Networking Technologies and Applications

Rolland Vida
BME TMIT

November 24, 2016

PIM

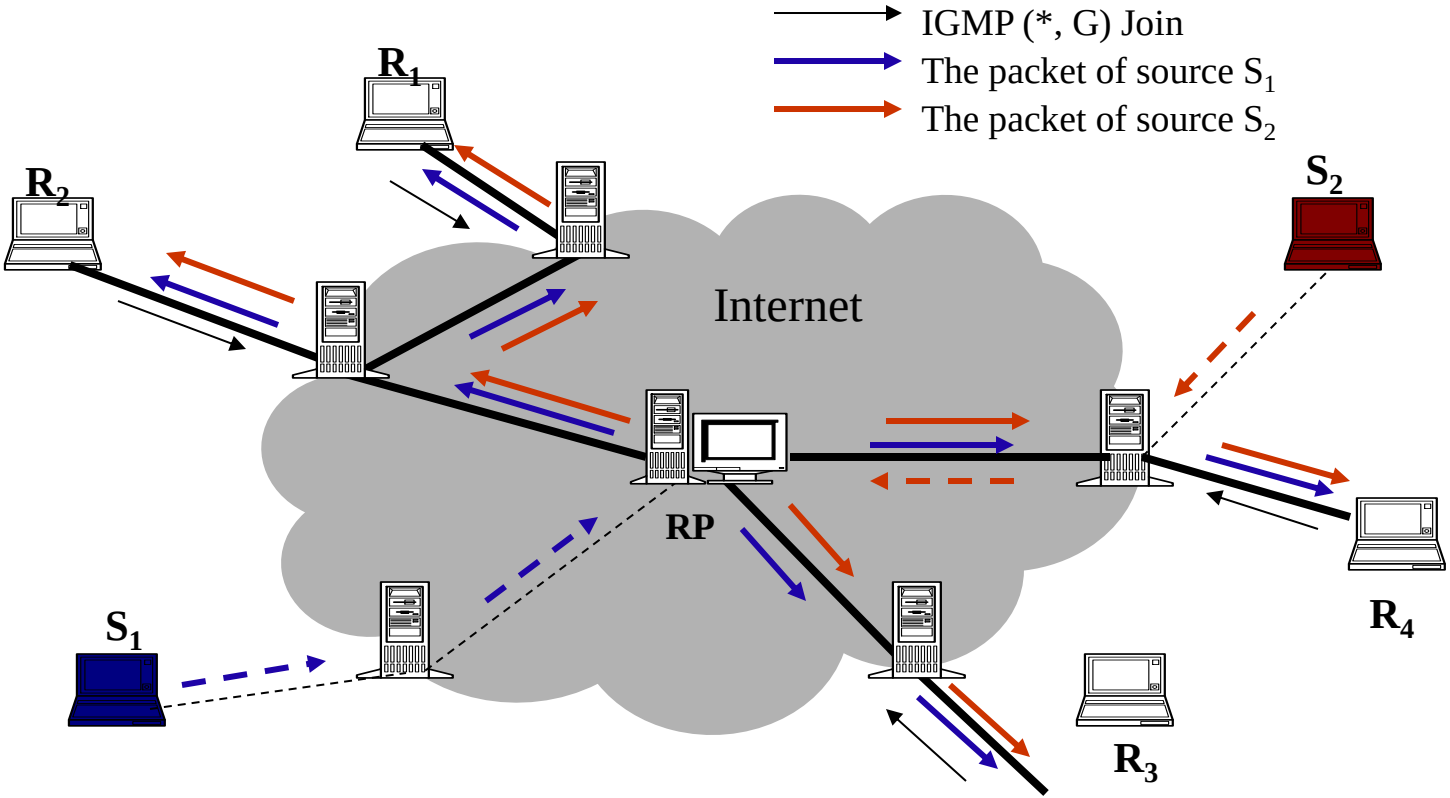
- Protocol Independent Multicast
  - PIM Dense Mode (PIM-DM)
  - PIM Sparse Mode (PIM-SM)

