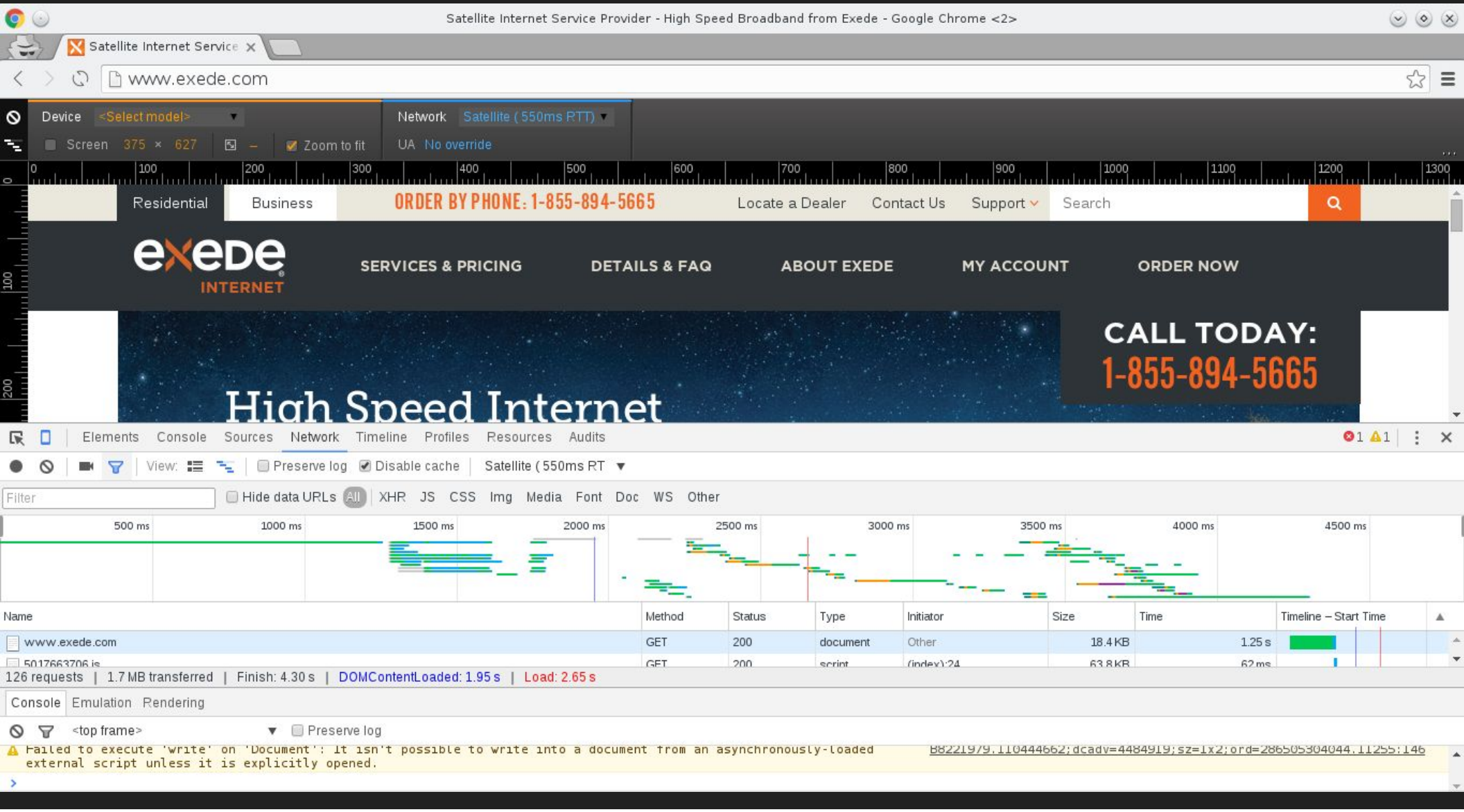
The Problem

- Potentially Lots of Internet Traffic
- ISP DNS Cache Good
- Local DNS Cache Limited or Nonexistent
- **Increased Page Load Times on High-Latency Networks**
  - Satellite Internet ~550ms Latency


