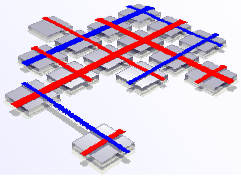
Telematics group
University of Göttingen, Germany

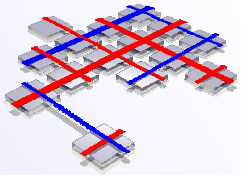


Dynamic Mesh-based overlay Multicast Protocol (DMMP)

< draft-lei-samrg-dmmp-00.txt >

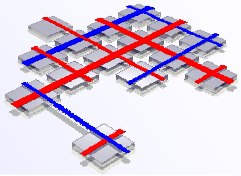
Jun Lei
Xiaoming Fu
Xiaodong Yang
Dieter Hogrefe

IETF#66 Montreal, Quebec, Canada



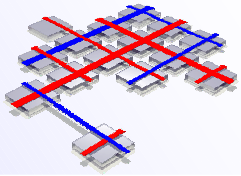
Acknowledgements

- Ruediger Geib
- Nicolai Leymann
- Jun-Hong Cui



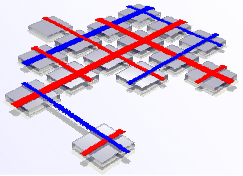
Overview

- Motivations
- Features of DMMP
- DMMP architecture overview
- DMMP messages
- Protocol details
- Security considerations
- Open issues



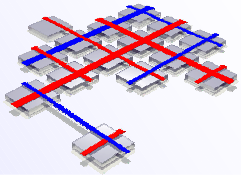
Motivation

- To support real-time media streaming applications, optimizing both the available bandwidth and the delay for group members
- To support large-scale groups without relying on any predetermined intermediate nodes, namely the overlay multicast is solely constructed by end hosts



Features of DMMP

- Support end hosts with heterogeneity
  - A small number of high-capacity end hosts are selected to construct the overlay mesh
- Dynamic mesh-based approach
  - Construction during the multicast initialization phase
  - The mesh structure subject to change when group member changes
- Efficient data distribution tree
  - Distribution of responsibilities to mesh members
- Adaptive and resilient to dynamic network changes
  - No single-node failure would lead to a catastrophe in any part of the overlay multicast tree



DMMP architecture overview (1/2)