- PIM-SM
  - W. Fenner et al., „Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)” , RFC 4601, August 2006
  - The most used multicast routing protocol today

PIM-SM

- Builds a **shared multicast tree**
- Chooses a rendez-vous point (RP)
  - The RP is the root of the shared tree
    - „Explicit join” – not everybody wants to listen to it
  - Each source sends its message to the RP
    - The RP forwards the messages along the shared tree
  - Optimization to switch after a while from the shared tree to a source-specific tree

PIM-SM operation



Drawbacks of the ASM model

- Several economic and technical issues delayed the large scale deployment of the ASM model
  - Complicated address allocation
    - Dynamic IP address allocation to the source
    - Complex address allocation solutions
      - GLOP (RFC 3180) – static assignment of multicast addresses to ASes
        - » Autonomous System – e.g., the network of an ISP
      - MALLOC - Multicast Address Allocation Architecture (RFC 2908)
        - » MADCAP – Multicast Address Dynamic Client Allocation Protocol
        - » AAP – Multicast Address Allocation Protocol
        - » MASC – Multicast Address Set Claim

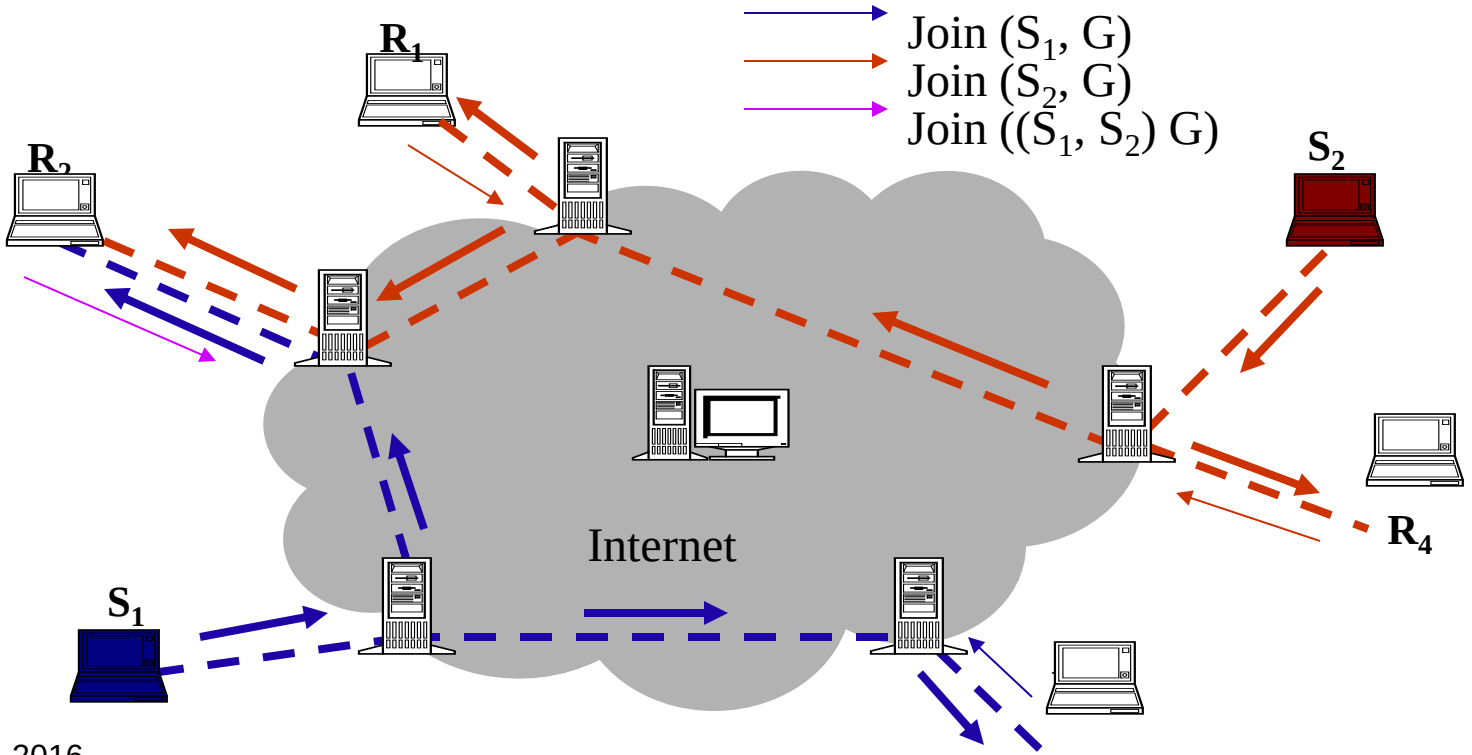
Drawbacks of the ASM model

- Too open model for service providers
  - No control over the sources and receivers
  - Difficult charging
- Not sclabale for inter-domain routing
  - PIM-SM only inside a domain
  - An ISP does not like if its traffic is controlled by an RP located in the network of another ISP
  - Other protocols for inter-domain routing
    - MSDP – Multicast Source Discovery Protocol
    - MBGP – Multicast Border Gateway Protocol

The SSM model

- Need for a simpler model
- **SSM - Source Specific Multicast**
  - Based on the Express model
  - H. Holbrook, D. Cheriton, "IP Multicast Channels: Express Support for Large-Scale Single-Source Application", in *Proceedings of ACM SIGCOMM'99*, Cambridge, MA, USA, Sept. 1999.
- The (*,G) multicast group is replaced by the (S,G) multicast channel
  - S the unicast address of the source
  - G the multicast address of the group
  - Only source S can send packets to the receivers of channel (S,G)
  - Traffic is forwarded along a source-specific tree

