Challenges, Design and Analysis of a Large-scale P2P-VoD System

Yan Huang*, Tom Z. J. Fu#, Dah-Ming Chiu#, John C. S. Lui and Cheng Huang*
*{galehuang, ivanhuang}@pplive.com, Shanghai Synacast Media Tech.
#{zjfu6, dmchiu}@ie.cuhk.edu.hk, The Chinese University of Hong Kong
cslui@cse.cuhk.edu.hk, The Chinese University of Hong Kong

ACM SIGCOMM 2008

1

Outline

- P2P overview
- An architecture of a P2P-VoD system
- Performance metrics
- Measurement results and analysis
- Conclusions

2

P2P Overview

- Advantages of P2P
  - Reduced server load
  - Robustness
- P2P services
  - P2P file downloading : BitTorrent and Emule
  - P2P live streaming : Coolstreaming, PPStream and PPLive
  - P2P video-on-demand (P2P-VoD) : Joost, GridCast, PFSVOD, UUSee, PPStream, PPLive...

3

P2P-VoD System Properties

- Less synchronous compared to live streaming
  - Peers may watch different parts of a viedo
- Requires more storage
  - each user contribute extra storage
- Requires careful design of mechanisms for
  - Content Replication
  - Content Discovery
  - Peer Scheduling

4

P2P-VoD system

- (Content) Servers
  - The source of content (e.g., movies)
- Trackers
  - Help peers connect to other peers to share the content
- Bootstrap server
  - Helps peers to find a suitable tracker
- Peers
  - Run P2P-VoD software
  - Some implement DHT(Dynamic Hash Table)
- Other servers
  - Log servers : log significant events for data measurement
  - Transit servers : help peers behind NAT boxes

5

Design Issues To Be Considered

- Segment size
- Replication strategy
- Content discovery
- Piece selection
- Transmission Strategy
- Others:
  - NAT and Firewalls
  - Content Authentication

6

Segment Size

- Segment is a piece of content
- What is a suitable segment size?
  - Small
    - More scheduling flexibility
    - But larger overhead
      - Header overhead
      - Bitmap overhead
      - Protocol overhead
  - Large
    - Smaller overhead
    - Limited by viewing rate
- Segmentation of a movie in PPLive's VoD system

Segment	Designed for	Size
movie	entire video	> 100MB
chunk	unit for storage and advertisement	2MB
piece	unit for playback	16KB
sub-piece	unit for transmission	1KB

Table 1: Different units of a movie

7

Replication Strategy

- Goal
  - To make the chunks as available to users as possible
- Considerations
  - Whether to allow multiple movies be cached
    - Multiple movie cache (MVC) – flexible; **PPLive uses MVC**
    - Single movie cache (SVC) – simple
  - Whether to pre-fetch or not
    - Improves performance
    - Unnecessarily wastes uplink bandwidth
    - **PPLive chooses not to pre-fetch**

8

Replication Strategy(Cont.)

- Remove chunks or movies when the disc cache is full
  - **PPLive marks entire movie for removal**
- Which chunk/movie to remove
  - Least recently used (LRU) –**Original choice of PPLive**
  - Least frequently used (LFU)
  - Weighted LRU—each movie is assigned a weight based on factors
    - How complete the movie is already cached locally?
    - How needed a copy of movie is ATD (Available To Demand)
      - ATD = c/n
    - It improves the server loading from 19% down to a range of 11% to 7% compared with LRU.

9

Content Discovery

- Goal
  - To discover the content a peer needs with the minimum overhead
- PPLive uses
  - Trackers
    - Used to keep track of which peer has what movie(s)
  - Gossip method
    - Used to discover which peers have the chunks needed
  - DHT
    - Originally used to assign movies to trackers for load balancing
    - Later, also implemented by peers to provide a non-deterministic path to trackers which are possibly blocked by ISPs

10

Piece Selection

- Which piece to download first
  - Sequential
    - Select the piece closest to the one needed
  - Rarest first
    - Select the rarest piece
  - Anchor-based
    - Select the closest anchor point to the missing piece
- **PPLive gives priority to sequential first and then rarest-first**
  - Anchor-based is not necessary
    - Users do not jump around much, only 1.8 times/movie observed
    - The initial buffering time is acceptable

