MQTT-ST: a Spanning Tree Protocol for Distributed MQTT Brokers

Edoardo Longo*, Alessandro E.C. Redondi*, Matteo Cesana*, Andrés Arcia-Moret† and Pietro Manzoni‡
*DEIB, Politecnico di Milano, Italy - Email: {edoardo.longo, alessandroenrico.redondi, matteo.cesana}@polimi.it
†Computer Laboratory, University of Cambridge, UK Email: andres.arcia@cl.cam.ac.uk
‡DISCA, Universitat Poltcnica de Valncia, Spain - Email: pmanzoni@disca.upv.es

***Abstract*—MQTT, one of the most popular protocols for the IoT, works according to a publish/subscribe pattern in which multiple clients connect to a single broker, generally hosted in the cloud. However, such a centralised approach does not scale well considering the massive numbers of IoT devices forecasted in the next future, thus calling for distributed solutions in which multiple brokers cooperate together. Indeed, distributed brokers can be moved from traditional cloud-based infrastructure to the edge of the network (as it is envisioned by the upcoming MEC technology of 5G cellular networks), with clear improvements in terms of latency, for example. This paper proposes MQTT-ST, a protocol able to create such a distributed architecture of brokers, organized through a spanning tree. The protocol uses in-band signalling (i.e., reuses MQTT primitives for the control messages) and allows for full message replication among brokers, as well as robustness against failures. We tested MQTT-ST in different experimental scenarios and we released it as open-source project to allow for reproducible research.**

***Index Terms*—MQTT, distributed pub/sub, Mobile Edge Computing**

I. Introduction

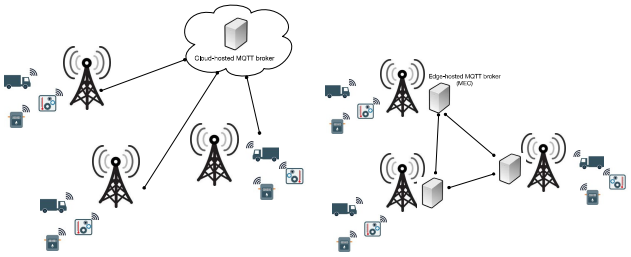
The advent of the 5th Generation (5G) of mobile cellular networks will boost the development and implementation of large scale, city-wide IoT applications. Indeed, two main 5G innovation pillars have been designed precisely for accommodating IoT requirements: massive Machine Type Communication (mMTC) and Multi-Access Edge Computing (MEC). On the one hand, mMTC will enable connection densities in the order of 10^6 low-power devices per square kilometre, with enormous implications on the amount of traffic generated and transmitted. On the other hand, MEC technology will bring computational power, storage resources and service infrastructures to the edge, alleviating the resources needed in the core network and reducing latency.

Such a dramatic change will also have a great impact on the communication protocols associated with the IoT ecosystem. While at the lower layers of the stack the plethora of short-range low-power protocols (e.g., the IEEE802.15.4 family) will have to compete with 5G-based solutions for survival, at the application layer the MQTT (Message Queuing Telemetry Transport) protocol is living its greatest period of popularity since its introduction and it can be considered the de-facto standard for IoT solutions[1].

[1]Not by chance, the four major cloud computing services up to date (Amazon AWS, Google Cloud Platform, IBM Cloud and Microsoft Azure) all adopt MQTT as protocol for connecting IoT devices to their endpoints.

MQTT is a lightweight publish/subscribe protocol designed around a central *broker*. Clients connect to the broker and subscribe to or publish data on specific *topics*. The broker is in charge of forwarding the data published to the clients interested in it, thus decoupling the process of data generation and consumption both in space and in time. This aspect, combined with the protocol simplicity at the client side and the support for reliability and quality of service (QoS), makes MQTT ideal for resource-constrained applications and motivates its great popularity. However, MQTT still remains a centralised protocol which nicely fits classical cloud-based architectures, in which all IoT devices connect to a single broker endpoint. This picture is partially at odds with the one envisioned by 5G, in which cloud services (including any broker instance) are moved to the edge, closer to the user devices. For MQTT this means moving from the current centralised, star-shaped, single-broker topology to a distributed, multi-broker topology which can cope with the massive numbers of devices envisioned to be served (see Figure 1).

