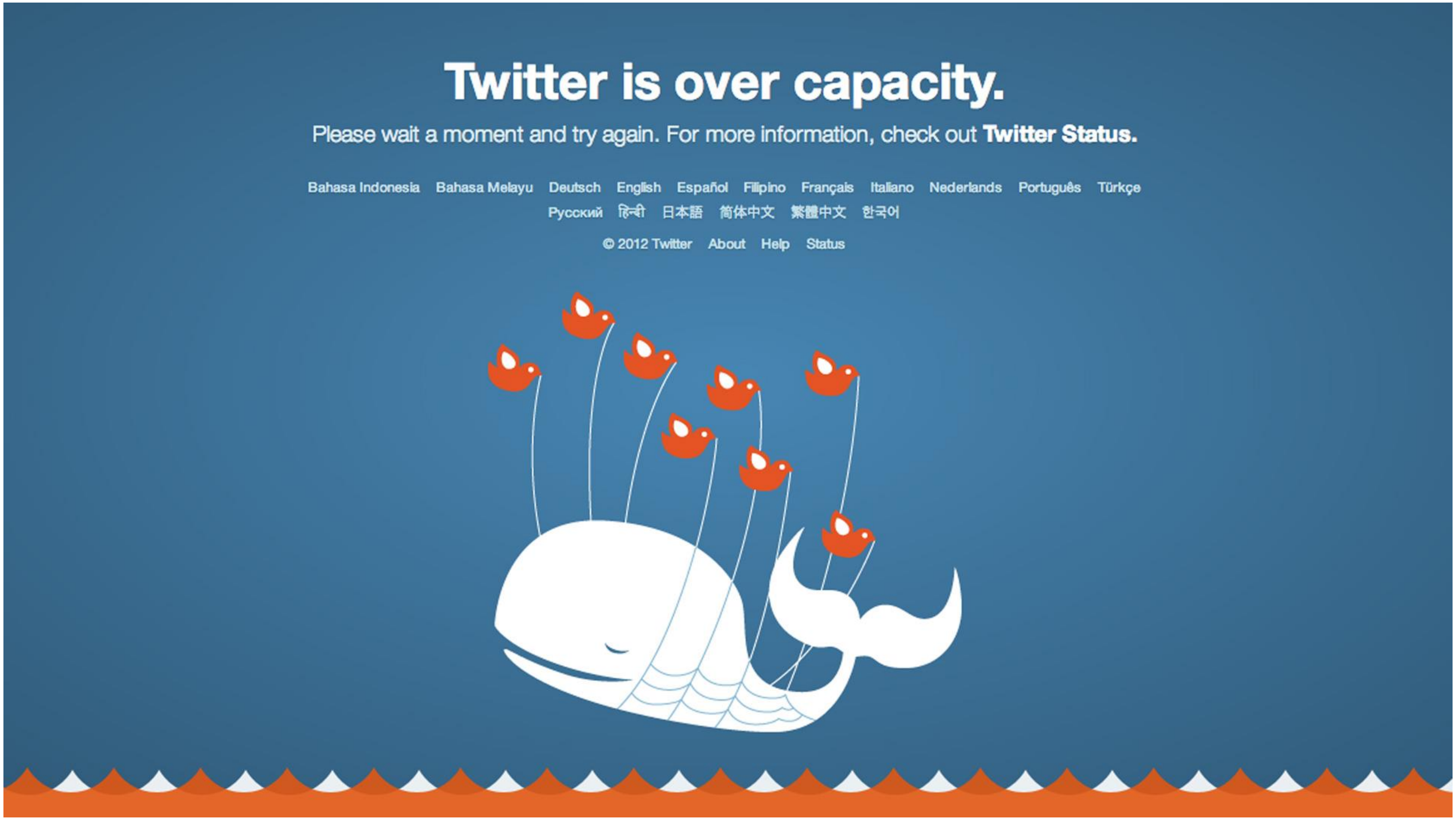
Performance Analysis of Overflow Loss Systems of Processor-Sharing Queues

