This item is the archived peer-reviewed author-version of:

Achieving end-to-end connectivity in global multi-domain networks

Reference:
Municio Esteban, Cevik Mert, Ruth Paul, Marquez-Barja Johann.- Achieving end-to-end connectivity in global multi-domain networks
Full text (Publisher's DOI): https://doi.org/10.1109/INFOCOMWKSHPS51825.2021.9484534
To cite this reference: https://hdl.handle.net/10067/1809500151162165141
Achieving End-to-End Connectivity in Global Multi-Domain Networks

Esteban Municio*, Mert Cevik†, Paul Ruth†, Johann M. Marquez-Barja*
*IDLab - imec, University of Antwerp, Belgium
†RENCI
{esteban.municio,johann.marquez-barja}@uantwerpen.be
{mert.cevik,paul.ruth}@renci.org

Abstract—Current trends on 5G network programmability evidence the need for end-to-end flexibility from the node and edge all the way to the cloud. Such multi-domain scenarios require realistic testbeds where different task-offloading algorithms, scheduling functions, and service orchestration techniques can be deployed and tested. While many of these research components can be often explored locally in small and isolated testbeds, new 5G demands are requesting for inter-operable platforms with a wider and a more global scope. The goal for these global platforms is that they can cope with multi-tier hierarchical architectures that are capable to face intense computational processes and heavy network traffic loads, while preserving dependability and keeping a low latency on the task executions and data transmission. In this paper we demonstrate a world-wide attempt to integrate different high-performance testing facilities, located in USA, Belgium, and The Netherlands, to enable experimentation on top of such large and complex architectures. In order to do this, we describe and deploy a multi-domain use case that can benefit from a global hierarchical infrastructure. Finally, we detail the performance characteristics of the deployment, discussing the experiences and technical challenges, and presenting the lessons learned we obtained when building and testing such experimental use case.

Index Terms—Orchestration, Control, Monitoring, Experimentation, Network Programmability, SDN, Industrial IoT

I. INTRODUCTION

New advances on Artificial Intelligence (AI) and 5G networks are pushing network management to its limits. A myriad of applications such as autonomous-driving, haptic sensing, and 3D holographic displaying are calling for very high computing power, while minimizing the energy consumption and latency communication. In order to deliver such demanding requirements, network management has the task to decide, given a limited amount of network, computing resources and data sources, where and when to execute each algorithm to offer the best possible performance/efficiency. While cloud infrastructures have a huge amount of computing capacity available, they usually have higher communication delay. On the other hand, edge infrastructures have low communications delay, but the computing capabilities are normally much more limited. Optimizing the task allocation between edge and cloud to fulfill the application requirements, and at the same reducing the energy consumption, is not a trivial challenge, and the research community is still investigating the best solutions to this problem.

Additionally, current network infrastructures are highly globalized, and network domains are usually federated in multi-tier architectures that allows them to scale as demand evolves. Since these architectures may involve different network technologies, topologies, and orchestration platforms, ensuring the end-to-end fulfillment of the service requirements adds an extra level of complexity to the edge/cloud interplay and task allocation problem. Fortunately, testbed facilities support researchers by allowing them to use very flexible and reconfigurable testing environments. There are currently several testbeds on different domains, such as cloud [1], Smart Cities [2], [3], Internet of Things [4] or Smart Highways [5]. These testbeds are excellent means to validate and reproduce the performance of new technologies in realistic conditions.

In this paper we present the integration of several of these testbeds. Specifically, we investigate on the research possibilities arising when Chameleon [1], Virtual Wall [6] and Citylab [2] testbeds are jointly integrated.
Our contributions are two-folded. First, we deploy a use case that can effectively leverage the testbed integration, presenting performance results in terms of computing capacity and network resources, and end-to-end task latencies for the joint solution. Secondly, we discuss the details and technical challenges of the integration, our experiences and lessons learned on favor of the reproducibility, and the opened research paths for such multi-domain testing architectures.

II. WHY A GLOBAL MULTI-DOMAIN NETWORK?

Consider a Smart City application that performs IoT-assisted and AI-based video processing to monitor car traffic flows to dynamically manage the traffic lights of a crossroad (Figure 1). Ideally the video processing and the decision making should be done as close as possible to the crossroad to keep the response time as low as possible. However, there are some reasons why it may be interesting to run these tasks somewhere else.

• **Tier 2 Local Cloud**: First, if we want to globally optimize the traffic we need to scale its management to the whole city. While sourced data streams may be filtered and aggregated in the edge, one (or more) local high-performance computing center can be required to process the information coming from all sources (e.g. crossroads). The same would apply in case data streams contain sensitive information and have to be centrally and securely stored for a period of time.

• **Tier 1 Cloud**: Secondly, we may want to cross-relate traffic information with other source types (e.g., pollution data) to augment the service offered. In that case, a high-level cloud could aggregate information from other cities or countries (e.g., satellite sources) to take better decisions. For the case of AI algorithms, a global cloud could aggregate data from other locations to train and improve the algorithms.