In our previous position paper [1] we analysed the main research challenges and possible solutions to scale up a pub/sub architecture for upcoming 5G networks, and we presented our view on system design and optimisation. In this paper we tackle the problem from an implementation perspective: we observe that due to the intrinsic nature of the pub/sub pattern, bridging brokers may result in potentially harmful message loops. Existing solutions solve the problem by imposing a static tree-based topology among brokers, which lacks adaptivity and robustness. Therefore, in this paper we propose and implement MQTT-ST, a protocol inspired by STP (Spanning Tree Protocol) for interconnecting MQTT brokers automatically in a loop-free topology, in order to distribute messages among them. The protocol uses in-band signalling (i.e., all control messages are embedded in MQTT native mechanisms) and allows to have full message replication as well as robustness against node failures. We test MQTT-ST in different experimental scenarios, focusing on latency, computational load and achievable throughput and we compare it with traditional cloud-based, single-broker approaches. Finally, we release MQTT-ST as open-source project based on the popular Eclipse Mosquitto MQTT broker, in order to allow for reproducible research.



(a) (b)

Figure 1. Centralised broker scenario (a) and distributed architecture (b).

II. System Overview

A. Motivation

Some existing MQTT broker implementations (e.g. Mosquitto, CloudMQTT, HiveMQ) allow the use of bridging, i.e., a direct connection between brokers. The feature allows a broker `B` to connect to another broker `A` as a standard client, subscribing to all or a subset of the topics published by clients to `A`. Unfortunately, such a procedure is prone to message loops among brokers: indeed, the existence of a cycle where a message is continuously republished by the participating brokers, can quickly deplete a broker's resources, ultimately making it unable to deliver meaningful traffic. Due to the enormous complexity of implementing duplicate detection in distributed scenarios, which would require to keep track of the original producer of every message received and forwarded by any broker, existing solutions require to manually configure the connections between brokers in a loop-free topology, i.e., a tree. However, such a manual configuration of MQTT bridges has two main drawbacks: first, similarly to wiring switches in small- or medium-size enterprises, it can become a very confusing task with a high chance of creating accidental duplicate connections, especially in large topologies. Second, by enforcing a loop-free static topology among brokers, adaptivity and robustness to failures are completely lost. Another option is to rely on an automatic way to set up a tree among brokers. In switched networks, this is obtained using the Spanning Tree Protocol (STP) and/or its following amendments: the purpose of this paper is to embed STP mechanisms into MQTT, ultimately allowing for the creation of a loop-free, dynamic and robust network of brokers.

B. Spanning Tree Protocol

The Spanning Tree Protocol (standardised as IEEE 802.1D) is a distributed protocol which creates a logical spanning tree over a meshed network of layer 2 switches. Such a spanning tree is obtained by electing a root switch and blocking some of the output ports of the other switches: blocked ports do not forward data frames, thus avoiding broadcast storms. In order to agree on the root node and on which ports should be blocked, switches exchange control packets known as BPDU (Bridge Protocol Data Unit).

In a nutshell, the main steps of STP are the following:

- At startup, each node sets itself as root and start broadcasting BPDU. Each BPDU contains among other parameters) the identifier of the node and the transmitting port, the identifier of the current root node selected by the transmitting node and the root patch cost. The node identifier, composed of both the node MAC address and a configurable priority value is used for root selection: the node with the lowest identifier is elected as root.
- Upon the reception of a BPDU, a node reconfigures its state by modifying the identifier of the (believed) root node and updating the root port (port that leads to the least cost path to the root). The rest of the active ports are labeled as either designated (used for forwarding traffic) or blocked. To avoid loops, nodes agree on which port should be designated or blocked, based again on the least-cost path to the root or the lowest identifier, in case of ties.
- BPDUs are periodically transmitted by the root, and forwarded by all other nodes, in order to keep the topology updated. Upon failures on active links, special BPDU known as Topology Change Notification (TCN) can be transmitted by non-root nodes to inform all the others.

Enhancements of this general scheme were proposed and standardised in several standards, including the Rapid Spanning Tree Protocol (RSTP, IEEE802.1D-2004), which provides significantly faster spanning tree convergence, the Multiple Spanning Tree Protocol (MSTP, IEEE802.1Q-2005), which allows the creation of multiple spanning trees for different VLANs or group of VLANs and recently the Shortest Path Bridging protocol (IEEE 802.1aq), which allows redundant links between switches to be active at the same time, thus increasing bandwidth.