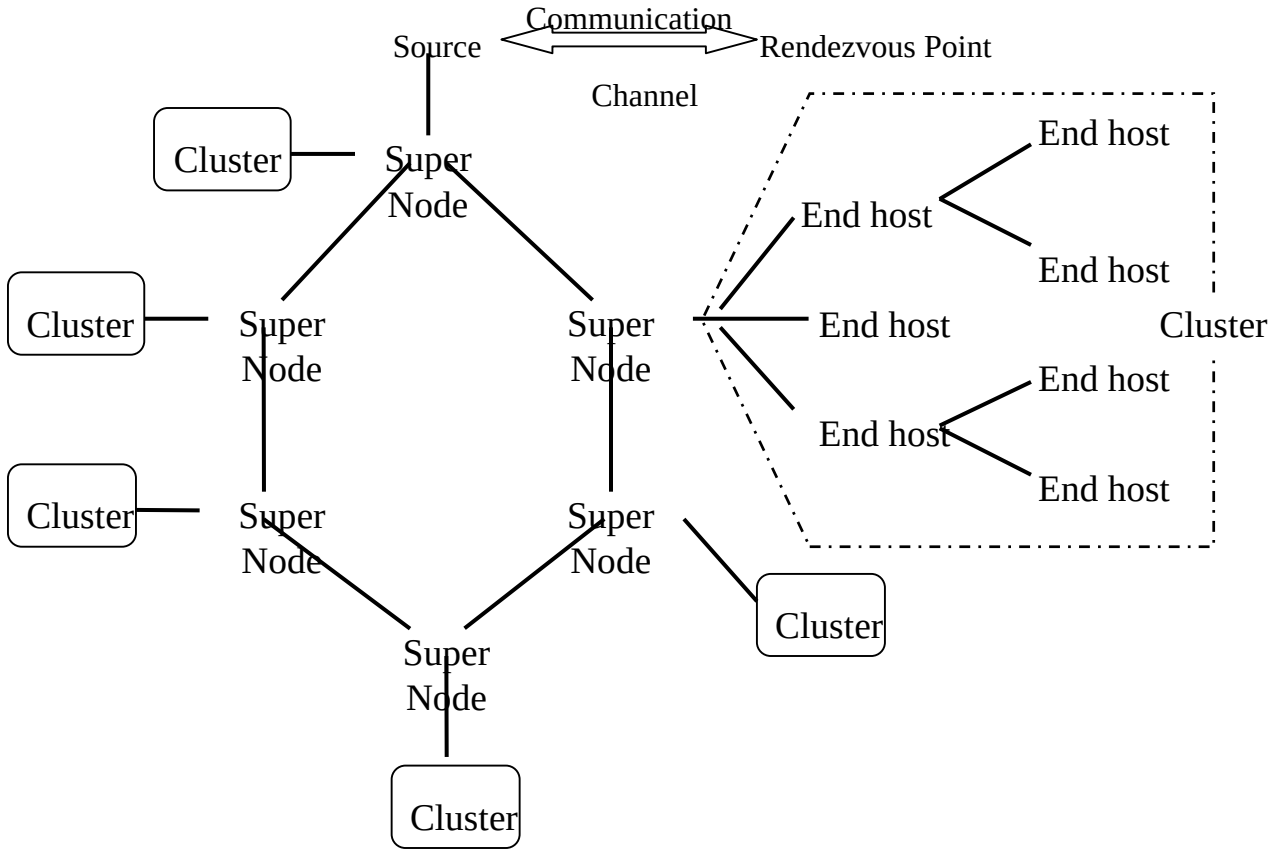
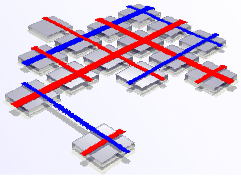
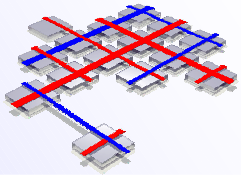


Figure 1 An example of DMMP overlay hierarchy



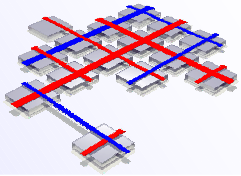
DMMP architecture overview (1/2)

- Control plane
  - Overlay mesh
  - Core-based clusters
  - Functionality: in charge of controlling the overlay hierarchy and completing the multicast tree configuration
- Data plane
  - Built on the top of the structured overlay hierarchy
  - Overlay mesh: Reverse Shortest Path First
  - Core-based clusters: parents -> children



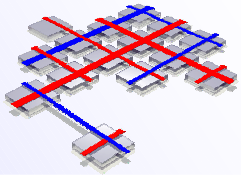
Control plane (1/2)

- Optimal metrics
  - Available bandwidth
  - Other possible criteria, e.g. end-to-end latency
- Super nodes keep the full knowledge among themselves
- Non-super nodes keep the knowledge of a small part of the group within each cluster
- Super nodes willing to contribute more to the network are likely to get better performances



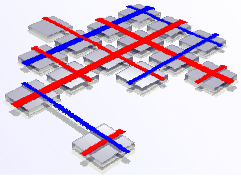
Control plane (2/2)

- Overlay mesh construction phrase
  - Rendezvous Point (RP) distributes all end hosts into two categories: leaf nodes & non-leaf nodes
  - Non-leaf nodes are placed in the order of their out-degree
  - The source selected some super nodes with higher capacity
  - Those selected super nodes self-organize into an overlay mesh
- Cluster construction phrase
  - Having received a list of super node candidates from the RP, each non-super node caches their capacities
  - Each end host chooses one super node who provides better service in terms of e2e latency
  - Non super nodes sharing the same super node will form a cluster
  - Within each cluster, higher capacity nodes are firstly selected to attach to the multicast tree



