Models and Tools for the High-Level Simulation of a Name-Based Interdomain Routing Architecture

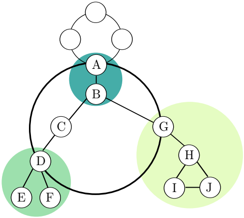
Kari Visala, Andrew Keating
Helsinki Institute for Information Technology HIIT /
Aalto University School of Science

Rasib Hassan Khan
University of Alabama

17th IEEE Global Internet Symposium, Toronto
April 27, 2014

PROBLEM: PURSUIT RENDEZVOUS ARCHITECTURE

- A hierarchical DHT [Canon] globally interconnecting *rendezvous networks* [DONA]
  - *Scopes* (containing publications) are advertised and previous query results are cached in the DHT nodes
  - Rendezvous networks are assumed to approximately evolve around neighboring stub ASes and Canon hierarchy to follow the structure of the AS graph
- Quantitative evaluation metrics
  - Distribution of latencies and overlay node and link resource usage, scalability, AS path stretch, determination of optimal cache size and number of overlay nodes



[Canon] Ganesan, P.; Gummadi, K., and Garcia-Molina, H. Canon in G Major: Designing DHTs with Hierarchical Structure Distributed Computing Systems. Proceedings, ICDCS'04, IEEE Computer Society, 2004, 263-272

[DONA] Koponen, T.; Chawla, M.; Chun, B.-G.; Ermolinskiy, A.; Kim, K. H.; Shenker, S., and Stoica, I. A Data-Oriented (and Beyond) Network Architecture SIGCOMM Comput. Commun. Rev., 2007, 37, 181-192

Problem: Approaches to Evaluation

- Complete architectures have many interfaces to the external world and require qualitative analysis, comparisons etc.
- Analytical results
  - Either too difficult or require simplifying assumptions in the case of complex, dynamic systems
- Prototyping and testing
  - PlanetLab overlay testbed: network conditions are not fully controllable, topology does not reflect the structure of the whole Internet, and the largest experiments may still not be feasible
  - NetFPGA, The Click Modular Router, OpenFlow..
- Simulation
  - Packet/router-level tools such as ndnSIM on top of ns-3: not scalable to Internet-wide scenarios
  - $\Rightarrow$ High-level approximate models

HIGH-LEVEL SIMULATION: OUR DESIGN PRINCIPLES

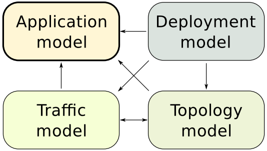
1. Construct models around known invariants, that have been empirically validated under many scenarios [Floyd and Paxson]
   - We also did not use algorithmically generated topologies that could leave out unnoticed features of the Internet
2. Tackle the scale by using aggregate models [Floyd and Paxson]
3. Parametrize the models for the uncertain variables
4. Modularize the different aspects of the simulation
5. Balance the level of detail of the different submodels
6. Use worst-case scenarios to increase confidence (datasets are incomplete etc.)

- High-level simulation can be thought as a hybrid between analytical results and a detailed simulation
  - Some aspects can be abstracted safely and the difficult parts are simulated
  - Relying on proofs leaves false negatives (too difficult to prove) and simulations allow some false positives (test cases cover the inputs only partially)

[Floyd and Paxson] Floyd, S. and Paxson, V. Difficulties in Simulating the Internet IEEE/ACM Transactions on Networking (TON), IEEE Press, 2001, 9, 392-403

Application-Driven Component Models

- The network and traffic models can be simplified by assuming a specific application
  - For example, we are only interested in the most important sources of control plane traffic
  - Problem 1: The models may not be reused without modifications
- PURSUIT is a clean-slate architecture
  - Problem 2: The invariants true for the current Internet may not hold anymore



NETWORK MODEL

- The global topology model should capture the Internet at least at the level of AS business relationships
  - Categorized in the datasets into *customer-to-provider* and peer-to-peer
  - Determine routing policies and rendezvous network formation
  - PoP-level models are still works-in-progress
- AS-level datasets contain mostly the same ASes and links but disagree about 34908 AS relationships
  - UCLA [Zhang et al.] dataset combined multiple sources: BGP route monitors, ISP route servers/looking glasses, and Internet routing registries
  - CAIDA [CAIDA] is another BGP-derived dataset
- 90% of the peering links may be missing because of the *valley-free* routing policies [Oliveira et al.]
  - IXP [Augustin et al.] identifies peering links by using a combination of IXP databases, Internet topology datasets, and traceroute-based measurements
  - We combined the UCLA and IXP datasets

