Bimodal Multicast

K Birman et al, ACM TOCS 1999

Presented by Yixiao Nie

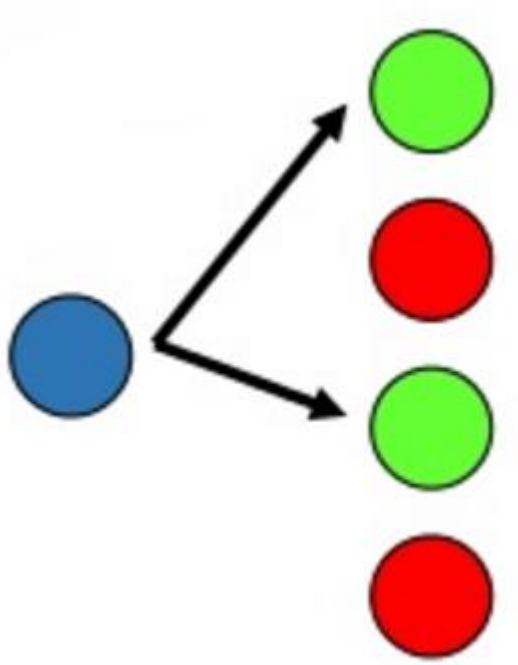
April 26, 2016

Introduction and Background

Multicast:

A method of sending information to a group of destination devices simultaneously

Different from broadcast and unicast.



Reliable Multicast Protocol:

A protocol providing a reliable multicast.

Traditional Reliable Multicast

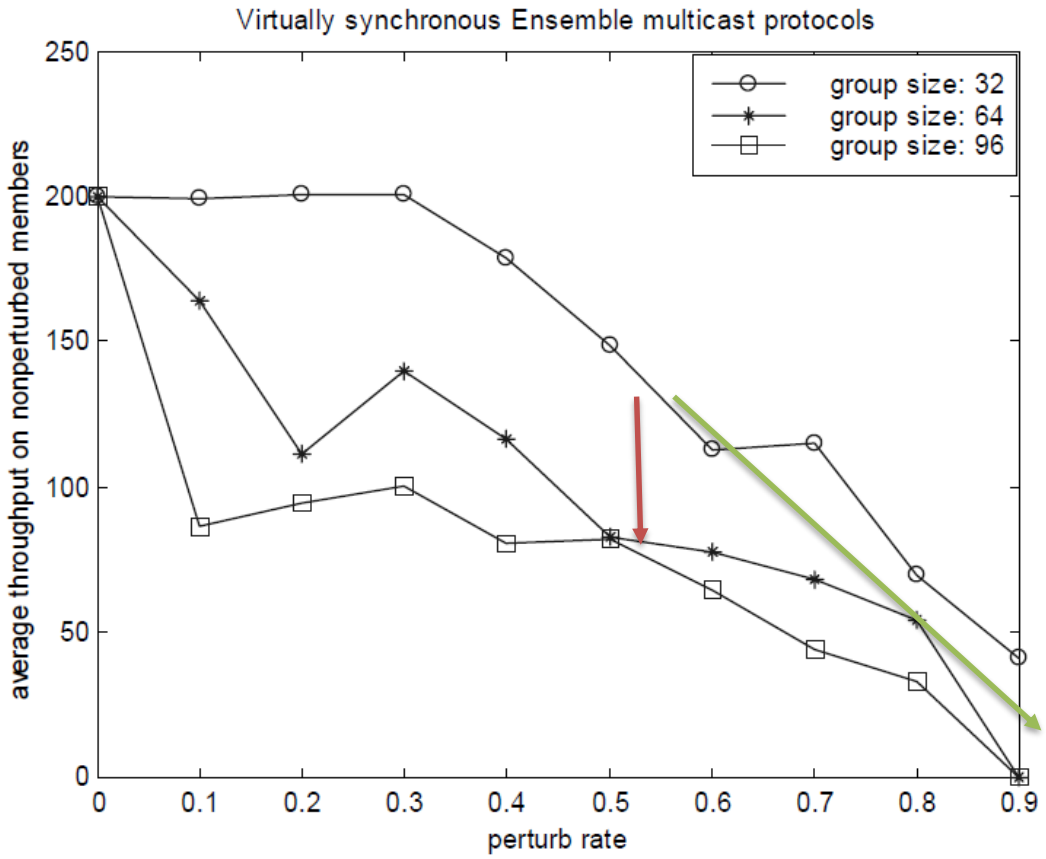
First class:

