Internet Traffic Modeling for Efficient Network Research Management

Prof. Zhili Sun, UniS

Zhiyong Liu, CATR

UK-China Science Bridge Workshop

13-14 December 2011, London

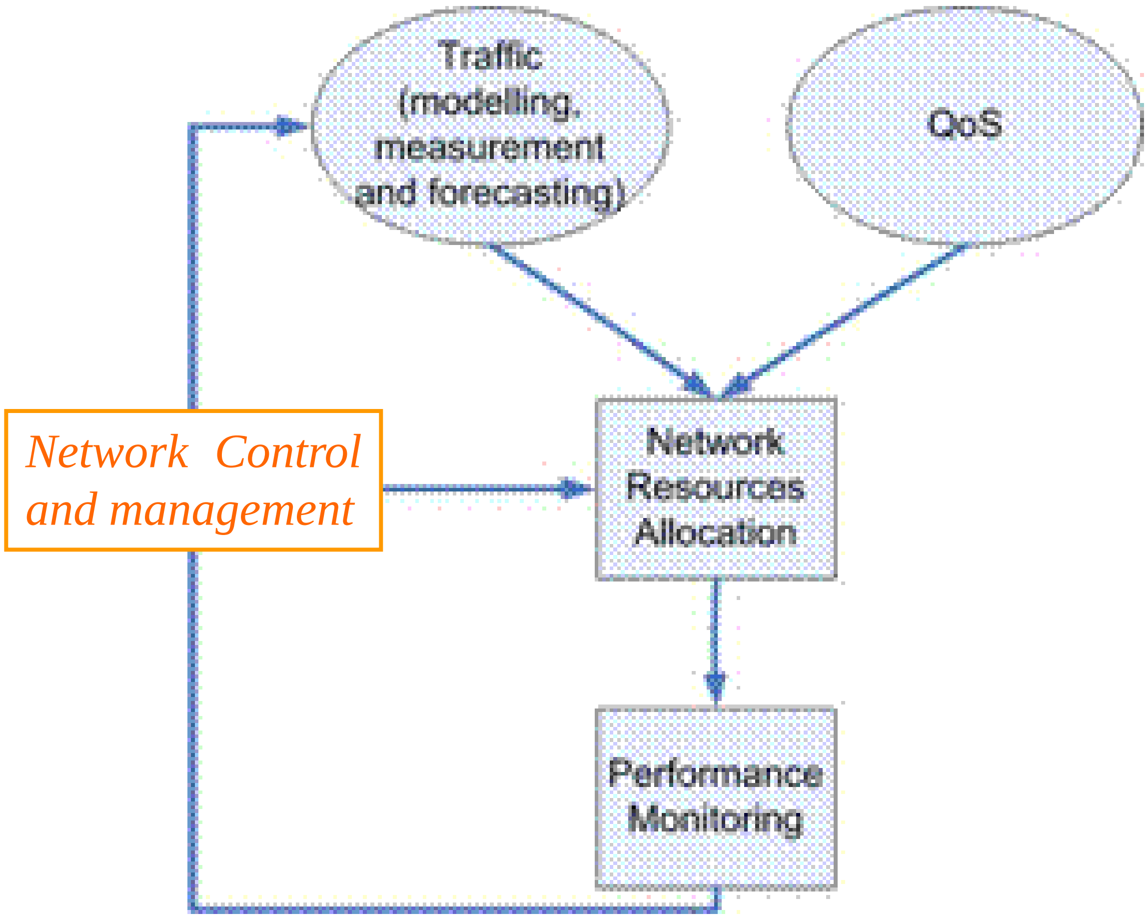
Outline

- Introduction
- Background
- Classical teletraffic engineering model
- Main parameters of the Internet traces
- Well known mathematical methods
- New approach to the Internet traffic dataset
- Conclusion and directions for further studies

Introduction

- Historically, teletraffic engineering successful in telecommunication networks (Poisson process)
- Internet traffic has grown significantly since 1990s; and over provision becomes impractical
- In 1990s, discovered that the Poisson function failed to model the Internet traffic.
- Many suggested Pareto and self-similar models but there is no conclusive confirmation due to the complexity of the Internet traffic.
- This leaves a big gap between the classical traffic engineering and the Internet traffic modeling
- This paper presented a new approach