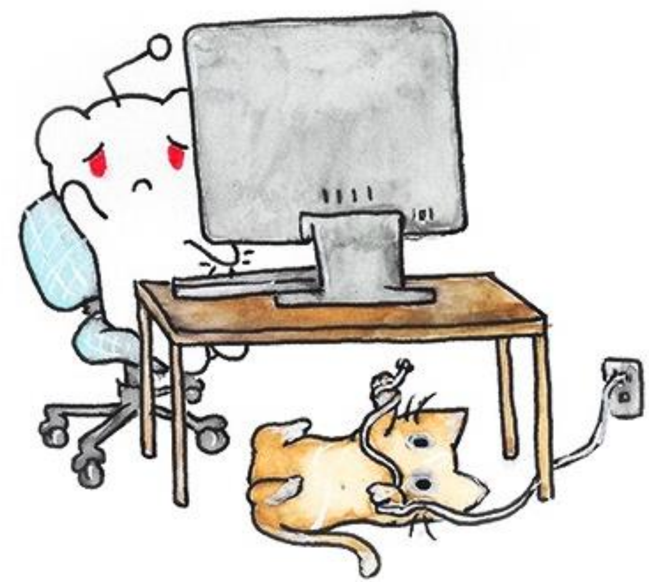
Presenter: Yin-Chi Chan
Collaborators: Jun Guo, Eric W. M. Wong, and Moshe Zukerman

Department of Electronic Engineering,
City University of Hong Kong

No one likes this.



Or this.



all of our servers are busy right now

please try again in a minute

(error code: 503)

Applications of overflow loss systems

- Content Delivery Networks (CDNs)
  - Choose closest available server with the requested file
  - All servers busy = <u>blocked request</u>
- Cloud computing
  - Consider speed, cost when assigning VMs to machines
  - All physical machines full = <u>blocked request</u>
- Cellular/wireless networks
  - Macro-cell/micro-cell layout
  - Channel borrowing between cells
  - No channel available = <u>blocked request/call</u>

Question

How to model and evaluate congestion in overflow loss systems?

– Represented by blocking probability of requests

Overflow Loss Systems

- Requests divided into classes, servers divided into groups
- Each server group serves some (or all) of the request classes
- Each request class has an ordered preference of server groups
- Requests overflow instantly from one server group to the next until an available server is found
- If all accessible servers busy, requests blocked and cleared

Terminology

- Blocking probability
  - Probability a request is blocked and cleared
- Availability
  - Proportion of server groups serving a particular request class (0 to 1)
- Processor sharing
  - Total service speed is constant, shared evenly among all active jobs in queue

Assumptions We Will Make

1. Arrivals follow a Poisson process
   - Assume time-homogeneity
2. Service durations are exponentially distributed
   - Effect of assumption demonstrated in our paper to be small

- Combined: system can be described using a Markov chain
  - # states: exponential in number of servers!
  - Need approximation method

History of Overflow Loss Systems

Classical Example: Grading System

- Originally invented for telephony switching
- Each input group connected to some of the output lines (servers)
- Request blocked if no output lines (servers) available

UNITED STATES PATENT OFFICE.

ERNEST A. GRAY, OF BOSTON, MASSACHUSETTS, ASSIGNOR TO AMERICAN TELEPHONE AND TELEGRAPH COMPANY, OF BOSTON, MASSACHUSETTS, A CORPORATION OF NEW YORK.

METHOD OF AND MEANS FOR CONNECTING TELEPHONE APPARATUS.

1,002,388. Specification of Letters Patent. Patented Sept. 5, 1911.

Application filed July 30, 1907. Serial No. 386,249.

Early Methods for Balanced Traffic

- Listed in A. Lotze's 1967 paper
  - A. Lotze, "History and development of grading theory", Proc. 5th International Teletraffic Congress, 1967
- Include:
  - Erlang Interconnection Formula (c. 1920)
  - Palm-Jacobaeus Formula (1940s)
  - Modified Palm-Jacobaeus Formula (1960s)
- Lotze 1967: these methods inadequate for unbalanced traffic

Classical Example: Grading System

- Alfred Lotze, 1967:

There is a large number of problems not yet solved, as for example:

a) Improved approximate methods for loss calculation, if unbalanced traffic is offered.

- Still true today, also relevant to more modern applications of overflow loss systems

A. Lotze, "History and development of grading theory," in Proc. 5th International Teletraffic Congress (ITC 5), 1967.

Fixed Point Approximation

- Classical approximation for telephony networks
- Idea:
  - Treat server groups as independent
  - Treat traffic (including overflows) as Poisson
- Adapted to systems of processor-sharing (PS) queues in 2012 (Muñoz-Gea *et al.*)
  - We shall call this PS-FPA

J. P. Muñoz-Gea, S. Traverso, and E. Leonardi, "Modeling and evaluation of multisource streaming strategies in P2P VoD systems," *IEEE Trans. Consum. Electron.*, vol. 58, no. 4, pp. 1202–1210, Nov. 2012.

