BANDWIDTH MODELING IN LARGE DISTRIBUTED SYSTEMS FOR BIG DATA APPLICATIONS

Bahman Javadi

School of Computing, Engineering and Mathematics

University of Western Sydney, Australia

Boyu Zhang and Michela Taufer

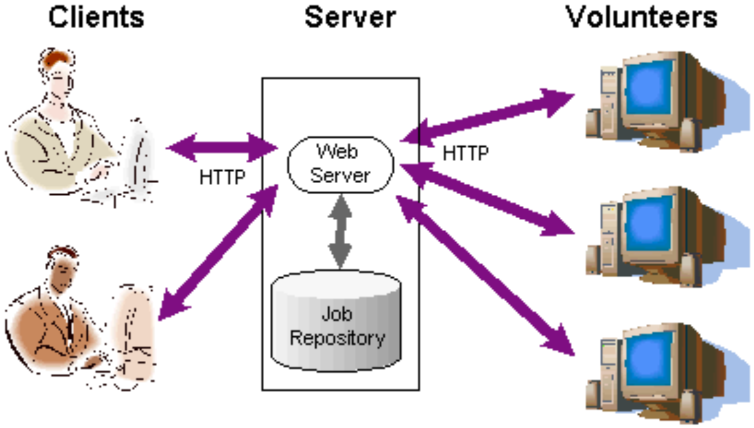
University of Delaware, USA

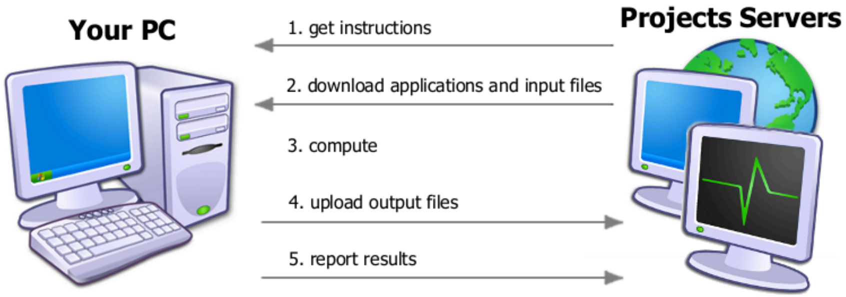
AGENDA

- Introduction
- Modeling Methodology
- Bandwidth Modeling
- Model Validation
- Conclusions

INTRODUCTION

- Volunteer computing (VC)
  - large-scale distributed paradigm that harnesses the computing power and storage capacity of thousands or millions of hosts owned by the public for scientific applications.
  - Can fully support Big Data generation





INTRODUCTION

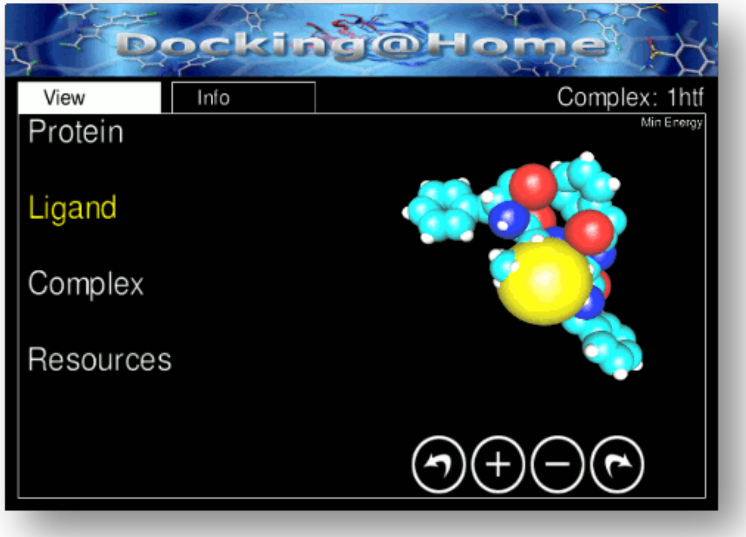
- Analyzing of data communication and host bandwidth in VCs have not been address yet.
- Goal: a general methodology to model the network bandwidth (download and upload rates) of a volunteer computing project.

BACKGROUND

- BOINC
  - Berkeley Open Infrastructure for Network Computing
  - Open-source system that harnesses the computing power and storage capacity of thousands or millions of hosts owned by the public for large-scale scientific projects
  - BOINC was originally developed to manage the SETI@home project.
  - It has been used for more than 70 scientific projects around the world.