Satellite Internet Service Provider - High Speed Broadband from Exede - Google Chrome <2>
Satellite Internet Service
www.exede.com
Device <Select model>
Screen 375 × 627
Zoom to fit
Network Satellite (550ms RTT)
UA No override
Residential
Business
ORDER BY PHONE: 1-855-894-5665
Locate a Dealer
Contact Us
Support
Search
exede INTERNET
SERVICES & PRICING
DETAILS & FAQ
ABOUT EXEDE
MY ACCOUNT
ORDER NOW
CALL TODAY:
1-855-894-5665
High Speed Internet
Elements
Console
Sources
Network
Timeline
Profiles
Resources
Audits
Preserve log
Disable cache
Satellite (550ms RT
Filter
Hide data URLs
All
XHR
JS
CSS
Img
Media
Font
Doc
WS
Other
500 ms
1000 ms
1500 ms
2000 ms
2500 ms
3000 ms
3500 ms
4000 ms
4500 ms
Name
Method
Status
Type
Initiator
Size
Time
Timeline – Start Time
www.exede.com
GET
200
document
Other
18.4 KB
1.25 s
126 requests | 1.7 MB transferred | Finish: 4.30 s | DOMContentLoaded: 1.95 s | Load: 2.65 s
Console
Emulation
Rendering
<top frame>
Preserve log
Failed to execute 'write' on 'Document': It isn't possible to write into a document from an asynchronously-loaded external script unless it is explicitly opened.

A Solution - Enter Cloud Computing Technologies

- Apache Kafka
- Spark Streaming
- Machine Learning
- Push-Based Local DNS Cache

Algorithms/Approaches

- Similarity Analysis - Starting Points
  - Streaming K-Means
  - Locality Sensitive Hashes
- Metrics to Examine
  - Query Source
  - Query Domain
  - Query Time

Related Work

- Accelerating Last-Mile Web Performance with Popularity-Based Prefetching (Sundaresan, Magharei, Feamster, Teixeira)
- Behavior-based tracking: Exploiting characteristic patterns in DNS traffic (Herrmann, Banse, Federrath)
- Google news personalization: scalable online collaborative filtering (Das, Datar, Garg, Rajaram)
- Streams and Funnels (Raghavan, AuYoung, Smith)
- Many Others - Focused Primarily on Spam Detection, Security, Tuning

Data Sets

- Toy Dataset: personal BIND server(s) pushing logs to Kafka
  - Basic Testing/Debugging
  - Unlikely to Yield Useful Results
  - No anonymity
- Anonymized Datasets: 24 hours of DNS Requests from Medium-Sized Enterprise Servers
  - Millions of Requests
  - Anonymized IP Addresses
  - Anonymized Domains
  - Anonymization Allows Clustering
- Possibly Others
  - Web Crawler
  - Traffic Simulator

Evaluation

- Testing Accuracy
  - Streaming Hit Rate
    - On-Line Calculation of Predictive Accuracy
  - Different Starting Points
- Testing Performance
  - Collect metrics on throughput, latency
  - Test in different environments
    - Local desktop
    - AWS Cluster
    - Azure Cluster
- Comparison to Caches?
  - Could do in-place analysis of data to analyze hit rates assuming LRU's of different sizes
    - # of Queries Answered Since This Domain Seen Last

Citations

"6.2. How DNS Works." *How DNS Works*. The Linux Documentation Project, n.d. Web. 27 Mar. 2016. <http://www.tldp.org/LDP/nag2/x-087-2-resolv.howdnsworks.html>.

Choi, Hyunsang, and Heejo Lee. "Identifying Botnets by Capturing Group Activities in DNS Traffic." *Computer Networks* 56.1 (2012): 20-33. *ScienceDirect*. Web. 29 Mar. 2016.

"Clustering - Spark.mllib." *Spark 1.6.1 Documentation*. Apache Software Foundation, n.d. Web. 29 Mar. 2016.