[Zhang et al.] Zhang, B.; Liu, R.; Massey, D., and Zhang, L. Collecting the Internet AS-level Topology ACM SIGCOMM Computer Communication Review, ACM, 2005, 35, 53-61

[CAIDA] The CAIDA AS Relationships Dataset, November 2009

[Oliveira et al.] Oliveira, R.; Pei, D.; Willinger, W.; Zhang, B., and Zhang, L. In Search of the Elusive Ground Truth: The Internet's AS-level Connectivity Structure SIGMETRICS Perf. Eval. Rev., 2008, 36, 217-228

[Augustin et al.] Augustin, B.; Krishnamurthy, B., and Willinger, W. IXPs: Mapped? Internet Measurement Conference (IMC), 2009, 336-349

Summary of the Datasets

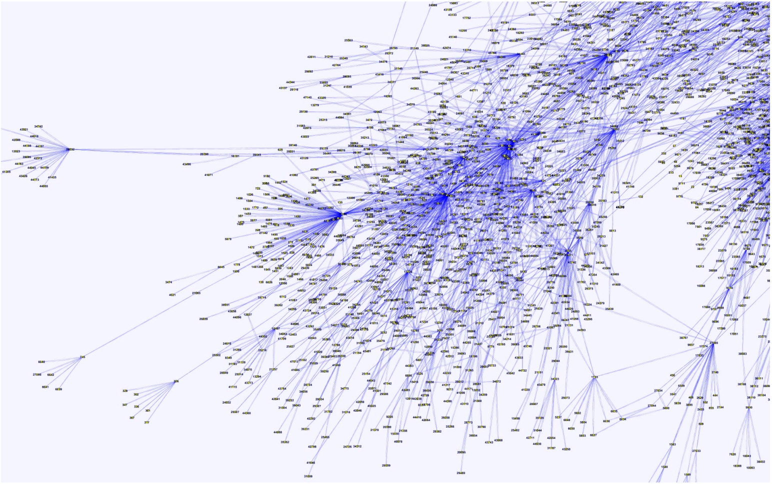
Table: Summary of CAIDA and UCLA datasets

Dataset	Unique ASes	Customer-Provider Links	Peer-to-Peer Links
CAIDA	36,878	99,962	3,523
UCLA	38,794	74,542	65,784

Table: Hybrid UCLA*-IXP topology

Dataset	Unique ASes	Customer-Provider Links	Peer-to-Peer Links
UCLA*	42,703	76,083	78,264
IXP	2,974	0	40,076
Hybrid	**43,018**	**75,421**	**105,772**

Part of the AS relationships dataset visualized



LATENCIES

- Underlay latencies (numbers derived form the findings in [Zhang et al.]
  - 34 ms for inter-AS hops
  - 2 ms for intra-domain router hops
  - The number of intra-domain router hops between the nodes in the same AS is

$$1 + \lfloor \log D \rfloor$$

, where D is the degree of the AS. There is a relationship between the degree of the AS and its size [Tangmunarunkit et al.].

[Zhang et al.] Zhang, B.; Ng, T. E.; Nandi, A.; Riedi, R.; Druschel, P., and Wang, G. Measurement-Based Analysis, Modeling, and Synthesis of the Internet Delay Space ACM SIGCOMM IMC'06, ACM, 2006, 85-98

[Tangmunarunkit et al.] Tangmunarunkit, H.; Doyle, J.; Govindan, R.; Jamin, S.; Shenker, S., and Willinger, W. Does AS Size Determine Degree in AS Topology? SIGCOMM Comput. Commun. Rev., 2001, 31, 7-8

Valley-Free Policy Routing

- ASes export routes based on the algorithm given below and prefer customer routes to peering and peering to provider routes and secondarily choosing the shortest AS-level path
- $\Rightarrow$ *valley-free routes* [Gao]
  - Every path concatenated from 0-n customer-to-provider links followed by 0-1 peering links and ending in 0-n provider-to-customer links

Algorithm 1 Export routes

for all $a \in AS, x \in neighbors(a)$ **do**
    if $x \in providers(a) \cup peers(a)$ **then**
        export all customer routes of a to x
    else if $x \in customers(a)$ **then**
        export all routes of a to x
    end if
end for

[Gao] Gao, L. On Inferring Autonomous System Relationships in the Internet IEEE/ACM Transactions on Networking, 2001, 9, 733-745