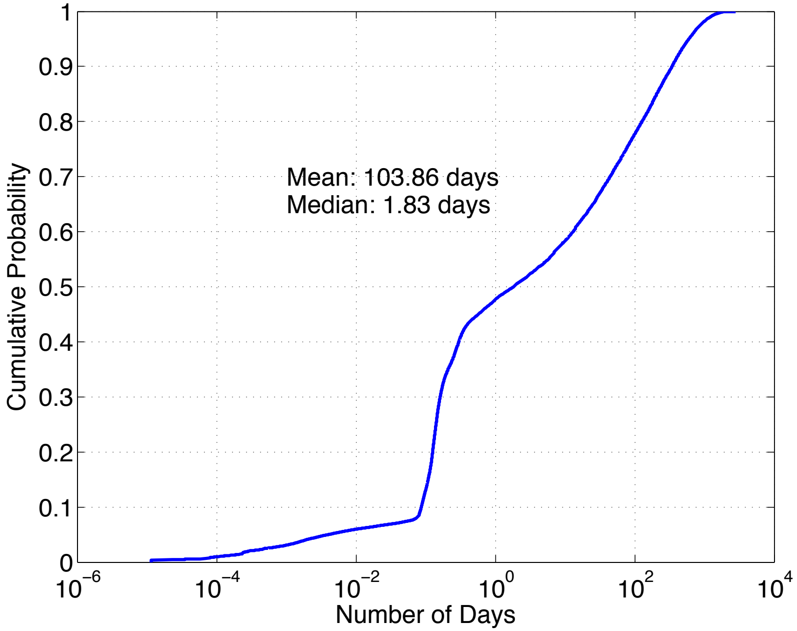
BACKGROUND

- Docking@home Project
  - Uses the BOINC Software
  - The Docking@Home project simulates the behavior of ligands when docking into the active site of a protein.
  - Job size: 1.2-1.9 MB
  - http://docking.cis.udel.edu/



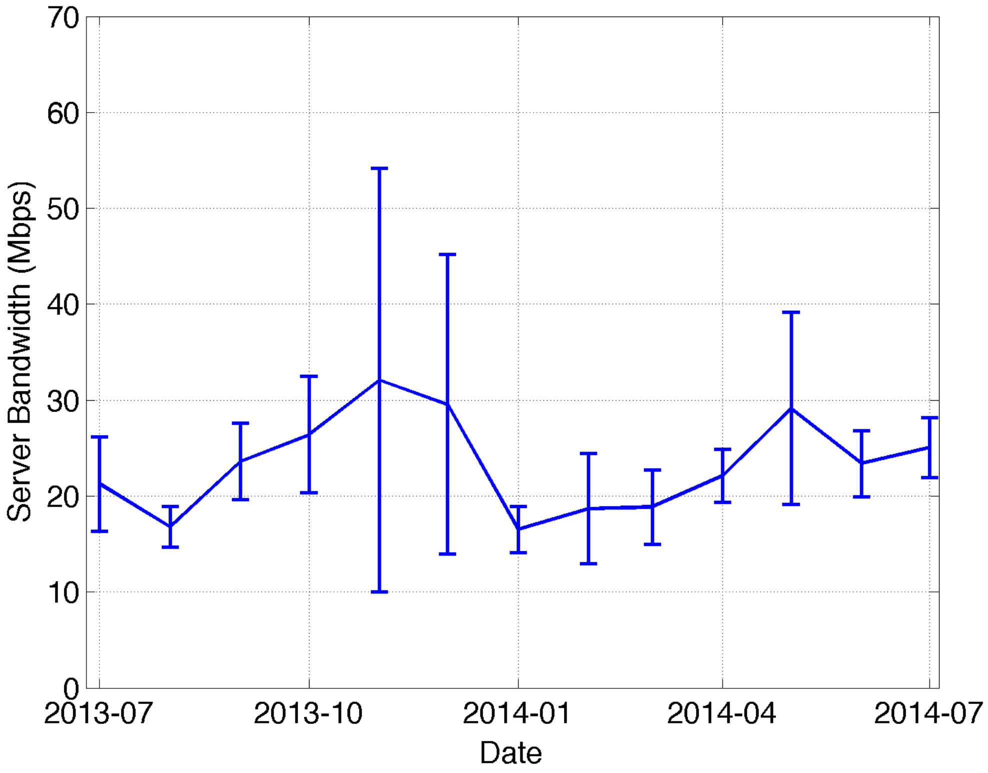
Modeling Methodology

- Real trace from the Docking@home project
  - ~280,000 hosts
  - Period: September 11, 2006 to May 5, 2014
  - Available on FTA: http://fta.scem.uws.edu.au/
  - Host life time: 103 days on average



