



清華大學
Tsinghua University

Understanding User Behavior in Large Scale Video-on-Demand Systems

Hongliang Yu, Dongdong Zheng, Ben Y. Zhao, Weimin Zheng
Tsinghua University and UC Santa Barbara
Eurosys Conference 2006

Motivation

- VOD: the future of media networks
 - Select your favorite movies as you like, any time, anywhere, impressive
 - In China, up to Jan. 2005, 8 million VOD users, 5 million of them use it frequently, increasing with a rate of 35% per year (China Telecommunication Newspaper).
 - In global view, 90 million VOD users in 2003, 138 million users in 2005, 327 million users estimated in 2010 (Information Media Group)
 - Most of current system are not True VOD: Business Reasons? Technical Reasons?

Motivation

- Characteristics of VOD
 - Multi data source
 - Asynchronous data stream
 - High interactivity, VCR
- Challenges
 - High Network Bandwidth
 - High Random I/O capacity
- Technical approaches
 - Caching Policies
 - Data replication
 - Distributed content delivery
- Providing VOD service to a huge number of clients in a scalable way still unsolved

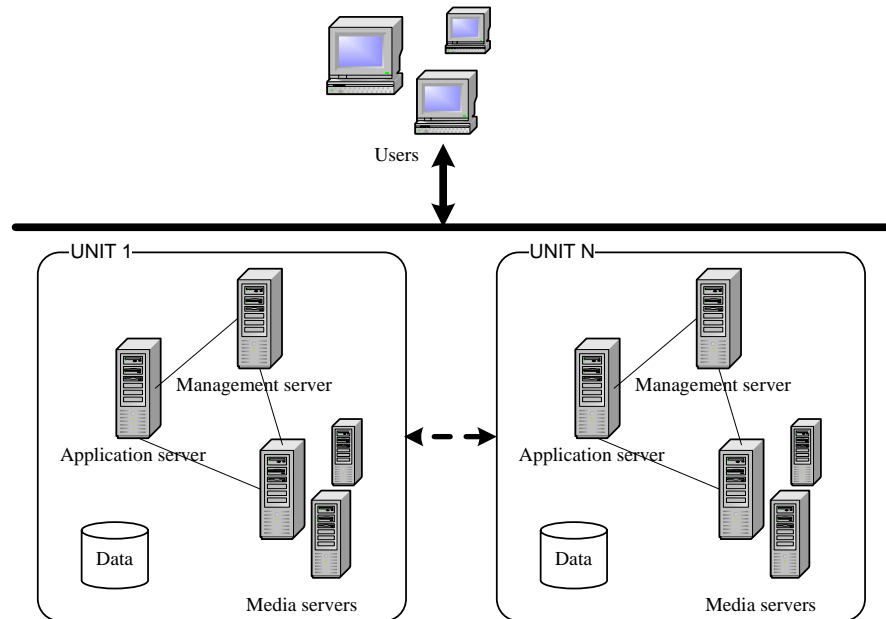
Motivation

- Challenging to address user behavior model for VOD system optimization
 - Little knowledge about the user behavior of deployed large scale VOD system, chicken and egg?
 - Current researchers based their studies on rental data from video stores, or small scale VOD systems, or web streaming services
 - ◆ Video rental: lack of enough video title, limited physical copy
 - ◆ Web streaming: narrow band service, smaller file size, bad video quantities, affects user behavior much

The focus of this paper

- Things useful to on demand video streaming system design and maintenance
 - How about the user-arrival rate in such a system
 - In what situation, people like to keep their patience
 - What part of content people tend to visit
 - How user interests change over time
 - What features should we keep in such services?

Source of Data



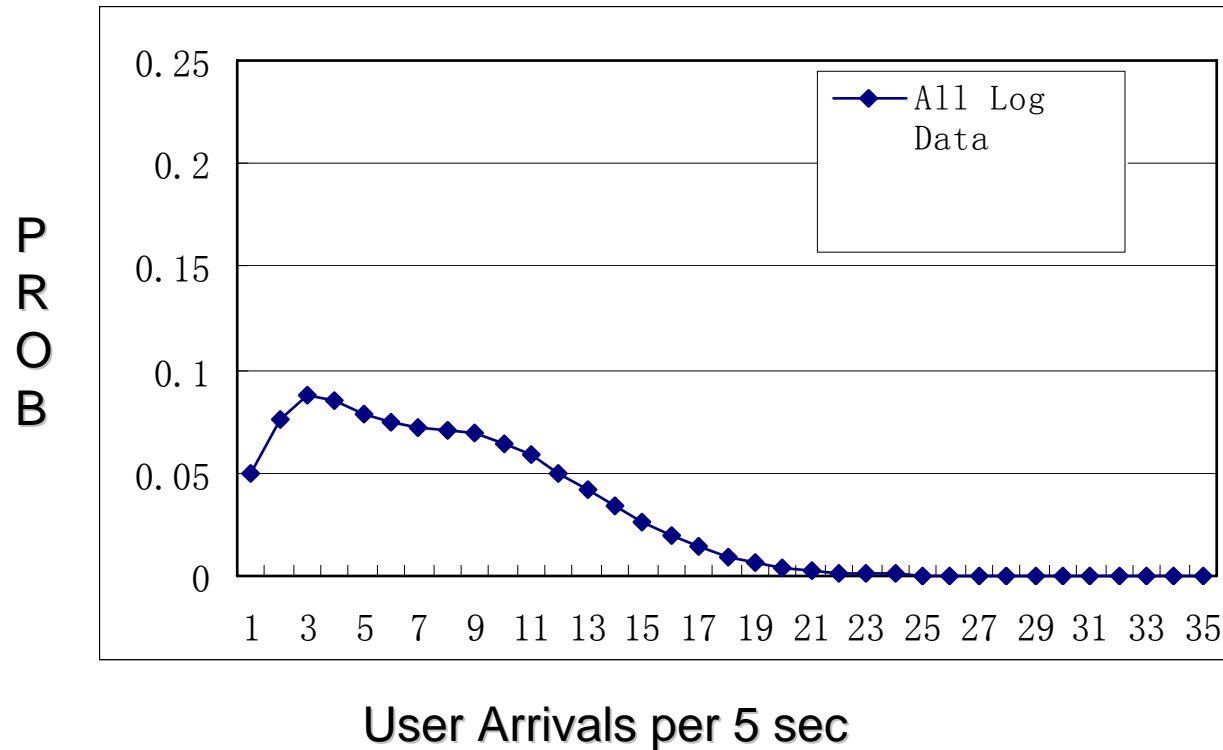
- Log data from an infrastructure based large scale VOD service deployed in China (<http://www.powerinfo.com.cn>)
- The total user of the system is over 1.5 million users, use a regional data contains about 150 thousand users

- 21,498,338 sessions in 219 days; 7,036 movies involved
- Movie length: 38.23%, >90min; 41.76%, 45-90min
- Average data rate is about 384Kbps (512K ADSL support)

