Management Substrate Structure for Adaptive Network Resource Management

Daphne Tuncer, Marinos Charalambides, Hisham El-Ezhabi, and George Pavlou
Department of Electronic and Electrical Engineering
University College London
Email: {d.tuncer, m.charalambides, h.elezaby, g.pavlou}@ee.ucl.ac.uk

***Abstract*—Centralized and offline network management functionality, traditionally deployed by operators, cannot easily deal with the traffic patterns of emerging services, which are becoming more dynamic and unpredictable. As such, decentralized solutions that are flexible and adaptive to traffic and network dynamics are of paramount importance. To this end, an in-network management approach in which an intelligent substrate allows the dynamic reconfiguration of resources according to network conditions has been developed. The set of nodes forming this logical structure are able to communicate with each other to coordinate their decisions. This paper investigates the use of three different topology models to organize the nodes in the substrate. Algorithms to compute the proposed structures that are described take into account important criteria such as minimizing the latency and the communication overhead among the substrate nodes. A quantitative and qualitative evaluation of the different structures in terms of cost and complexity is performed based on real network topologies.**

I. Introduction

Network resource management approaches traditionally deployed by operators rely on offline functionality that cannot easily deal with the traffic patterns of emerging services, which are becoming more dynamic and unpredictable. As such, solutions that are flexible and adaptive to traffic and network dynamics are of paramount importance. Furthermore, network resource management normally relies on centralized managers that periodically compute new configurations according to dynamic traffic behaviors. These centralized approaches have limitations especially in terms of scalability (i.e. communication overhead between the central manager and devices at runtime) and lag in the central manager reactions that may result in sub-optimal performance. To meet the requirements of emerging services, network resource management functionality that is decentralized, flexible, reactive and adaptive to traffic and network dynamics is necessary.

To overcome the limitations of current approaches, this paper presents a new in-network management framework for dynamic resource reconfiguration in fixed backbone networks. According to the proposed framework, the decision-making process is distributed across nodes in the network, so that each node is responsible for deciding on reconfiguration actions to take based on local feedback regarding the state of the network. Nodes are equipped with the necessary logic that can allow them to perform reconfigurations, so that the network resources can be better utilized. In order to avoid inconsistencies between several independent decisions, the network nodes cooperatively decide on the most suitable changes to apply depending on network characteristics and conditions. The network nodes participating in the resource management process belong a *management substrate*, which is a logical structure used to facilitate the exchange of information between distributed decision-making points. Such a framework was used in our previous work [1] [2] [3] for the purpose of adaptive traffic engineering, energy efficiency and in-network cache management, respectively. However, due to the distributed nature of the decision-making process, the performance of the proposed management scheme in terms of communication overhead can be affected by the structure of the management substrate. In this paper, three different topology structures are investigated to connect management substrate nodes. A set of methods to compute the proposed topology structures is described and a quantitative and qualitative comparison of the three topologies according to different parameters is presented.

The rest of the paper is organized as follows. Related work is presented in section II. Section III introduces in more details the in-network management substrate framework developed to perform adaptive resource management. Sections IV, V and VI present the three different topology structures. The characteristics of the different structures are evaluated and discussed in section VII. The contributions of this work are finally summarized in section VIII.

II. Related Work

Interaction and communication between autonomic elements have been described as fundamental architectural features of autonomic computing systems in [4]. The authors highlight in particular that autonomic elements can establish relationships between each other in order to request or offer some service. A generic model based on negotiation is proposed to drive the interaction between autonomic elements. Other generic interaction models have been considered in [5], where four types of behavior that can be exhibited by an autonomic element towards other autonomic elements are described, i.e. the cooperative behavior, the selfish behavior, the punishment behavior and the mixed behavior. Communication models between network entities to support management tasks have also been considered by [6] in the context of autonomic networks. In [6], the interaction between the decision elements relies on a hierarchical structure in which the decisions taken by each decision element are orchestrated by one or more ”arbiter”

elements that are in charge of detecting potential overlapping or contracting actions and configurations.

Some research efforts have also investigated the use of generic hierarchical architectures inspired by multi-agent systems to support the interaction and the cooperation between nodes [7] [8]. The use of gossip-based protocols to distribute information across distributed decision-making points was considered in [9] [10] [11]. According to gossip-based approaches, the interaction between nodes relies on a random process, so that at regular time intervals, one node in the network initiates a communication with a randomly selected neighbor in order to exchange information. The work in [9] focused on the development of scalable and adaptive mechanisms for calculating aggregates in a pro-active manner. A gossip-based approach was used in [10] for dynamic resource allocation in cloud environments and in [11] for development of decentralized self-adaptive aggregation mechanisms.

The design of logical infrastructures to connect a set of nodes has received a lot of attention from the research community over the last decade, especially in the context of peer-to-peer networks [12] [13]. While research efforts in this area have focused on developing scalable systems through optimized logical topologies and overlay routing protocols, the purpose of the work presented in this paper is not to investigate features and techniques to support overlay systems. It focuses, instead, on the design of topology structures that can offer good performance in terms of communication cost and management overhead for supporting the interaction between network reconfiguration entities.

III. In-Network Management Substrate

A. Notations and Definitions

The following notations are used in this paper. Let $\mathcal{L}$ be the set of network links and $\mathcal{N}$ the set of network nodes. The later is further divided into the set of network edge nodes, i.e. network nodes generating and absorbing traffic, and the set of network core nodes. The network infrastructures considered here correspond to intra-domain fixed backbone networks.

B. Decentralised Resource Management Framework

In the proposed in-network resource management framework, network edge nodes are embedded with a level of intelligence that allows them to react to network conditions in a decentralized and adaptive fashion based on periodical feedback information received from the network. Compared to centralized offline solutions, where reconfigurations are decided by a centralized management system that has a global knowledge about the network, reconfiguration decisions are directly taken by the network edge nodes that coordinate among themselves in order to decide upon the best sequence of actions to perform to satisfy a common objective. These can for instance consist in the adjustments of routing parameters for load-balancing purposes. In order to support this decentralized decision-making process, the network edge nodes are organized into a *management substrate*, which is a logical structure used to facilitate the exchange of information between decision-making entities. The management substrate is used by the edge nodes for coordination purposes, in particular, since it provides a means through which nodes can communicate. It is worth mentioning that the substrate is only used for signalling, and not for direct traffic routing/forwarding.

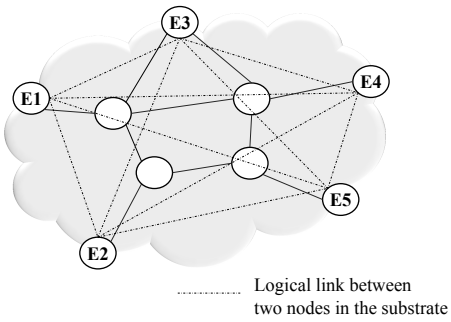

Fig. 1. In-network management substrate overview.

A management substrate structure example is depicted in Fig. 1, where each network edge node E is logically connected to a set of other network edge nodes (neighbor nodes in the management substrate (MS)). Any MS node can directly communicate only with its neighbors, which are defined by the topological structure used. The choice of the substrate topology can be driven by different parameters related to the physical network, such as its topology, the number of edge nodes, but also by the constraints of the coordination mechanism between the nodes and the associated communication protocol. The overhead incurred by the communication protocol in terms of delay and number of messages exchanged, for example, is a key factor that can influence the choice of the topology.