11

Transmission Strategy

- Goals
  - Maximize downloading rate
  - Minimize the overheads
- Strategies—a peer requests
  - the content from a neighbor at a time
  - the same content from multiple neighbors simultaneously
  - different contents from multiple neighbors simultaneously; **PPLive uses this scheme**
    - E.g., playback rate = 500Kbps, 8~20 neighbors is the best
    - E.g., playback rate = 1Mbps, 16~32 neighbors is the best
    - The content server can always be used to supplement data need, when peers cannot supply sufficient downloading rate

12

Other Design Issues

- NAT
  - Discovering different types of NAT boxes
    - *Full Cone* NAT, *Symmetric* NAT, *Port- restricted* NAT…
  - About 60%-80% of peers are found to be behind NAT
- Firewall
  - Proper upload rate and request rate
- Content authentication
  - Chunk level authentication
  - A weaker form of piece level authentication

13

Performance Metrics

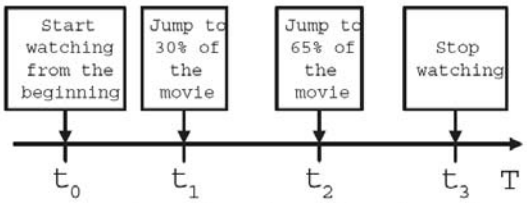
- User behavior
  - User arrival patterns
  - How long they stay to watch a movie
  - How they jump from one position to another in a movie
- External performance metrics
  - User satisfaction
  - Server load
- Health of replication
  - Measures how well a P2P-VoD system is replicating a content

14

User Behavior-MVR (Movie Viewing Record)

User ID	Movie ID	Start time	End time	Start pos.

MVR format



	UID	MID	ST	ET	SP
MVR1:	U1	M1	t_0	t_1	0%
MVR2:	U1	M1	t_1	t_2	30%
MVR3:	U1	M1	t_2	t_3	65%

Example to show how MVRs are generated

15

User Satisfaction

- Fluency
  - Fraction of time a user spends watching a movie out of the total viewing time (waiting and watching time for that movie)
  - Fluency *F(m,i)* for a movie *m* and user *i*

$$F(m,i) = \frac{\sum_{r \in R(m,i)} (r(ET) - r(ST) - r(BT))}{\sum_{r \in R(m,i)} (r(ET) - r(ST))}. \qquad (1)$$

R(*m*, *i*) : the set of all MVRs for a given movie *m* and user *i*

n(*m*, *i*) : the number of MVRs in *R*(*m*, *i*)

r : one of the MVRs in *R*(*m*, *i*)

BT : Buffering Time, ST : Starting Time, ET : Ending Time, and SP : Starting Position

16

User Satisfaction (Cont.)

- User satisfaction index
  - Considers the quality of the delivery of the content

$$S(m,i) = \sum_{k=1}^{n(m,i)} W_k r_k(Q). \qquad (3)$$

r(Q) : a grade for the average viewing quality for an MVR *r*

$$W_k = \frac{(r_k(ET) - r_k(ST) - r_k(BT))}{\sum_{r \in R(m,i)} (r(ET) - r(ST))}$$

  - In reality, it is not possible to get explicit user feedbacks for each MVR; PPLive simply uses fluency as user satisfaction

17

Health of Replication

- Health index : to reflect the effectiveness of the content replication strategy of a P2P-VoD system.
- The health index (for replication) can be defined at 3 levels:
  - Movie level
    - The number of active peers holding part of the movie—collected by the tracker
  - Weighted movie level
    - Considers the fraction of chunks a peer has in computing the index
  - Chunk bitmap level
    - The number of copies of each chunk stored by a peer
    - Used to compute other statistics, such as average number of chunks

18

Measurement

- All data were collected from 12/ 23/2007 to 12/29/2007
- Log server : collect various sorts of measurement data from peers.
- Tracker : aggregate the collected information give it to the log server
- Peer : collect data and do some amount of aggregation, filtering and pre-computation before passing them to the log server
- To determine the most popular movie, only MVRs starting from zero are counted—among the 3 typical movies

