PSIRP
PUBLISH-SUBSCRIBE
INTERNET ROUTING
PARADIGM

Peer Assisted Content Distribution over Router Assisted Overlay Multicast

Euro-NF Future Internet Architecture Workshop

George Xylomenos

PSIRP
PUBLISH-SUBSCRIBE
INTERNET ROUTING
PARADIGM

Outline

- Context
- Motivation
- The BitTorrent application
- The BitTorrent model
- Applying multicast to BitTorrent
- Multicast incentives
- Router assisted overlay multicast

PSIRP
PUBLISH-SUBSCRIBE
INTERNET ROUTING
PARADIGM

Context

- The ICT PSIRP Project
  - The Internet mostly disseminates data
  - Publish-Subscribe Internet Routing Paradigm
  - Clean slate approach to Future Internet
  - Pub-Sub at application and network levels
- Why Multicast?
  - The PSIRP architecture is not yet complete
  - Multicast data delivery seems to be set in stone
  - Will it replace peer assisted content distribution?

PSIRP
PUBLISH-SUBSCRIBE
INTERNET ROUTING
PARADIGM

Motivation

- Why BitTorrent?
  - Hugely popular content dissemination application
  - Not just a substitute for native multicast!
    - Asynchronous distribution of very large files
    - No need for sender/receiver rendez vous in time
- BitTorrent over Multicast
  - Exploit multicast as much as possible
  - Use overlay multicast for the time being
  - Maybe learn some things for PSIRP on the way

PSIRP
PUBLISH-SUBSCRIBE
INTERNET ROUTING
PARADIGM

The BitTorrent application

- Preparation for file exchange
  - Organize files as a sequence of bytes
  - Logically split the sequence into equal size pieces
  - Calculate the checksum of each piece
  - Locate a server to host the exchange (tracker)
  - Create a metafile: checksums, piece size,tracker
- Client initialization
  - Connect to the indicated tracker
  - Ask for a list of participating hosts (swarm)

PSIRP
PUBLISH-SUBSCRIBE
INTERNET ROUTING
PARADIGM

The BitTorrent application

- Client operation
  - Maintain a bitmap of locally available pieces
    - Shows what we have and what we miss
  - Semi-randomly contact peers
  - Select peers with which to exchange pieces
    - Must have useful data (check their bitmaps)
    - Should offer good download speeds
  - Piece exchange proceeds in a tit-for-tat fashion
    - Occasionally give out pieces for free to help new peers
    - Punish (blacklist) misbehaving peers

PSIRP
PUBLISH-SUBSCRIBE
INTERNET ROUTING
PARADIGM

The BitTorrent model

- Key decision: the exchange is based on pieces
  - Everything else follows from that
- The file exchange is asynchronous
  - A client can join and leave the swarm at will
- No trusted third parties
  - Each piece can be independently verified
  - Peers serving bad or no pieces are punished
- Choose your peer for yourself
  - The criteria are up to the implementation

PSIRP
PUBLISH-SUBSCRIBE
INTERNET ROUTING
PARADIGM

The BitTorrent model

- The tracker is a bottleneck with many peers
  - Only a limited number of peers is returned
- The exchange may be inefficient
  - Nearby peers may be downloading the same pieces
- Peer selection is very expensive
  - A peer may not be available (left the swarm)
  - A peer may be unwilling (too many peers already)
  - A peer may not be useful (no pieces to exchange)
  - A peer may not be good (low download speed)

PSIRP
PUBLISH-SUBSCRIBE
INTERNET ROUTING
PARADIGM

Applying multicast to BitTorrent

- Retain the key decision of BitTorrent: pieces!
  - Distribute each piece over a separate group
    - Use the piece checksum or name as a group identifier
  - A receiver joins the groups for its missing pieces
  - A sender asks the RV points before sending
    - Have any receivers joined the group?
  - The RV point should delay consecutive replies
    - Avoid multiple transmissions of the same piece
  - Send your bitmap along with each piece
    - The receivers automatically learn what you need

PSIRP
PUBLISH-SUBSCRIBE
INTERNET ROUTING
PARADIGM

Multicast incentives

- BitTorrent incentive model
  - Each client talks with a specific peer
  - The exchange proceeds on a tit-for-tat basis
- Why multicast is not the same?
  - The sender-receiver relationship is one to many
    - The sender may not even know the receivers
  - The sender transmits its bitmap along with the data
  - How can we punish receivers that do not send back?
- Solution: partially encrypt each piece
  - Exchange keys after the pieces but in unicast mode