In the proposed management framework, it is assumed all MS nodes are network edge nodes. As such, unless otherwise stated, the term node is used to refer to a network edge node. In addition, this work assumes that the networks considered are reliable in terms of node failures.

C. Substrate Characteristics

Three different topology structures are investigated to connect the MS nodes. The proposed structures differ in terms of degree of connectivity (i.e. number of neighbors of each node in the substrate) and the number of hierarchy levels. The degree of connectivity of a topology defines the visibility of each node in the substrate and can thus affect the volume of information that needs to be maintained at each MS node. In addition, the number of hierarchy levels in the structure drives the number of modes of communication required between MS nodes, which may influence the complexity of the communication protocol. The main objective considered for the design of each structure is to minimize the communication overhead incurred by the coordination process. This is defined by the volume of signalling messages and the delay, which is driven by the communication cost between MS nodes.

The communication cost between two MS nodes i and j is defined as the cost of the logical link between the two nodes.

This cost, denoted $C_{LL}(ij)$, is driven by the cost of the path between node i and node j in the underlying physical network topology, where the cost of a path is equal to the sum of the cost of the links involved in the path, i.e.:

$$C_{LL}(ij) = \sum_{l \in \mathcal{L}} \delta_{ij}^{l} \cdot c(l) \quad (1)$$

where δ_{ij}^{l} is a $\{0-1\}$ binary variable equal to 1 if link l is included in the path between nodes i and j, and $c(l)$ is the cost of link l.

The cost $c(l)$ of link l can be defined, for instance, according to the administrative cost (i.e. link weight) which is the metric used to compute the shortest paths. Administrative costs are usually assigned based on the characteristics of the underlying physical network topology and on traffic engineering requirements. A common practice is to set link weights equal to the inverse of the link capacities [14]. These costs may not, however, be sufficient to account for the communication cost in terms of delay between two nodes since the delay is also influenced by the geographical distance between the nodes (i.e. propagation delay). In order to take the geographical distance into account, an additional metric, called link distance factor (c_{φ}), is defined for each link. This represents the relative distance between two nodes in the network and is defined as the ratio between the geographical distance d_l (e.g. in kilometers) obtained for each link l divided by the smallest geographical distance observed in the network, i.e.

$$c_{\varphi} = \frac{d_l}{\min_{l \in \mathcal{L}}(d_l)}$$

The cost of a link l is then defined as the product of the link administrative cost c_{α} and the link distance factor c_{φ}, i.e.

$$c(l) = c_{\alpha} \cdot c_{\varphi}$$

It is assumed that the path used between two nodes is the shortest-path and that all network links are bidirectional, so that for any pair of nodes i and j, $C_{LL}(ij) = C_{LL}(ji)$. Finally, it is worth noting that, for reliability purposes, the proposed framework relies on the Transmission Control Protocol (TCP) [15] as the underlying transport protocol.

IV. Full-mesh Management Substrate Structure

A. Topology Structure

MS nodes are connected according to a full-mesh topology as shown in Fig. 2, where each node is logically connected to every other node.

The full-mesh model is a flat structure (i.e. not hierarchical) and has a high degree of connectivity. The total number of logical links is equal to $\frac{N \cdot (N-1)}{2}$, where N is the total number of MS nodes. In this model, each MS node has a global view about other MS nodes. This provides a greater flexibility in the choice of neighbors with which to communicate since all MS nodes belong to the set of neighbors. Each MS node, however, needs to locally maintain information about every other MS node, which may raise some limitations with an increasing number of substrate nodes.

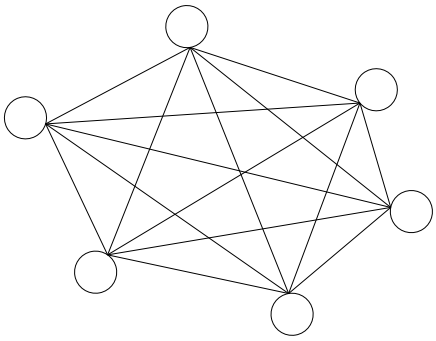

Fig. 2. Full-mesh topology structure.

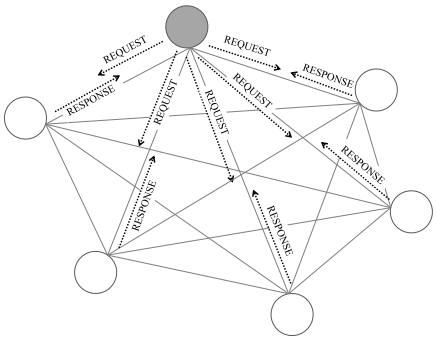

Fig. 3. Communication model in the full-mesh topology structure.

B. Communication Protocol For Management Operations

1) Communication Model: As explained in section III-B, the MS facilitates the communication between reconfiguration entities (i.e. edge nodes). These can for instance exchange information related to the reconfigurations to perform or request assistance from each other when local reconfigurations are not possible. The communication model used in the full-mesh topology follows a star structure centered on the node that initiates the communication as depicted in Fig. 3. The initiator (represented as a gray disc in the figure) sends a $REQUEST$ message to all its neighbors in the management substrate. The $REQUEST$ message can be used, for instance, to request some information from the other nodes in the substrate. Upon receiving a $REQUEST$ message, each neighbor node analyzes the content of the message to decide whether it can provide a satisfactory reply to the request. In case it can, the node appends the required information to a $RESPONSE$ message and forwards it back to the initiator. Otherwise, a negative $RESPONSE$ message is returned. A similar communication protocol can also be used to notify other nodes in the substrate about local changes. In order to minimize the communication overhead due to signalling, the size of the messages needs to be small.

2) Communication Overhead: Intuitively, the total number of messages sent by the initiator node is equal to the number of neighbors it has, i.e. $(N-1)$ where N is the number

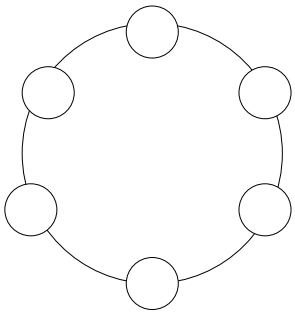

Fig. 4. Ring topology structure.

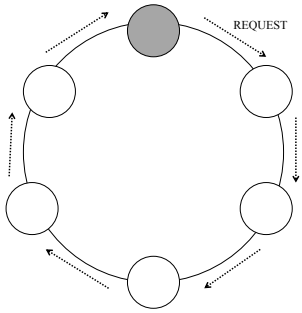

Fig. 5. Communication model in the ring topology structure.

of nodes in the management substrate (MS). The volume of signalling messages increases linearly with the number of MS nodes. Due to the structure of the full-mesh topology, the communication cost in terms of delay mainly depends on the maximum round-trip time (RTT) between the initiator and any other MS node. As shown in previous studies [16] [17], the values of the RTT are mainly influenced by the geographical distribution of network nodes. It can therefore be inferred that the communication cost in terms of delay in the full-mesh model is driven by the maximum geographical distance between MS nodes.

V. Ring Management Substrate Structure

A. Topology Structure

MS nodes are connected according to a ring topology as shown in Fig. 4, where each node is logically connected to two other nodes only.