19

Statistics on video objects

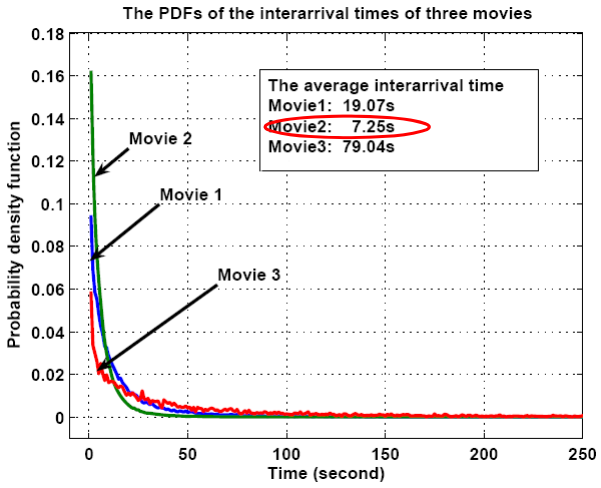
- Overall statistics of the 3 typical movies

Movie Index:	Movie 1	Movie 2	Movie 3
Total Length (in sec)	5100s	2820s	6600s
No. of Chunks	121	67	151
Total No. of MVRs	56157	322311	15094
Total No. of MVRs with Start Position = 0 (or # of unique viewers)	35160	95005	8423
Ave. # of Jump	1.6	3.4	1.8
Ave. viewing Duration for a MVR	829.8s	147.6s	620.2s
Normalized viewing Duration (normalized by the movie duration)	16.3%	5.2%	9.4%

Table 3: Overall statistics of the three typical movies.

20

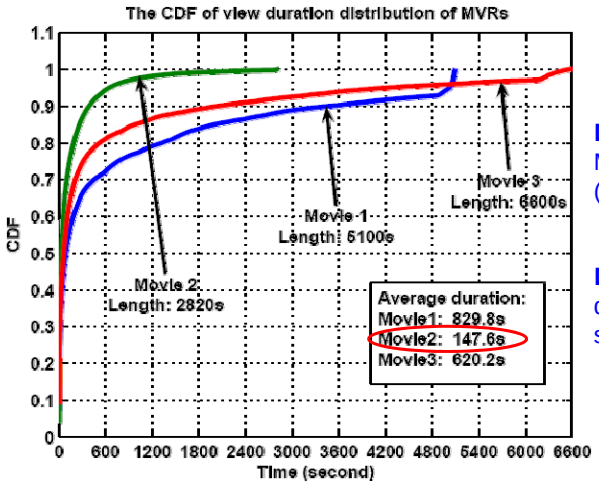
Statistics on user behavior (1) :
Interarrival time distribution of viewers



Interarrival times of viewers : the differences of the ST fields between two consecutive MVRs which start at zero position

21

Statistics on user behavior (2) :
View duration distribution



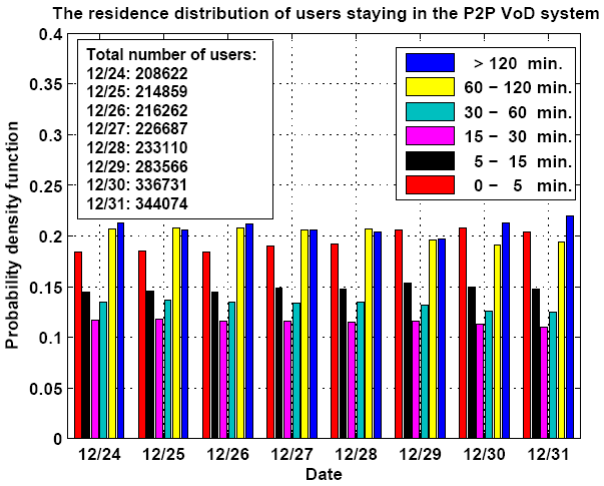
- Very high percentage of MVRs are of short duration (less than 10 minutes).
- Shorter view duraion may due to users actions like quick scan for an interesting movie

CDF: cumulative distribution function

22

Statistics on user behavior (3) :