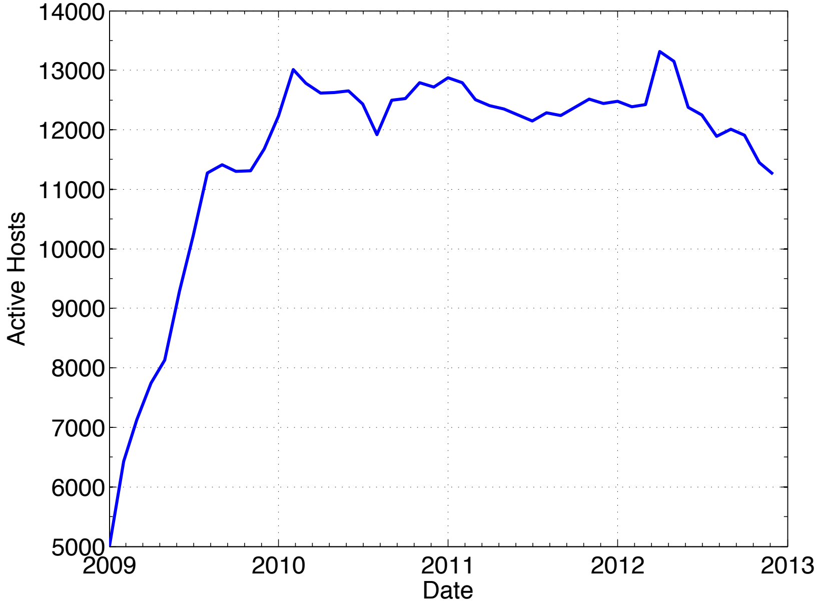
SERVER ANALYSIS

- Server Bandwidth could be a system bottleneck.
- Server Bandwidth: 1Gbps



HOST ANALYSIS

- Distribution of active hosts in Docking@Home from 2009 to 2012
    - An active host at time T is a host that had connected to the server before time T and whose last connection to the same server takes place after time T
    - ~10,000 hosts are active on average