- "strong" reliability properties
- Atomicity
  - All or None
  - Security properties
  - Real-time guarantees

- Unstable or unpredictable performance under stress
- tolerates limited scalability

Traditional Reliable Multicast

- Evaluation on virtually synchronous multicast groups (traditional)
  - Group size: 32, 64, 96
  - One injects 7KB multicast messages at 200 message/second
  - Some member is “perturbed” to sleep for varying amounts of time



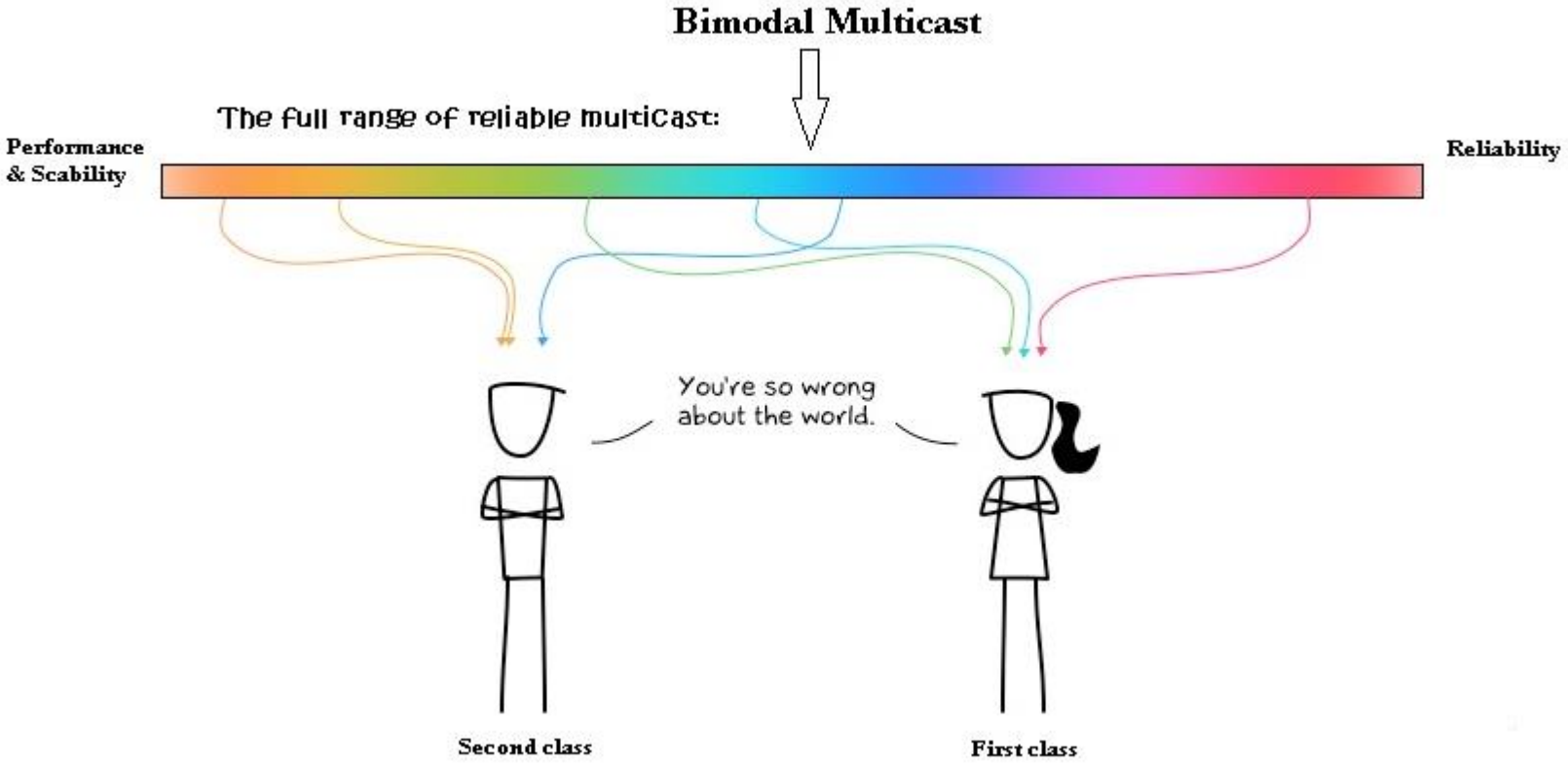
- Result:
  - Throughput drops as perturb rate increases
  - Throughput drops as group size increases