Inadequacies of FPA

1. Poisson error: overflow traffic is not actually Poisson
   – Propagation of error means error is worse when availability is high
2. Independence error: server groups not actually independent

- <u>IESA designed to address these two error sources</u>

Information Exchange Surrogate Approximation

The Surrogate in IESA

- Evaluate blocking probability of original model via a surrogate (different) model
- Model dependencies via use of information exchange between requests
- Use information to preemptively discard requests that are likely to be blocked
  - Reduce Poisson and independence error when approximation is applied to surrogate model

The Surrogate in IESA

- Approximate original (non-hierarchical) overflow structure with a hierarchical one
  - →Less error when assuming server independence + Poisson input as dependencies are one-way only
  - →IESA terminates in bounded number of iterations (avoid fixed-point solution)
- Traffic to a server at a given tier may depend on traffic to any server, but only for lower tiers

The Surrogate in IESA

Countered by reduced approximation error when applying approximation to surrogate

Surrogate system slightly alters blocking probability of system

IESA is a Framework

- **ICC 2005**: IESA1 (originally OPCA)
- **ITC 2013**: IESA2
- **<u>INFOCOM 2015</u>**: two approximations for systems of processor-sharing nodes
  - PS-IESA1
  - PS-IESA2

E. W. M. Wong, M. Y. M. Chiu, **M. Zukerman**, Z. Rosberg, S. Chan, and A. Zalesky, “A novel method for modeling and analysis of distributed video on demand systems,” in Proc. 2005 IEEE International Conference on Communications (ICC), vol. 1, May 2005, pp. 88–92.

E. W. M. Wong, J. Guo, B. Moran, and **M. Zukerman**, “Information exchange surrogates for approximation of blocking probabilities in overflow loss systems,” in Proc. 25th International Teletraffic Conference (ITC 25), Sep. 2013.

IESA2

- Request properties
  - Δ, set of already attempted server groups
  - Ω, congestion estimate (scalar)
- If server group is full:
  - Find request in service with largest Ω (request A)
  - If Ω of incoming request < Ω of request A, swap Ω values
- Ω also increases by 1 for every overflow
- Overflowing requests are preemptively discarded with a probability based on Ω and $|\Delta|$

Example

Ω = 7	Ω = 9
~~47~~	~~18~~
~~6~~	~~41~~
~~8~~	~~19~~
~~9~~	~~97~~
20	20
1	4
2	7

Server group 20: capacity = 2

Ω = 8
~~89~~
~~6~~
~~4~~
~~44~~
20
16
55

Incoming

Example

Ω = 7	Ω = 9	Ω = 8
~~47~~	~~18~~	~~89~~
~~6~~	~~41~~	~~6~~
~~8~~	~~19~~	~~4~~
~~9~~	~~97~~	~~44~~
20	20	20
1	4	16
2	7	55

Server group 20: capacity = 2

Incoming

Example

Ω = 7

~~47~~
~~6~~
~~8~~
~~9~~
20
1
2

Ω = 8

~~18~~
~~41~~
~~19~~
~~97~~
20
4
7

Server group 20: capacity = 2

Ω = 910

~~89~~
~~6~~
~~4~~
~~44~~
~~20~~
16
55

Overflow

NUMERICAL RESULTS

Processor-Sharing Model

- Inspired by peer-to-peer VoD
  - Peers: store movies, serve and generate requests
  - Leeches: generate requests only
  - Fixed content allocation
- Each peer is a single processor-sharing server
  - Finite number of slots
- Popularity of each movie follows a Zipf distribution: $\lambda_c \propto c^{-z}$ for some $z \geq 0$
- Availability of each movie directly proportional to its popularity

J. P. Muñoz-Gea, S. Traverso, and E. Leonardi, "Modeling and evaluation of multisource streaming strategies in P2P VoD systems," *IEEE Trans. Consum. Electron*., vol. 58, no. 4, pp. 1202–1210, Nov. 2012. [Model used with minor modifications]