The ring topology is also a flat structure. Unlike the full-mesh topology, however, it has a low degree of connectivity. The total number of logical links is equal to the number of nodes in the substrate. The view of each MS node is limited to its two direct neighbors only, and it is thus not possible to directly communicate with any other nodes in the substrate. In order to communicate, a message needs to be sent over the ring until it reaches its destination. Given that the total communication cost (i.e. delay) can be defined as the sum of the cost between all successive nodes, it is affected by the order according to which the nodes are connected. Based on the definition of the cost provided in section III-C, the next subsection presents a heuristic to connect a set of nodes according to a ring topology, so that the total cost (i.e. sum of the cost between each successive node) is minimized.

B. Ring Model Construction

This problem is similar to the Travelling Salesman Problem (TSP) [18]. The TSP is a well-know NP-Hard combinatorial optimization problem that consists in determining, given a list of locations and their pairwise distances, the shortest possible route that visits each location exactly once and that returns to the starting location. Although a number of approaches with near-optimal performance exists in the literature to solve the TSP, they are computationally expensive. In order to keep the complexity of the construction algorithm low, an approach based on the simple *Nearest Neighbors* tour construction heuristic [19] has been developed. This has a time complexity $O(N^2)$, with N being the number of nodes to consider.

The principle of the proposed approach is as follows. Given a node i, node j is selected as the successor of i such that the cost $C_{LL}\,(ij)$ is the lowest. The *Nearest Neighbor* algorithm considers each node i iteratively and selects, among other MS nodes that have not already been considered, the successor of i, i.e. the node with the lowest logical link cost to i. The algorithm terminates when all nodes have been considered and the successor of the last node is set to be the initial node.

C. Communication Protocol For Management Operations

1) Communication Model: The communication between MS nodes in the ring model relies on a hop-by-hop mechanism as depicted in Fig. 5. Communication is unidirectional, which means that a node can only pass information to its immediate neighbor in the ring. To communicate with any other node a message needs to be sent over the ring until it reaches its destination. The communication direction followed in the ring must be fixed but it can be either anticlockwise or clockwise.

The initiator node (represented as a gray disc in the figure) sends a $REQUEST$ message to one of its neighboring nodes according to the communication direction followed. The initiator node then enters a listening period where it waits for the message to travel hop-by-hop through the ring until it reaches the initiator again. Upon receiving the $REQUEST$ message, the next hop node analyzes the content of the message, appends the required information as well as its identity to the message and forwards it to the next hop node.

2) Communication Overhead: In the case of the ring model, the number of neighbors of each MS node is independent of the total number of nodes in the substrate. In order to communicate, the initiator node sends a message to one of its direct neighbors and the message is forwarded hop-by-hop to the other nodes in the ring. As such, the number of messages sent by each MS node is independent of the total number of nodes in the substrate. In contrast, due to the characteristics of the communication model, it can be inferred that the communication cost in terms of delay will be driven by the number of nodes in the ring, and as such, there may

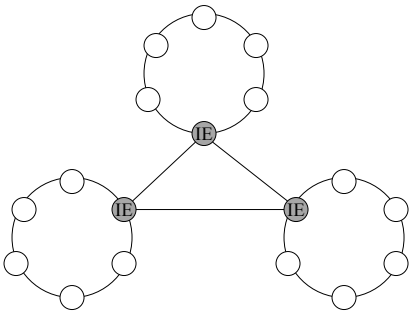

Fig. 6. Hybrid topology structure.

be some scalability limitations as the number of nodes in the substrate increases.

VI. Hybrid Management Substrate Structure

A. Topology Structure

Due to their characteristics, the full-mesh and ring models present some scalability limitations when the number of MS nodes increases. In the case of the full-mesh model, this can incur a significant increase in the volume of substrate information to be maintained locally at each MS node and in the case of the ring model it can significantly affect the total communication delay. In order to overcome the limitations of these two simple structures, the design of a more sophisticated model to organize the MS nodes is investigated. This model, referred to as a hybrid topology, is a combination of the ring and full-mesh structures, as depicted in Fig. 6.

The hybrid topology consists of a set of rings inter-connected in a fully-meshed fashion through *Intermediate Entity* (IE) nodes, so that there exists exactly one IE node in each ring. More specifically, MS nodes are partitioned into at least two clusters, so that nodes in each of the clusters are connected according to a ring topology. One node is then selected in each cluster to be the IE, i.e. to act as an interface to the other clusters. It is worth noting that each MS node belongs to one cluster only. One of the incentives for using a hybrid structure is to provide a trade-off in terms of performance between the message overhead and the delay incurred when two MS nodes need to exchange information. Such a trade-off raises some requirements when deciding how to connect nodes according to the hybrid model. The next section presents a set of methods that define how to partition the MS nodes into clusters and how to select the IE node in each cluster.

B. Hybrid Model Construction

1) Constructing Multiple Rings: One key challenge when forming the hybrid structure is to determine which metric to use in order to partition the MS nodes into clusters. A natural choice is to use the logical link cost metric defined in section III-C, which is a function of the link administrative cost and the geographical distance. As such, nodes are clustered based on their proximity with respect to the logical link cost.

In order to reduce the communication delay compared to the ring structure, the total communication cost permitted in each sub-ring of the hybrid structure needs to be less than an upper bound threshold θ. The value of the threshold is a key factor since it can influence whether a node should be considered as a member of a specific ring or not, and, as such, directly affects the size of each sub-ring. To set the appropriate threshold value, it is also essential to take into account the fact that nodes located in different sub-rings can communicate. Given that nodes can directly communicate between each other in the full-mesh model, it can be inferred that the communication cost in this model is less than the cost in the ring model. As explained in section IV-B, the communication cost of the full-mesh structure is driven by the maximum geographical distance between MS nodes and, as such, by the maximum logical link cost in the MS. This can therefore be used as a reference metric to derive the value of the threshold to apply to the total cost in each sub-ring. Two cases are investigated:

1) θ is equal to $\theta_{HalfMax}$, i.e. to half of the maximum logical link cost obtained if MS nodes were connected in a full-mesh fashion.
2) θ is equal to θ_{Avg}, i.e. to the average logical link cost obtained between all possible pairs of nodes if MS nodes were connected in a full-mesh fashion.

An approach has been designed to partition the MS nodes into the different clusters according to θ, and compute the resulting sub-rings. The proposed algorithm follows an iterative process where all MS nodes are considered one-by-one. The number of clusters is not determined a priori. One cluster is initially formed by the algorithm and nodes are successively added to this cluster until the threshold condition is violated. In this case, the initial cluster is said to be complete and a new cluster is formed to accommodate the remaining nodes. The different clusters are thus formed successively according to the threshold value θ and the order in which nodes are considered. To ensure that each node belongs to one cluster only, the algorithm maintains the list of MS nodes that have not been considered yet. The list contains initially all MS nodes and is updated at each iteration by removing the node selected by the algorithm. The output of the algorithm is a set of rings.

The various steps of the algorithm are as follows. N_{curr} is the node considered by the algorithm at each iteration and $\mathcal{N}_{wait}$ the list of the MS nodes that have not been considered yet. N_{ini} is the initial node and $\mathcal{S}_{RINGS}$ is the set of constructed rings.