C. MQTT-ST

We develop MQTT-ST starting from the latest MQTT protocol specifications [2]. The main changes and modifications to the protocol are reported below.

1) Connection phase: At startup, a broker willing to create a bridge with another broker transmits a MQTT CONNECT message. The address and port of the broker (or brokers) to connect to is specified in a configuration file. To inform a broker that the connection request comes from another broker and not from a client, we set the most significant bit of the Protocol Version byte in the CONNECT header. Upon the reception of a CONNECT message with such a bit set, a broker performs the following operations:

- First, to allow bidirectional communication, the broker transmit back a modified CONNECT message to the originating node, using the standard MQTT port 1883. Note also that this allows a node with no configuration file set to be part of the broker network, if contacted by an already connected broker.
- The broker stores the IP addresses of all directly connected brokers in a local table, which is used to keep track of the state of each connection marked as root, designated or blocked. For each connection, the table also stores the

average Round Trip Time (RTT) and a value C, which summarises the resource capability of the endpoint broker as detailed later.
- Finally, the broker sets itself as root and start transmitting signalling messages towards all connected brokers. Instead of creating a new specific message, we reuse MQTT PINGREQ messages.

2) Signalling phase: Standard MQTT specifies a Keep Alive parameter, which defines the maximum time interval permitted to elapse after the last client transmission. In case the timer expires, the broker closes the connection with the client. Therefore, to maintain the connection alive, a client transmits periodical PINGREQ messages. MQTT-ST reuses such messages, which play the role of STP BPDUs. In details, the following information is appended to PINGREQ messages: IP address of the current root broker for the transmitter, the broker capability value C, and a root path cost P. The latter two fields are used for root selection and path computation, respectively.

3) Root selection: The root broker plays a crucial role in the broker tree, as it is the relay node for all traffic and it is therefore subject to an increased computational load. Indeed, selecting a broker with poor or overloaded resources may result in poor overall performance. In STP the root is selected based only on its identifier, which does not suit well the scenario under consideration. In MQTT-ST, instead, the root broker is selected according to the capability value C, defined as:

$$C = \alpha L + \beta M \tag{1}$$

where L is the broker CPU speed, R is the amount of RAM memory and α, β are tuneable conversion parameters[2]. In case of tie, the broker with the lowest IP address is selected as root.

4) Path computation: In STP, each node selects the best path to the root according to a bandwidth-related criteria, in order to avoid the use of reduced capacity links in the tree which may slow down an entire network. For MQTT-ST we observe that latency, rather than bandwidth, plays a critical role. Each broker therefore continuously monitor the RTT to other brokers and uses that value for updating the root path cost P. In order to do this, we leverage the request/response mechanism already present in MQTT through the PINGREQ/PINGRESP. A timer is started when the client transmits the PINGREQ message and it is stopped when the corresponding PINGRESP message is received by the broker, providing an estimate of the current RTT. Upon reception of a PINGREQ message, the connection providing to the lowest latency path towards the root is marked as root connection. In case of ties in the cumulative latency to the root, the connection passing through the broker with the highest amount of resources is selected as root connection. All the other connections are labeled following the same logic of the STP protocol: (i) all connections of the root broker as marked as designated, (ii) the non-root connections of other brokers are marked as designated if the broker as a better path cost (or a better value of C in case of tie) compared to their neighbouring broker and (iii) all other ports are labeled as blocked.

5) Runtime behaviour: At runtime, a MQTT-SN broker works exactly like a MQTT broker from the perspective of the connected clients. Moreover, for every published message on any topic, the broker forwards it on its non-blocked connections (root and designated), while any message incoming on a blocked connection is discarded. The forwarding is performed like a standard MQTT PUBLISH message. We highlight that:

- To allow full replication of the messages published at one broker to all other brokers, message forwarding is performed with the highest MQTT QoS available (QoS = 2). On the one hand, this guarantees that each message is received only once by the intended recipients, while on the other hand it requires a four-part handshake with a non-negligible associated overhead.
- Since forwarding is implemented through a standard MQTT PUBLISH, all other features of the latest MQTT specification are conserved (e.g., retain, topic alias, message expiry interval, etc.)
- MQTT-ST also forwards Last Will and Testament (LWT) messages, which are automatically generated by a broker upon ungraceful disconnection of a client.