PSIRP
PUBLISH-SUBSCRIBE
INTERNET ROUTING
PARADIGM

Multicast incentives

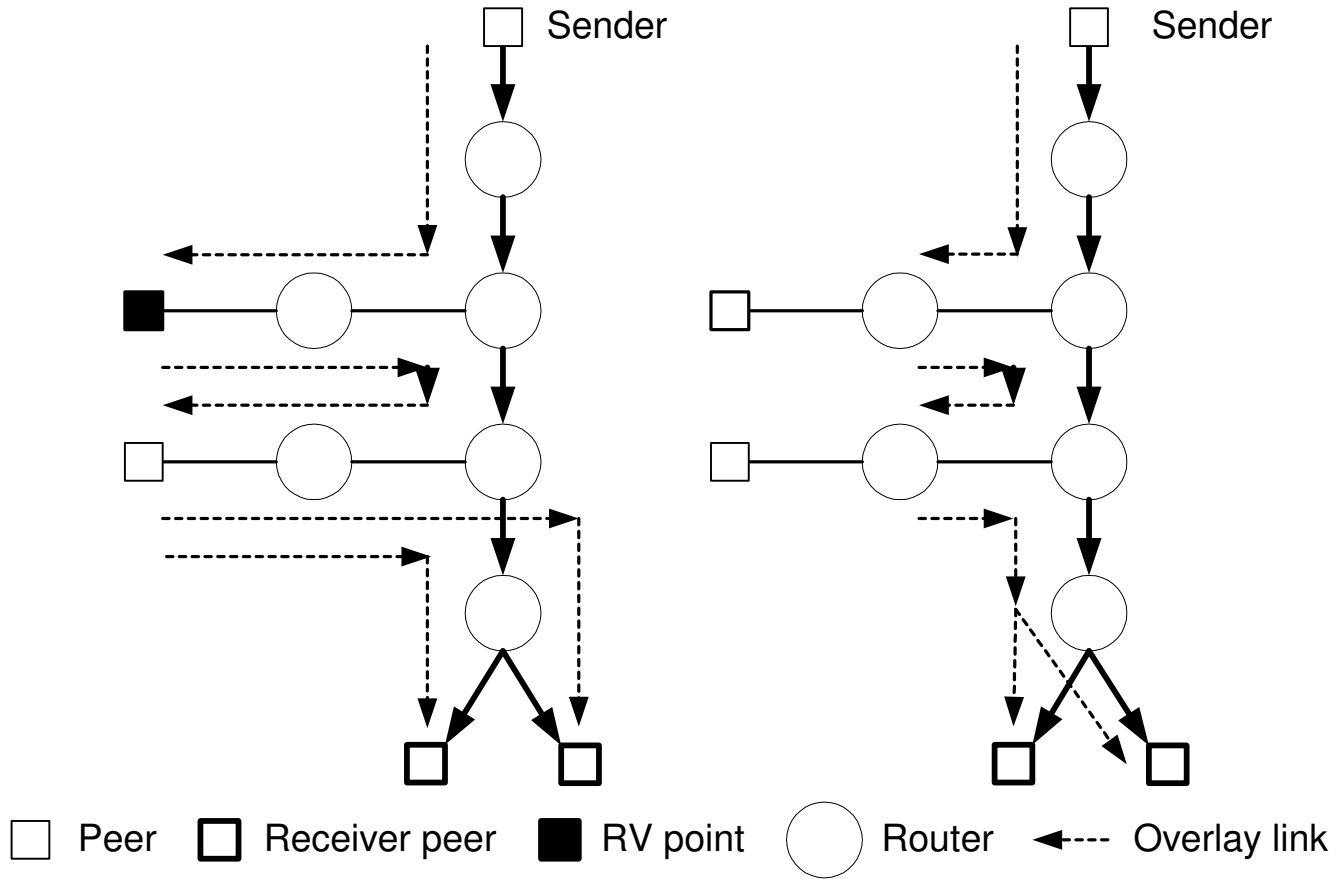
- Partial piece encryption
  - The sender transmits a piece and waits
    - Encrypt n+k bits for an n bit hash
  - Each receiver “returns” an encrypted piece
    - We know what the sender needs from its bitmap
    - Cannot guess the n+k bits and verify with the hash
  - The sender transmits the key to obliging receivers
  - The receivers return their keys to the sender
  - Clients that do not return keys are blacklisted
    - Same for bad keys or bad pieces

PSIRP
PUBLISH-SUBSCRIBE
INTERNET ROUTING
PARADIGM

Router assisted overlay multicast

- What kind of multicast are we talking about?
  - IP multicast is unlikely to get going any time soon
  - End System Multicast is not scalable
- DHT based multicast (e.g. Scribe/Pastry)
  - The group name is mapped to a node
    - This node is the RV point for the group
  - Receiver join: send a message to the RV point
    - Reverse path forwarding state is established
  - Senders simply send their data to the RV point
    - The path may be quite suboptimal

Router assisted overlay multicast



PSIRP
PUBLISH-SUBSCRIBE
INTERNET ROUTING
PARADIGM

Router assisted overlay multicast

- Why router assisted overlay multicast?
  - Scribe relies on end hosts only
  - Data must cross many access links twice
    - The uplink is normally the bottleneck
  - Ask your access router to be your proxy
    - Data only crosses the downlink for receivers
  - Multicast trees are shortened
    - Path stretch: 3 for regular Scribe, 1.8 for our approach
  - Incentives for access routers
    - Independent performance improvement for local hosts

PSIRP
PUBLISH-SUBSCRIBE
INTERNET ROUTING
PARADIGM

Conclusion

- Multicast is not directly applicable to BitTorrent
  - BitTorrent is asynchronous, unlike IP multicast
- Use a separate group per piece
  - Reduced peer searching overhead
  - More economical data distribution
- Issues to be resolved
  - Incentive model: multicast tit-for-tat
  - Sender policy: who to query, when to send
  - Need for an efficient overlay multicast scheme