1) Select an initial node N_{ini}, set N_{curr} to N_{ini} and remove N_{ini} from $\mathcal{N}_{wait}$.
2) Create a new cluster $\mathcal{C}$ with N_{curr}.
3) Compare the cost of the logical links from N_{curr} to all nodes in $\mathcal{N}_{wait}$. Select the pair (i.e. logical link) with the lowest cost and mark the relevant peer node as N_{test}.
4) Apply the ring construction algorithm described in section V-B to the set of nodes formed by the union of the set of nodes in cluster $\mathcal{C}$ and N_{test}. Determine the total ring cost C_{ring}.

5) Compare C_{ring} to the threshold value θ. If $C_{ring} \leq \theta$, add N_{test} to cluster $\mathcal{C}$, remove N_{test} from $\mathcal{N}_{wait}$ and set N_{curr} to N_{test}. Go back to step 3. If $C_{ring} > \theta$, apply the ring construction algorithm to nodes in cluster $\mathcal{C}$ and add the resulting ring to $\mathcal{S}_{RINGS}$. Set N_{curr} to N_{test}, remove N_{test} from $\mathcal{N}_{wait}$ and go back to step 2.
6) Continue until $\mathcal{N}_{wait}$ is empty.

Two criteria to select the initial node are compared. In the first case, the node connected to the logical link with the lowest cost in the MS is selected as the initial node, while, in the second case, the node connected to the logical link with the highest cost in the MS is selected. Given that logical links are bidirectional, the node with the lowest identifier is selected by default.

There may be cases where some of the sub-rings obtained contain one element only, which is not acceptable by definition. The structure of each sub-ring is therefore analyzed at the end of the algorithm. If single node sub-rings are found, the algorithm disregards them and assigns the involved nodes to other sub-rings, so that the selected rings are those for which the addition of an extra node leads to the lowest increase in terms of cost.

2) Intermediate Entity Selection: Another key issue raised by the design of the hybrid topology is the selection of the most appropriate IE in each sub-ring, so that these can be efficiently inter-connected in a full-mesh. In a similar fashion to the method used to select successor nodes in the ring construction algorithm, IE nodes are chosen according to their proximity, in terms of logical link cost, to other rings. This can be formally described as the problem to determine which node to select in each sub-ring so that the maximum logical link cost between all pairs of the IE nodes is minimized. In order to simplify the Intermediate Entity Selection procedure, a heuristic to select the node in each sub-ring that is the closest on average to every other remote node in the substrate (i.e. to the nodes in other sub-rings) has been investigated. The proposed approach relies on an iterative process, where sub-rings are considered one-by-one, so that at each iteration, one node in the considered sub-ring is selected as the IE. In order to select the appropriate IE in each ring, the algorithm computes, for each node in the ring, the average logical link cost to every other remote node. The selected IE in each ring is the one with the lowest average cost, i.e. the node that is closest on average to every other node in the substrate.

C. Communication Protocol For Management Operations

1) Communication Model: The protocol for the communication between the MS nodes organized into a hybrid structure supports two modes of communication as depicted in Fig. 7 and described below.

Local Sub-ring Communication: The first mode concerns the communication between nodes located in the same ring. This mode corresponds to the case where the node that initiates the communication (represented by a gray disc in the figure) needs to exchange some information with another (other) node(s) in the local sub-ring. In that case, the hop-by-hop mechanism described in section V-C for the ring model is used. More specifically, the initiator node sends a local request in the form of a $LOCAL\_REQ$ message to one of its neighboring nodes according to the communication direction followed. The message then travels hop-by-hop through the ring until it reaches the initiator node again.

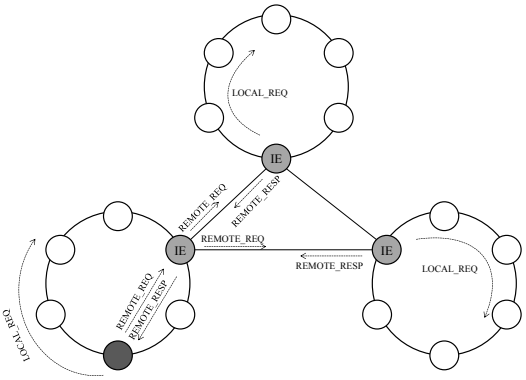

Fig. 7. Overview of the hybrid topology communication model.

Remote Sub-ring Communication: The second mode concerns the communication between nodes located in different rings, when for example the initiator node needs to retrieve information from a node located in a remote sub-ring. To do this, the initiator needs to first communicate with its local IE since this node acts as an interface to the other rings. It is assumed that the address of the IE in a given sub-ring is known by all the nodes of that ring. The initiator starts by sending a remote request ($REMOTE\_REQ$) message directly to its IE node, which then forwards it to all the other IE nodes of the MS. Each IE is subsequently responsible for circulating a $LOCAL\_REQ$ message in its local ring. Upon receiving this message back, each IE analyzes its content and creates a remote response ($REMOTE\_RESP$) message that contains information about potential satisfactory replies from its ring. This is sent back to the original requesting IE, which forwards it to the initiator.

2) Communication Overhead: In the full-mesh MS topology model, the communication overhead incurred when a node requests information is proportional to the number of nodes in the MS, since a message is exchanged with every other node (section IV-B). According to the communication protocol used in the hybrid model, the total number of messages exchanged depends on the communication mode considered. For the local sub-ring communication case, only one message needs to be sent by each node: a $LOCAL\_REQ$ message to its direct neighbor. For the remote sub-ring communication case, however, with r being the number of sub-rings, one $REMOTE\_REQ$ message is sent by the initiator node to the local IE and $(r-1)$ $REMOTE\_REQ$ messages are sent from the local IE to other IE nodes in the substrate. As such the communication overhead in terms of number of messages in the hybrid model is, in the worst case, proportional to the number of sub-rings. Compared to the full-mesh topology, the performance of the hybrid model improves as the size of each ring increases,

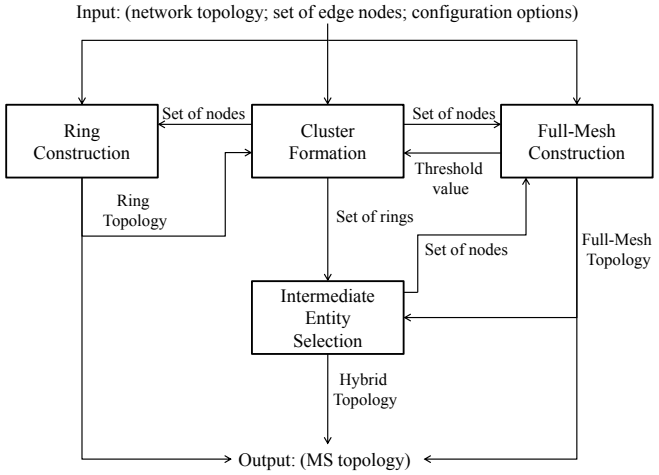

Fig. 8. Overview of the Management Substrate computation software architecture.

and consequently, as the number of rings decreases.

Given the hybrid nature of the model, it can be deduced that the communication cost in terms of delay will be driven by the characteristics of the full-mesh and ring structures. It can therefore be inferred that the total delay will be influenced by the size of the largest sub-ring and the maximum distance between IE nodes. In addition, it is expected that better performance in terms communication cost will be achieved with the hybrid model than with the ring model but that these may not outperform the performance obtained with the full-mesh approach. An quantitative evaluation of the communication cost obtained with each topology model is presented in section VII-C.

