Overlay Networks for Multimedia Contents Distribution

Vittorio Palmisano
vpalmisano@gmail.com

26 gennaio 2007

Outline

1. Mesh-based Multicast Networks
   - Narada (Carnegie Mellon University, 2000)
2. Tree-based Multicast Networks
   - Overcast (Cisco, 2000)
   - P2Cast (University of Massachusetts Amherst, 2003)
   - Peercast (Stanford, 2002)
   - CoopNet (Microsoft Research, CMU, 2002)
3. DHT-based Multicast Networks
   - Bayeux (Berkeley, 2001)
   - Scribe (Microsoft Research, 2002)

Outline

1. Mesh-based Multicast Networks
   - Narada (Carnegie Mellon University, 2000)
2. Tree-based Multicast Networks
   - Overcast (Cisco, 2000)
   - P2Cast (University of Massachusetts Amherst, 2003)
   - Peercast (Stanford, 2002)
   - CoopNet (Microsoft Research, CMU, 2002)
3. DHT-based Multicast Networks
   - Bayeux (Berkeley, 2001)
   - Scribe (Microsoft Research, 2002)

Narada[1]

- Suitable for live streaming (low latency)
- Multi-source multicast, but...
- ...data delivery from 1 source at time
- Mesh topology:
  - spanning trees routed at sources in order to support Multi-source applications
  - trees optimised for each source, rather than a specific group
- It uses a variant of the standard distance vector routing algorithm (DVMRP)
  - utility function: $quality \Leftrightarrow \{delay, bandwidth\}$

Narada

Group management:

- every member maintains a list of all other members in the group (medium sized groups)
- every member periodically sends a message ($< k_i, s_{k_i} >$) to his neighbors with a list of group members addresses (k_i) and a sequence number (s_{k_i}), and it receives an ack
- Member join: bootstrap from a known node (out-of-band)
- Member leave:
  - leave notification
  - timeouts on message/ack exchange
  - partition recovery

Narada

Mesh performance:

- links between members are added or removed depending on *utility* (e.g. latency, bandwidth)
- addition of links: every member periodically probes some random non-neighbor member and evaluates the *utility* of adding a link to this member
- dropping of links: every member periodically calculates the *consensus cost* to a neighbor member and drops links with cost below a threshold
- stability of adding/removing links is assured by some heuristics

Narada

Conferencing applications with Narada:

- It uses **TFRC** as transport protocol
- Routing is a function of both **bandwidth** and **latency** (*shortest widest path algorithm*):
  - first, find the highest bandwidth path
  - if there are multiple paths with the same bandwidth, pick the path with the lowest latency
- Latency is estimated with raw pings between neighbors every 200 ms
- Bandwidth is estimated by monitoring data flows

Considerations

Narada's issues:

- Messages overhead in large sized groups
- Nodes informations propagate slowly $\Rightarrow$ poor transient performances
- Organize members into hierarchies?

Outline

1. Mesh-based Multicast Networks
   - Narada (Carnegie Mellon University, 2000)
2. Tree-based Multicast Networks
   - Overcast (Cisco, 2000)
   - P2Cast (University of Massachusetts Amherst, 2003)
   - Peercast (Stanford, 2002)
   - CoopNet (Microsoft Research, CMU, 2002)
3. DHT-based Multicast Networks
   - Bayeux (Berkeley, 2001)
   - Scribe (Microsoft Research, 2002)

Overcast[2]

- Suitable for Video on Demand (VoD)
- Supports large scale multicast groups
- Proxy-based (HTTP)
- Single-source multicast:
  - simple
  - optimizated
  - enough to many applications (e.g. only one node is active at a time)
- Bandwidth optimizated (not latency)
- Nodes are dedicated machines

Overcast

Management:

- Overlay network is build in the form of a **distribution tree** with the root (source) node
- Root node manages the joining of new nodes
- New nodes are placed as far away from the root in order to save the root bandwidth
- Trees may have many levels (=**increasing latency**)
- Every nodes stores his children **full** topology
- 10 Kbytes transfers are used in order to evaluate bandwidth

Outline

1. Mesh-based Multicast Networks
   - Narada (Carnegie Mellon University, 2000)
2. Tree-based Multicast Networks
   - Overcast (Cisco, 2000)
   - P2Cast (University of Massachusetts Amherst, 2003)
   - Peercast (Stanford, 2002)
   - CoopNet (Microsoft Research, CMU, 2002)
3. DHT-based Multicast Networks
   - Bayeux (Berkeley, 2001)
   - Scribe (Microsoft Research, 2002)

P2Cast[3]

- Suitable for Video on Demand (VoD)
- Tree overlay network approach (as Overcast)
- Clients that arrive within a time T constitute a *session*
- In a same session, clients that arrives later obtain the current **and** the initial part of the video (**patch**) from the server or other clients
- Clients need a storage space for patches
- Shifted forwarding in case of node failures (=interval caching)

Outline

1. Mesh-based Multicast Networks
   - Narada (Carnegie Mellon University, 2000)
2. Tree-based Multicast Networks
   - Overcast (Cisco, 2000)
   - P2Cast (University of Massachusetts Amherst, 2003)
   - Peercast (Stanford, 2002)
   - CoopNet (Microsoft Research, CMU, 2002)
3. DHT-based Multicast Networks
   - Bayeux (Berkeley, 2001)
   - Scribe (Microsoft Research, 2002)

PeerCast[4]

- Designed for:
  - Short lifetime nodes
  - Hundreds of nodes
  - Unreliable transport UDP/RTP
- Tree overlay network approach (as Overcast)
- GPL code available (based on Gnutella)

Management:

- A peer only knows **only his local topology** (the parent and immediate children)
- `Redirect` primitive from one node to his children in case of failure
- Join policies:
  - a unsaturated node accepts requests
  - a saturated node redirects the requests to his children

Outline

1. Mesh-based Multicast Networks
   - Narada (Carnegie Mellon University, 2000)
2. Tree-based Multicast Networks
   - Overcast (Cisco, 2000)
   - P2Cast (University of Massachusetts Amherst, 2003)
   - Peercast (Stanford, 2002)
   - CoopNet (Microsoft Research, CMU, 2002)
3. DHT-based Multicast Networks
   - Bayeux (Berkeley, 2001)
   - Scribe (Microsoft Research, 2002)

CoopNet[5]

- Tree overlay network approach (as Overcast), but using **multiple distribution trees**
- A central server handles the requests. It stores the entire topology of the tree
- In the case of live streams, it employs **Multiple Description Coding** (**MDC**):
  - Audio/video signal encoded into separate streams (*descriptions*), **not** layers
  - More descriptions received $\Rightarrow$higher quality (redundancy)
  - e.g. M *descriptions* $\Rightarrow M \frac{frames}{s}$
- In the case of on-demand streaming, it uses *distributed streaming*:
  - a stream is divided into a number of *substreams*, each of which may be server by a different peer
  - each substream corresponds to a *description*
  - centralized server sends a list of peers who have the requested part of stream
- Centralized approach is more subject to failures

Considerations

- Tree-based routing can lead to many levels (=more latency)
- Centralized systems:
  - not fault-tolerant
  - scalability issues

Outline

1. Mesh-based Multicast Networks
   - Narada (Carnegie Mellon University, 2000)
2. Tree-based Multicast Networks
   - Overcast (Cisco, 2000)
   - P2Cast (University of Massachusetts Amherst, 2003)
   - Peercast (Stanford, 2002)
   - CoopNet (Microsoft Research, CMU, 2002)
3. DHT-based Multicast Networks
   - Bayeux (Berkeley, 2001)
   - Scribe (Microsoft Research, 2002)

Bayeux[6]