6) Reaction to failures: Upon a broker failure, MQTT-SN handles the corresponding socket error to re-establish the forwarding tree. In details, the broker detecting the socket error transmits a special PINGREQ message used to restart the tree construction from scratch. The broker sets itself as root and append to the message an additional Topology Change (TC) field set, similarly to what happens in STP. Any broker receiving such a message restart the root selection procedure, which eventually will converge to a new tree.

III. Experimental results

MQTT-ST has been developed in C language starting from the open-source project Eclipse Mosquitto [3], which offers both broker and client capabilities. In order to test its functionality and evaluate its performance, two simulation environments have been created.

A. Local environment

The first simulation environment is aimed at performing stress tests on the processes implementing the brokers without looking at network-related factors (e.g. link delay or link capacity). For this reason, the stress test simulation consists in a script which automatically creates a given number of MQTT-ST broker processes, all running on the same machine (Intel i7-3770 with 8 CPUs @ 3.40GHz with 16 GB of RAM, running Ubuntu 16.04.6) and connected together in a fully meshed network, i.e., each broker has connections to all other brokers. To generate client publications and subscriptions, we use a simple open-source benchmark tool[3]. Such a tool, which is executed on a different machine in the same private LAN

[2]Operatively, L and R are read by the broker at startup using the `/proc/cpuinfo` and `/proc/meminfo` system files available on Linux

[3]https://github.com/krylovsk/mqtt-benchmark

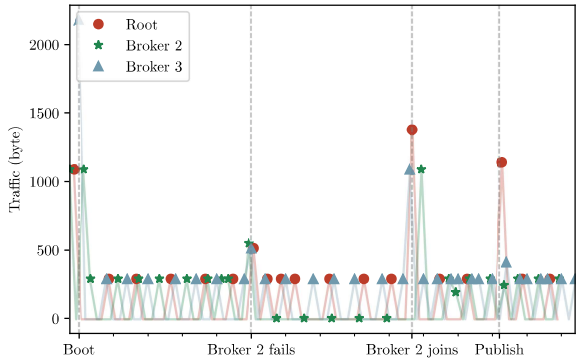


Figure 2. Traffic trace relative to different phases of MQTT-ST operating on three brokers.

of the one running the brokers, generates a given number of clients (both publishers and subscribers), allowing to tune several parameters (e.g., messages size, number of topics, etc.) and to measure the associated broker performance.

1) Signalling overhead: MQTT-ST requires brokers to exchange information periodically, therefore consuming additional resources compared to the standard MQTT. As explained in section II, we modified the PINGREQ messages in order to add BPDU information to them. This adaptation increases the size of the PINGREQ message from 78 to 114 bytes, which is acceptable especially compared to the massive amount of data traffic expected in common IoT scenarios. To better show the resource utilisation, we display in Figure 2 the overall traffic (in bytes) consumed by three MQTT-ST brokers when the Keep Alive value is set to 10s; the red line represents the root broker while the other two lines corresponds to brokers connected to the root. We highlight the start up event (Boot), the fail and the subsequent restore of the Broker 2 which triggers the execution of the STP algorithm. Finally, we show a publish event from a client to the root broker, which in turns forwards the publication to all other brokers.

2) Publication throughput: A performance measure generally used in related works [4], [5] is the publication throughput, which is the maximum speed at which a client can publish messages to a broker when using publication QoS = 2 (that is, waiting for the MQTT four way handshake acknowledgment to take place). Since the transmission of a new message needs to wait for the ACK coming from the broker, such a performance measure gives an indication of the current broker workload. We evaluate such a measure in three different scenarios:

- Benchmark: N publishers and M subscribers connected to a single, centralised standard MQTT broker. This corresponds to a traditional, cloud-based scenario.
- Distributed: N publishers and M subscribers evenly connected to K MQTT-ST brokers (i.e., each broker is connected to M/K publishers and N/K subscribers). This scenario corresponds to the envisioned distributed architecture.
- 100% Locality: N publishers and M subscribers connected to only one of the K MQTT-ST brokers. This is an extreme unfavourable scenario in which messages are distributed to all brokers, but consumed at the same broker where they are produced.