VII. Implementation and Evaluation

A. Software Architecture

To evaluate the proposed topology models, a Java program that computes the MS topology structure corresponding to any physical network topology has been developed. The program takes as input the network topology, the identifiers of the network edge nodes and a set of configuration parameters. The latter allow the user to control the type of MS structure to compute (ring, full-mesh, hybrid), the logical link cost model, the threshold value and the initial node selection criterion. An overview of the main components of the program is depicted in Fig. 8.

B. MS Construction Algorithms Performance Evaluation

1) Experiment Settings: The impact of key parameters associated with the construction process has been evaluated and analyzed using two real PoP (Point of Presence)-level network topologies, Abilene [20] and GEANT [21]. The PoPs, in each topology, are mapped to cities, which allows to determine the geographical distance between every pair of nodes (Google Maps was used). While the full 11-node topology is used in the case of the Abilene network, a reduced GEANT topology, which excludes the two non-European PoPs, i.e. 21 instead of 23 nodes, is considered.

2) Ring Construction: To evaluate the performance of the ring construction algorithm described in section V-B, the total ring cost obtained for a set of nodes is compared to the cost of the optimal ring structure [18]. The optimal ring structure is computed using the GLPK (GNU Linear Programming Kit) linear programming/mixed integer programming solver [22]. The performance obtained by a method that randomly connects the nodes in a ring is also considered. To analyze the influence of the number of nodes on the performance of the algorithm, experiments using different number of nodes in both the Abilene and GEANT networks are performed. A subset of nodes is randomly selected and connected into a ring according to the three approaches considered. To obtain a better approximation of the cost in the random case, the number of executions of the algorithm is proportional to the number of nodes and the results are averaged. The logical link cost is equal to the product of the administrative link weight and the geographical distance.

The deviation of the total ring cost obtained with the proposed and the random algorithms from the optimum, for different number of nodes, is depicted in Fig. 9. The deviation increases linearly with the number of nodes for both algorithms. Given that the proposed approach follows an iterative process where nodes are iteratively added to the ring structure, the error made at each iteration in the choice of a successor node incurs a cost penalty to the total ring cost. The penalty increases as the number of nodes to consider increases, and, as a result, the deviation from the optimum increases. It can be noticed, however, that the proposed algorithm outperforms the random one since the deviation is significantly lower in all cases.

3) Multiple Rings Construction: This subsection provides an analysis of how the logical link cost C_{LL}, the threshold θ and the initial node selection criterion can influence the structure of the sub-rings (i.e. number and size) computed according to the algorithm described in section VI-B1. A comparison of the structures obtained when using the threshold values $\theta_{HalfMax}$ and θ_{Avg}, and the initial node selection criterion lowest C_{LL} and highest C_{LL}, as explained in section VI-B1, is performed. In addition, two different cases for the logical link cost are considered: a) the administrative link weight of each network link is set to 1 (i.e. in this case, the C_{LL} is mainly driven by the geographical distance between the nodes), and, b) the administrative link weights are the original ones. These are denoted as $Model_{1*D}$ and $Model_{W*D}$, respectively. The sub-ring structures obtained in 8 different scenarios in both the 11-node Abilene network and the 21-node GEANT network are analyzed.

To compare the different scenarios, three metrics are defined to describe the characteristics in terms of size of the different sub-rings obtained in each scenario. The $MinDeviation$ metric is equal to the ratio of the minimum total ring cost among the different sub-rings to the threshold value θ. In a similar fashion, the $MaxDeviation$ metric is defined as the ratio of the maximum total ring cost to the threshold value θ. Finally the $AvgDeviation$ metric is defined as the ratio of the average

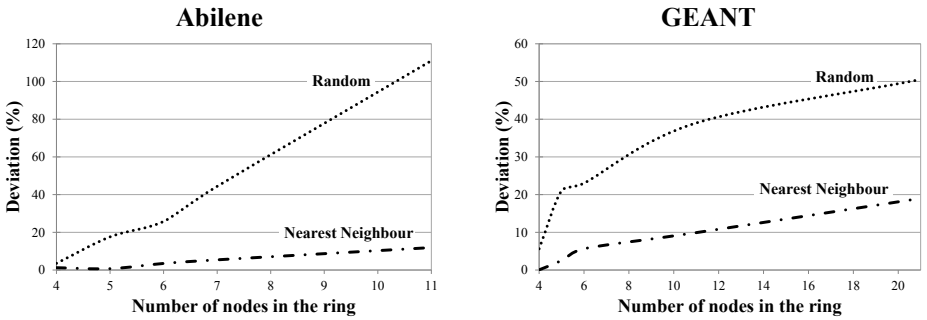

Fig. 9. Evolution of the deviation from the optimum in case of the Abilene and GEANT networks.

total ring cost between the different sub-rings to the threshold value θ. The results obtained for each scenario in the case of Abilene and GEANT are reported in Table I.

As observed, using $Model_{1*D}$ to set the logical link cost leads on average to the formation of more sub-rings in both Abilene and GEANT, although the difference is smallest in the Abilene case given the small size of the network. In the case of GEANT, it can be noticed that using the threshold value $\theta_{HalfMax}$ results in sub-rings that are more balanced in terms of size, especially when $Model_{W*D}$ is used. More precisely, when comparing the ratio between $\theta_{HalfMax}$ and θ_{Avg} in the case of $Model_{W*D}$, it can be observed that its value is around 3 for GEANT, whereas it is equal to 1.15 for Abilene. As a result, the structure of the sub-rings obtained are strongly affected by the value of θ in the case of GEANT. In the case of $Model_{1*D}$, the ratios are 1.18 and 1.04 for GEANT and Abilene, respectively, and as shown in the table, the structure of sub-rings is less affected by the choice of the threshold value. In addition, it can be noticed that the structure of the sub-rings is not significantly affected by the initial node selection criterion in all the cases.

4) Intermediate Entity Selection: An analysis of how the logical link cost C_{LL} and the threshold θ can influence the intermediate entity selection in each sub-ring is finally presented. Here, the highest C_{LL} is used as the initial node selection criterion. In a similar fashion to previous experiments, a metric is defined to compare the different scenarios. This, denoted $Deviation$, represents the ratio of the maximum logical link cost between the different IE nodes to the value of the average ring cost obtained in the corresponding scenario in Table II.

As observed, the value of $Deviation$ increases with the number of sub-rings. A larger number of sub-rings means that more clusters were formed during the multiple rings construction process. As such, the distance between clusters tends to increase. In addition, it can be observed that the value of $Deviation$ is higher in the case of the Abilene network, which shows that the total cost in the sub-rings is on average smaller than the cost between the different sub-rings.