Outline

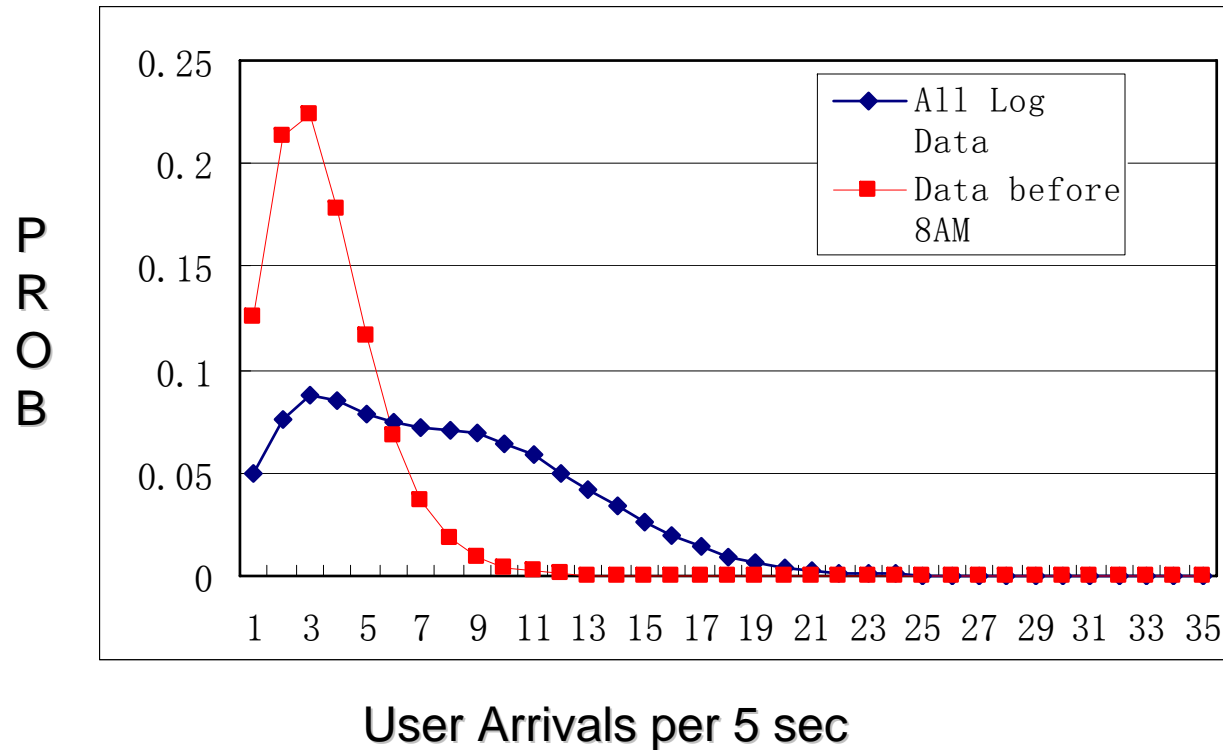
- Motivation
- Source of Data
- **Poisson Distribution**
- Session Length
- User Interests
- Summary

User arrival rate



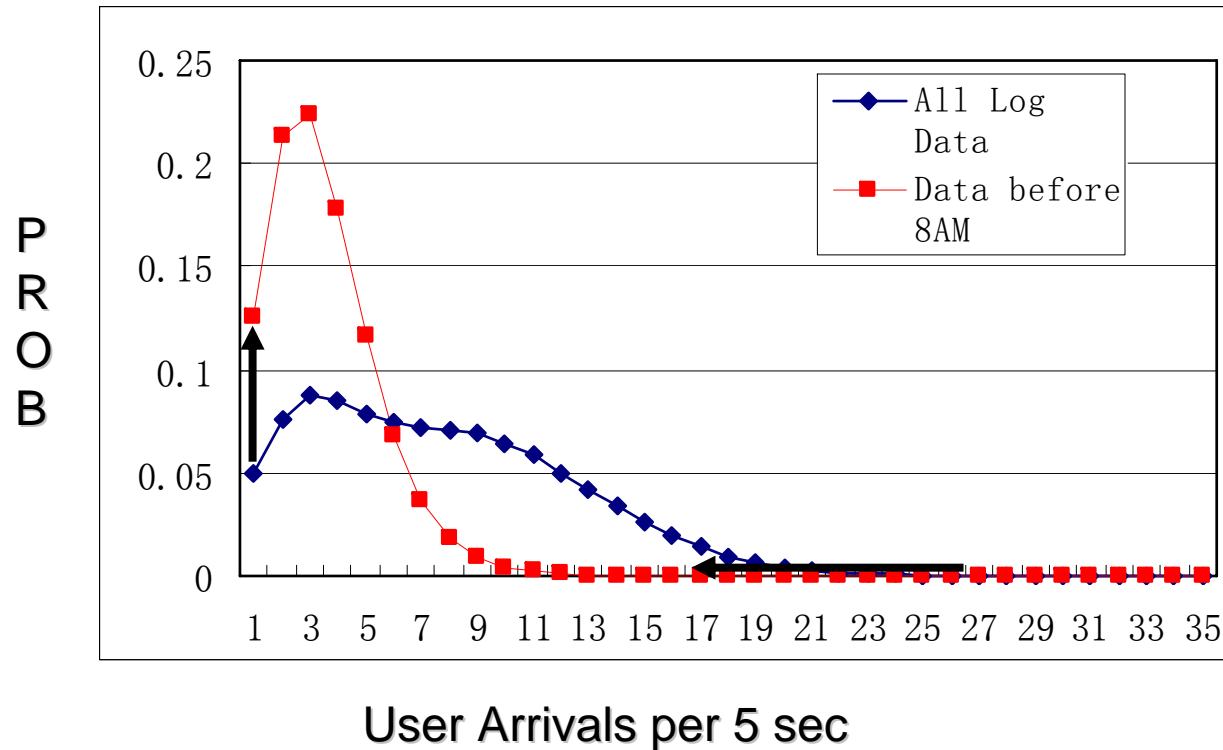
- 0-27 arrivals per 5 seconds, do not match the Poisson

User arrival rate



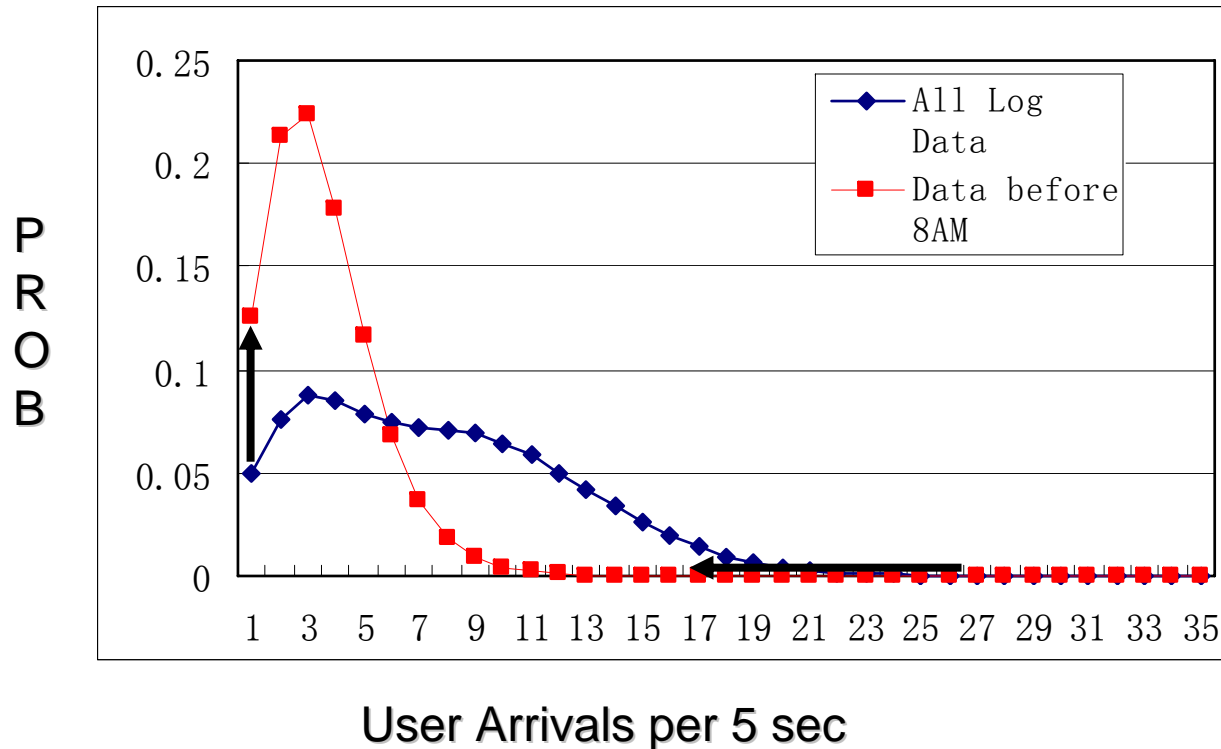
- 0-27 arrivals per 5 seconds, do not match the Poisson