Traditional Reliable Multicast

Second class:

- "Best-effort" reliability
  - if a participating process discovers a failure, a reasonable effort is made to overcome it. But it may not always be possible to do so.
- Higher scalability

- No "end to end" reliability guarantee
- Impossible to reason about the system behavior when things go wrong



* The original picture is from http://thedoghousediaries.com/4564

Bimodal Multicast

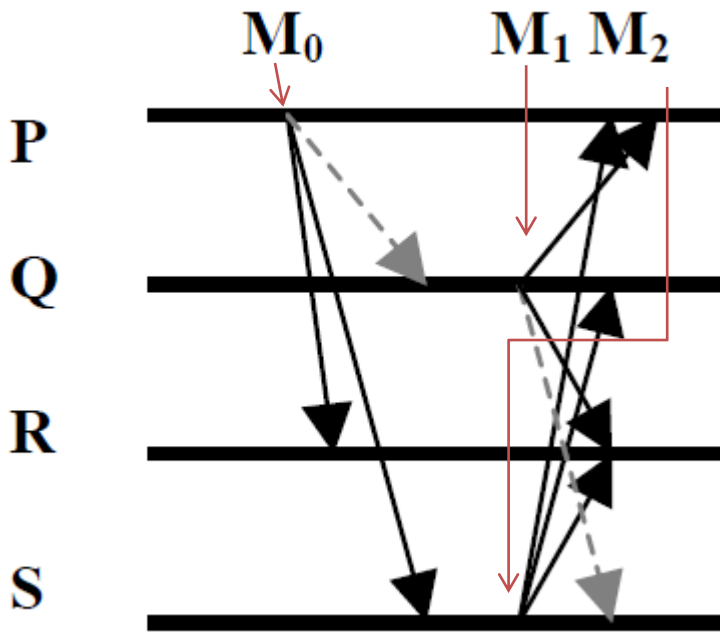
- Also called Probability Broadcast (pbcast), composed of two subprotocols.
- Main version was implemented within Cornell University's Ensemble system. (There is a second version developed called Spinglass)

- Atomicity
  - Almost all or almost none.
- Throughput stability
  - Expected variation in throughput is low in comparison to traditional multicast.
- Scalability
  - Costs are constant or grow slowly as a function of the network size.
- Ordering
  - Messages delivered in FIFO order.

Protocol Details

First stage: Optimistic Dissemination Protocol

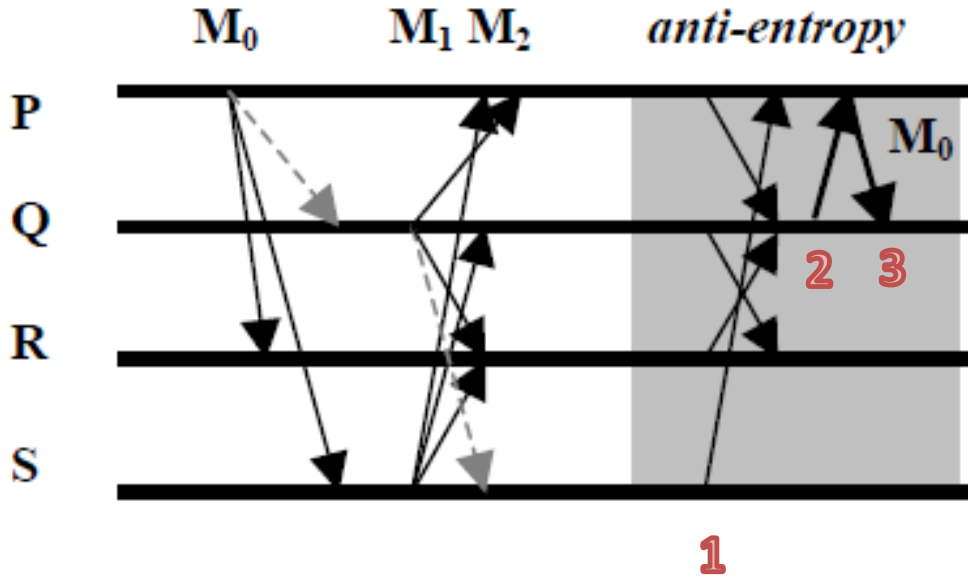
- An unreliable, hierarchical, and best-effort-attempt broadcast
  - Broadcast a message by using a randomly selected spanning tree.
  - Attach a tree identifier to the message. Upon receipt, those members deliver the message and then forward it using the tree identifier.