5) Time Complexity Analysis: The time complexity of the proposed construction algorithms is influenced by the number of nodes in the management substrate, i.e. by the number of network edge nodes. It is necessary for each algorithm to determine the distance between each pair of MS nodes, which is equivalent to obtaining the shortest paths between all pair of nodes. This can be achieved with the Floyd-Warshall's algorithm [18] that can be implemented with a running time of $O(N^3)$, where N is the number of nodes in the substrate. The time complexity of the ring construction algorithm is $O(N^2)$ [19]. In the case of the multiple rings construction algorithm, the time complexity depends both on the number of nodes in the substrate and the number of sub-rings that are formed. The latter is driven by the factor θ, i.e. by the geographic distribution of edge nodes in the network. For each MS node, the algorithm executes two main methods: it determines the closest neighbor of the node in the list of the MS nodes that have not been considered yet and it applies the ring construction algorithm. Given that the minimum number of nodes allowed in each sub-ring is equal to 2, the maximum number of sub-rings that can be obtained is equal to $\frac{N}{2}$. In that case, the time complexity of the ring construction algorithm executed at each iteration is constant and equal $O(2)$, and the time complexity of the multiple rings construction algorithm is dominated by the size of the list of nodes to compare against at each iteration, i.e. $\frac{N(N+1)}{2}$ which gives a time complexity of $O(N^2)$. On the contrary, in the case where one sub-ring only is formed, the running time of the algorithm is mainly driven by the complexity the ring construction algorithm, which is affected by the size of the sub-ring at each iteration. This gives a time complexity of $O(N^3)$. The average time complexity can be obtained by considering that there is an equal number of nodes in each sub-ring. In that case, the time complexity of the algorithm can be defined as $O(\frac{N^3}{r^2}+N^2)$ with r the number of sub-rings. Finally the intermediate entity selection algorithm consists in determining and comparing the distance between every nodes in the substrate, and as such has a running time of $O(N^2)$.

The network environment considered in this work consists of intra-domain fixed backbone network for which the maximum number of edge nodes are usually in the order of hundreds [23] [24]. In addition the construction of the management substrate is an offline process that is executed during the setup phase of the network. As such, the proposed algorithms can provide

TABLE I
MULTIPLE RINGS EVALUATION.

Cost C_{LL}	Threshold θ	Initial Node	Rings Size	Min Deviation	Max Deviation	Avg Deviation
GEANT Topology						
$Model_{1*D}$	θ_{Avg}	Lowest	4,6,3,2,2,2,2	0.71	1.92	1.20
$Model_{1*D}$	θ_{Avg}	Highest	5,2,3,3,3,3,2	0.53	2.47	1.16
$Model_{1*D}$	$\theta_{HalfMax}$	Lowest	4,6,3,2,2,2,2	0.60	1.62	1.01
$Model_{1*D}$	$\theta_{HalfMax}$	Highest	5,2,2,3,2,5,2	0.56	2.08	0.90
$Model_{W*D}$	θ_{Avg}	Lowest	9,6,3,3	1.00	8.98	4.53
$Model_{W*D}$	θ_{Avg}	Highest	2,5,7,5,2	0.80	8.10	3.61
$Model_{W*D}$	$\theta_{HalfMax}$	Lowest	15,6	2.39	3.01	2.70
$Model_{W*D}$	$\theta_{HalfMax}$	Highest	15,6	2.39	3.01	2.70
Abilene Topology						
$Model_{1*D}$	θ_{Avg}	Lowest	3,2,3,3	0.28	1.68	1.05
$Model_{1*D}$	θ_{Avg}	Highest	3,2,3,3	0.67	0.96	0.82
$Model_{1*D}$	$\theta_{HalfMax}$	Lowest	3,2,3,3	0.27	1.60	1.00
$Model_{1*D}$	$\theta_{HalfMax}$	Highest	3,2,2,2,2	0.22	1.36	0.71
$Model_{W*D}$	θ_{Avg}	Lowest	5,3,3	0.78	1.90	1.26
$Model_{W*D}$	θ_{Avg}	Highest	3,2,3,3	0.78	1.12	0.95
$Model_{W*D}$	$\theta_{HalfMax}$	Lowest	5,3,3	0.67	1.63	1.09
$Model_{W*D}$	$\theta_{HalfMax}$	Highest	3_j,2,3,3	0.67	0.96	0.82

TABLE II
INTERMEDIATE ENTITY SELECTION.

Cost C_{LL}	Threshold θ	Rings Size	Selected IE Nodes	Deviation
GEANT Topology				
$Model_{1*D}$	θ_{Avg}	5,2,3,3,3,3,2	16,21,1,20,17,7,2	1.310
$Model_{1*D}$	$\theta_{HalfMax}$	5,2,2,3,2,5,2	10,16,21,1,20,17,2	1.28
$Model_{W*D}$	θ_{Avg}	2,5,7,5,2	21,17,1,7,20	0.120
$Model_{W*D}$	$\theta_{HalfMax}$	15,6	9,8	0.013
Abilene Topology				
$Model_{1*D}$	θ_{Avg}	3,2,2,2,2	2,6,11,7,10	2.753
$Model_{1*D}$	$\theta_{HalfMax}$	3,2,2,2,2	2,6,11,7,10	2.753
$Model_{W*D}$	θ_{Avg}	3,2,3,3	2,6,7,11	1.678
$Model_{W*D}$	$\theta_{HalfMax}$	3,2,3,3	2,6,7,11	1.678

satisfactory performance in terms of time complexity and can be applied in practice by the network operators.

C. Communication Cost Analysis

This section investigates how the number of nodes in the substrate, as well as the geographical distribution of the nodes, can affect the performance of each topology structure in terms of communication cost (delay). The ring structure considered in this section is obtained using the ring construction algorithm described in section V-B, and the hybrid structure is obtained with the following parameters: θ_{Avg}, $Model_{1*D}$ and highest C_{LL} as the initial node selection criterion, as defined in section VII-B.

1) Round Trip Time Measurements: In order to evaluate the delay, the value of the round trip time (RTT) between pairs of nodes in the Abilene and the GEANT networks have been measured using the looking glass service available for these networks [25] [26]. The evolution of the values of the RTT according to the geographical distance between the nodes is shown in Fig. 10. As observed, the value of the RTT linearly increases with the distance, which is consistent with previous studies [16] [17].

2) Comparison of the MS Communication Costs: Based on the values of the RTT, experiments are performed to compare the communication cost in terms of the maximum delay incurred in each topology structure. The communication cost is defined according to the communication model used in each structure as follows. In the case of the full-mesh model, the communication cost is given by the maximum delay between any pair of nodes in the substrate. In the case of the ring model, this corresponds to the total time required for a message to travel around the ring. In the hybrid model, the communication cost is defined as the sum of the maximum delay between IE nodes and the maximum sub-ring cost. As such, the values considered in each model represent an upper bound in terms of delay for each of the structures.

The experiments consist of a set of substrate configurations defined for each network topology (i.e. Abilene and GEANT),

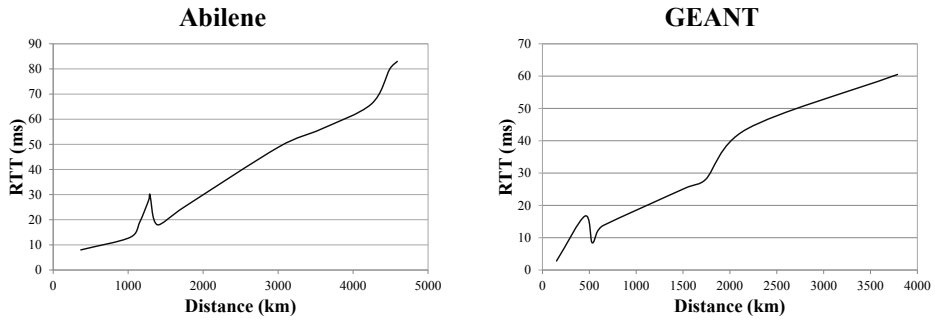

Fig. 10. Evolution of the RTT as a function of the distance in case of the Abilene and GEANT networks.

so that for each configuration, a number of nodes (N) is randomly selected and connected according to the proposed structures. N varies from 4 to the total number of nodes in the network. In the case of 3 nodes and less, the different models lead to identical structures. In order to represent various geographical distributions of nodes in the substrate, a large number of configurations (i.e. choice of network nodes) is considered for each value of N. The results of the experiments are shown in Fig. 11 and Fig. 12, where the communication cost (maximum delay) obtained in each topology structure for different number of substrate nodes is presented for the Abilene and GEANT networks, respectively.