The lefthand side pictures in Figures 3, 4 and 5 show the publication throughput of the three different broker configurations (where the green curve is the average across the K distributed brokers), at different number of subscribers and for 10, 100 and 1000 publishers, respectively. As one can see, the publication throughput of the benchmark scenario rapidly decreases as the number of subscribers and publishers increases, while MQTT-ST always shows higher throughput. The 100% locality case shows always the worst performance, as it corresponds to the case where one single broker manages all the clients, in addition to forwarding their messages to all other brokers.

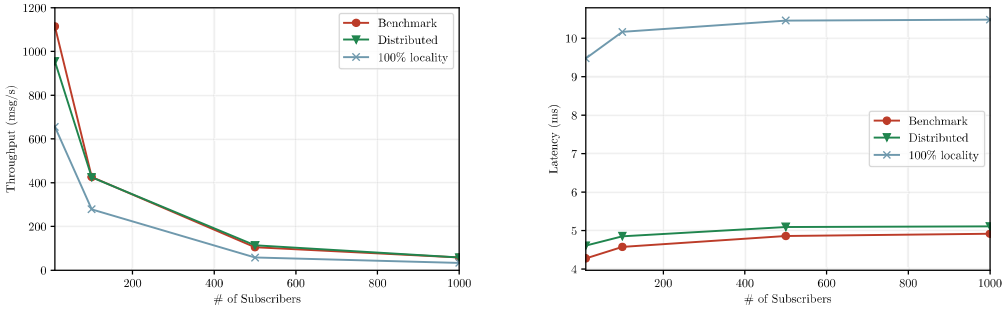
3) End-to-end delay: We also evaluate the end-to-end delay (i.e., the average amount of time elapsed between the publication of a message and its reception at a subscriber) in the same aforementioned scenarios, which is reported in the righthand side of Figures 3, 4 and 5. We can observe that (i) the end-to-end delay is sensible to the number of publishers much more than the number of subscribers; (ii) the 100% locality scenario always shows the worst performance and (iii) MQTT-ST outperforms the benchmark scenario when the number of publishers increases.

B. Networked environment

The second simulation environment is aimed at demonstrating the functionality of MQTT-ST in a networked scenario compared to a centralised cloud architecture. To this end, we leverage the capabilities of Amazon Web Services EC2 to create several brokers which run on virtual machines (VM) located in different regions of the world. We focus on a scenario with 3 brokers executed on t2.micro VM instances (1 vCPUs @ 2.40GHz, 1 GB RAM) and deployed in the following locations: US West (Oregon), US East (N. Virginia), EU (Ireland). When running the centralised scenario, only one of the brokers is chosen, and the other two are shut off. Clients and subscribers are run on two separate VMs, located in US West (N. California) and EU (Germany). For both the centralised and distributed scenarios, the following tests are performed:

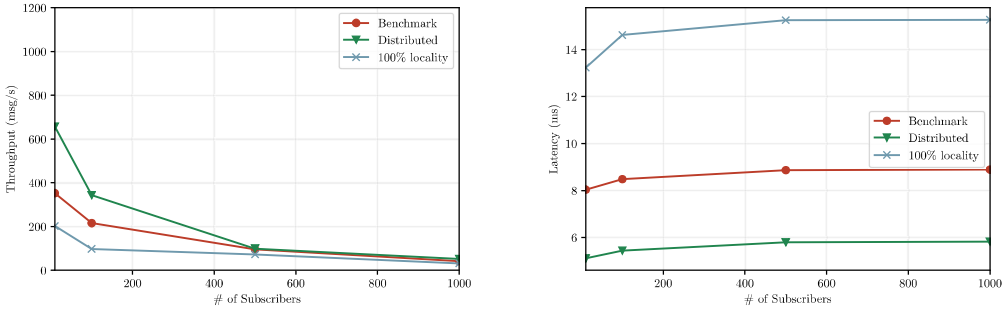
- 0% locality: 100 publishers in N. California and 100 subscribers in Germany
- 50% locality: 50 publishers and 50 subscribers in both N. California and Germany
- 100% locality: 100 publishers and 100 subscribers in N. California

Figure 6 shows the average end-to-end latency, where the cloud scenario is averaged over all possible locations of the centralised broker. As one can see, MQTT-ST allows to obtain significant latency improvement compared to centralised