Protocol Details

Second stage: Two-Phase Anti-Entropy Protocol

- Detect and correct inconsistencies in a system by continuous gossiping.
  - Gossiping: Randomly choose other members to send a summary of their message histories.
  - Solicitation: Solicit copies of any messages they discover themselves to be lacking to converge toward identical histories.
  - Retransmission: Upon receipt, the member retransmits some of the messages.



Protocol Optimization

1. Soft-Failure Detection

   Retransmission requests are only serviced if they are received in the same round for which the original solicitation was sent.

2. Round Retransmission Limit

   The maximum amount of data (in bytes) that a process will retransmit in one round is also limited.

3. Cyclic Retransmissions

   Processes responding to retransmission requests cycle through their undelivered messages.

4. Most-Recent-First Retransmission

   Messages are retransmitted in the order of most recent first.

Protocol Optimization

5. Independent Numbering of Rounds

   The protocol allows each process to maintain its own round numbers and to run them asynchronously.

6. Random Graphs for Scalability

   Create hierarchy, develop a membership service, and separate WAN and LAN gossips.

7. Multicast for Some Retransmissions

   In certain situations, the protocol employs multicast to retransmit a message.

Integration with Ensemble

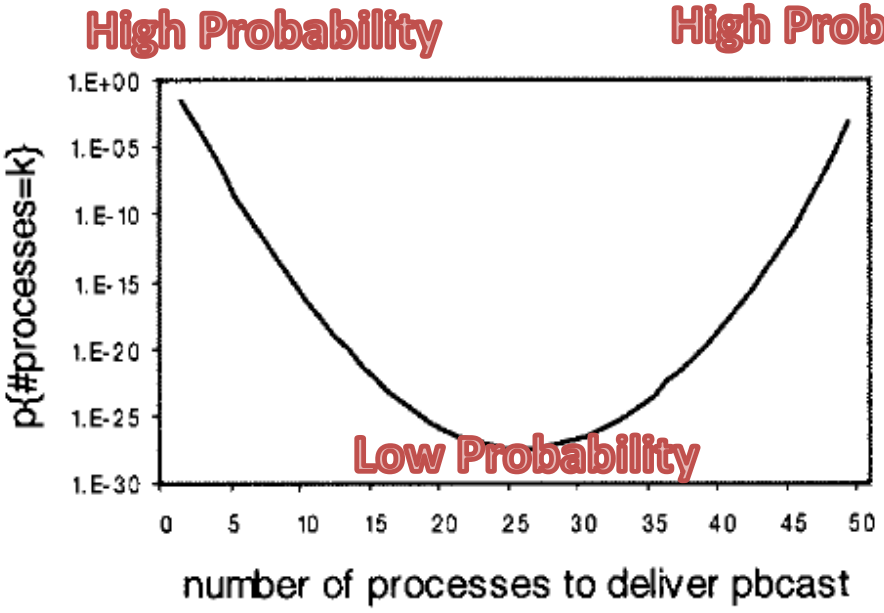
- Flow control
  - Combined pbcast with a form of flow control tied to the number of buffered messages active within the protocol itself.
  - For the experiments reported here, they employed application-level rate limitations.

- Recovery from Delivery Failures
  - In Spinglass, they were exploring the possibility of varying the amount of buffering used by each pbcast participant.
  - In Ensemble, a pbcast participant that falls behind uses the Ensemble state transfer as a recovery mechanism.