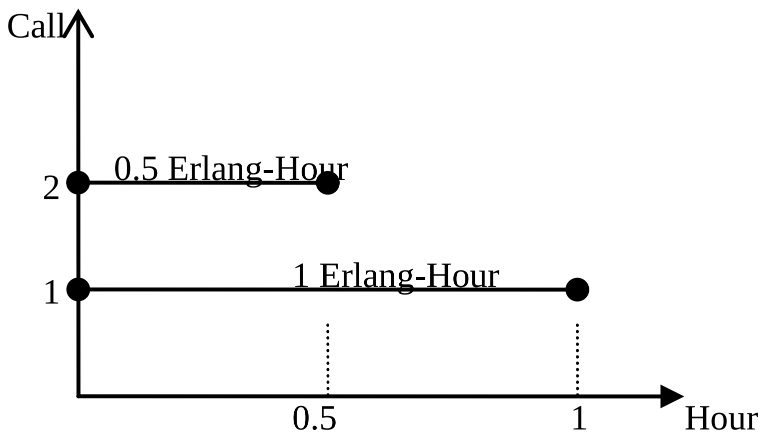
Teletraffic engineering components



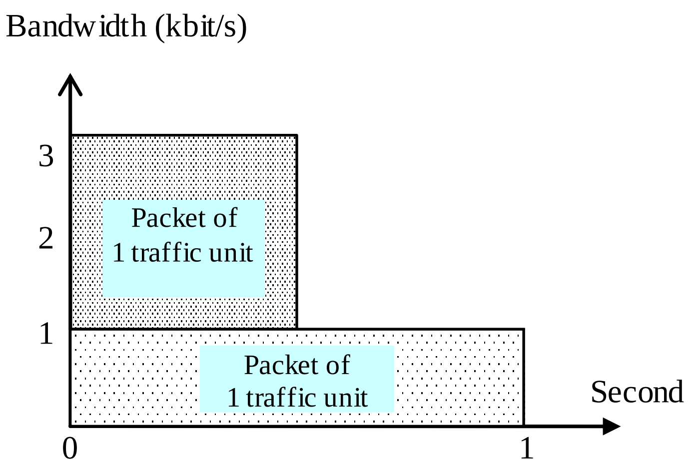
Basic concepts

Traffic load:

$$E = \frac{\lambda}{\mu} = \lambda \bullet s = \lambda \frac{p}{b} = \frac{\lambda}{b/p} = \left(\frac{p}{b}\right) \Big/ \left(\frac{1}{\lambda}\right)$$



(a) Traffic for telephony networks



(b) Traffic for packet networks

Arrival process

- **Arrival time of the i'th packet is at T_i as the following:**

$$0 = T_0 < T_1 < T_2 < \ldots < T_i < T_{i+1} < \ldots .$$

For simplicity, we can assume that the observation takes place at time $T_0 = 0$.

- **The number of calls in the interval [0, t) is denoted as N_t. Here N_t is a random variable with continuous time parameters and discrete space. When t increases, N_t never decreases.**

- **The time distance between two successive arrivals is:**

$$X_i = T_i - T_{i-1},\ i = 1, 2, \ \ldots.$$

This is called the inter-arrival time, and the distribution of this process is called the interarrival time distribution.

Number and Interval representations

- **Corresponding to the two random variables N_t and X_i, the two processes can be characterized in two ways:**
  - **Number representation N_t: time interval t is kept constant to observe the random variable N_t for the number of IP packets in t.**
  - **Interval representation T_i: number of arriving IP packets is kept constant to observe the random variable T_i for the time interval until there are n arrivals.**
- **The fundamental relationship between the two representations is given by the following simple relation:**

$$N_t < n, \text{ if and only if }, n = 1, 2, \ldots$$

- **This is expressed by Feller-Jensen's identity:**
  - $Prob\{N_t < n\} = Prob\{T_n \geq t\},$

Exponential and Poisson distributions

Three assumptions were made to model the arrival process using exponential distribution and Poisson distribution:

- **<u>Stationary</u>: For any arbitrary $t_2>0$ and $k\geq 0$, the probability that k calls arrival in $[t_1, t_2)$ is independent of t_1.**
- **<u>Independence</u>: The probability of k calls arrival taking place in $[t_1, t_2)$ is independent of calls before t_1.**
- **<u>Simplicity</u>: it is call simple process if the probability that there is more than one calls arrival in a given point of time is 0.**