Das, Abhinandan S., Mayur Datar, Ashutosh Garg, and Shyam Rajaram. "Google News Personalization: Scalable Online Collaborative Filtering." *Google News Personalization*. Proc. of Google News Personalization: Scalable Online Collaborative Filtering, New York, New York. ACM, 2007. Web. 29 Mar. 2016. <http://dx.doi.org/10.1145/1242572.1242610>.

Citations

"DomainHelp: DNS and Domain Name FAQs » The Three Basic Types of Nameservers."*DomainHelp DNS and Domain Name FAQs RSS*. Thirteen Ventures, 2010. Web. 28 Mar. 2016.

Herrmann, Dominik, Christian Banse, and Hannes Federrath. "Behavior-based Tracking: Exploiting Characteristic Patterns In DNS Traffic." *Computers & Security* 39 (2013): 17-33. *ScienceDirect*. Web. 29 Mar. 2016.

Ishibashi, Keisuke, and Kazumichi Sato. "Classifying DNS Heavy User Traffic by Using Hierarchical Aggregate Entropy." *IEEE Xplore*. NTT Corporation, 2012. Web. 29 Mar. 2016. <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6170436>.

Jung, Jaeyeon, Emil Sit, Hari Balakrishnan, and Robert Morris. "DNS Performance and the Effectiveness of Caching." *IEEE Xplore*. IEEE, 2002. Web. 29 Mar. 2016. <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1041066>.

Citations

"Latency- Why Is It a Big Deal for Satellite Internet?" *Satellite Internet Latency*. VSAT SYSTEMS, 2013. Web. 28 Mar. 2016.

Leskovec, Jure, Anand Rajaraman, and Jeff Ullman. "Mining of Massive Datasets." *Mining of Massive Datasets*. Stanford University, n. d. Web. 29 Mar. 2016. <http://mmds.org/>.

"Monitoring and Instrumentation." *Spark 1.6.1 Documentation*. Apache Software Foundation, n.d. Web. 29 Mar. 2016.

"Netnod." *What Are Root Name Servers?* NetNod, Mar. 2013. Web. 28 Mar. 2016.

Plonka, David, and Paul Barford. "Context-aware Clustering of DNS Query Traffic."*Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement Conference - IMC '08* (2008): n. pag. Web.

Raghavan, Barath, Alvin AuYoung, and Andrew Smith. "Streams and Funnels." (n.d.): n. pag. Web. 29 Mar. 2016. <http://cseweb.ucsd.edu/~atsmith/cse201_project.pdf>.

Citations

Romaña, Dennis Arturo Ludeña, and Yasuo Musashi. "Entropy Based Analysis of DNS Query Traffic in the Campus Network." *ResearchGate*. ResearchGate, Dec. 2006. Web. 29 Mar. 2016. <https://www.researchgate.net/publication/239809551_Entropy_Based_Analysis_of_DNS_Query_Traffic_in_the_Campus_Network>.

"StreamingListener." *Spark API*. Apache Software Foundation, n.d. Web. 29 Mar. 2016.

Sundaresan, Srikanth, Nazanin Magharei, Nick Feamster, and Renata Teixeira. "Accelerating Last-mile Web Performance with Popularity-based Prefetching."*Accelerating Last-mile Web Performance with Popularity-based Prefetching*. ACM, Oct. 2012. Web. 28 Mar. 2016. <http://dx.doi.org/10.1145/2377677.2377742>.

Yadav, Sandeep, Ashwath Kumar Krishna Reddy, A. L. Narasimha Reddy, and Supranamaya Ranjan. "Detecting Algorithmically Generated Domain-Flux Attacks With DNS Traffic Analysis." *IEEE/ACM Transactions on Networking IEEE/ACM Trans. Networking* 20.5 (2012): 1663-677. *IEEE Xplore*. Web. 29 Mar. 2016.

Citations

Yang, Yinghui (Catherine). "Web User Behavioral Profiling for User Identification."*Decision Support Systems* 49.3 (2010): 261-71.

ScienceDirect. Web. 29 Mar. 2016.