Residence distribution of users

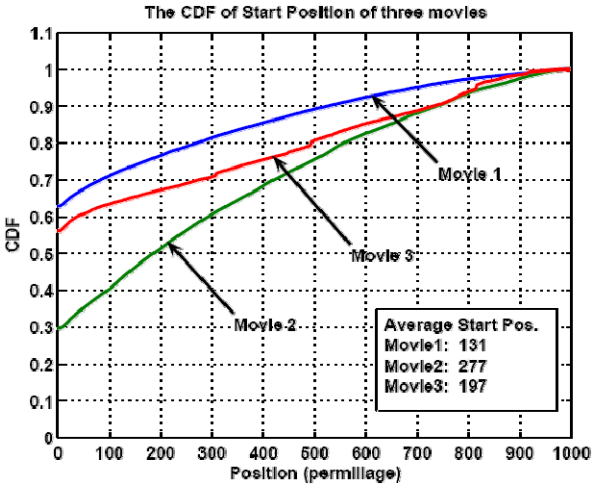


There is a high fraction of peers (over 65%) which stays in the P2P-VoD system for over 15 minutes, and these peers provide upload services to the community.

23

Statistics on user behavior (4):

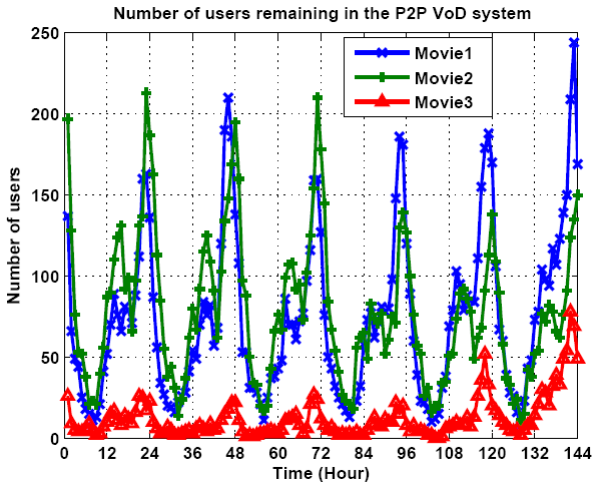
Start position distribution



- Users who watch Movie 2 are more likely to jump to some other positions than users who watch Movie 1 and 3
- The landing points of various jump operations is uniformly distributed, which implies that one can uniformly space the anchor points

24

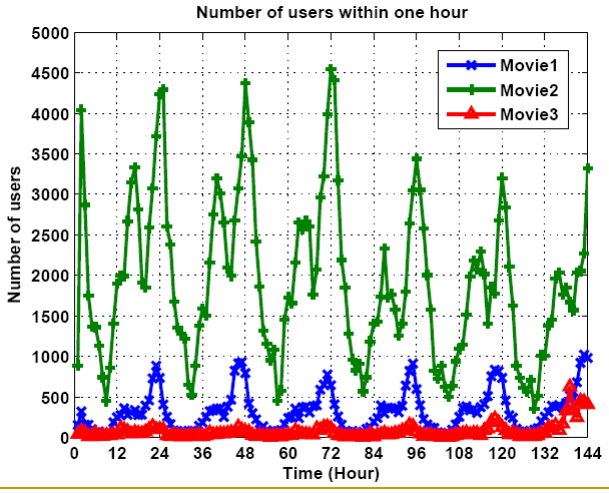
Statistics on user behavior (5):
Number of viewing actions



- The total number of viewing activities (or MVRs) at *each* sampling time point
- There are two daily peaks, which occur at around 2:00 P.M. and 11:00 P.M.

25

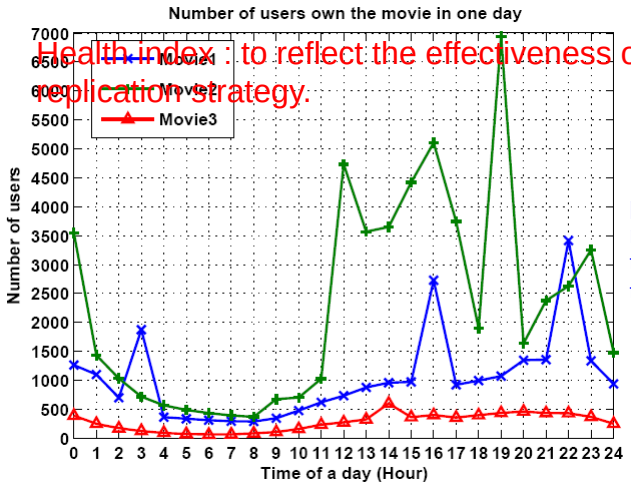
Statistics on user behavior (5):
Number of viewing actions(Cont.)