Data plane (1/2)

- In accordance with the control plane
  - Overlay mesh: the reverse shortest path first
    - super node B receives the packet from the source through its neighbor A only if A is the next hop on the shortest path from B to the source
    - Having received the data, super nodes replicate and forward data to its children in the local cluster
  - Core-based clusters: from higher level to lower level
    - Data are firstly forwarded from the super node to its immediate children
    - Receivers will replicate the data and forward them to its children at the lower level



Data plane (2/2)

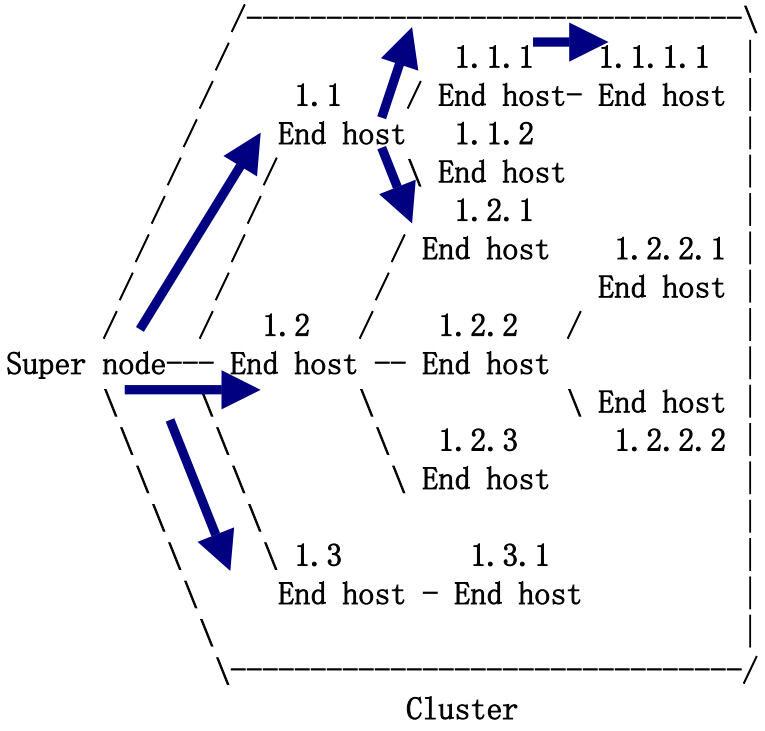
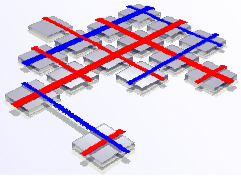


Figure 2 An example of local Cluster

As shown in the figure, data are firstly replicated into three copies, respectively delivered from the super node to its direct children 1.1, 1.2 and 1.3 using unicast.

Similarly, 1.1 replicates copies of the data according to the number of their children (e.g. two copies), sending separately to 1.1.1 and 1.1.2.

In the next iteration, the receiver will similarly make copies and deliver to its children (i.e. 1.1.1.1).

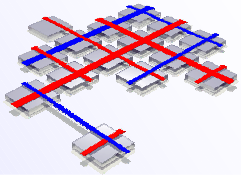


DMMP messages

Messages	Operation	From	To
Subscription Rq	Initializ-ation	Group Member	DNS server
Subscription Res		DNS server	Group Member
Ping_RP Request	Bootstrap	Group Member	RP
Ping_RP Response		RP	Group Member
Source Request	Member Join	new End Host	RP
Source Response		RP	new End Host
Cluster Request	Construct	Cluster Mem.	Super Node
Cluster Response	Clusters	Super Node	Cluster Mem.
Join Request	Member	End Host	Cluster Mem.
Join Response	Join	Cluster Mem.	End Host