AS Utility-Based Traffic Model

- ASes are modelled as points in a three dimensional utility space based on their business model [Chang et al.]
  - Each utility follows a Zipfian distribution with different exponents
  - The rank correlations between different utilities were measured
- The traffic is roughly categorized into the following three utilities:
  - Web hosting U_{web}
  - Residential access U_{ra}
  - Business access U_{ba}
    - $=$ the cumulative transit provided by the AS (in case of multihoming, the utility is divided equally between all providers)
- We assume that the locations of rendezvous networks hosting the scope in a query are distributed to ASes proportional to $U_{\text{web}} + \alpha U_{\text{ra}}$, where α is a parameter
- Subscriptions originate from ASes proportional to the U_{ra}

[Chang et al.] Chang, H.; Jamin, S.; Mao, M., and Willinger, W. An Empirical Approach to Modeling Inter-AS Traffic Matrices ACM SIGCOMM IMC'05. Proceedings, 2005, 139-152

APPLICATION TYPE-BASED TRAFFIC MODEL

- Application models are based on total *throughput* (parameter)
  - Projected to be 37,000 PB/month (14.6 TB/sec) in 2013 [Cisco]
- Two most popular types of traffic: web and P2P (BitTorrent)
  - *WebMix* and *P2PMix* parameters determine the share of each type of the throughput
  - Web traffic observed to be nearly 60% and P2P contributing about 14% [Maier et al.]
- Labovitz et al. observed that over 50% of interdomain traffic was originated by just 150 ASes [Labovitz et al.]
  - We model this spatial locality of generated traffic by parametrizing the share of each AS for each traffic type
  - Problem: Can conflict with the popularity distribution!

[Cisco] Cisco Visual Networking Index: Forecast and Methodology, 2010-2015 Cisco, 2011

[Maier et al.] Maier, G.; Feldmann, A.; Paxson, V., and Allman, M. On Dominant Characteristics of Residential Broadband Internet Traffic Internet Measurement Conference (IMC), 2009, 90-102

[Labovitz et al.] Labovitz, C.; Iekel-Johnson, S.; McPherson, D.; Oberheide, J., and Jahanian, F. Internet Inter-Domain Traffic SIGCOMM'10, 2010

APPLICATION TYPE-BASED TRAFFIC MODEL (2)

- Web traffic parameters
  - *WebReqsPerObj* determines the number of rendezvous requests per page
    - Empirical study shows median 12 embedded objects per page [Ihm and Pai]
  - *WebObjSize*
    - Median page size of 133KB [Ihm and Pai]
  - Popularity distribution assumed to follod Zipf's law [Breslau et al.]
- P2P traffic parameters
  - *P2PReqsPerObj* determines the number of rendezvous requests per unit time per object
  - *P2PShareRatio* is the percentage of objects republished after *P2PShareDelay* seconds after they are subscribed
  - Popularity distribution is Zipf-Mandelbrot [Hefeeda and Saleh]
  - We collected information about the content size of torrents by crawling The Pirate Bay:

Min.	Max.	Q1	Median	Mean	Q3	Std. Dev.
0B	641.40GB	93.33MB	350.47MB	1.05GB	883.39MB	3.60GB

[Ihm and Pai] Ihm, S. and Pai, V. S. Towards Understanding Modern Web Traffic Internet Measurement Conference (IMC), 2011, 295-312

[Breslau et al.] Breslau, L.; Cao, P.; Fan, L.; Phillips, G., and Shenker, S. Web Caching and Zipf-like Distributions: Evidence and Implications INFOCOM'99, 1999, 1, 126-134

[Hefeeda and Saleh] Hefeeda, M. and Saleh, O. Traffic Modeling and Proportional Partial Caching for Peer-to-Peer Systems IEEE/ACM Transactions on Networking, 2008, 16, 1447-1460

EVENT GENERATION

- The traffic generator produces rendezvous request events, that are 4-tuples of type

$$< Timestamp, RequestType, RId, ASN > .$$

- The number of objects is huge
  - In 2008 Google reported that their web crawlers had indexed 10^{12} unique URLs
  - $\Rightarrow$ we cannot store per-object state
- Approximate Zipf/Zipf-Mandelbrot laws by using their continuos power law equivalents and use the constant time *inverse transform method* for generating samples by solving the integral (for Zipf)

$$\int_1^x \frac{1}{z^\alpha}\ \mathrm{d}z = \frac{z^{1-\alpha}}{1-\alpha}\bigg|_1^x = \frac{x^{1-\alpha}}{1-\alpha} - \frac{1}{1-\alpha}$$