- The total number of viewing activities (or MVRs) within one hour
- There are two daily peaks, which occur at around 2:00 P.M. and 11:00 P.M

26

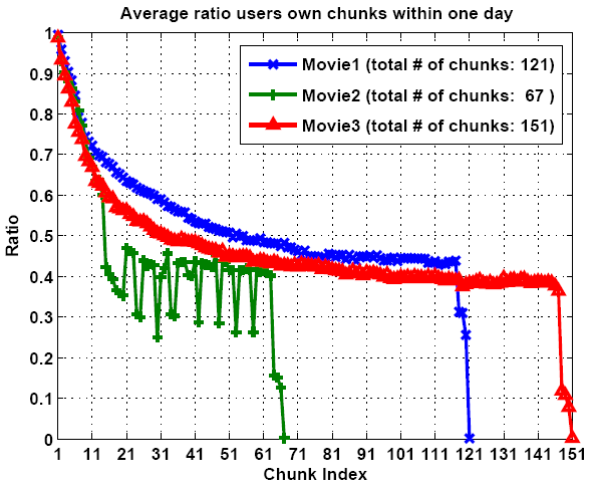
Health index of Movies (1)



Health index : to reflect the effectiveness of the content replication strategy.

- Movie 2 is the most popular movie, the number of users owning the movie is higher than that of movie 1 and 3

Health index of Movies (2)



unks

- The health index for "early" chunks is very good.
- Many peers may browse through the beginning of a movie.
- The health index is still acceptable since at least 30% of the peers have those chunks.

$\frac{\text{at time } t}{\text{time } t}$

y of chunk i in

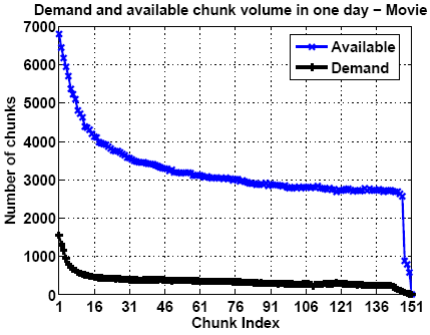
Health index of Movies (3)

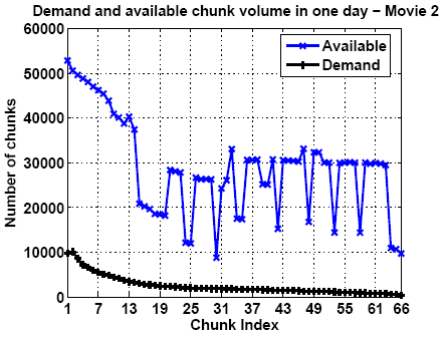
(a) The health index for these 3 movies are very good

(b) Movie 2's large fluctuation of chunk availability is due to users' high interactivity

(c) Users tend to skip the last chunk of the movie

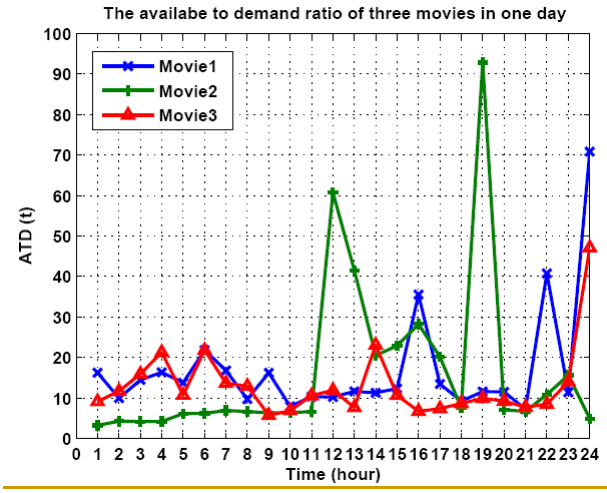
Chunk availability and chunk demand





29

Health index of Movies (4):
ATD (Available To Demand) ratios



$$\frac{\text{nk } i \text{ at } t}{\text{unk } i \text{ at } t}$$

- To provide good scalability and quality viewing, *ATDi*(*t*) has to be greater than 1.
- *ATDi*(*t*) ≥ 3 for all time *t* in this figure
- There are two peaks for Movie 2 at 12:00 or 19:00.