As it can be observed, the communication cost obtained for a given number of nodes N forms, in all the cases, a vertical line parallel to the y-axis. This confirms that the maximum delay is influenced by the geographical distribution of the nodes. It can be noted, however, that in the case of the full-mesh model, there exists an upper bound in terms of the maximum value that the delay can reach, which is independent of the number of nodes in both the Abilene and GEANT networks. As a result, it can be deduced that the communication cost is mainly driven by the maximum distance between MS nodes in the case of the full-mesh model. In the case of the ring model, it can be observed that the maximum delay does not only depend on the geographical distribution of nodes but is also driven by the number of nodes in the substrate. The value of the communication cost increases as the number of substrate nodes increases. While the cost in the hybrid model is also affected by the number of nodes, the influence of this factor is much lower compared to the ring model.

3) Average Deviation Analysis: The deviation between the communication cost obtained with (i) the hybrid model and the full-mesh model, and (ii) the ring model and the full-mesh model is finally investigated. Fig. 13 represents the average deviation in percentage obtained in each case for the Abilene and the GEANT networks.

As observed, the average deviation is always positive[1], which shows that the full-mesh model performs better than the two other approaches in terms of communication cost. In addition, it can be noted that the performance obtained by the three models is similar when the number of nodes is small, especially in the case of the Abilene network where the average deviation is below 10%. As the number of nodes increases, however, the full-mesh model outperforms the two other models. The performance of the ring model, in particular, becomes much worse than that of the full-mesh when the number of substrate nodes increases. A deviation of more than 100% is obtained in the case of the GEANT network with 21 MS nodes. In contrast, the difference in terms of performance between the hybrid model and the full-mesh model is much less significant. A maximum of 30% deviation can be observed in the case of Abilene and around 25% in the case of GEANT. It is finally interesting to note that the average deviation between the hybrid and the full-mesh model is not significantly affected by the number of nodes in the substrate.

[1]The average deviation obtained between the hybrid and full-mesh models is close to zero when the number of nodes is smaller than 7 and as such, these results are not visible in the figure.

D. MS Structure Comparison Summary

As shown in the previous section, the performance of the ring model, in terms of communication cost, is very sensitive to the number of nodes in the substrate, i.e. to the number of decision-making points, and as such, this model presents some scalability limitations. In contrast, the performance of the full-mesh and the hybrid models are mainly affected by the geographical coverage of the network. In order to efficiently support the adaptive resource management scheme, the delay incurred by the communication between the decision-making points needs to be kept small. Therefore, the choice of a ring structure is not recommended when the number of substrate nodes exceeds 5 or 6.

In addition to the performance in terms of delay, the topological characteristics of the structures can also influence the choice of the model to apply. Due to its high degree of connectivity, the full-mesh topology model provides a greater flexibility to select which nodes to communicate with. The full-mesh structure, however, requires that every node in the substrate maintains locally information about every other MS node, which may also raise some scalability limitations. The hybrid model offers a trade-off between the full-mesh and the ring models. It outperforms the ring model in terms of communication cost and

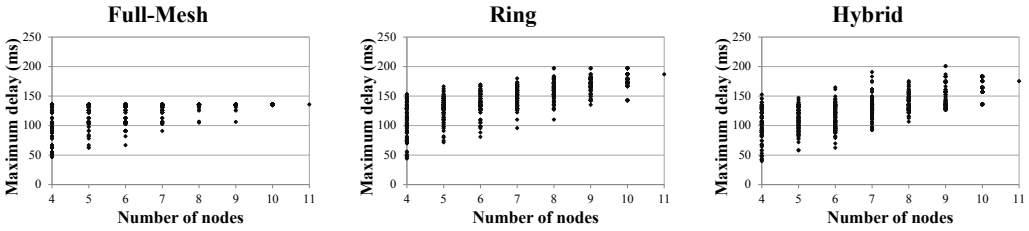

Fig. 11. Evolution of the maximum delay according to the number of MS nodes in case of the Abilene network.

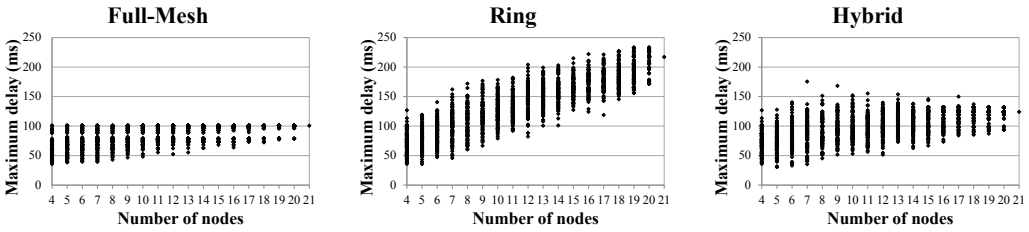

Fig. 12. Evolution of the maximum delay according to the number of MS nodes in case of the GEANT network.

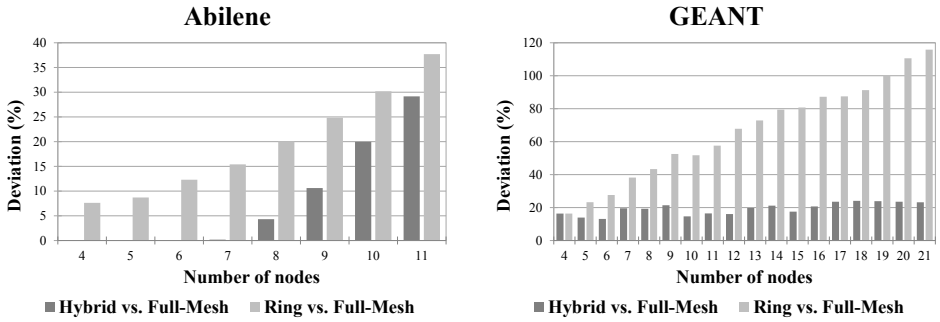

Fig. 13. Evolution of the deviation between the communication cost obtained with the hybrid model and the full-mesh model, and the ring model and the full-mesh model in case of the Abilene and GEANT networks.

can also offer competitive performance compared to the full-mesh approach. Furthermore, it can overcome the limitations of the full-mesh model by restricting the volume of substrate information that needs to be maintained at each MS node.

The characteristics of the three models investigated in this chapter are summarized in Table III. The total number of nodes in the substrate is noted N and the total number of sub-rings in the hybrid model is noted r.