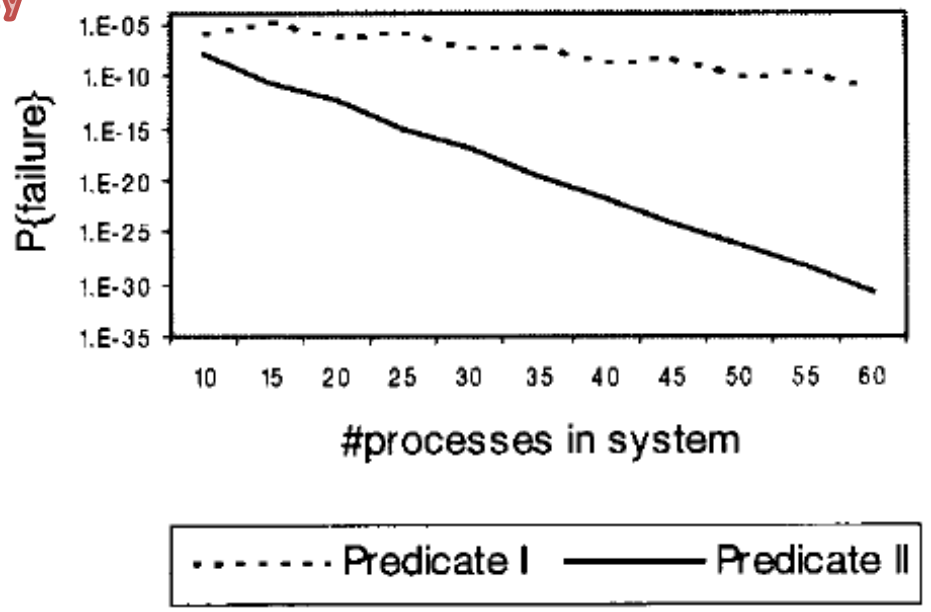
Graphic Results

Pbcast's bimodal delivery distribution

- Atomicity: almost all or almost none



Scalability of pbcast reliability



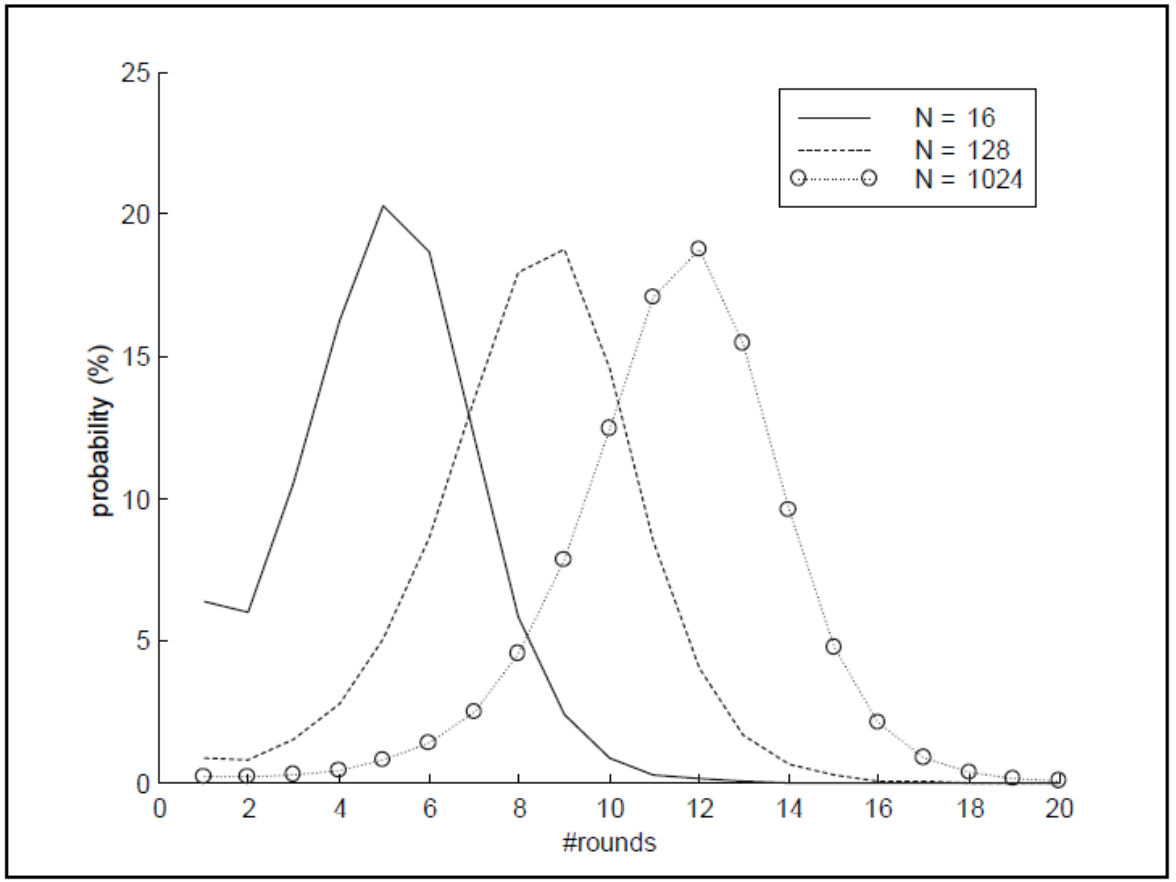