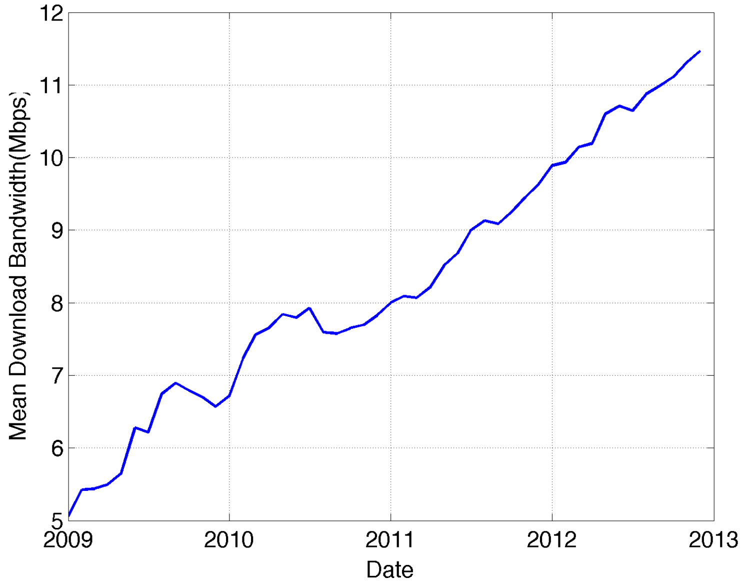


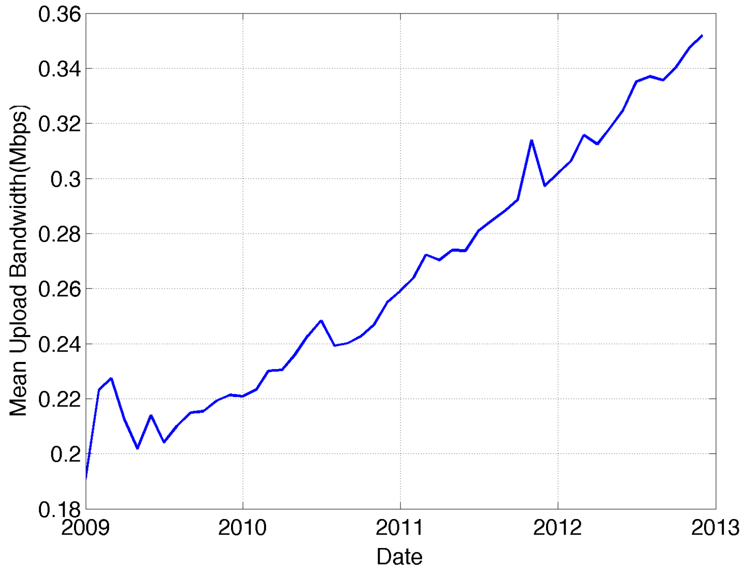
Modeling: 2009-2012

Evaluation: 2013

BANDWIDTH ANALYSIS

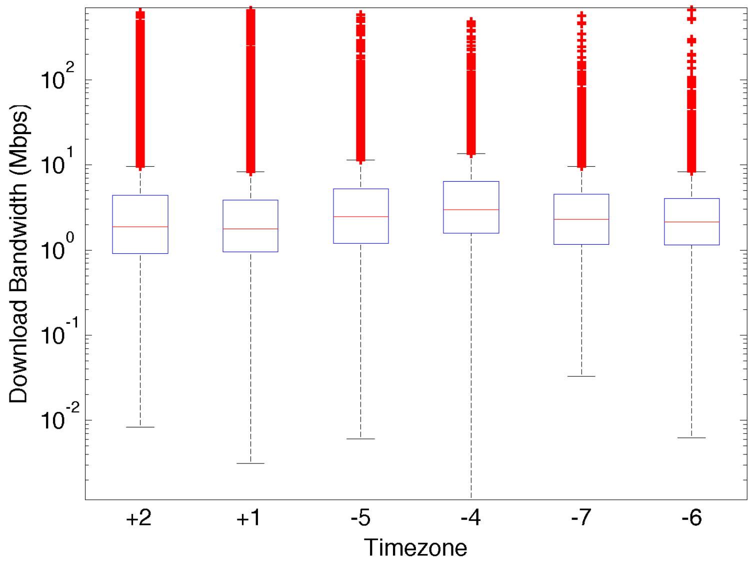
- Average download and upload bandwidth for active hosts in Docking@Home
  - The mean download bandwidth is about 30 times more than the mean upload bandwidth for each year.

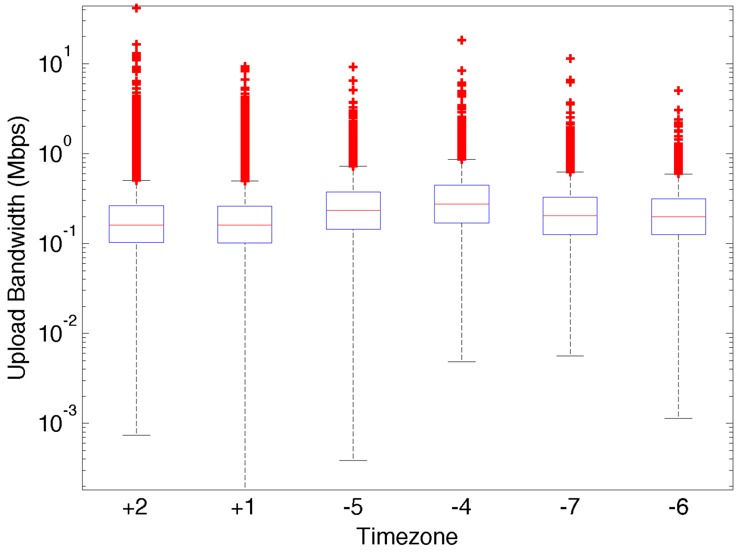




BANDWIDTH MODELING

- Host Bandwidth Correlations
  - Test for host time zone





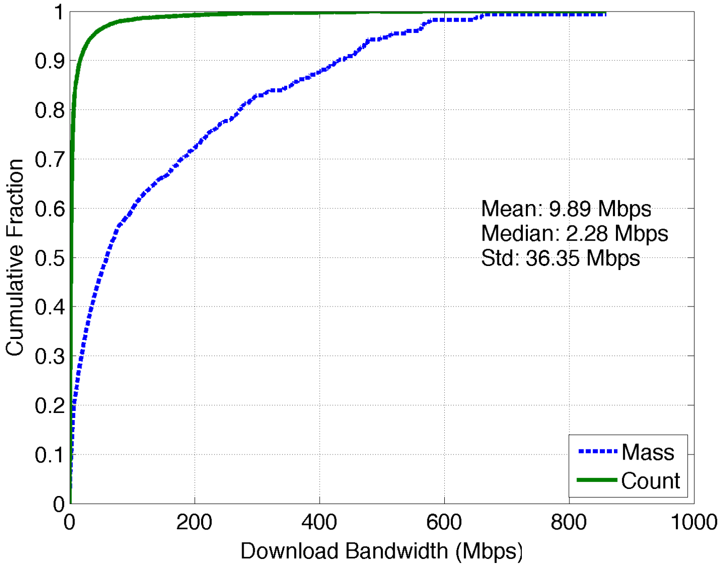
BANDWIDTH MODELING (CONT.)