User arrival rate



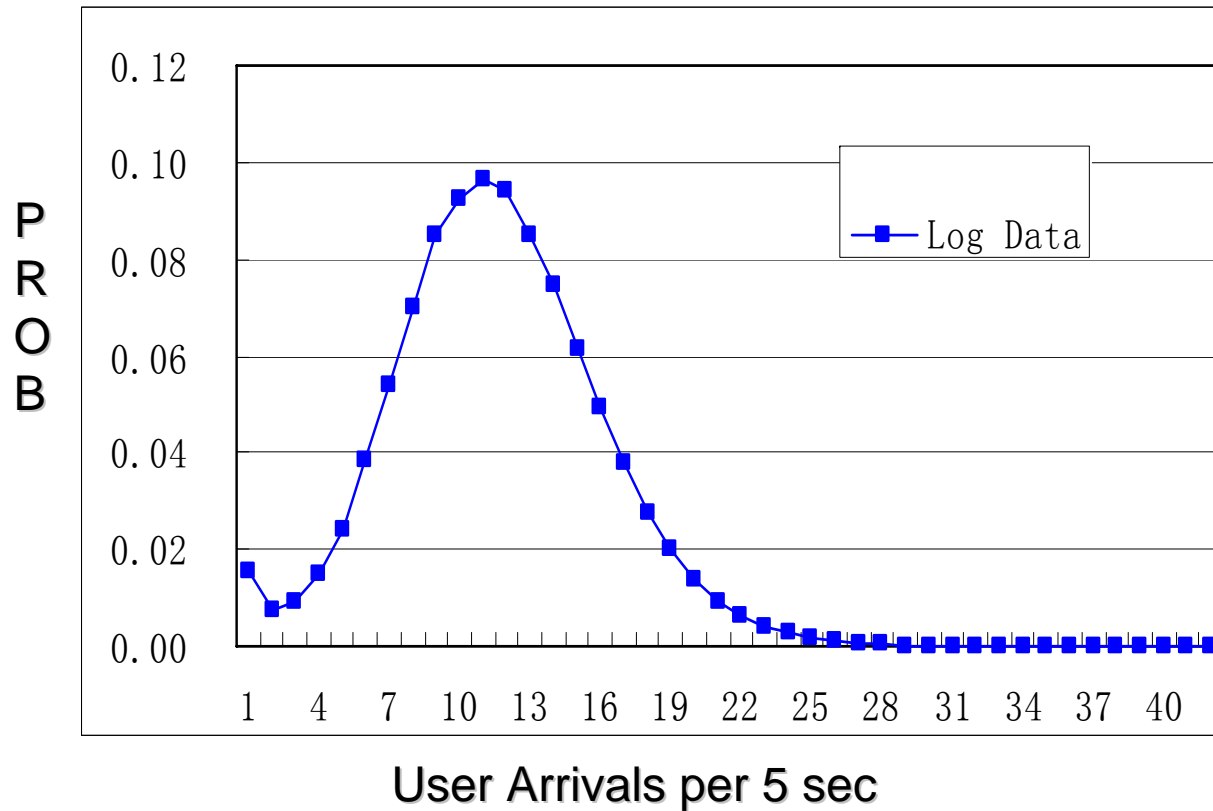
- 0-27 arrivals per 5 seconds, do not match the Poisson

User arrival rate



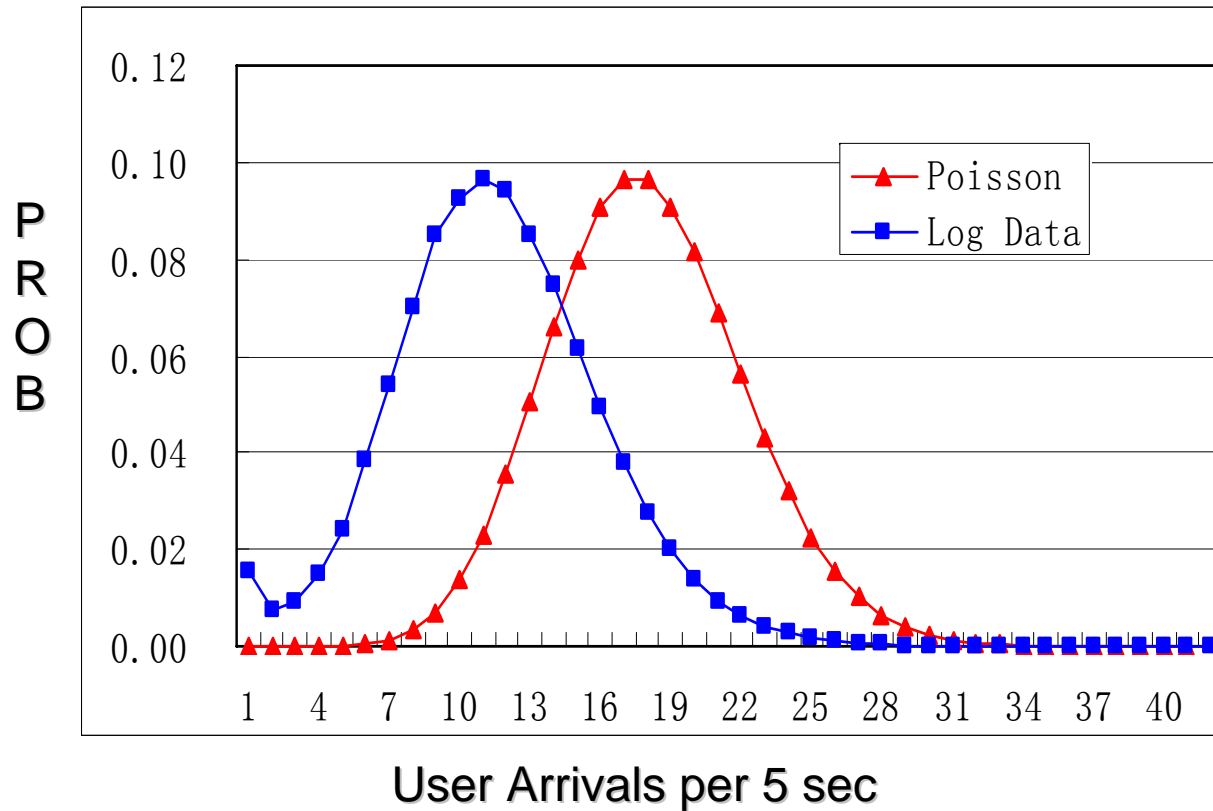
- 0-27 arrivals per 5 seconds, do not match the Poisson
- Guess: System Idle time may be responsible for the failure of Poisson