Failure: 10%<p<90%

Failure: p≈50%

Graphic Results

Probability and latency for receiving a phcast in a particular round

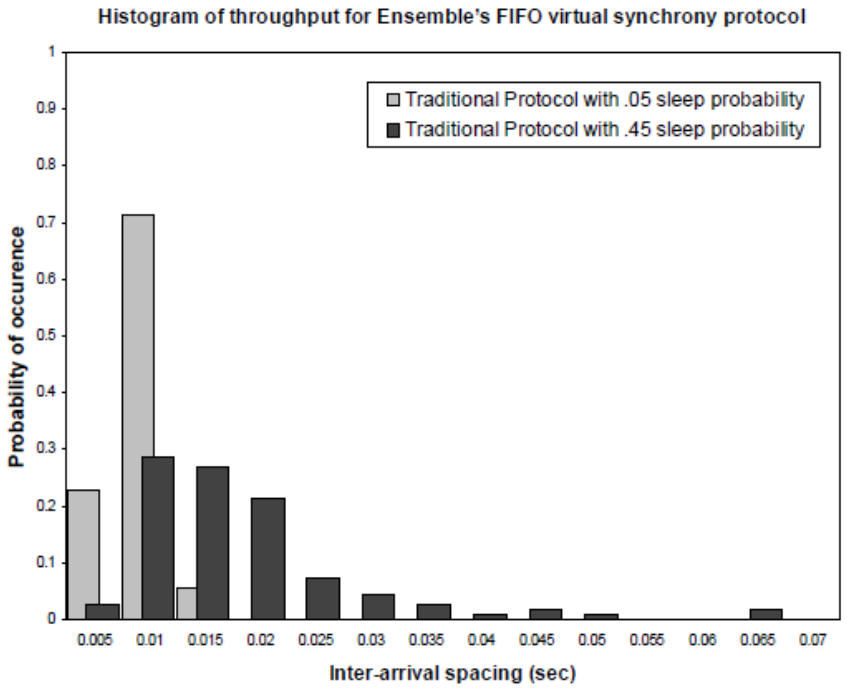
- Normal distribution
- Centered at log(n)