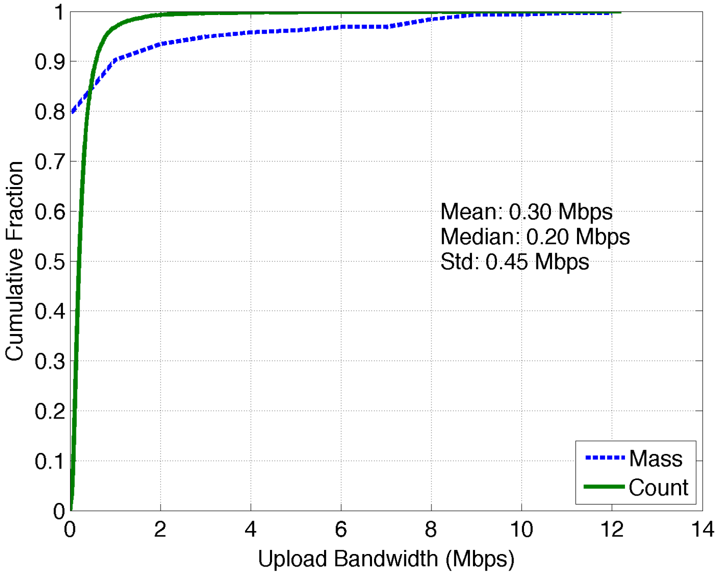
- Host Bandwidth Correlations
  - Test for host time zone
  - Test for upload and download
    - NO Obvious Correlation

- The absence of obvious correlations in the host bandwidth drives our modeling approach towards the design of an **independent** statistical model that predicts the download and upload rate for a given host and at a specific time.

BANDWIDTH MODELING (CONT.)

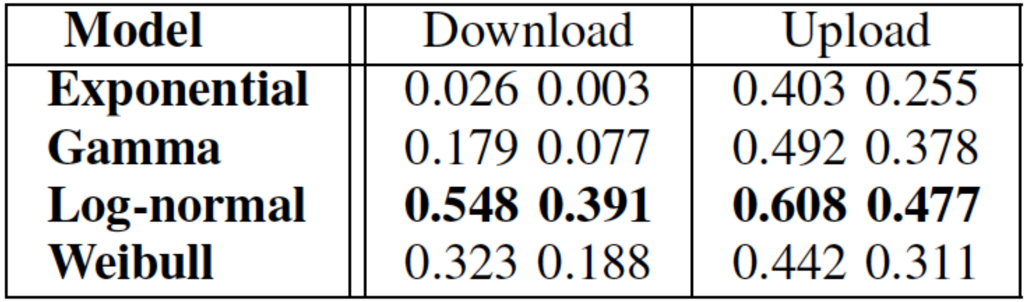
- Mass-Count Plot
  - For 2012





BANDWIDTH MODELING (CONT.)

- Statistical Modeling
  - Model nomination: Weibull, Log-normal, Gamma and Exponential
  - Goodness of fit tests results (p-values): **Log-normal** is the best fit.



Model	Download		Upload	
Exponential	0.026	0.003	0.403	0.255
Gamma	0.179	0.077	0.492	0.378
Log-normal	**0.548**	**0.391**	**0.608**	**0.477**
Weibull	0.323	0.188	0.442	0.311

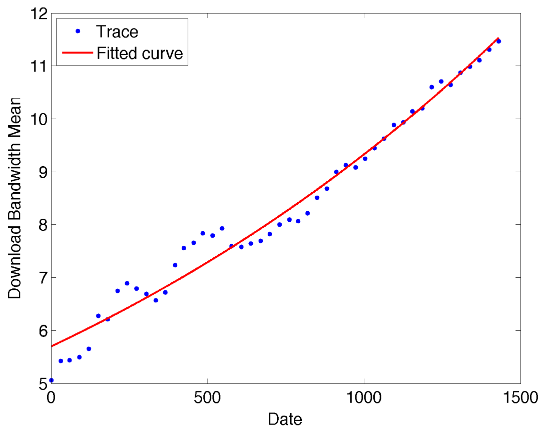
BANDWIDTH MODELING (CONT.)