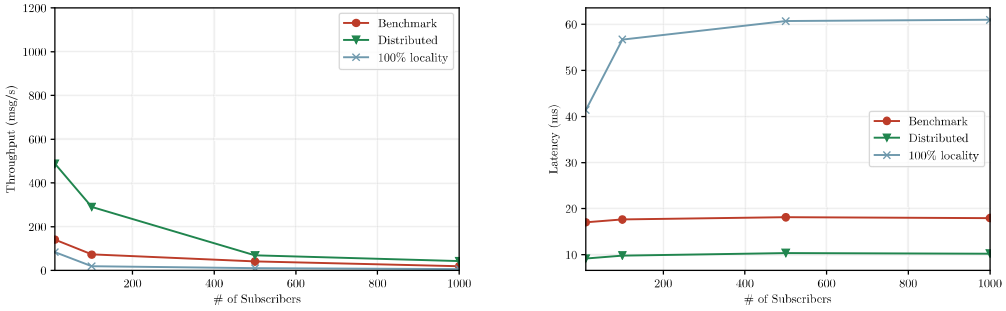
(a) (b)

Figure 3. Publication throughput (a) and end-to-end delay (b) with 10 publishers.



(a) (b)

Figure 4. Publication throughput (a) and end-to-end delay (b) with 100 publishers.



(a) (b)

Figure 5. Publication throughput (a) and end-to-end delay (b) with 1000 publishers.

solutions when publishers and subscribers are colocated. Moreover, the additional cost for full message replication in case publishers and subscribers are far from each other (0% locality) is limited to few tens of milliseconds.

IV. Related Work

The research area of message dissemination in distributed generic pub/sub system has been very active in the last 20 years. Most works focus on the development of efficient and scalable routing algorithms to create topic-based dissemination trees (in the form of multicast groups) that cover only the subscribers matching a particular topic [6]–[10]. In such works, no specific broker implementation is considered and the overlay broker topology is assumed to be known. Only very recently, motivated by the protocol popularity, some attention has been given to the problem of interconnecting MQTT-specific brokers [11]. Some works focused primarily on vertical clustering, where the single broker is replaced by

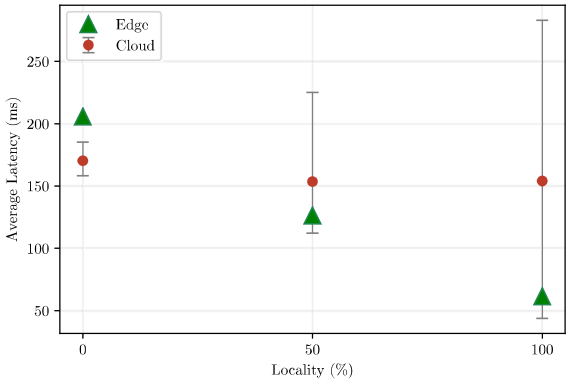


Figure 6. End-to-End latency in a networked scenario.

many virtualized broker instances running behind a single end point, typically a load balancer [5] [12]. These approaches introduce the concept of multiple brokers cooperating with each other, although the broker cluster is seen as a single centralized entity from the perspective of clients. Pure MQTT broker distribution is introduced in Banno et al. in [4]: authors propose ILDM (Internetworking Layer for distributed MQTT brokers), where heterogeneous brokers are connected with each other through specific nodes, placed between clients and brokers. Similar to our work, message distribution is obtained with publication flooding, but the underlying network of ILDM nodes is assumed to be already loop-free. Also no automatic mechanisms for broker failure recovery are present. In [13] and [14] authors also propose to interconnect MQTT brokers, with the possibility of dynamically changing the topology configuration at run time through specific MQTT messages transmitted by a centralised trusted entity. On the same line, the work in [15] creates a broker network and uses an external monitoring agent to check the status of each broker. Clients are connected to brokers through local gateways: upon any change in the broker configuration (broker failure, increase in latency, etc.) the gateway reconnects client to a new broker, according to the information retrieved by the monitoring agent. Such an approach enables client mobility, dynamic broker provisioning, and broker load balancing. Finally, an example of tree-based MQTT broker topology is given in [16], where authors propose the use of Software Define Networking (SDN) to create per-topic multicast groups in order to minimise data transfer delay. The SDN controller gathers information about clients and relative pub/sub topics from all the edge brokers through a master broker, which acts as root of the multicast tree. However, the paper assumes a static topology and no details are given on how such a root broker should be elected.