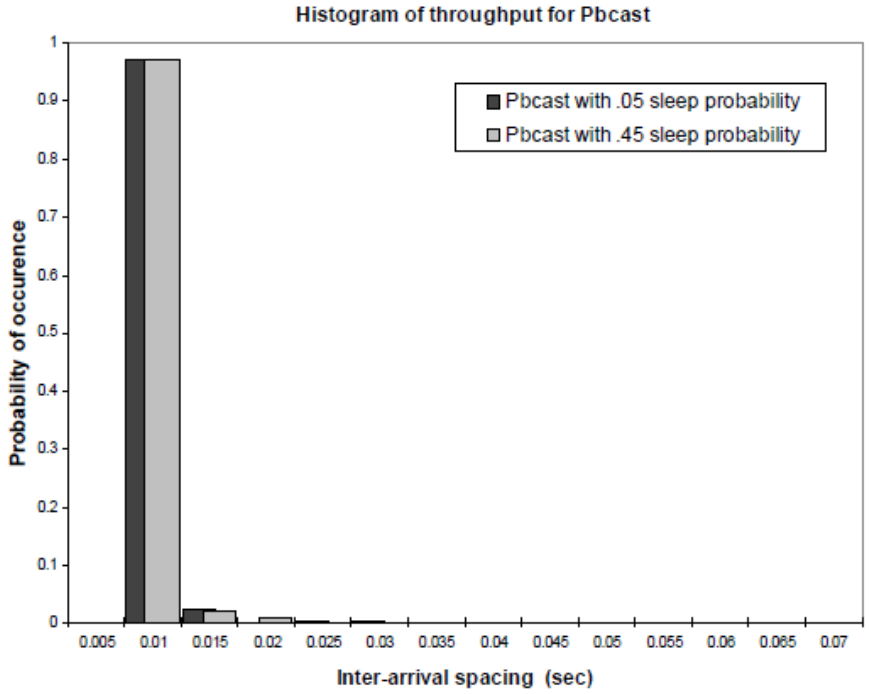


Graphic Results

Steady throughput under perturbation

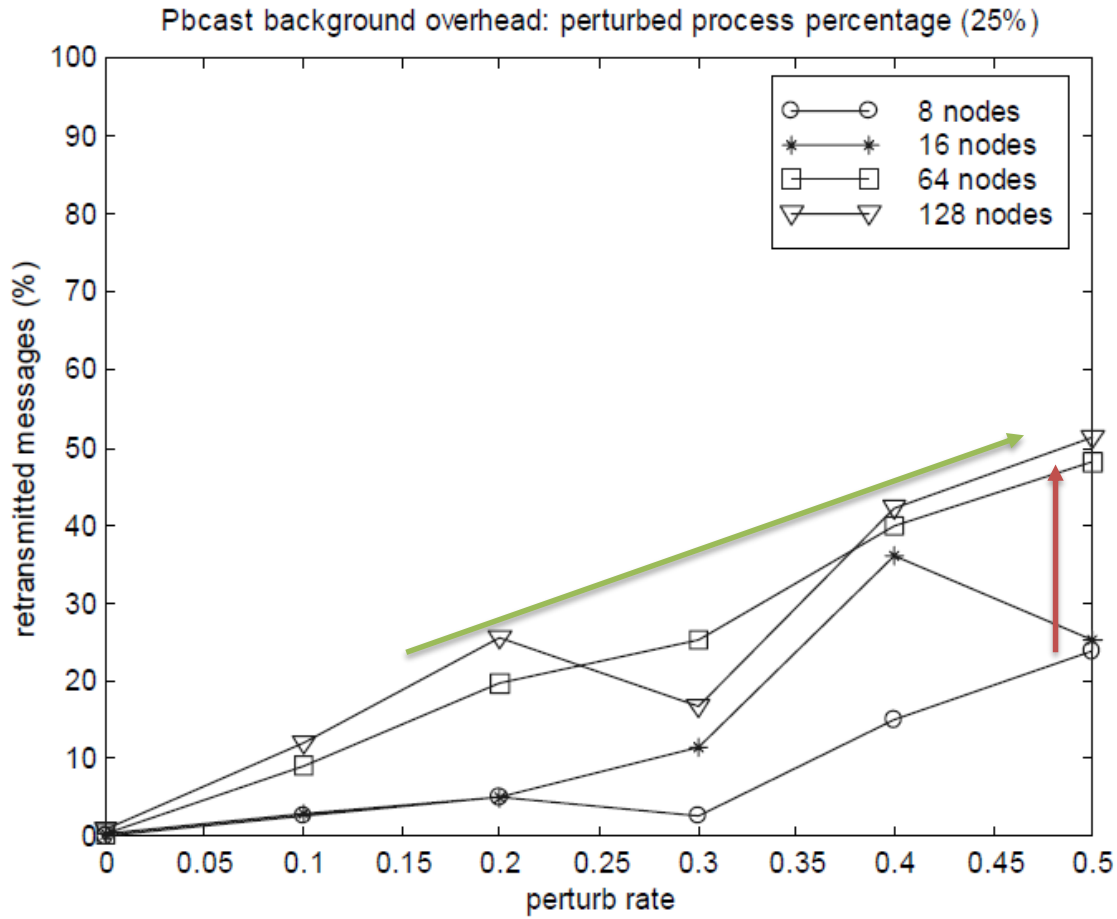
- Traditional (left) vs pbcast (right)