User Arrival Pattern



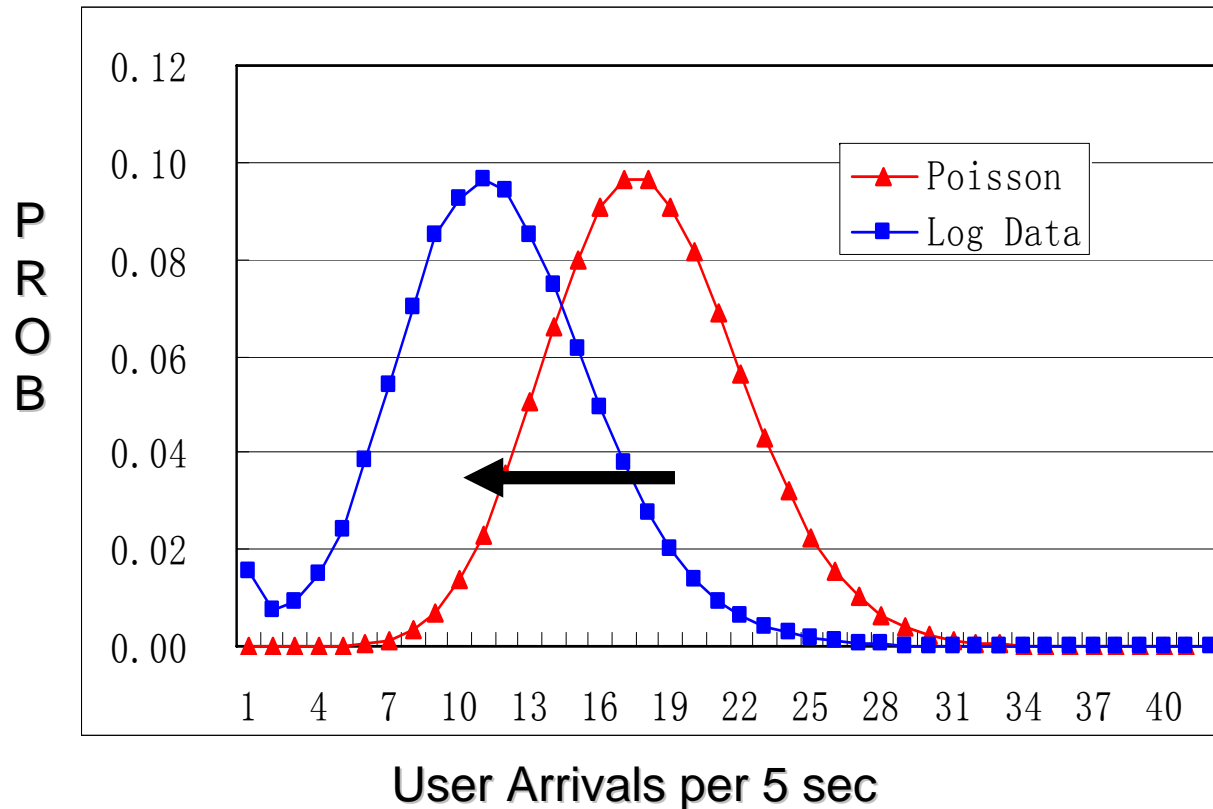
- Using data from rush hour(6PM to 9PM), similar shape with Poisson

User Arrival Pattern



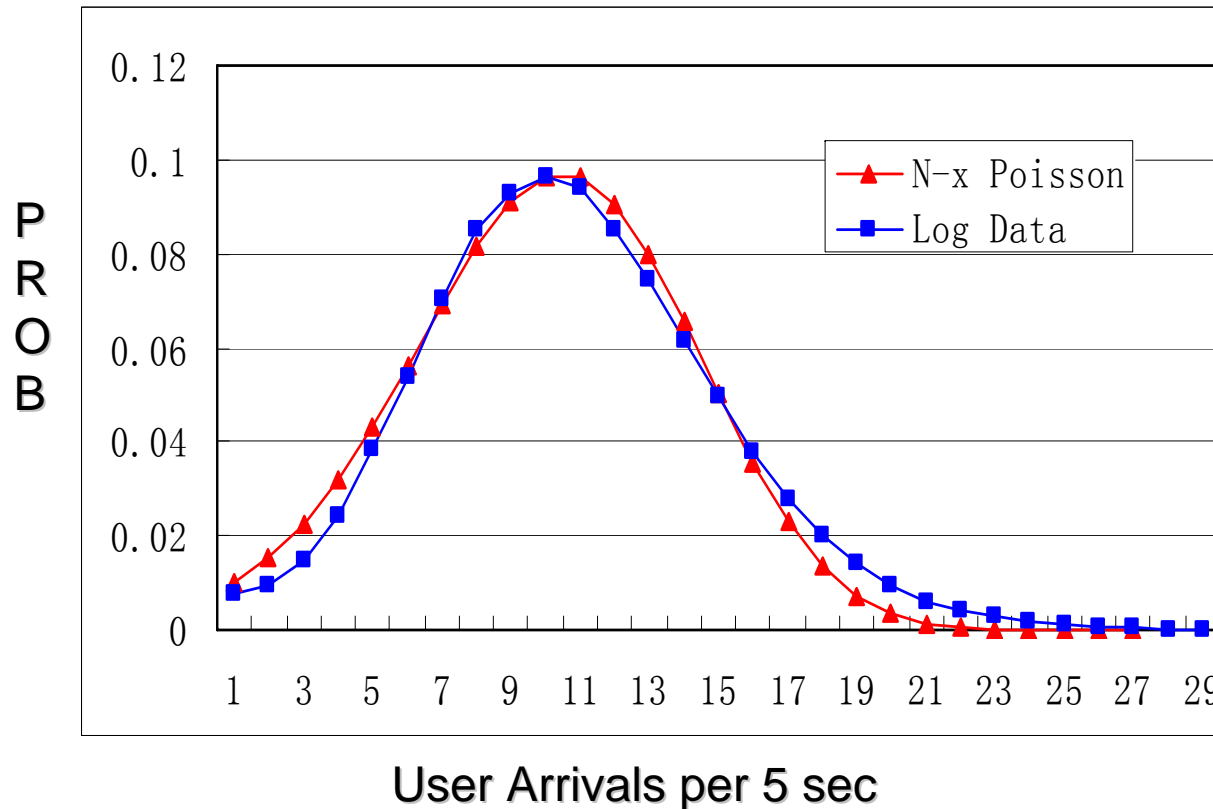
- Using data from rush hour(6PM to 9PM), similar shape with Poisson

User Arrival Pattern



- Using data from rush hour(6PM to 9PM), similar shape with Poisson
- Large arrivals overestimated

User Arrival Pattern



- Using data from rush hour(6PM to 9PM), similar shape with Poisson
- Modified version of Poisson fit well with real workload:

$$P(X) = \frac{\lambda^{N-X} e^{-\lambda}}{(N-X)!}, \quad X=0,1,2,\dots$$

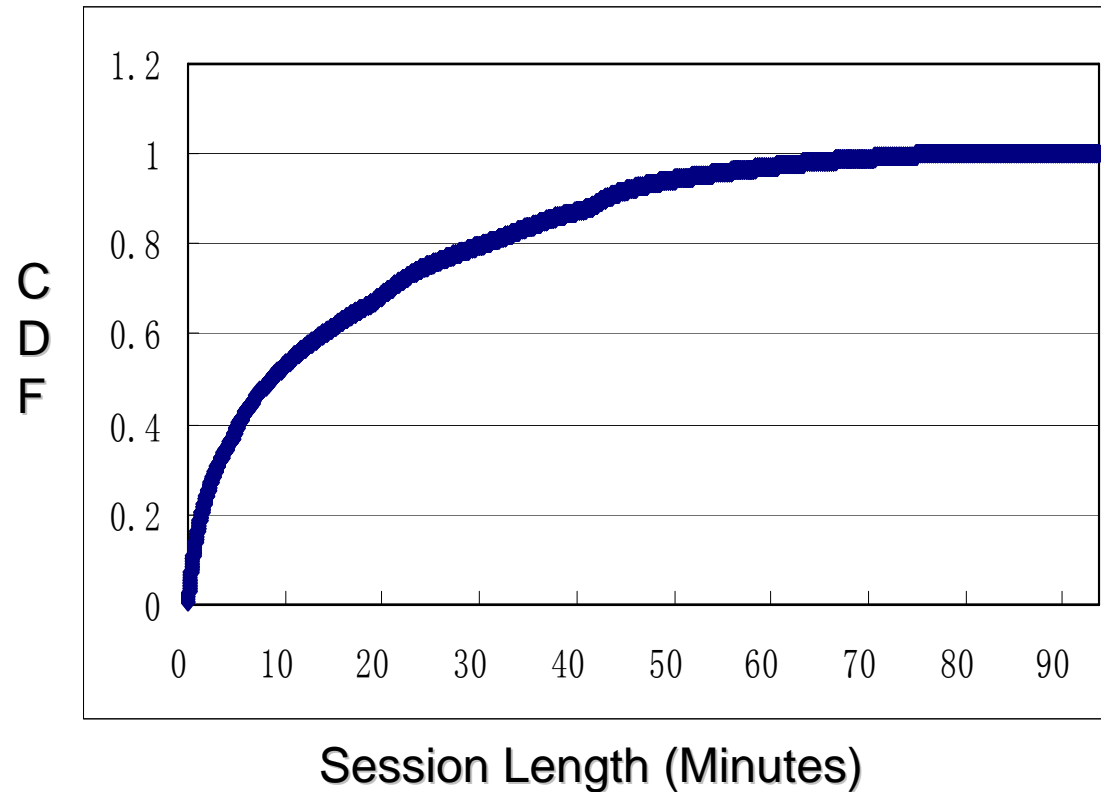
Indication

- The Poisson distribution underestimates the possibility of small arrival cases and it over-estimates the probability of large arrivals, inefficient resource reservation
- With modified model, you can “design” the maximum user arrival rate (N) according to user requirement and investment plan

Outline

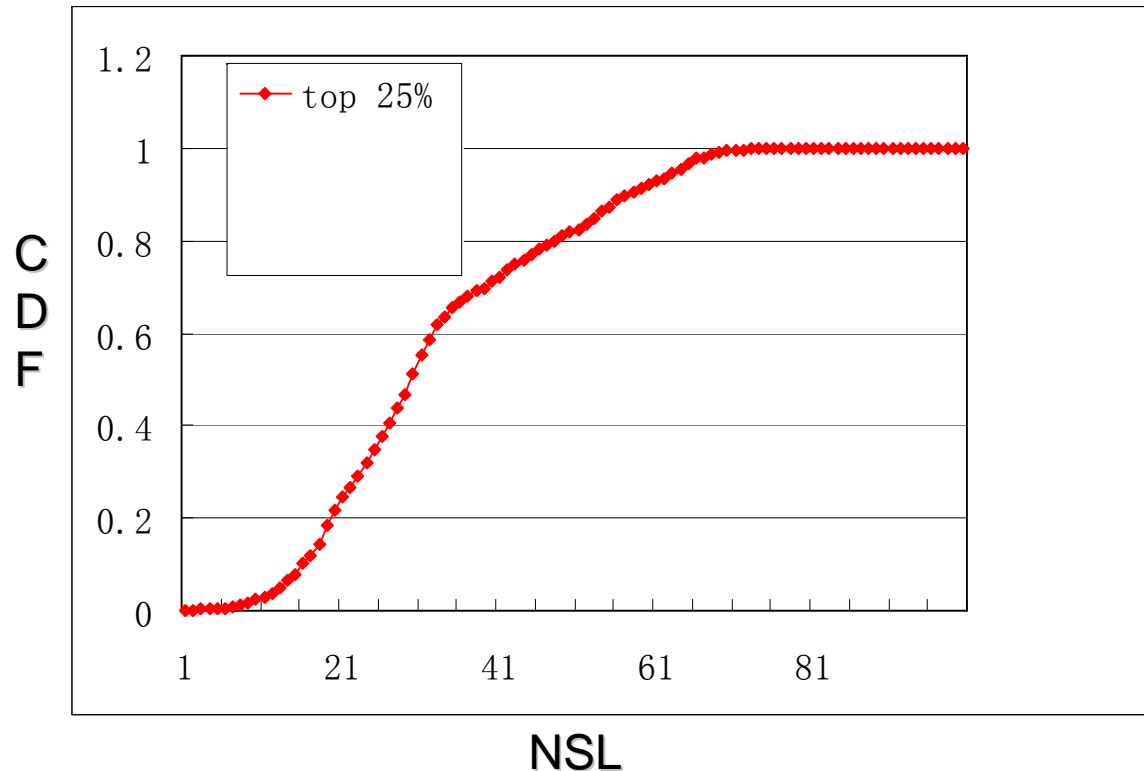
- Motivation
- Source of Data
- Poisson Distribution
- **Session Length**
- User Interests
- Summary

Session length: impatient audience



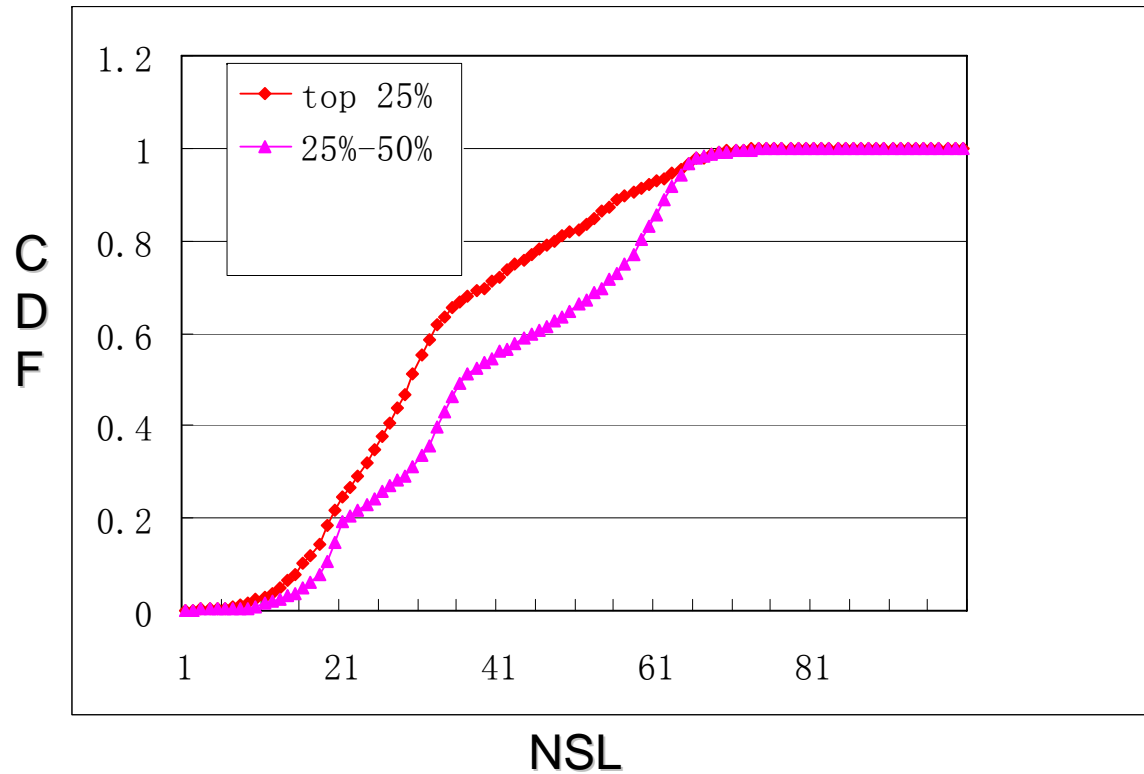
- 37% users terminate their session in the first 5 minutes
- 52.55% in 10 minutes
- 75% in 25 minutes

Session length: related with popularity?