V. Conclusion

This paper proposes MQTT-ST, a system based on the Spanning Tree Protocol, which is able to create a distributed network of MQTT brokers. The brokers in the network generate a tree-based topology in a distributed way, which is able to fully replicate messages in every broker and to react to failures. The system has been tested in different scenarios, comparing the obtained performances with the legacy centralised solution. Future work direction will explore the integration of more complex routing strategies, besides message flooding, that can further increase the system performance. The MQTT-ST project is available for download at https://github.com/ANTLab-polimi/mosquitto.

References

[1] A. E. C. Redondi, A. Arcia-Moret, and P. Manzoni, "Towards a scaled iot pub/sub architecture for 5g networks: the case of multiaccess edge computing," in *2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*, April 2019, pp. 436–441.

[2] A. Banks, E. Briggs, K. Borgendale, and R. Gupta, "Mqtt version 5.0," *OASIS Standard*, 2019.

[3] R. A. Light *et al.*, "Mosquitto: server and client implementation of the mqtt protocol." *J. Open Source Software*, vol. 2, no. 13, p. 265, 2017.

[4] R. Banno, J. Sun, M. Fujita, S. Takeuchi, and K. Shudo, "Dissemination of edge-heavy data on heterogeneous mqtt brokers," in *2017 IEEE 6th International Conference on Cloud Networking (CloudNet)*. IEEE, 2017, pp. 1–7.

[5] P. Jutadhamakorn, T. Pillavas, V. Visoottiviseth, R. Takano, J. Haga, and D. Kobayashi, "A scalable and low-cost mqtt broker clustering system," in *2017 2nd International Conference on Information Technology (INCIT)*. IEEE, 2017, pp. 1–5.

[6] R. Baldoni, R. Beraldi, L. Querzoni, and A. Virgillito, "Efficient publish/subscribe through a self-organizing broker overlay and its application to SIENA," vol. 50, no. 4, pp. 444–459.

[7] A. Majumder, N. Shrivastava, R. Rastogi, and A. Srinivasan, "Scalable content-based routing in pub/sub systems," in *IEEE INFOCOM 2009*. IEEE, 2009, pp. 567–575.

[8] J. L. Martins and S. Duarte, "Routing algorithms for content-based publish/subscribe systems," *IEEE Communications Surveys & Tutorials*, vol. 12, no. 1, pp. 39–58, 2010.

[9] G. Siegemund, V. Turau, and K. Maâmra, "A self-stabilizing publish/subscribe middleware for wireless sensor networks," in *2015 International Conference and Workshops on Networked Systems (NetSys)*. IEEE, 2015, pp. 1–8.

[10] V. Turau and G. Siegemund, "Scalable routing for topic-based publish/subscribe systems under fluctuations," in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2017, pp. 1608–1617.

[11] A. Al-Fuqaha, A. Khreishah, M. Guizani, A. Rayes, and M. Mohammadi, "Toward better horizontal integration among iot services," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 72–79, 2015.

[12] S. Sen and A. Balasubramanian, "A highly resilient and scalable broker architecture for iot applications," in *2018 10th International Conference on Communication Systems & Networks (COMSNETS)*. IEEE, 2018, pp. 336–341.

[13] A. Schmitt, F. Carlier, and V. Renault, "Dynamic bridge generation for iot data exchange via the mqtt protocol," *Procedia computer science*, vol. 130, pp. 90–97, 2018.

[14] ——, "Data exchange with the mqtt protocol: Dynamic bridge approach," in *2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring)*. IEEE, 2019, pp. 1–5.

[15] T. Rausch, S. Nastic, and S. Dustdar, "Emma: Distributed qos-aware mqtt middleware for edge computing applications," in *2018 IEEE International Conference on Cloud Engineering (IC2E)*. IEEE, 2018, pp. 191–197.

[16] J.-H. Park, H.-S. Kim, and W.-T. Kim, "Dm-mqtt: An efficient mqtt based on sdn multicast for massive iot communications," *Sensors*, vol. 18, no. 9, p. 3071, 2018.