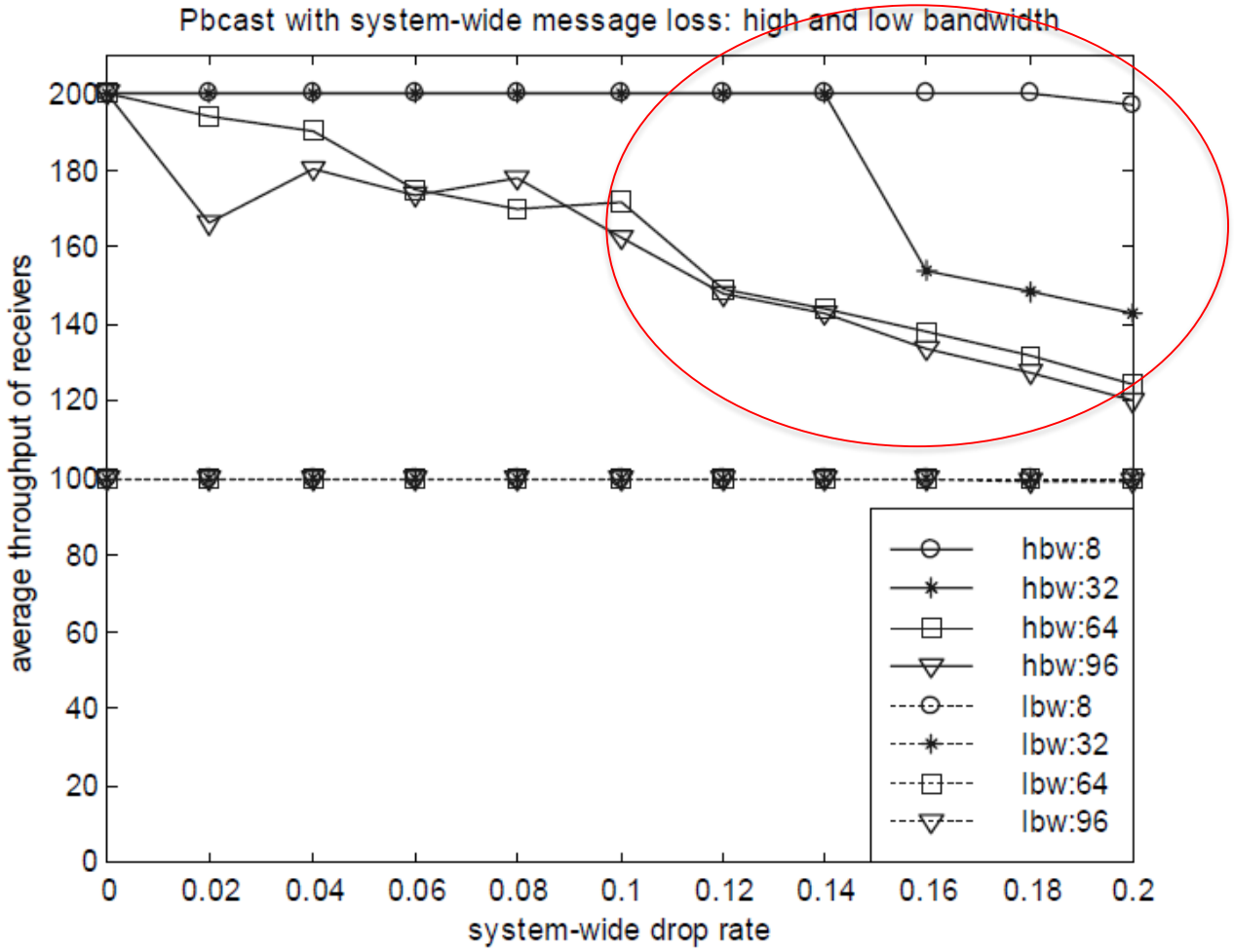
Graphic Results

Retransmission solicitations received at health process



Graphic Results

Pbcast with high message loss



Discussion

- At high data rates, reliability of pbcast drops as the network data loss rate increases
- If message injection rate is low, pbcast sends unnecessary gossip messages.
- If IP multicast is reliable, pbcast may waste a lost of messages.
- Management of the multicast dissemination routes for pbcast is a topic for future study.

Conclusion

- Bimodal Multicast is a protocol that falls between two traditional classes of reliable multicast.
  - Reliable by providing bimodal delivery guarantee
  - Provides remarkably stable delivery throughput
  - Scalable

- Potential applications:
  - Applications that send media, such as radio, television, or teleconferencing data over the internet.
  - In a stock market or equity-trading environment.
  - In an air traffic control setting.
  - In a health care setting.

Questions?

Thank you!