- Adjusting for normalization, we define our invertible approximation of the Zipf distribution's CDF as:

$$F(x;\alpha,N) = \frac{\alpha - x^{1-\alpha}}{\alpha - N^{1-\alpha}}$$

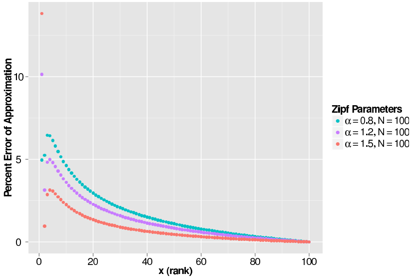
Event Generation (2)

- We can now draw random popularity ranks for requests via the inverse

$$F^{-1}(y;\alpha,N) = \left((N^{1-\alpha} - \alpha) \left(y - \frac{\alpha}{\alpha - N^{1-\alpha}} \right) \right)^{\frac{1}{1-\alpha}}$$

, where y is a random number from the uniform distribution in the interval $[0,1]$ and N is the total number of objects.

- We precalculated 10^6 first values and use binary search to find the exact value.
- The percent error of the approximation is plotted below:



DEPLOYMENT MODEL

- The rendezvous networks were formed by
  1. Extracting a transit hierarchy from the AS topology (in case of multihoming we preferred smaller provider)
  2. Joining ASes in this tree top-down starting from tier-1 domains and offering a rendezvous network service at AS x to its customer y if the number of y's transitive customers is smaller than predefined limit or y and its customers do not host much more content than x and its customers transitively
- The Canon hierarchy formation
  1. Each rendezvous network forms a Chord ring with enough nodes to store the hosted scopes
  2. By traversing the transit tree bottom-up by creating a new layer in the Canon when 5 sub-rings were transitively collected

RENDEZVOUS SYSTEM MODEL

- The Canon overlay routing algorithm is fully simulated
- Network failures were not modeled
- The main limitation: *linearity* assumption for requests by simulating them independently
  - Minimizes the amount of needed memory and allows us to generalize from a small sample size of requests
- Each node contains βk amount of storage for caching the most recent scope pointers queried via them.
  - k is the amount of storage used for storing scopes at the node
- An analytical model of the cache performance in steady state
  - If each node perfectly caches the n most popular scopes, a scope with a popularity rank pr is found cached at a node x on level a of the Canon hierarchy when

$$pr < \left(\frac{\beta \cdot s \cdot (A/N)}{(A_{x+1,a} - A_{x,a}) mod A}\right)$$

  , where $A_{i,j}$ is the Canon node identifier of ith node at level j and N is the total number of nodes, s is the total number of scopes and A is the size of the whole address space.

SIMULATOR

- The simulator environment is writen in Python and was also ported into GNU Octave with optimizations for the experiments in [Rajahalme et al.]
- Some example outputs of the simulator:

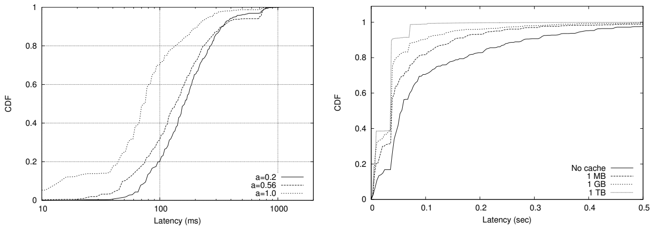


Figure: The graphs on the left show CDFs for the delay caused by the rendezvous phase with different popularity power-law exponents when the number of scopes is fixed to 10^{11}. On the right, the effect of the node cache size on the rendezvous latency distribution is plotted.

- More results in [Rajahalme et al.]

[Rajahalme et al.] Rajahalme, J.; Srel, M.; Visala, K., and Riihijrvi, J. On name-based inter-domain routing Computer Networks Journal: Special Issue on Architectures and Protocols for the Future Internet, 2011, 55, 975-986

Lessons Learned

- Efficiency and scalability are paramount in large simulations
  - Aggregate algorithms over an AS-level graph
  - Application-specific models can simplify the problem
- Too much detail in the submodels may cause unintentional correlations of variables
- Future work
  - Massive distributed simulation would remove the need for analytical model for caches and linearity assumption, but would probably be very slow
  - Flash crowds etc.
  - PoP-level topology

Thank You! Questions?