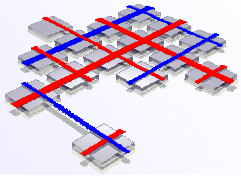
Messages	Operation	From	To
Setup Request	Mesh Management	Super Node	Super Node
Setup Response		Super Node	Super Node
Status Report	Cluster Member Monitoring	Group Member	Group Member
Status Response		Group Member	Group Member
Probe Request	Probe Members	Group Member	Group Member
Probe Response		Group Member	Group Member
Leave Report	Member Leave	Leaving Node	Group Member
Leave Response		Group Member	Leaving Node
Refresh Request	Update Information	Group Member	Group Member
Refresh Response		Group Member	Group Member

Legend: SN - Super Node
Cluster Mem. - Cluster Member



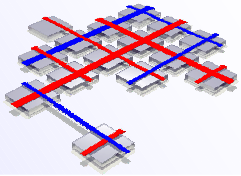
DMMP details

- Initialization
- Super node selection
- Member Join
- Data delivery control
- Refresh information
- Capacity specification
- Member leave
- Failure recovery



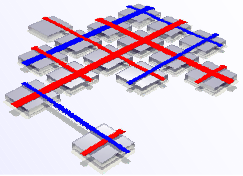
Initialization/assumptions

- Assume
  - DMMP is supported in selected nodes: source, RP, end hosts; and
  - Use of out-of-band channel between the RP and the source
  - Group members using out-of-band bootstrapping mechanism get necessary information



Super node selection

- Requirements
  - Availability: higher power and reliability
  - Number: no more than one hundred
  - Downstream: to satisfy the bandwidth requirement
  - Additional conditions
    - Heterogeneity
    - Resilience
    - Security
- Capacity considerations
  - Out-degree: to speed up the convergence of the overlay tree and to satisfy the bandwidth requirements
  - Uptime: to strengthen the stability of the overlay hierarchy by switching long-term node into the high levels of the tree



Member join

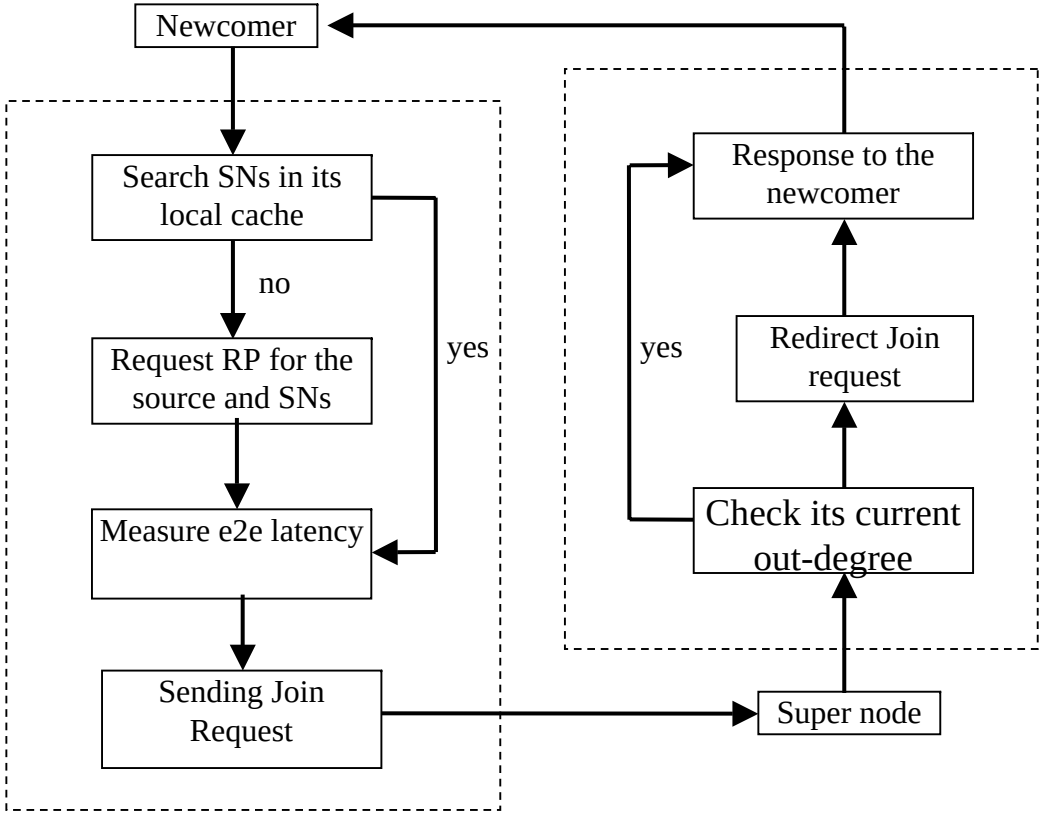
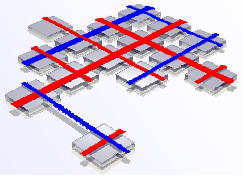