- It uses **Tapestry** DHT routing:
  - Nodes have a *nodeID* uniformly and randomly distributed using SHA-1 hashing
  - Each node has a multi-level neighbors map:

nodeID	**4**	**2**	**2**	**7**
1st	*	*	*	*
2nd	4	*	*	*
3th	4	2	*	*
4th	4	2	2	*

  - Provides multiple paths to every destination (*fault tolerant*)
  - Scalable routing overhead: $n.\ of\ hops \leq n.\ of\ digits$
- A **root node** stores all session members nodeIDs
- It uses dedicated servers as Tapestry nodes
- There are multiple root nodes in order to be fault-tolerant (*Tree Partitioning*)
- Delivery protocol: each node sends packets to the shortest outgoing link that shows best rate/latency. Nodes use small UDP packets in order to evaluate rate/latency

Outline

1. Mesh-based Multicast Networks
   - Narada (Carnegie Mellon University, 2000)
2. Tree-based Multicast Networks
   - Overcast (Cisco, 2000)
   - P2Cast (University of Massachusetts Amherst, 2003)
   - Peercast (Stanford, 2002)
   - CoopNet (Microsoft Research, CMU, 2002)
3. DHT-based Multicast Networks
   - Bayeux (Berkeley, 2001)
   - Scribe (Microsoft Research, 2002)

Scribe[7, 8]

- Based on **Pastry**'s DHT
- Each node may create a *group*, messages are multicasted within the same group
- Multicast trees are created joining each group member

Results:

- Allocation of nodes based on bandwidth constrains lead to:
  - large depth
  - non-DHT links
- Solution:
  - to assign ID based on nodes bandwidth
  - non-DHT links problems persist

Considerations

- Routing:
  - DHT routing reduces the number of hops from source to destinations
  - ...but QoS constrains force to have non-DHT links
  - Find hash keys based on QoS parameters (e.g. IP locations) ?
- Fragmentation:
  - In the case of file download, more fragments can be received and reordered
  - In the live stream case, fragmentation is impossible
  - Use multiple description coding?
  - Use cache fragmentation?

References

[1] Y. Chu, S. Rao, S. Seshan, and H. Zhang, "A case for end system multicast," *Selected Areas in Communications, IEEE Journal on*, vol. 20, no. 8, pp. 1456–1471, 2002.

[2] J. Jannotti, D. Gifford, K. Johnson, M. Kaashoek, and J. O'Toole, "Overcast: Reliable multicasting with an overlay network," *Proc. OSDI 2000*.

[3] Y. Guo, K. Suh, J. Kurose, and D. Towsley, "P2Cast: peer-to-peer patching scheme for VoD service," *Proceedings of the twelfth international conference on World Wide Web*, pp. 301–309, 2003.

[4] H. Deshpande, M. Bawa, and H. Garcia-Molina, "Streaming Live Media over a Peer-to-Peer Network," *Submitted for publication*, 2002.

[5] V. Padmanabhan, H. Wang, P. Chou, and K. Sripanidkulchai, "Distributing streaming media content using cooperative networking," *Proceedings of the 12th international workshop on Network and operating systems support for digital audio and video*, pp. 177–186, 2002.

[6] S. Zhuang, B. Zhao, A. Joseph, R. Katz, and J. Kubiatowicz, *Bayeux: an architecture for scalable and fault-tolerant wide-area data dissemination*. ACM Press New York, NY, USA, 2001.

[7] M. Castro, P. Druschel, A. Kermarrec, and A. Rowstron, "Scribe: a large-scale and decentralized application-level multicast infrastructure," *Selected Areas in Communications, IEEE Journal on*, vol. 20, no. 8, pp. 1489–1499, 2002.

[8] A. Bharambe, S. Rao, V. Padmanabhan, S. Seshan, and H. Zhang, "The Impact of Heterogeneous Bandwidth Constraints on DHT-Based Multicast," 2005.