SSM model



Source filtering

- The SSM model needs source filtering
  - The host specifies not only which group it wants to listen to, but also which source that sends to that group
- IPv4 – IGMPv3
  - B. Cain, et. Al, "Internet Group Management Protocol, Version 3", RFC 3376, October 2002.
    http://www.ietf.org/rfc/rfc3376.txt
- IPv6 – MLDv2
  - R. Vida, L. Costa, „Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, June 2004.
    http://www.ietf.org/rfc/rfc3810.txt

Message types

- IGMP/MLD Query
  - General Query
    - Who listens what?
  - Group Specific Query
    - Does anybody listen this specific group?
  - Group and Source Specific Query
    - Does anyone listen to this specific source that sends to this specific group?
- IGMP/MLD Report
  - Current State Record
    - What do I listen to – e.g. Include (A) or Exclude (B)
      - A and B are source address sets
  - Filter Mode Change Record
    - Changing the filter mode (Include or Exclude)
  - Source List Change Record
    - Allow (A) or Block (B)

IP Multicast

- Considered for several years the „revolutionary technology of the future”
- Advantages
  - Efficient data transfer
    - Usually over the shortest path (DVMRP, MOSPF, PIM-SSM)
    - Taking into accoun the physical topology
  - Efficient use of resources
    - One packet is sent just once over a specific link
  - Scalable for handling the communication of large groups
    - The group is identified by a virtual group address
      - One routing table entry for a very large group
    - Nobody tracks who is part of the group, and how large is the group

IP Multicast

- Still not deployed at large scale
  - Technical and economic reasons
- Technical reasons
  - Complicated addressing
  - No sclabale inter-domain multicast routing
  - Does not scale to a large number of groups
    - The router has to keep one entry per multicast group
    - Multicast addresses are hard to aggregate
  - Lack of support for higher layer services
    - IP multicast is a *best-effort (multi)point-to-multipoint* data transfer service
    - End users are responsible for handling higher layer services
    - Difficult congestion control and reliablility handling