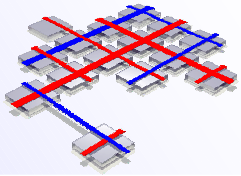
Figure 3 Join procedure in DMMP

Note: Suppose that the newcomer fails to find an appropriate position in any cluster to satisfy application requirements/local policies, it can sell itself as a potential super node and report its own capacities to the RP.



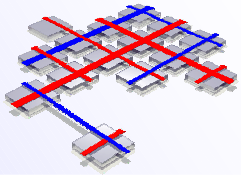
Data delivery control

- After joining the multicast tree, the newcomer
  - Asks its immediate parent to send the data
    - If the parent still holds the data, the newcomer can get data from it
    - If the parent has not received the data yet
      - It waits until the parent forwards the data after receiving (prefer)
      - It directly requires the super node to transfer the data
- On receiving the data, the newcomer forwards
  - to its parent if its parent still has not received the data
  - to its siblings on the condition its PLNs haven't received the data
- Joining as a super node, the newcomer could
  - ask it neighbor in the overlay mesh to transfer the data
  - receive data from existing children
  - directly require the source to send the data



Refresh information

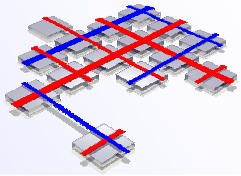
- Periodically sending refresh message to maintain the overlay hierarchy
- Refresh mechanism: active & passive models
  - Overlay mesh
    - Each super node sends update messages to all mesh members including the source
    - Once stopping receiving refresh message exceeds a certain time, a probe message will be initiated
  - Clusters
    - Each end host exchanges refresh message with its relatives (PLNs , siblings and CLNs)
    - End host is able to request refresh message from their relatives



Capacity specification

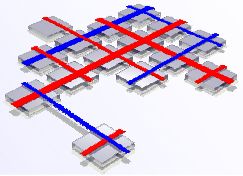
Metric	Operation
Out-degree	Differentiation non-leaf nodes from leaf nodes
	Super nodes selection
	Tree construction within clusters
	New member joins the group
	Failure recovery mechanism
	Self-improving mechanism
E2E latency	Non-super nodes attach to super nodes to form clusters
	New member joins the group
Uptime	New member joins the group
	Self-improving mechanism

- Out-degree is the main criterion
- Out-degree, e2e delay and uptime are all taken into considerations when regarding the member joining procedure
- The combination of out-degree and uptime is chosen as a comparison metric to self-improve the overlay multicast tree