The reasons for the failures of Poisson

- The main reasons are the nature of Internet traffic and properties of TCP on which many applications are based including WWW, FTP, email, P2P, Telnet, etc.
- These break the assumptions made for classic teletraffic engineering model, due to acknowledgement, flow control and congestion control mechanisms.
- To investigated the features of the Internet traffic to find alternative traffic models and to show that the Internet traffic showed properties of long tail and self-similarity
- But still can not fully model the real Internet traffic, Due Internet applications and their complexity,

Traffic parameters

- The traffic traces contain information on each packet captured on the Internet networks.
- The flows of packets depend on the user activities and the applications used.
- The information in each packet captured includes:
  - Time stamp when the packet is captured
  - Media Access Control (MAC) frame header
  - IPv4 or IPv6 Header
  - Transmission control protocol (TCP) header with application protocols such as HTTP, SMTP, FTP, etc.
  - User datagram protocol (UDP) header with application protocols such as DNS, RTP, etc.

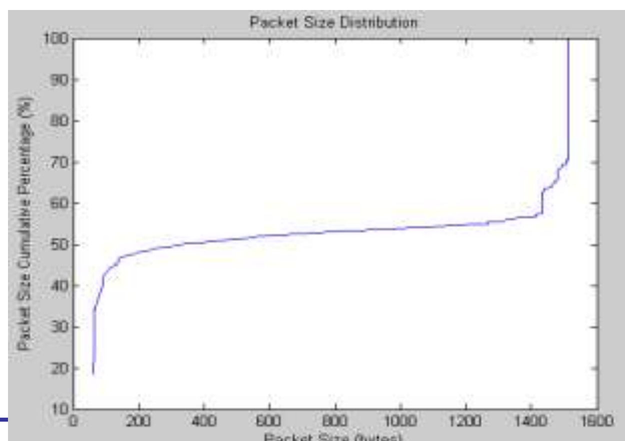
Traffic traces

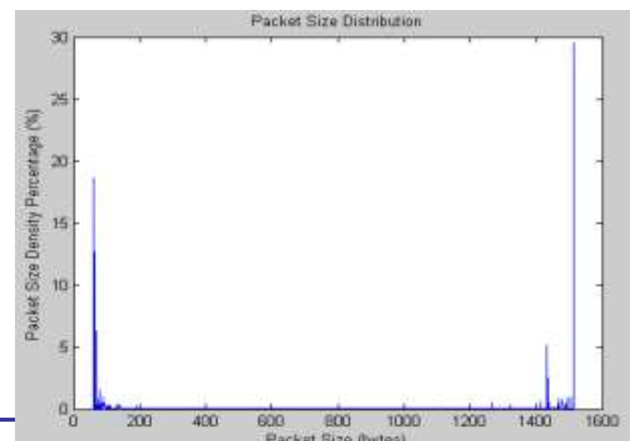
Traffic traces observations on 1st August 2011, at a trans-Pacific line (150Mbps link) in operation since 1st J uly2006:

- IPv4 packets counts 99.57% of the total IP packets (99.6% in bytes),
- Only 0.43% for IPv6 (0.4% in bytes); it showed clearly that the usage of IPv6 is still very low,