VIII. Conclusion

This paper presents a new adaptive in-network management framework. The proposed framework relies on a decentralized decision-making process distributed over the network edge nodes. These decide in a coordinated fashion which configuration changes to apply, according to current conditions in the network, so that network resources can be better utilized. In order to support this decentralized and coordinated decision-making process, the concept of management substrate to organize the nodes participating into the reconfiguration process is introduced. The management substrate is defined as a logical infrastructure that facilitates the exchange of information between decision-making points. Three different topology structures to organize the nodes in the management substrate are investigated. The characteristics of each model are described and compared and a set of offline algorithms that can be used in practice to construct the proposed structures are designed. The choice of the different design parameters are discussed and evaluated. Finally, a thorough quantitative and qualitative evaluation of the impact of key parameters (i.e. the number of nodes in the substrate and their geographical distribution) on the performance of the different models in terms of communication overhead is performed. The analysis of the results of the evaluation provides useful indications about the use of a particular structure for specific network settings and characteristics. The results show that the use of a ring model to connect the substrate nodes can lead to poor performance in terms of communication cost.

TABLE III
COMPARISON OF THE PROPOSED MANAGEMENT SUBSTRATE TOPOLOGY MODELS.

Model	Full-mesh	Ring	Hybrid
Level of hierarchy	Flat structure	Flat structure	One level hierarchy
Number of neighbours	N-1	2	$\min = 2$ $\max = 2 + (r - 1)$
Communication model	Star fashion mechanism	Hop-by-hop mechanism	Star & hop-by-hop mechanisms
Communication cost	Driven by the geographical distance	Driven by the number of nodes	Driven by both factors

REFERENCES

[1] D. Tuncer, M. Charalambides, G. Pavlou, and N. Wang, "DACoRM: A coordinated, decentralized and adaptive network resource management scheme," in *the proceedings of the IEEE Network Operations and Management Symposium (NOMS'12)*, 2012, pp. 417–425.

[2] M. Charalambides, D. Tuncer, L. Mamatas, and G. Pavlou, "Energy-aware adaptive network resource management," in *the proceedings of the IFIP/IEEE International Symposium on Integrated Network Management (IM'13)*, 2013, pp. 369–377.

[3] D. Tuncer, M. Charalambides, R. Landa, and G. Pavlou, "More Control Over Network Resources: An ISP Caching Perspective," in *the proceedings of the 9th IEEE/IFIP International Conference on Network and Service Management (CNSM'13)*, 2013.

[4] S. White, J. Hanson, I. Whalley, D. Chess, and J. Kephart, "An architectural approach to autonomic computing," in *the proceedings of the International Conference on Autonomic Computing (ICAC'04)*, 2004, pp. 2–9.

[5] N. Samaan, "Achieving self-management in a distributed system of autonomic but social entities," in *Modelling Autonomic Communications Environments*, ser. Lecture Notes in Computer Science, S. Meer, M. Burgess, and S. Denazis, Eds., 2008, vol. 5276, pp. 90–101.

[6] N. Tcholtchev, M. Grajzer, and B. Vidalenc, "Towards a unified architecture for resilience, survivability and autonomic fault-management for self-managing networks," in *the proceedings of the 2009 International Conference on Service-Oriented Computing (ICSOC/ServiceWave'09)*, 2009, pp. 335–344.

[7] E. Lavinal, T. Desprats, and Y. Raynaud, "A generic multi-agent conceptual framework towards self-management," in *the proceedings of the 10th IEEE/IFIP Network Operations and Management Symposium, (NOMS'06)*, April 2006, pp. 394 –403.

[8] D. Gracanin, S. Bohner, and M. Hinchey, "Towards a model-driven architecture for autonomic systems," in *the proceedings of the 11th IEEE International Conference and Workshop on the Engineering of Computer-Based Systems*, 2004.

[9] M. Jelasity, A. Montresor, and O. Babaoglu, "Gossip-based aggregation in large dynamic networks," *ACM Transactions on Computer Systems*, vol. 23, no. 3, pp. 219–252, Aug. 2005.

[10] F. Wuhib, R. Stadler, and M. Spreitzer, "Gossip-based resource management for cloud environments," in *the proceedings of the 2010 International Conference on Network and Service Management (CNSM'10)*, 2010, pp. 1–8.

[11] R. Makhloufi, G. Doyen, G. Bonnet, and D. Gaiti, "Situated vs. global aggregation schemes for autonomous management systems," in *the proceedings of the 4th IFIP/IEEE Workshop on Distributed Autonomous Network Management Systems*, 2011, pp. 1135–1139.

[12] E. K. Lua, J. Crowcroft, M. Pias, R. Sharma, and S. Lim, "A survey and comparison of peer-to-peer overlay network schemes," *Communications Surveys Tutorials, IEEE*, vol. 7, no. 2, pp. 72 – 93, quarter 2005.

[13] J. Risson and T. Moors, "Survey of research towards robust peer-to-peer networks: search methods," *Computer Networks*, vol. 50, pp. 3485–3521, December 2006.

[14] S. Uhlig, B. Quoitin, J. Lepropre, and S. Balon, "Providing public intradomain traffic matrices to the research community," in *the proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM'06)*, vol. 36, January 2006, pp. 83–86.

[15] J. Postel, "Transmission Control Protocol," RFC 793 (Standard), Internet Engineering Task Force, Sep. 1981, updated by RFCs 1122, 3168, 6093. [Online]. Available: http://www.ietf.org/rfc/rfc793.txt

[16] S. Agarwal and J. R. Lorch, "Matchmaking for online games and other latency-sensitive P2P systems," vol. 39, no. 4, pp. 315–326, Aug. 2009.

[17] Y. Zhu, C. Dovrolis, and M. Ammar, "Combining multihoming with overlay routing (or, how to be a better isp without owning a network)," in *the proceedings of the 26th IEEE International Conference on Computer Communications (INFOCOM'07)*, may 2007, pp. 839 –847.

[18] C. H. Papadimitriou and K. Steiglitz, *Combinatorial optimization: algorithms and complexity*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1982.

[19] Y. Crama, A. Kolen, and E. Pesch, "Local search in combinatorial optimization," in *Artificial Neural Networks*, ser. Lecture Notes in Computer Science, P. Braspenning, F. Thuijsman, and A. Weijters, Eds. Springer Berlin Heidelberg, 1995, vol. 931, pp. 157–174.

[20] "The Abilene Internet 2 Topology," http://www.internet2.edu/pubs/200502-IS-AN.pdf.

[21] "The GEANT topology," 2004, http://www.dante.net/server/show/nav.007009007.

[22] "GNU Linear Programming Kit (GLPK)," http://www.gnu.org/software/glpk/.

[23] N. Spring, R. Mahajan, D. Wetherall, and T. Anderson, "Measuring ISP topologies with rocketfuel," *IEEE/ACM Transactions on Networking*, vol. 12, no. 1, pp. 2–16, Feb. 2004.

[24] S. Knight, H. Nguyen, N. Falkner, R. Bowden, and M. Roughan, "The Internet Topology Zoo," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 9, pp. 1765–1775, 2011.

[25] "Abilene: the Internet2 Router Proxy Service," http://routerproxy.grnoc.iu.edu/internet2/.

[26] "The Backbone GEANT Looking Glass," https://tools.geant.net/portal/links/lg/index.jsp.