30

User Satisfaction Index (1)

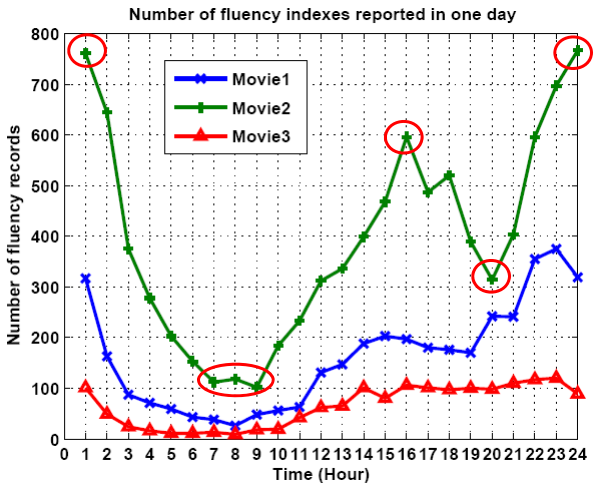
- User satisfaction index is used to measure the quality of viewing as experienced by users
  - A low user satisfaction index implies that peers are unhappy and these peers may choose to leave the system

$$F(m,i) = \frac{\sum_{r \in R(m,i)} (r(ET) - r(ST) - r(BT))}{\sum_{r \in R(m,i)} (r(ET) - r(ST))}. \quad (1)$$

- Generating fluency index
  - The client software reports all MVRs and the fluency *F*(*m*, *i*) to the log server when-
    - The STOP button is pressed
    - Another movie is selected
    - The user turns off the P2P-VoD software

31

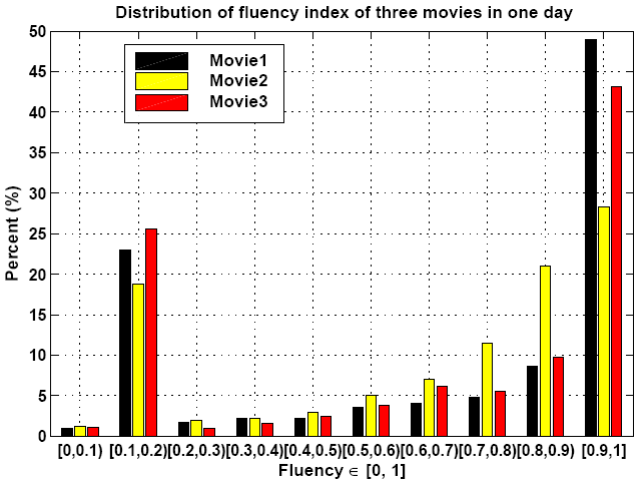
User Satisfaction Index (2)



The number of viewers in the system at different time points.

32

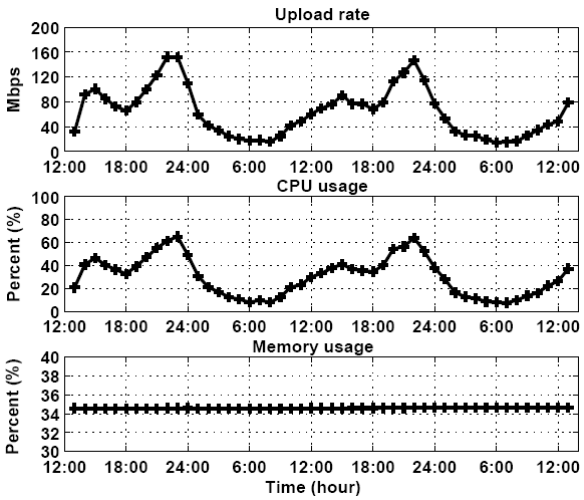
User Satisfaction Index (3):
The distribution of fluency index



- High percentage of fluency indexes whose values are greater than 0.7
- Around 20% of the fluency indexes are less than 0.2, because there is a high buffering time (which causes long start-up latency) for each viewing operation.