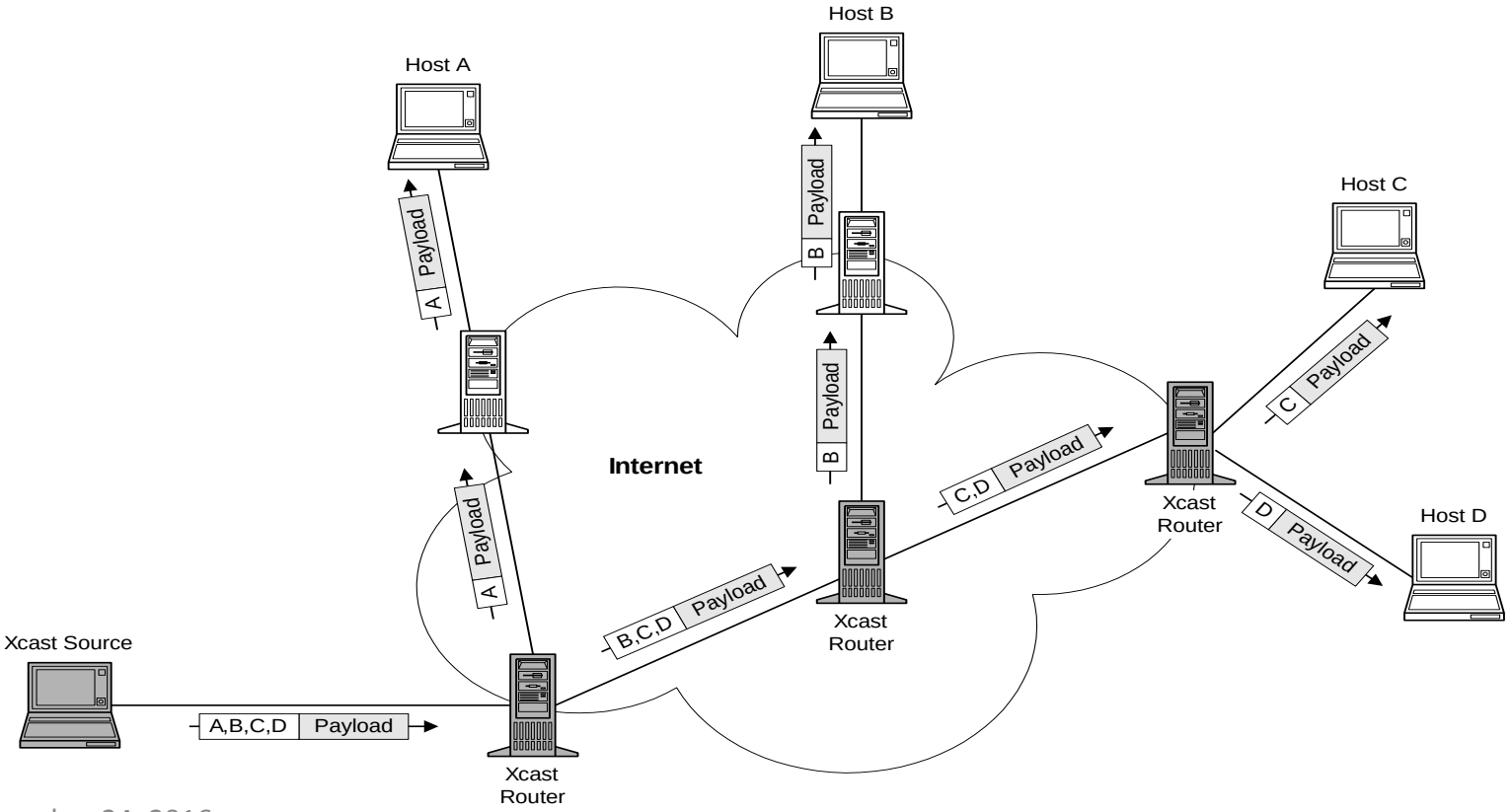
IP Multicast

- Economic reasons
  - Slow and difficult deployment in the network
    - Even though all the routers „speak” today the most important multicast protocols, the ISPs sometimes do not activate them on their networks
    - Really efficient only if used in the entire network
    - Otherwise tunneling is needed
  - „Chicken-egg” problem
    - ISPs do not support it, not enough multicast applications, no need for it
    - A szoftware cégek nem fejlesztenek multicast alkalmazásokat, mert nincs hálózati támogatás, nem lehet majd őket eladni
  - No convenient economic model behind it
    - ISPs have difficulties in controlling the use of networking resources
    - The content provider has difficulties in controlling who uses the service
    - No convenient charging solution behind it

Explicit Multicast (Xcast)

- Network layer multicast solution
- Does not use multicast addresses
  - The source puts in the paket header the unicast IP address of all the group members
- Intermediate Xcast routers duplicate the packets if needed, based on their own internal unicast routing tables
  - The router checks which are the outgoing interfaces for each of the group members, based on its routing table
  - Duplicates the packets if needed, and prepares the corresponding headers

Explicit Multicast (Xcast)



Explicit Multicast (Xcast)

- Not scalable for large groups
  - If many group members, the header becomes too large
- Scales very well for many small groups (for which IP multicast is not good)
  - Routers do not need multicast routing tables