Processor-Sharing Model

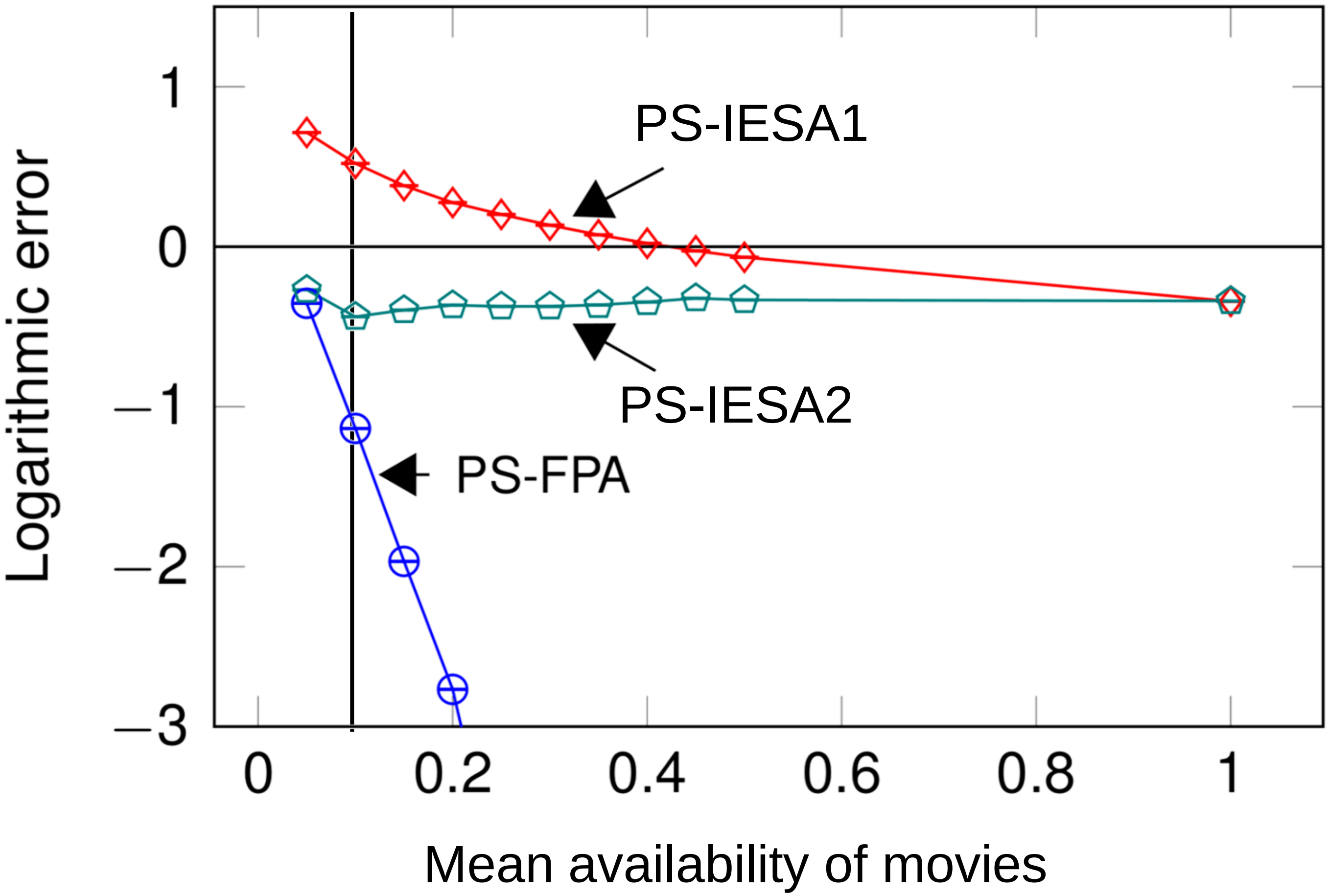
- Requests from each peer/leech form a Poisson process
  - If peer requests a movie it can serve itself, no resources used up
- Assume each request requires an exponentially distributed service time with unit mean

To Compare

- Processor-sharing version of EFPA
  - PS-FPA
- Processor-sharing version of IESA1
  - PS-IESA1
- Processor-sharing version of IESA2
  - PS-IESA2
- Simulation