- Embedding time into the model
  - Best Fit: Log-normal
  - Modeling Mean and Variance for the Log-normal distribution
  - Best model is following exponential function:
    - a,b: function metric
    - t: date (*t=date-Rdate*)
    - *date*: given date
    - *Rdate*: reference date (January 2009 for our traces)

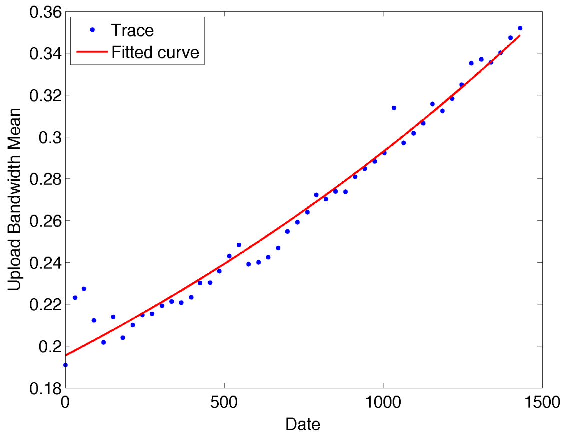
$$f(t) = ae^{bt}$$

BANDWIDTH MODELING (CONT.)

- Fitting Results for Mean Download bandwidth

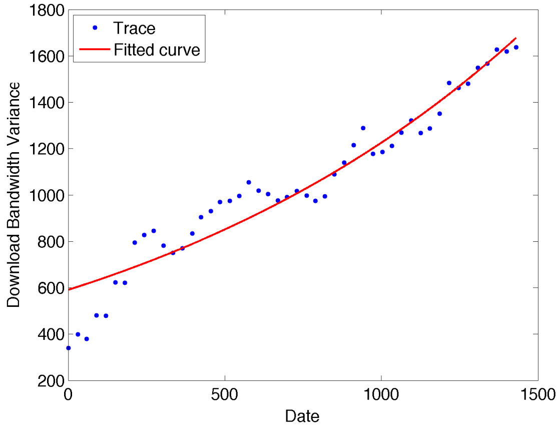


- Fitting Results for Mean Upload bandwidth



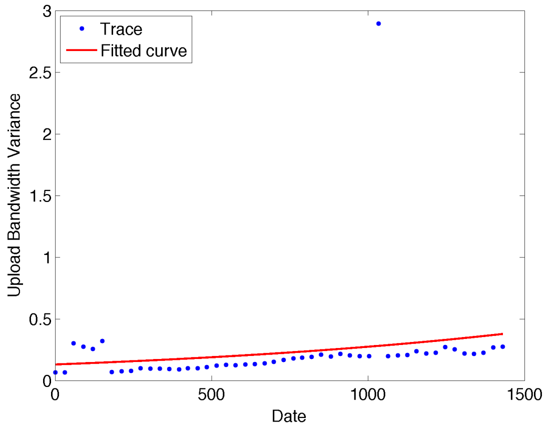
BANDWIDTH MODELING (CONT.)