- R. Boivie, N. Feldman, C. Metz, "Small Group Multicast: A New Solution for Multicasting on the Internet", *Internet Computing*, vol. 4, no. 3, May/June 2000, pp. 75-79.

Alternative multicast solutions

C. Diot et al, "Deployment Issues for the IP Multicast Service and Architecture", *IEEE Network Magazine*, *Special Issue on Multicasting*, vol. 14, no. 1, January/February 2000, pp. 78-88.

Can we imagine a group communication service where we do not need network layer support from ISPs?

ALM – Application Layer Multicast

or...

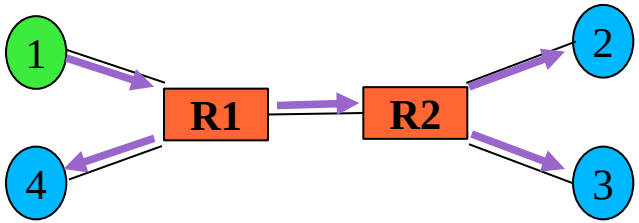
ESM – End System Multicast

or..

HBM – Host-based Multicast

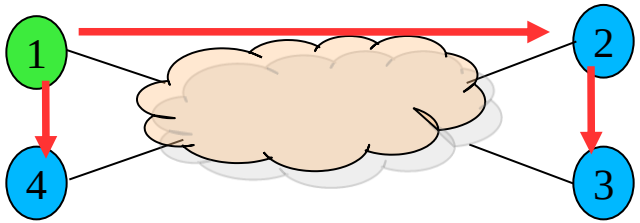
IP multicast - ALM

- IP multicast



- Duplication in the routers
  - Network support
- The topology depends on...
  - The routing tables
  - The physical topology

- ALM
-



- Duplication at the end hosts
  - No network support needed
- Virtual topology
  - The physical topology is a „black box”

ALM: motivation

- Data transfer
  - No IP multicast support needed
    - Uses only unicast communications
  - Small groups
    - IP multicast is not always the best solution
  - Actively using the data
    - Data can be modified/analyzed during transmission
    - Topology can be modified on the fly, based on the content
- Control
  - Aggregation of control data (reliable multicast)

ALM: drawbacks

- Efficiency
  - End-to-end “branches”
    - Delay might be very large
    - Inefficient use of resources

- Scalability
  - Continuous evaluation of the connections between peers
    - Complete graph: n*(n-1) virtual connections in a group with n members

ALM: drawbacks (2)