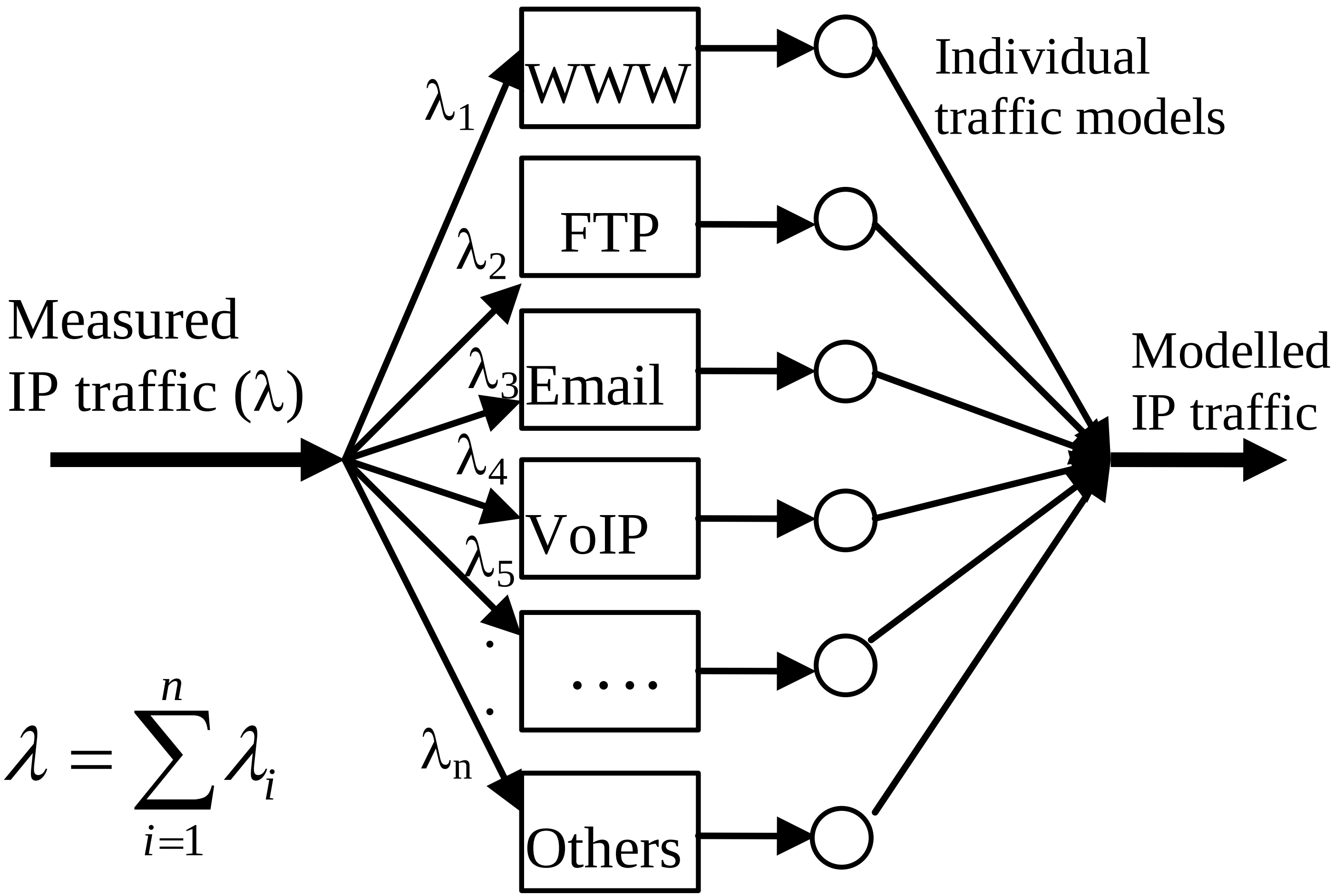
TCP/UDP

- **UDP for 16.81% (11.79% in bytes)**
  - For voice over IP, there is a constant stream of packets with 14 bytes of MAC header, 20 of IP, 8 of UDP and 12 of RTP;
  - Plus payload of 160 bytes for ITU-T G.711 codec as an example that it has 64 kbps, 20 ms sample period and 1 frames per packet (20 ms) [11].
- **TCP counts for 79.76% in packets (85.1% in bytes);**
  - HTTP server counts for 35.04% in packets (64.78% in bytes);
  - HTTP client counts for 20.36% in packets (7.2% in bytes);