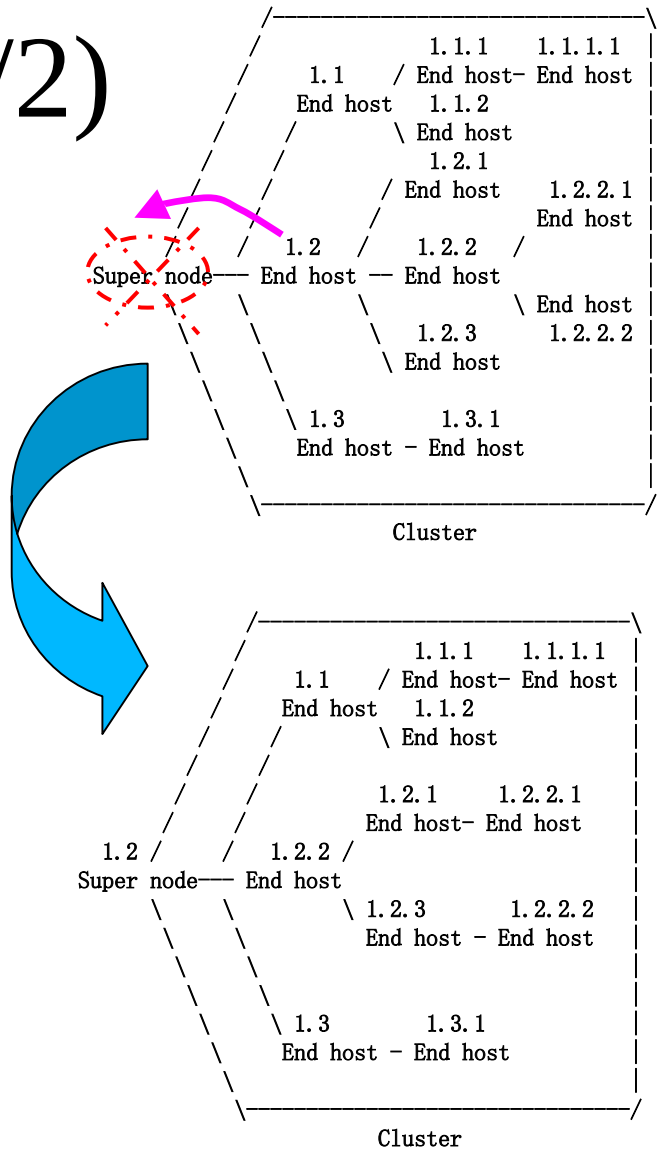
Member leave (1/2)

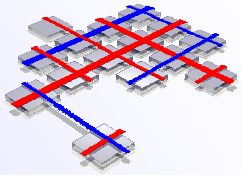
- Two situations: gracefully or ungracefully
  - Clusters
    - Graceful leaving
      - Leaving member needs to send a Leave Request to its parent or one of its children
      - Notified member will propagate the Leave message to its relatives
    - Ungraceful leaving
      - Detected by periodically exchanging refresh messages
      - May cause the crash of the whole multicast tree, which is handled by the failure detection and recovery mechanism



Member leave (2/2)

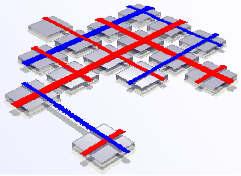
- Mesh
  - Graceful leaving
    - The leaving super node must elect a replacement leader and inform the other super nodes
  - Ungraceful leaving
    - Depending on refresh message, DMMP detects unannounced leavings
    - Source selects one of the victim's children with largest out-degree as the new super node
    - Correspondence information will be updated to the RP
    - The neighbors in the same cluster adjust their positions





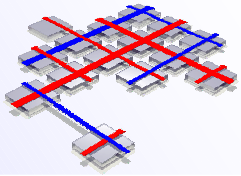
Failure recovery

- Failure detection
  - By noticing missing periodical Refresh/Update message
- Failure recovery mechanisms
  - Proactive approach: used in overlay mesh
    - Backup parent for the immediate children of each super node
    - After super node leaving the group, each child tries to contact with alternative parent
  - Active approach: in each local cluster
    - Each end host periodically estimates their relatives
    - Possible solution: Randomized Forwarding with Triggered NAKs



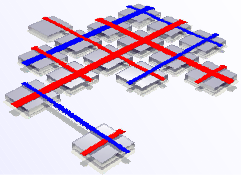
Security considerations

- Super node selection
  - Authority center (AC) to qualify the trust level of end hosts
  - end host can be selected as a super node only if it obtains a security certificate from the AC
- Within clusters
  - Cluster key
  - Group key
  - Private key



Open issues

- Large scale efficiency
- Security
- NAT and firewall traversal
- E2e QoS provision?



Questions and comments appreciated!

For further information, please contact:

{lei,fu}@cs.uni-goettingen.de