- Fitting Results for Variance Download bandwidth



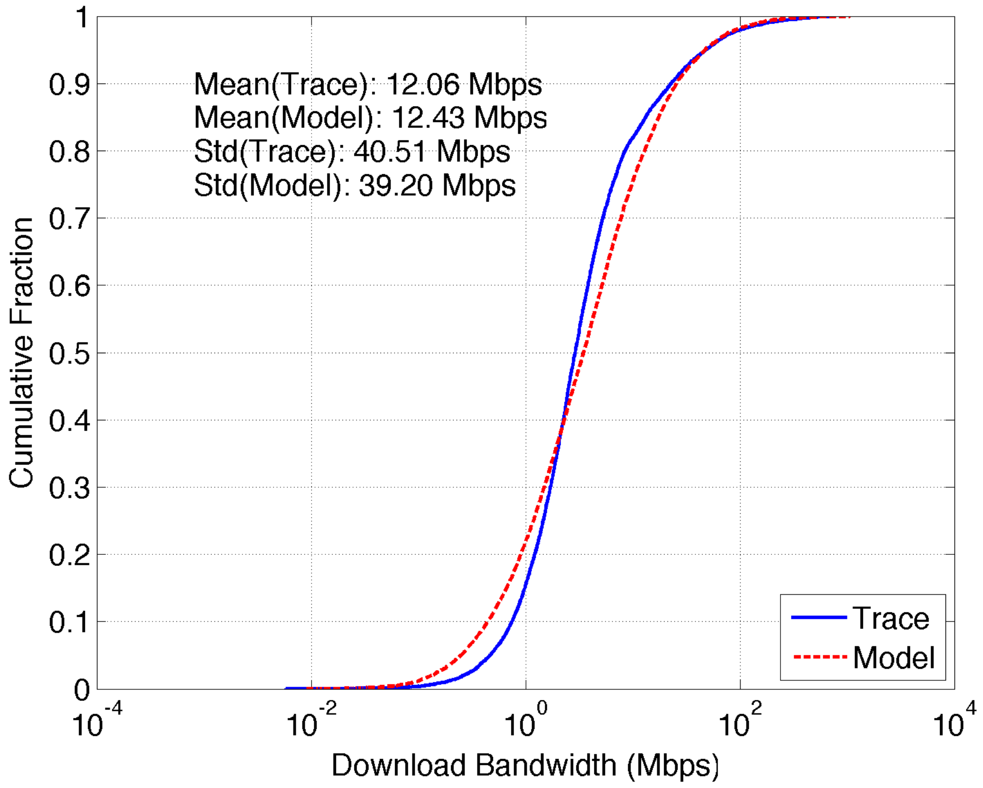
Model	a	b	R^2
Mean download	5.698	$0.493e^{-3}$	0.9748
Variance download	590.7	$0.729e^{-3}$	0.9227
Mean upload	0.1955	$0.404e^{-3}$	0.9735
Variance upload	0.129	$0.745e^{-3}$	0.0376

- Fitting Results for Variance Upload bandwidth



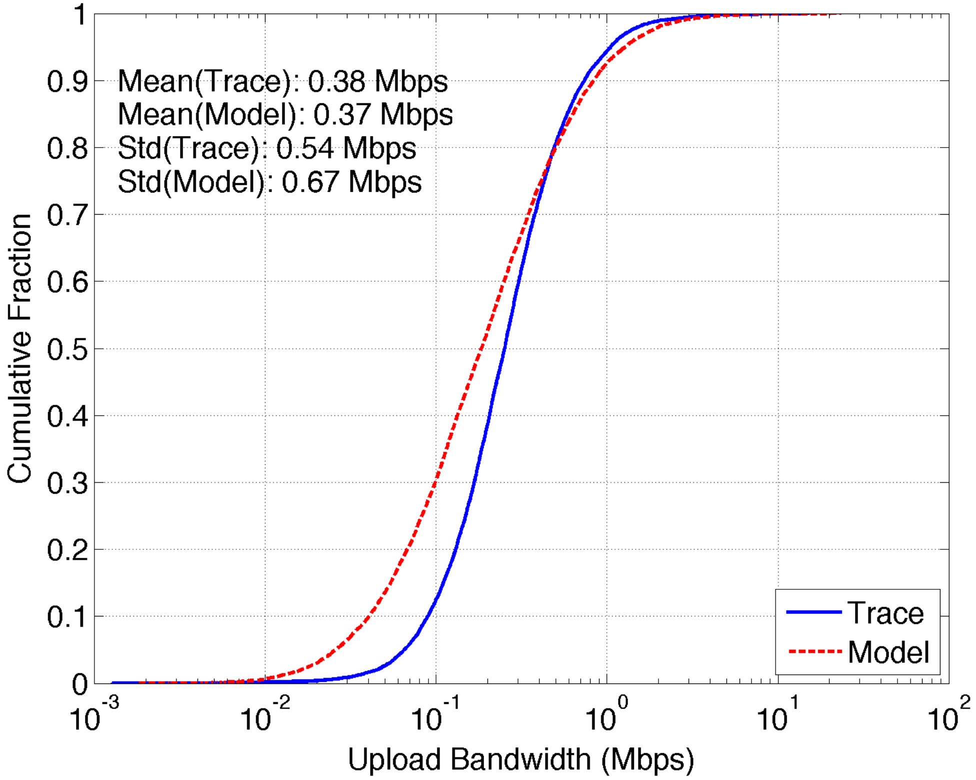
Model Validation

- Model-based validation
  - Download bandwidth for May 2013
  - 3% relative error



MODEL VALIDATION (CONT.)