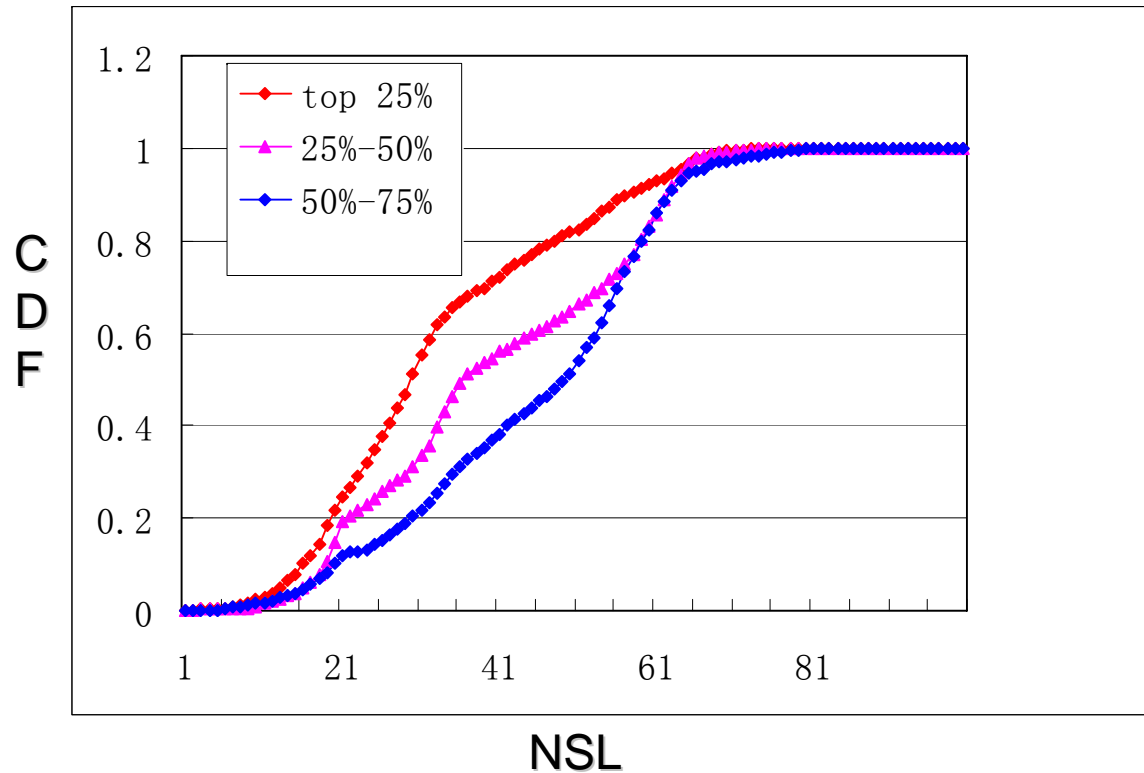
- NSL: a ratio of SessionLength / VideoLength
- Expected: Movies with higher popularity have longer session length.

Session length: related with popularity?



- NSL: a ratio of SessionLength / VideoLength

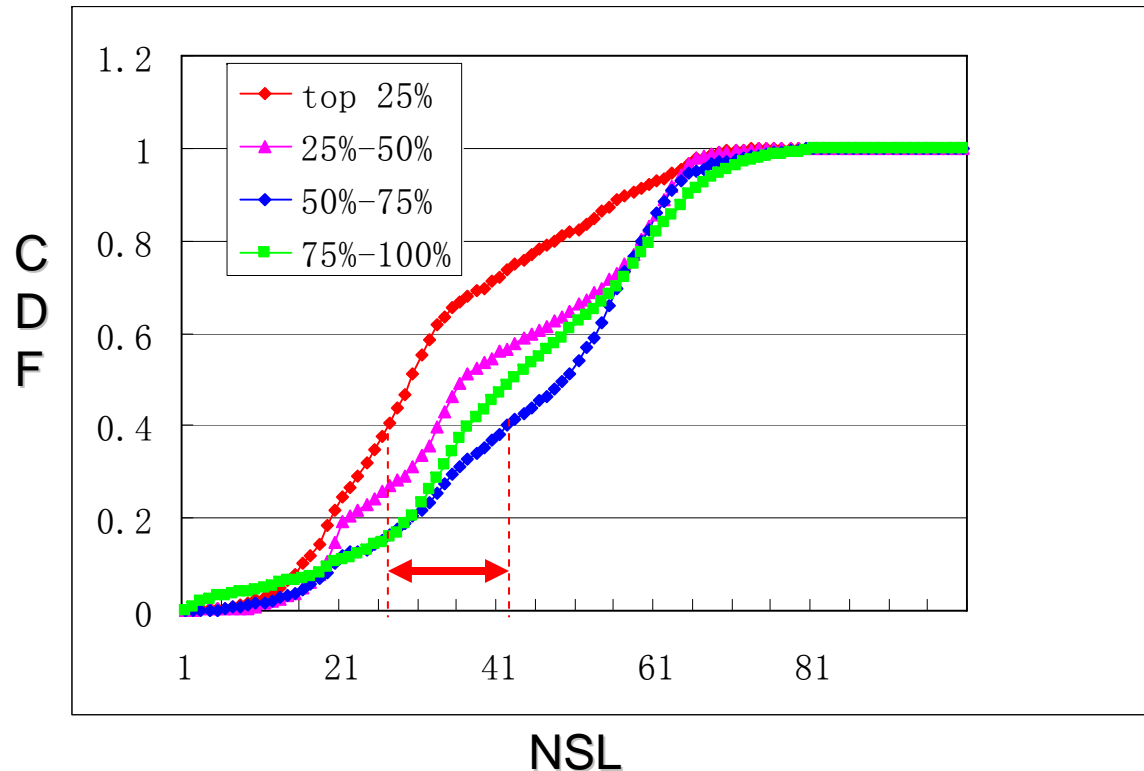
Session length: related with popularity?



- NSL: a ratio of SessionLength / VideoLength
- Movies with HIGHER popularity tend to have SHORTER session length!

Surprise!!!

Session length: related with popularity?



- NSL: a ratio of SessionLength / VideoLength
- The relation between movie popularity and session length does exist, but not so strong

Example: caching optimization

Movie A

A ₀	A ₁	A ₂	A ₃
----------------	----------------	----------------	----------------

Movie B

B ₀	B ₁	B ₂	B ₃
----------------	----------------	----------------	----------------

Movie C

C ₀	C ₁	C ₂	C ₃
----------------	----------------	----------------	----------------

- Movie A is the most popular movie, movie B second, Movie C last

Caching Priority:

A ₀	A ₁	A ₂	A ₃	B ₀	B ₁	B ₂	B ₃	C ₀	C ₁	C ₂	C ₃
----------------	----------------	----------------	----------------	----------------	----------------	----------------	----------------	----------------	----------------	----------------	----------------