Accuracy vs. Availability

2000 M/M/1/5-PS peers, 2000 contents, Zipf(0.271), target blocking = 0.005

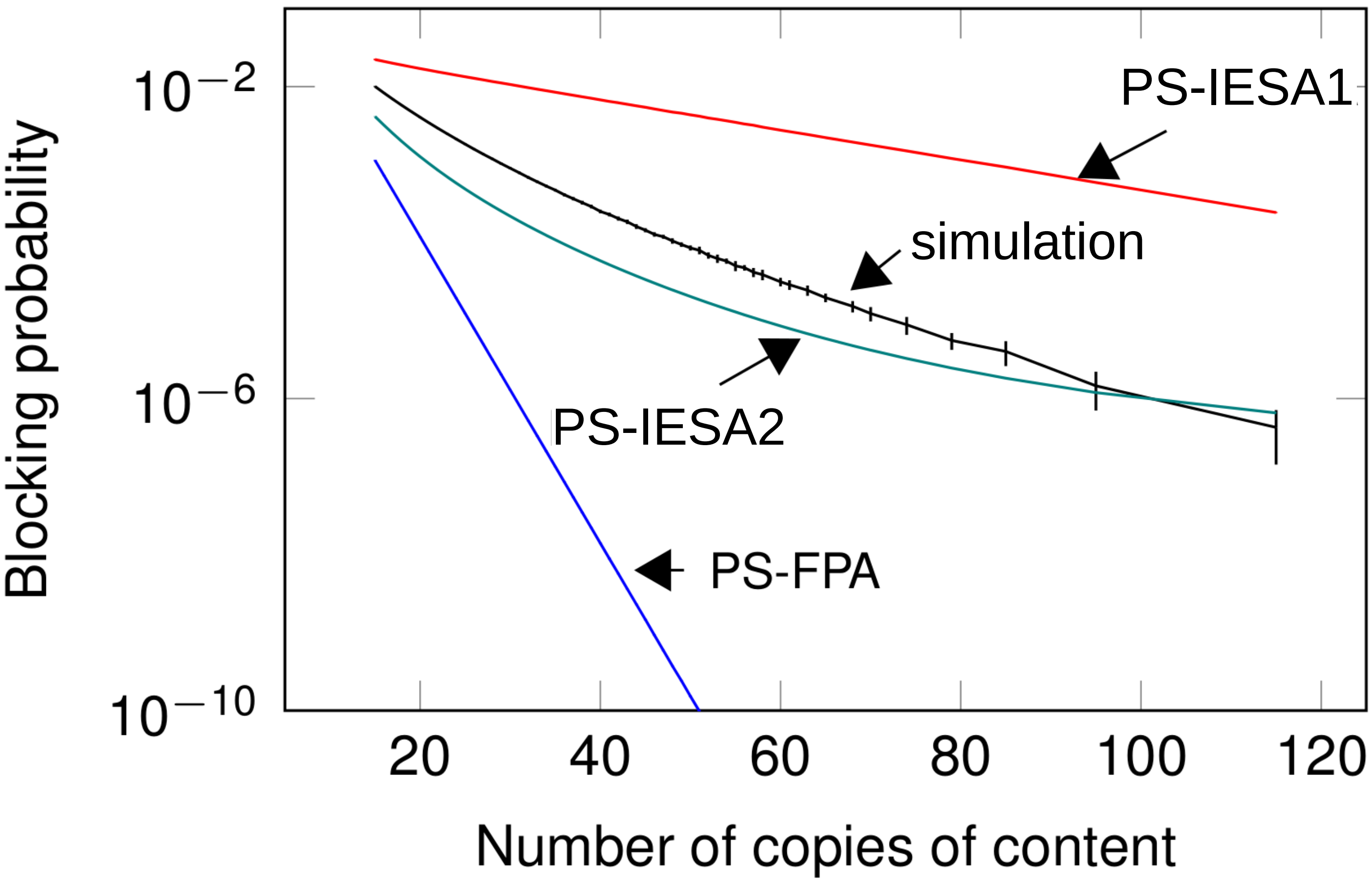


Both PS-IESA1 and PS-IESA2 are much more accurate than PS-FPA

PS-IESA2 is only approximation to be (mostly) insensitive to availability

Accuracy for each Content Type

2000 M/M/1/5-PS peers, 0.1 availability, 2000 contents, Zipf(0.271), target blocking = 0.005



PS-IESA2 captures blocking probability of <u>individual</u> content types more accurately than PS-FPA and PS-IESA1

Concluding Remarks

- More data is available in the conference paper
  - Demonstrates that PS-IESA2 is generally more accurate and robust than PS-FPA and PS-IESA1
- Further work
  - Better surrogates for IESA
  - Moment matching
  - Application to other models
    - IESA1 has been applied to optical networks and cellular networks
    - Can do same for IESA2

E. W. M. Wong, J. Baliga, **M. Zukerman**, A. Zalesky, and G. Raskutti, "A new method for blocking probability evaluation in OBS/OPS networks with deflection routing," J. Lightwave Technology, vol. 27, no. 23, pp. 5335–5347, Dec 2009.

J. Wu, **J. Guo, E. W. M. Wong, and M. Zukerman**, "Performance analysis of channel borrowing in mobile networks," in Proc. 14th IEEE HK/Macau AP/MTT Postgraduate Conference, Oct. 2013

Q&A

Thank you