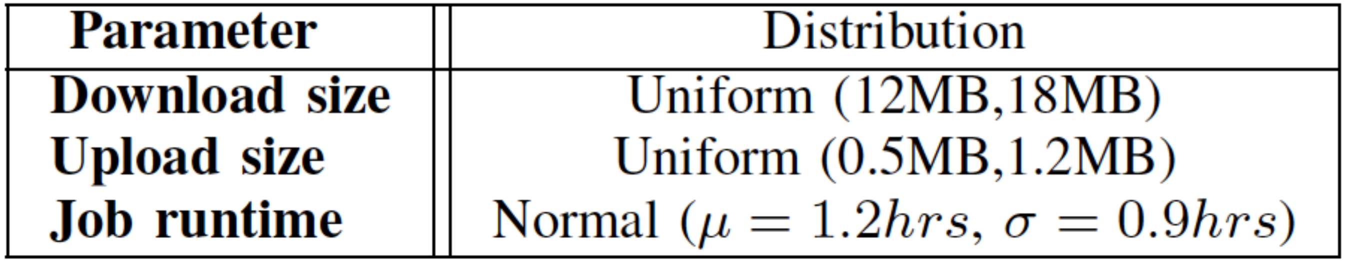
- Model-based validation
  - Upload bandwidth for May 2013
  - 15% relative error



Model Validation (Cont.)

- Simulation-based validation
  - Simulating a queue with the variable service time
  - Using **model** and **trace** to generate the service time
    - M/Model/1
    - M/Trace/1

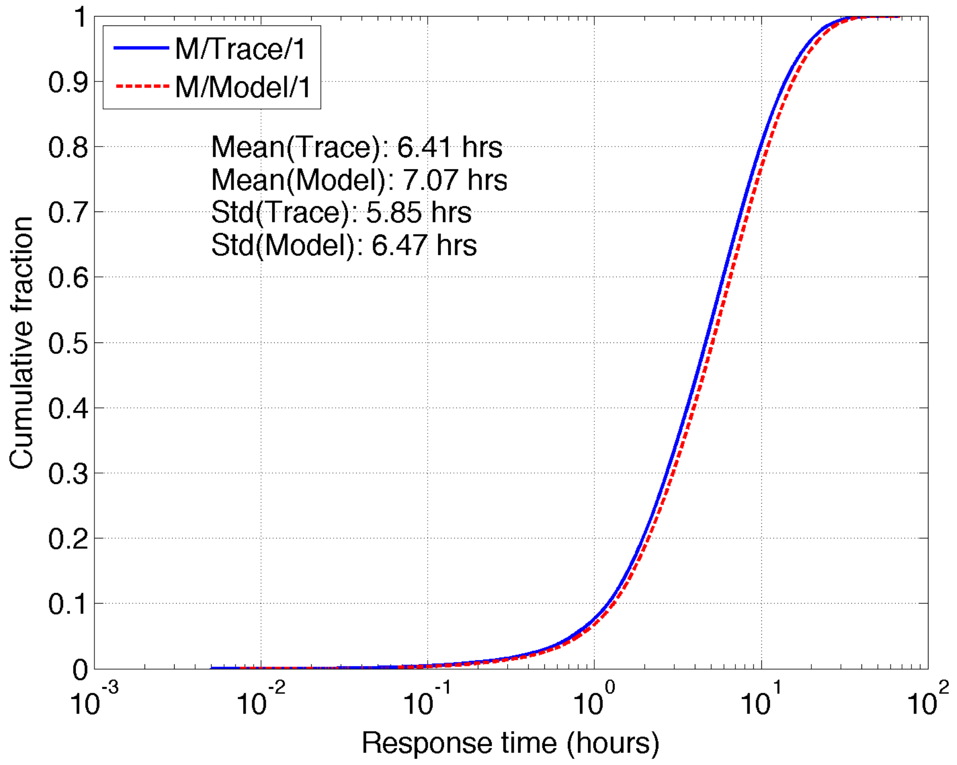
INPUT PARAMETERS FOR THE SIMULATION.



Parameter	Distribution
Download size	Uniform (12MB,18MB)
Upload size	Uniform (0.5MB,1.2MB)
Job runtime	Normal ($\mu = 1.2hrs$, $\sigma = 0.9hrs$)

Model Validation (Cont.)

- Simulation-based validation
  - Queue response time (metric)
  - High accuracy of the model



CONCLUSIONS

- A new methodology to analyze and model the host bandwidths of volunteer computing projects
  - 5-year trace with 280,000 hosts
- Proposing a bandwidth model using the Log-normal distribution in combination of an exponential model to predict the mean and variance.
- Future Work
  - the study of scenarios in which our model is used for the prediction of in-situ and in-transit analysis of data generated in Docking@Home and other volunteer computing projects.

Thank You