Bruno Ribeiro
Umass Amherst

A lightweight on-line flow size distribution estimator

Introduction

- Accurate flow size distribution estimates with
  - One hash calculation *per* packet
  - Few bits of memory *per* flow
- Desired characteristics
  - Be the fastest on-line estimator in the literature
  - High accuracy
- E.g.: 2.5 bits *per* flow
  - 12.8 million flows fit in 4 MB

Inner workings

- E.g.: 4MB memory divided into 6.4 million 5 bit counters
  - Yes, there is a tradeoff btw number of counters and their size
- Upon packet arrival
  - Computes hash of packet
    - All packets from the same flow hash to the same value
    - Increments corresponding counter
- Hash function property
  - **Uniformly** at random associates a newly arrived flow to a counter
- Collects flows during a fixed time interval
- Computes the flow size distribution of that interval

Don't dismiss Poisson

- Target average number of flows:
  - E.g.: 12.8 million
- Total number of flows of size i
  - Poisson distributed (for number of flows $>>$ 1)
    - Constant, Poisson, all the same for large numbers.
- Number of flows of size i assigned to each counter
    - Approximated by the Poisson distribution
    - Comes from the design of the hash function

Counters and the flow size distribution

- λ_i — Normalized avg # arrivals of flows of size *i* *per* counter
- A counter
  - Has average value $\sum_{i=1}^{\infty} i\lambda_i$
- Probability that counter c_k has value 3 is

$$P[c_k = 3] = (\lambda_1^3 e^{-\Lambda})/3! + \lambda_2\lambda_1 e^{-\Lambda} + \lambda_3 e^{-\Lambda},$$

where $\Lambda = \sum_{i=1}^{\infty} \lambda_i$.

- Thus

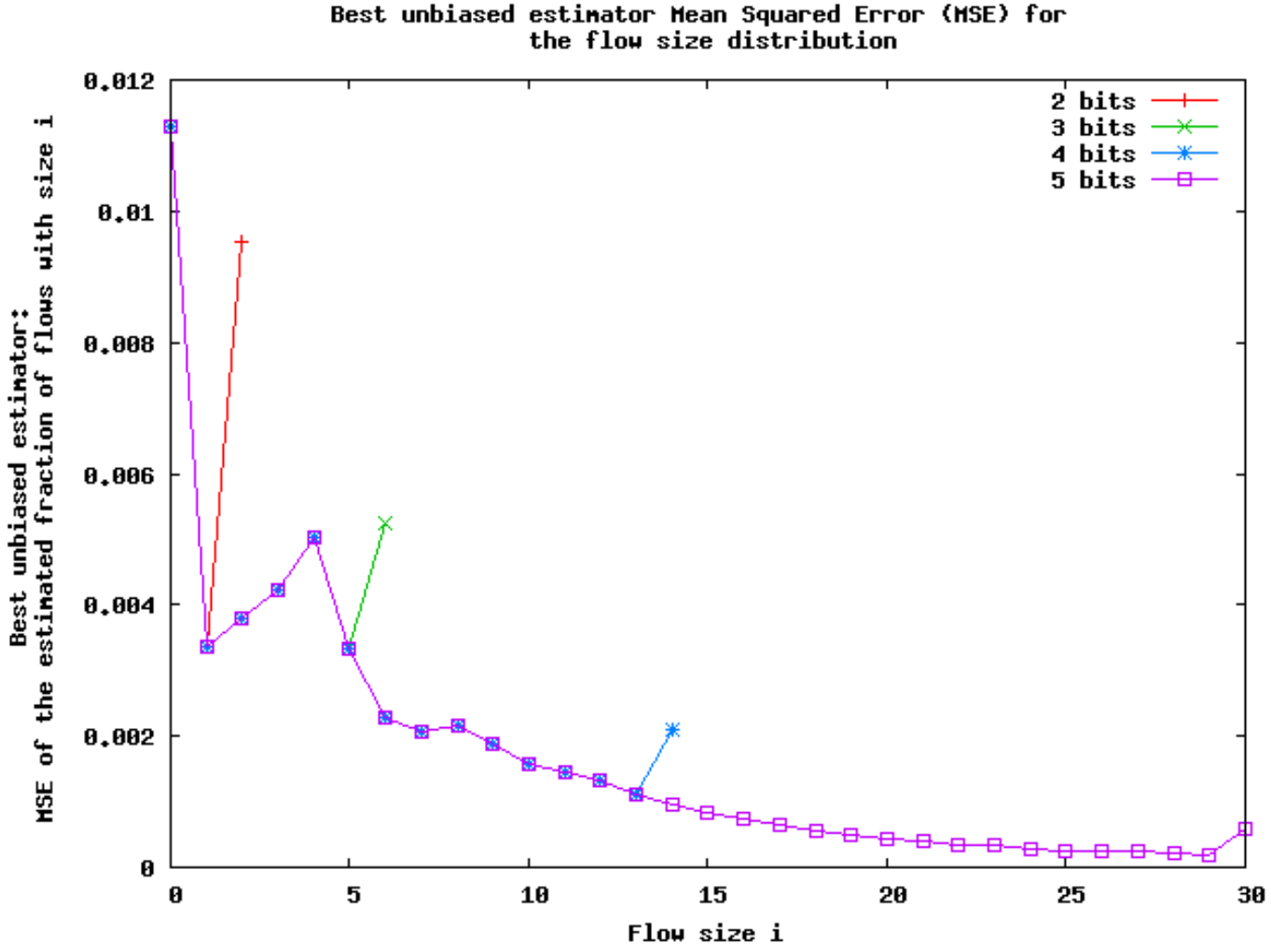
$$\lambda_3 = \frac{P[c_k = 3] - (\lambda_1^3 e^{-\Lambda})/3! + \lambda_2\lambda_1 e^{-\Lambda}}{e^{-\Lambda}}$$

In the example counters only count until 32

- In the example c_k only counts until 32
- Flow sizes can be much bigger
  - And there are, in average, 2 flows *per* at c_k
- Solution: Semi-probabilistic counter
  - Probabilistic counters introduced by Robert Morris '78
  - Sequential counting until counter reaches 14
  - After that, add one with probability $2^{(c_k-13)}$

Flow size distribution estimator accuracy

- *Per* counter upper bound on estimator accuracy
  - Counter width: 2 to 5 bits
    - (without semi-probabilistic counting)



Best estimator

- Maximum likelihood estimator is an efficient estimator
  - But too resource intensive
    - May take days to compute all flow sizes!

FAST flow size distribution estimator

- g_j — fraction of counters with value j
- If $j < 14$

$$\Lambda \approx -\ln(g_0)$$
$$\lambda_1 \approx g_1 / e^{-\Lambda}$$
$$\lambda_2 \approx (g_2 - \lambda_1^2 e^{-\Lambda}/2)/e^{-\Lambda}$$
$$\lambda_3 \approx \frac{g_3 - (\lambda_1^3 e^{-\Lambda})/3! + \lambda_2 \lambda_1 e^{-\Lambda}}{e^{-\Lambda}}$$
$$\vdots$$

- If $j \geq 14$

$$\lambda_{(12+2^{(j-13)})} \approx g_j / 2^{(j-13)}$$

Missing: Normalize by Λ

- Interpolate the remaining λ_i

FAST estimator, preliminary results

- 5-bit counters
- 2×10^6 flows
- Sequential until $c_k = 14$
- 625KB of memory

confidence intervals very small