- Stability
  - Stability of the nodes
    - In the overlay network the participants („routers”) are end hosts
      - Not as reliable as a real router
      - **High churn** - Hosts might join and leave the group quite often
  - Stability of the measurements
    - The efficiency of the overlay depends also on the stability of the chosen metric
      - RTT, bandwidth, etc.
    - Trade-off between the efficient data transfer and the signalling overhead

ALM – general concept

- ALM solutions group the participants in two topologies
  - Control topology („mesh”)
    - Nodes in the control topology periodically refresh their neighbor information
      - Detect and handle errors
  - Data transfer topology („tree”)
    - Part of the control topology, containing the links that are used for data transfer
- Based on the order in which these topologies are built, we have ….
  - Mesh-first ALM: Narada
  - Tree-first ALM: Yoid, HMTP, TBCP, Overcast, ALMI

Narada

Hindu mythological figure

http://en.wikipedia.org/wiki/Narada

Y. Chu, S. Rao, H. Zhang, “A case for End System Multicast”, Proceedings of ACM Sigmetrics, June 2000

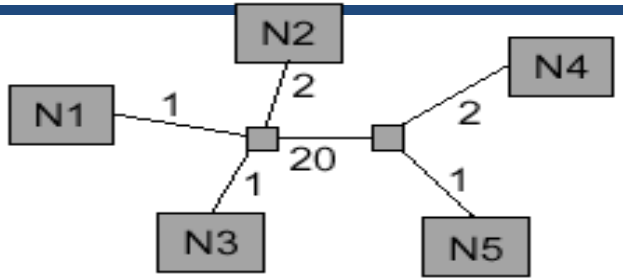
http://www.cs.cmu.edu/~srini/Papers/2002.Chu.jsac.pdf

Narada

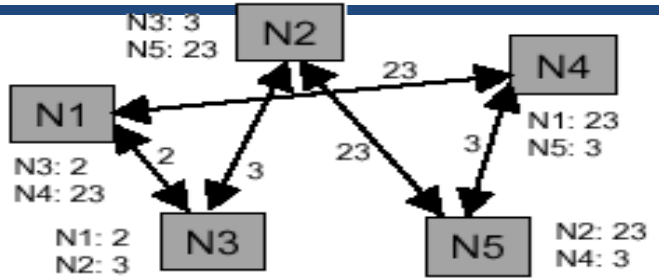
- Distributed, self-managing and self-optimizing overlay solution

- Mesh-first algorithm
  - First builds a bi-directional mesh between participants
  - Then cuts out a Shortest Path Tree (SPT) from the mesh to build the forwarding topology
- Consequences:
  - The quality of the multicast tree will depend on the quality of the mesh
  - Distributed tree building
  - Builds one-directional source-specific trees

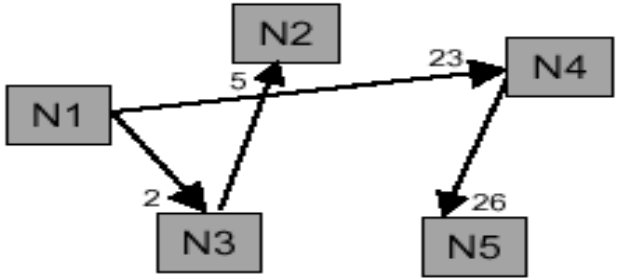
Narada



1. Physical topology



2. Overlay mesh (max 2 neighbors)



3. Data transfer tree

- The mesh is bi-directional
- Separate one-directional tree for each source
- If N1 is the source, then N2 will not send data towards N5, as the shortest path from N1 to N5 is through N4

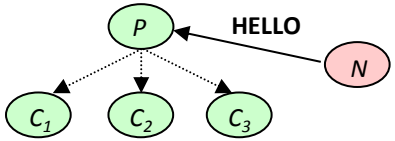
TBCP

- Tree Building Control Protocol
- L. Mathy, R. Canonico, D. Hutchison. *An overlay tree building control protocol*. In Proceedings of International Workshop on Networked Group Communication (NGC), London., 2001.
- http://citeseer.ist.psu.edu/mathy01overlay.html