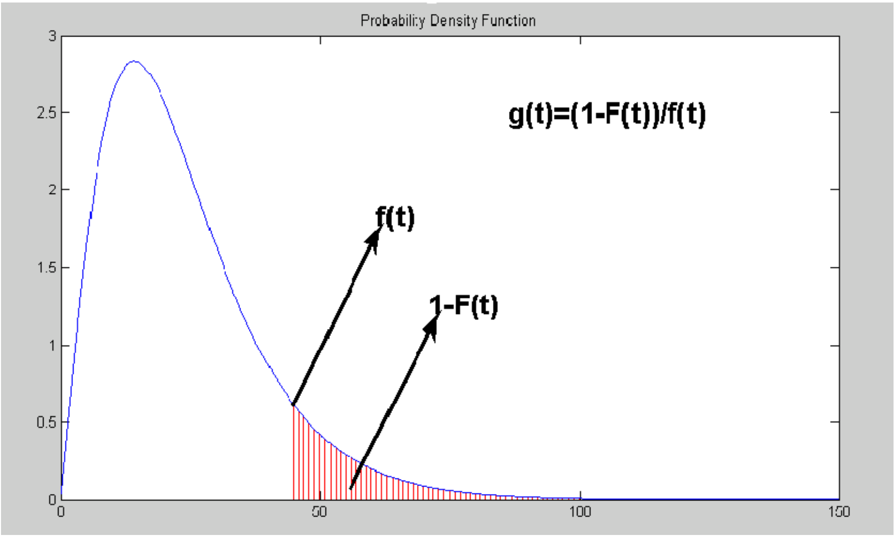


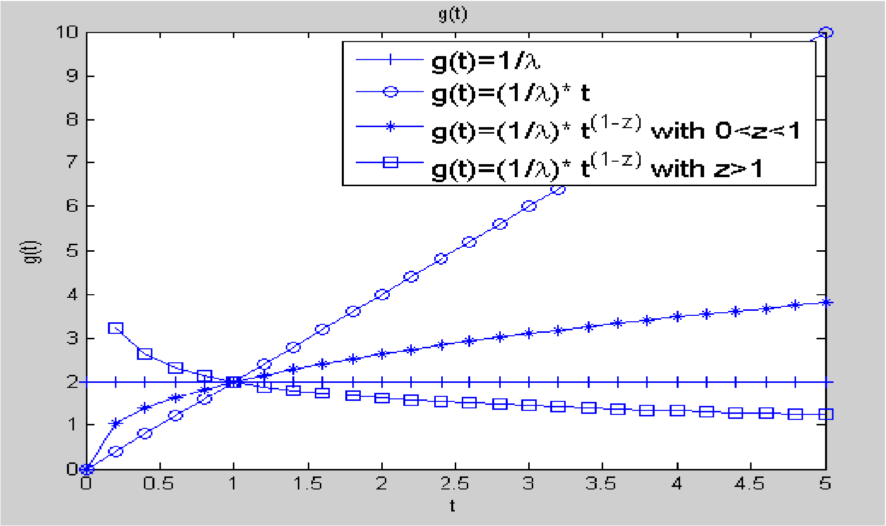
IP Traffic decomposition



$$\lambda = \sum_{i=1}^{n} \lambda_i$$

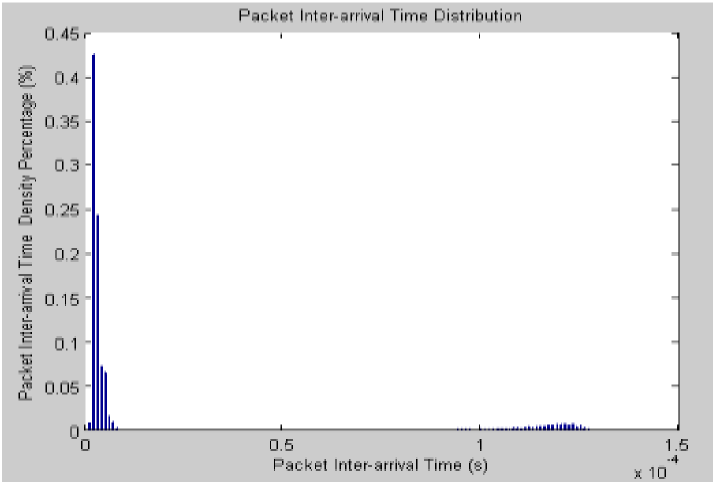
Candidate Mathematical functions



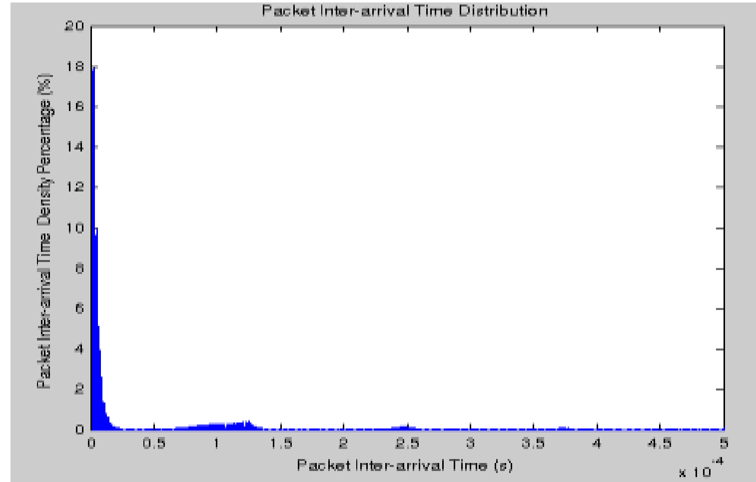


	Exponential	Pareto	Weibull
Distribution function	$F(t)=1-e^{-\lambda t}, \quad \lambda>0, t\geq 0$	$F(t)=1-\left(\frac{t_m}{t}\right)^{\alpha}, \quad t>t_m, \alpha>0$	$F(t)=1-e^{-(t/\lambda)^k}, t\geq 0, and\ F(t)=0\ for t<0$
Density function	$f(t)=\lambda e^{-\lambda t}, \quad \lambda>0, t\geq 0$	$f(t)=\frac{\alpha t_m^{\alpha}}{t^{\alpha+1}}, \quad t>t_m, \alpha>0$	$f(t)=\left(\frac{k}{\lambda}\right)\left(\frac{k}{t}\right)^{k-1} e^{-(t/\lambda)^k}, t\geq 0, and F(t)=0\ for t<0$

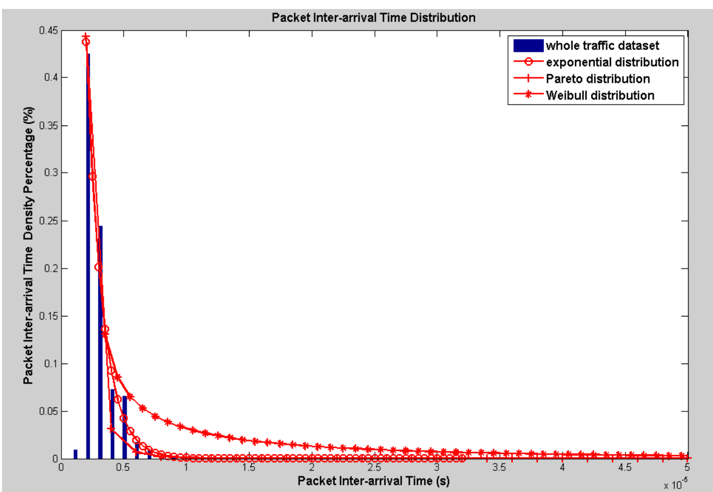
Fitting results



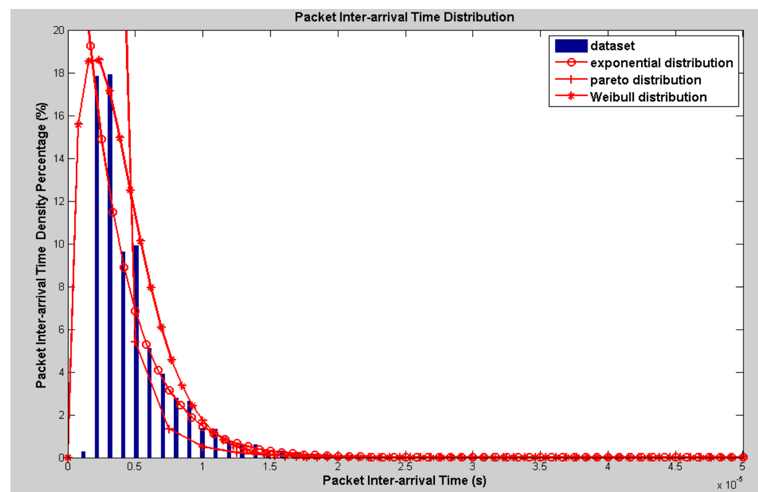
➢Original whole traffic



➢HTTP downlink



➢Modeling the whole traffic



➢Modeling for HTTP downlink traffic

Conclusion

- Due to the limitation of classical techniques, difficult to to model the Internet traffic.
- We introduced a new approach to classify the mathematical functions using the reference function of $g(t)=t^{(1-z)}/\lambda$, and
- Apply the function on the decomposed subset of the Internet traffic rather than the complete dataset
- Weibull distribution gives a better fitting than Pareto and exponential functions.
- Therefore, we can conclude that the rang of distribution functions with $g(t)=t^{(1-z)}/\lambda$, where $0 <= z <= 1$, provided the choices for modeling the decomposed Internet traffic.

Directions for further studies

- **The results show that there is a great potential for the new approach in the Internet traffic engineering with decomposition of Internet traffic.**
- **In future work, the new approach has yet been further validated for modeling on the decomposed components of traffic, such as HTTP, FTP, Email, VoIP and Streaming media, etc.**
- **The important issue remains: there is a new comprehensive model exist that it is simple enough like the classical teletraffic engineering but accurate enough for modeling the future Internet traffic**
- **This paper presented a new approach to resolve the issues, hence am important topic for further studies.**

Any Question?

Z.Sun@surrey.ac.uk