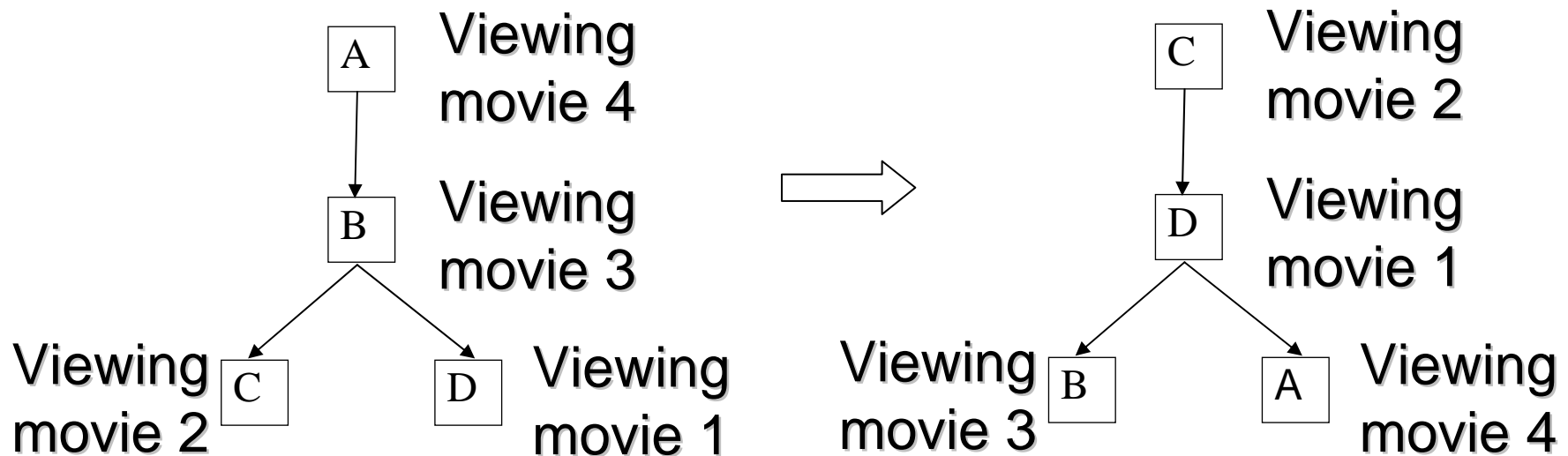


A ₀	A ₁	B ₀	C ₀	B ₁	C ₁	A ₂	B ₂	B ₃	C ₂	A ₃	C ₃
----------------	----------------	----------------	----------------	----------------	----------------	----------------	----------------	----------------	----------------	----------------	----------------

- The latter priority list is more reasonable
- Not all part of the most popular movie should be stressed

Example: ALM optimization

- Movie 1,2,3,4 from least popular to most popular



The right ALM tree has a better chance to be stable

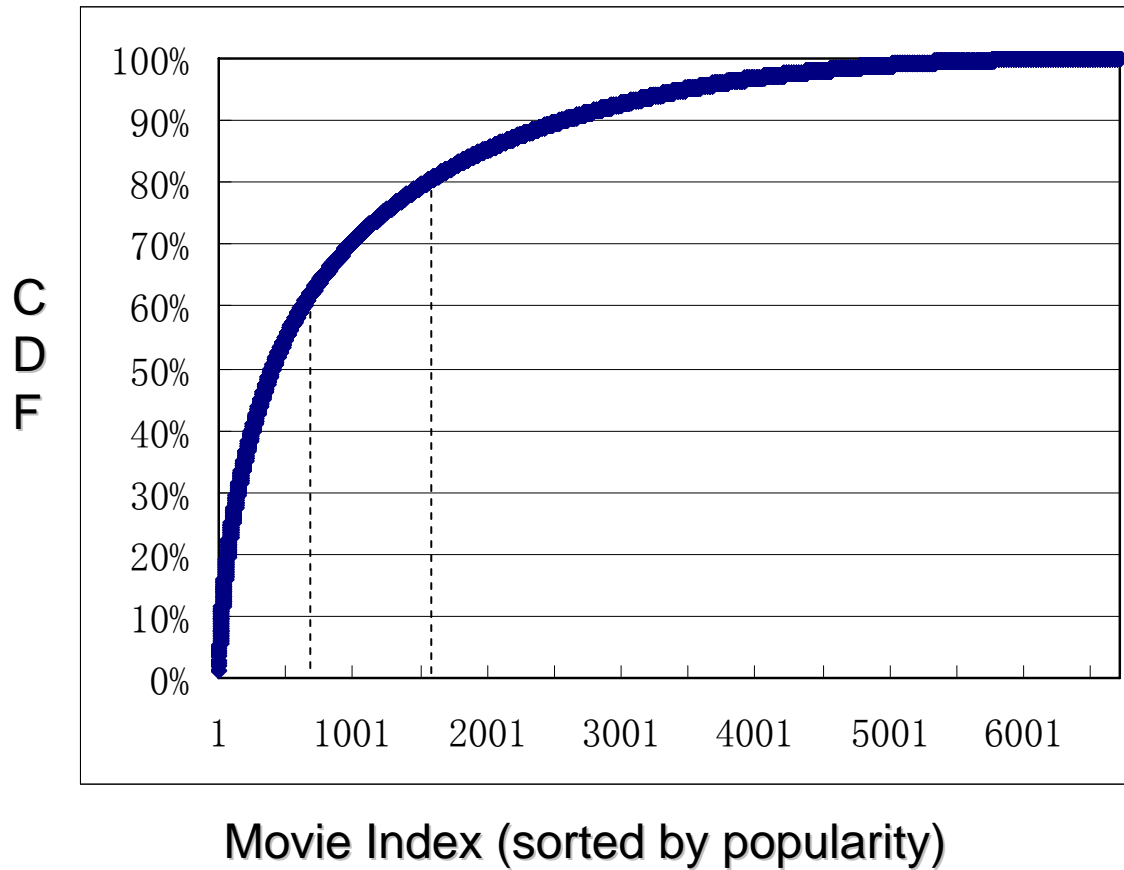
Indication

- Caching the prefix is effective
- Popularity may not reflect the potential of the content
 - In Ebay, high reputation user concede much higher reputation in latter time than they owned
 - In Powerinfo VOD, high reputation movies are not always so attractive, people only attracted by its reputation
- Caching policy based on content segment popularity counting is more effective
- Set the node viewing relative “colder” contents to the position near the root of ALM tree will be effective

Outline

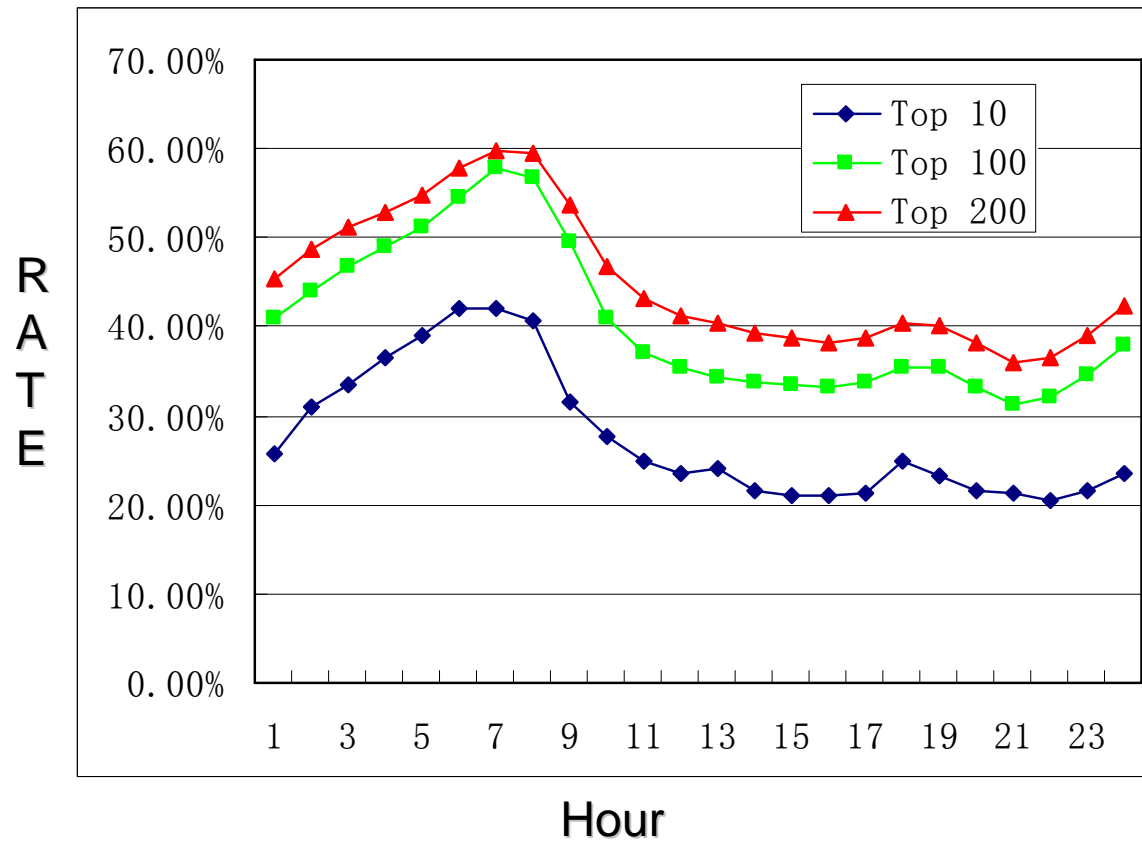
- Motivation
- Source of Data
- Poisson Distribution
- Session Length
- **User Interests**
- Summary