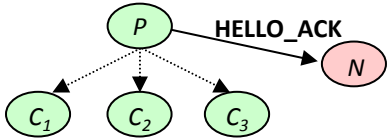
- Tree-first protocol
- Based on measurements between peer nodes
- The data transfer tree is built based on a series of decisions that analyse local full mesh topologies

TBCP algorihm

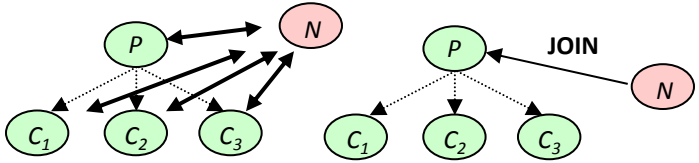
- Basic idea:
  - Each node sends his first join request to the root
  - Peers fall „like dominos” along the tree



1) N sends **HELLO** message to the root P

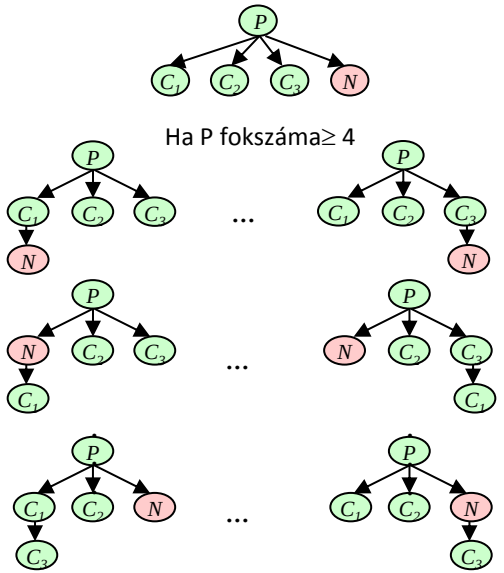


2) P answers with a **HELLO_ACK** sending the list of his immediate children (C_i)



3) N measures its distance to P and each of the children C_i, and sends back the results in a **JOIN** message

TBCP algorithm (2)

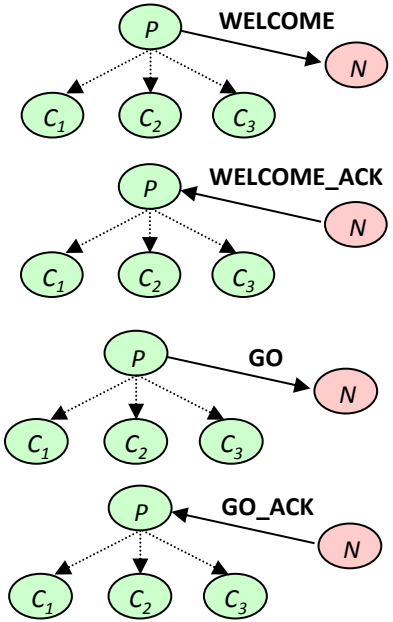


- We use a quality measure to compare the different possible configurations
- P analyses all the possibilities, and chooses the „best” option (local decision)
- We could use different metrics
- Different metrics → different trees
- Which tree is the best? Depends on what we want to do with it…

Advantage/drawback:

the tree is built after a series of local decisions

TBCP algorithm (3)



- If P accepts N as its own child, it sends him a **WELCOME** message

- P might decide to send N, or one of its children, to a lower level down the tree, with a **GO(Ck)** message

- P, or the child Ci that received the **GO(Ck)** message restarts the algorithm by sending a **HELLO** message to Ck

Analysis

- It scales quite well
  - No information needed on the physical topology
  - No need to know all the group members
  - Distributed solution
  - Several peers might join the tree in the same time
- Easy to deploy
  - Builds a relatively good tree relatively rapidly
- Implementation hack to increase efficiency
  - If a measurement towards a node takes too long, that value set to infinity