33

Server Load



- The server upload rate and CPU utilization are correlated with the number of users viewing the movies
- The server has implemented the memory-pool technique which makes the usage of the memory more efficient. (The memory usage is very stable)

34

Server Load(Cont.)

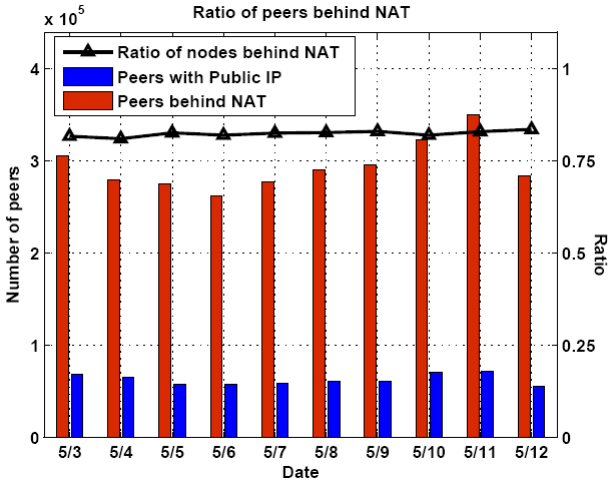
Upload (Kbps)	# of Peers (%)	Download (Kbps)	# of Peers (%)
[0, 200)	65616(35.94%)	[0, 360)	46504(25.47%)
[200, 360)	51040(27.96%)	[360, 600)	118256(64.78%)
[360, 600)	45368(24.86%)	[600, 1000)	14632(8.01%)
[600, 1000)	9392(5.14%)	[1000, 2000)	3040(1.67%)
> 1000	11128(6.10%)	> 2000	112(0.07%)
Total	182544	Total	182544

Table 4: Distribution of average upload and download rate in one-day measurement period.

- Measure on May 12, 2008.
- The average rate of a peer downloading from the server is 32Kbps and 352Kbps from the neighbor peers.
- The average upload rate of a peer is about 368Kbps.
- The average server loading during this one-day measurement period is about 8.3%.

35

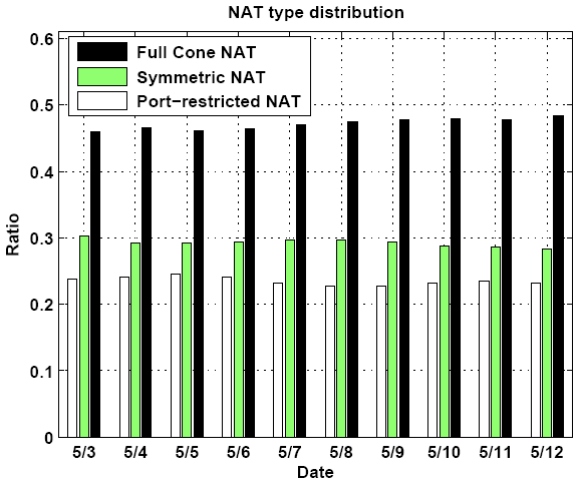
NAT Related Statistics



- The ratio of peers behind NAT boxes remains stable, around 80%.

36

NAT Related Statistics(Cont.)



- Full Cone NAT has the largest proportion(47%)
- Symmetric NAT is the second (30%)
- the Port-restricted NAT is the third (23%)
- There is no Restricted Cone NAT

37

Conclusions

- A general architecture and important building blocks of realizing a P2P-VoD system are presented
  - Performing dynamic movie replication and scheduling
  - Selection of proper transmission strategy
  - Measuring User satisfaction level
- This work is the first to conduct an in-depth study on practical design and measurement issues deployed by a real-world P2P-VoD system—PPLive
- The data is measured and collected from PPLive with totally 2.2 million independent users

38

Thank You!