User interests distribution

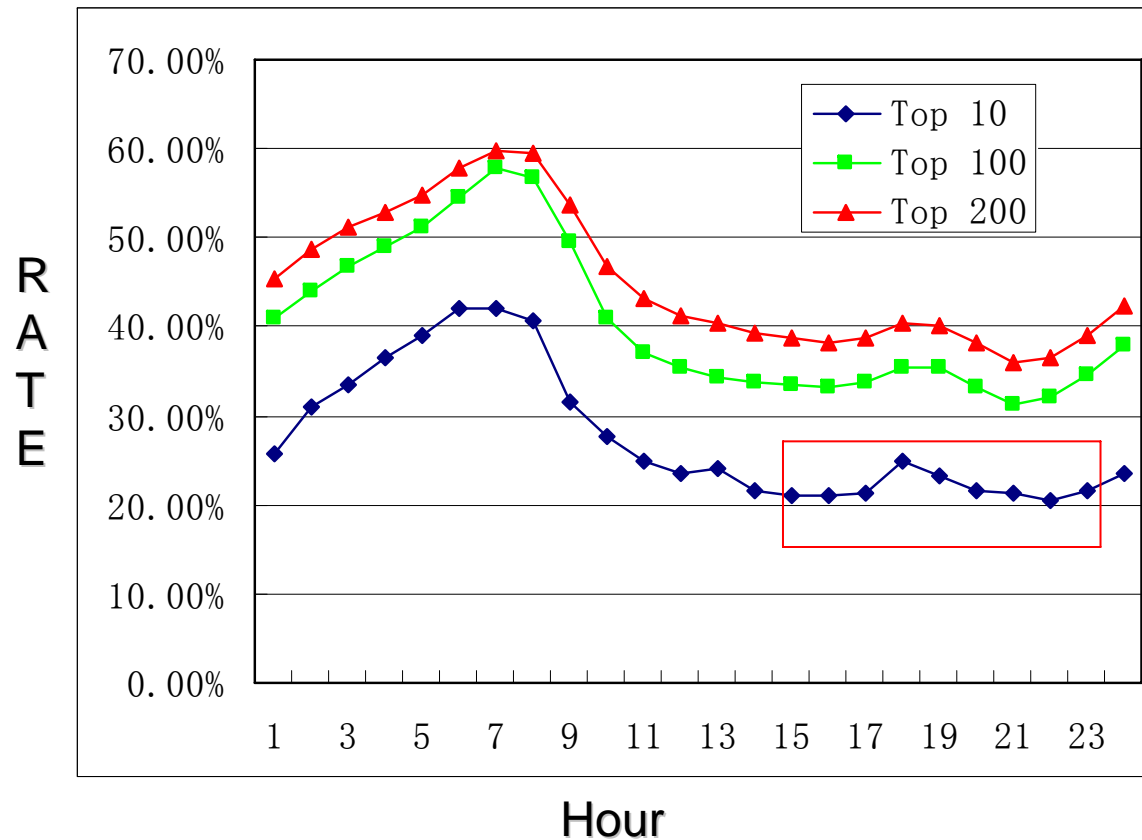


- 10% objects covering 60% of accesses
- 23% objects got 80% of the hits

User interests transferring

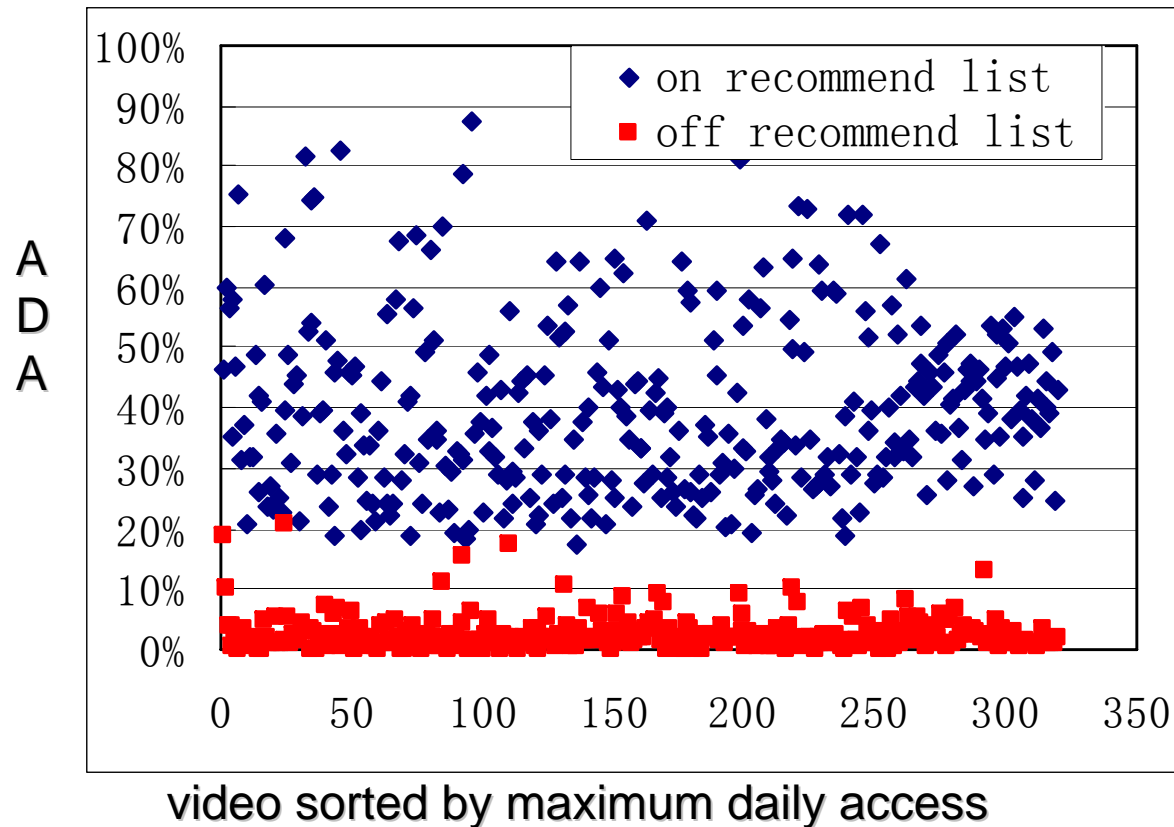


User interests transferring



- User interest changes slowly
- Caching replacement overhead will be low

Understanding popularity: recommendation



- ADA: average daily access / maximum daily access
- Caching according to recommendation is good

Indication

- User interests change slowly, caching benefit
- Interest inducement: user interests can be “induced”, with mechanisms like movie recommendation
- Features like movie recommendation are performance benefit

Summary

- **Indications**
 - Poisson over-estimates the probability of large arrivals, inefficient resource reservation
 - Caching and forwarding with regards to content popularity will be necessary
 - Use features like content recommendation benefits caching policy much
- **Ongoing work**
 - VCR studies
 - Feasibility of P2P VOD
 - Optimization deployment
 - Data set open



清華大學
Tsinghua University

Thanks!!!

Any Questions?