• **Tier 1 vs Tier 2 vs Edge**: However, there is not a one-size-fits-all approach. Application and data policies can change, and demand is expected to be dynamic. Also, hardware specifications in each site have different power consumption and may be optimized for a certain type of data (e.g., CPUs vs GPUs). For these reasons, a flexible multi-tier architecture with dynamic task offloading is needed to cope with a wide range of services.

Of course, the former discussion applies to any other “smart” domain. Because each use case is different, it is important to have a flexible tool common to all domains to be able to test different solutions, but yet in a controlled and reproducible manner. Multi-domain testbeds have the potential to support dynamic and evolving experiments, and the research community has been relying on such infrastructures for deploying and testing purposes for the last 10 years [7]–[13].

Many studies have already used testbeds for research on SDN [14], [15] service chaining [16], [17], IoT-cloud integration [2], [18], security [19], [20], etc. Additionally, works on testbed integration [21]–[25] have allowed researchers to interconnect testbeds of different domains to improve the scalability, flexibility and scope of their solutions.

What follows is one such attempt to leverage current testbed infrastructures to further extend the testing possibilities towards a global scale.

III. DEPLOYMENT

In order to characterize the overall performance of the integrated testbed, in this section we will describe the details of the previously described Smart City use case.

A. Testbeds

The complete architecture is shown in Figure 2. It consists of 3 heterogeneous domains, each of them located in a different geographical location in the world.

• **Citylab**: In order to instantiate edge nodes and wireless connectivity in a real Smart City scenario, we use the Citylab1 testbed [26]. Citylab is federated within Fed4FIRE+ [2] and it is intended for wireless networking experimentation in the unlicensed spectrum. Citylab is also used for Smart City edge computing deployments. It is located in the city of Antwerp (Belgium) and nodes are distributed along the city center over an area of about 0.5 x 0.5 km². These nodes receive data streams from multiple sensors, using different wireless technologies, i.e., WiFi, 6TiSCH, DASH7, Zigbee and Bluetooth, filtering and processing the bulk data stream, and only offloading such data processing to higher levels clouds when CPU intensive tasks are required.

• **Virtual Wall**: For the tier 2 local cloud we use Virtual Wall3 testbed. It is also federated within Fed4FIRE+ and consists of more than 550 servers to be used as bare metal hardware or as virtual resources through XEN virtual machines or docker containers. It is located in Gent (Belgium). We use these nodes to form a SDN network composed by OpenVSwitch switches and controlled through OpenFlow by an ONOS controller.

• **Chameleon** As tier 1 cloud we use the Chameleon4 testbed. Chameleon consist of two operating sites: University of Chicago (UC) and Texas Advanced Computing Center (TACC), and also supports both bare metal instantiation or virtualization through a KVM cloud. It consists of more than 15000 cores and numerous hardware flavors can be configured. Because of size of the testbed infrastructure, Chameleon perfectly fits as large-scale testing platform. An extended description can be found in [27].

B. Connectivity

The per-link connectivity characterization can be found in Table I. We have connected the Citylab nodes through IPv6 GRE tunnels to the OVS switches deployed in Virtual wall. While these links are not completely dedicated to each experiment, the available bandwidth (864 Mbps for UDP and 756 Mbps for TCP) is enough for a wide range of experiments.

1https://doc.lab.cityofthings.eu/
2List of Fed4FIRE+ testbeds https://www.fed4fire.eu/testbeds
3https://doc.ilabt.imec.be/ilabt/virtualwall
4https://www.chameleoncloud.org/
In Virtual Wall, the OVS network receives traffic from the edge nodes in Citylab and either serves the traffic to the local cloud servers or forwards some parts of it to the tier 1 Chameleon cloud. The connection between Virtual Wall and Chameleon is done through a level-2 stitched link [28], which means that resources are privately allocated to the experiment through VLANs. The link is formed by two segments: one between Virtual Wall and Universiteit van Amsterdam, and the other between Universiteit van Amsterdam and Chameleon. The switching rack is in an ExoGENI site [12] located at Universiteit van Amsterdam that can relay traffic from the Fed4FIRE+ testbeds to the ExoGENI network, which is the network that orchestrates the federation of independent cloud sites located across the US, including Chameleon. There are two possible entry points to ExoGENI, through UC or through TACC. Both possibilities have available throughputs of almost 1 Gbps, however the experienced latency is 115 ms and 140 ms for TACC and UC respectively.

Finally, in Table II we compare the end-to-end performance from Citylab to TACC and UC with the performance obtained from using the public Internet network. When using the testbed infrastructure, the achieved end-to-end throughput is \( \sim 860 \) Mbps and \( \sim 720 \) Mbps for UDP and TCP respectively in both UC and TACC. However, many limitations appear when using the public Internet. First, most of the ports are forbidden. While we can tackle this with a reverse SSH connection, this is only practically possible for TCP. Despite this, the available end-to-end throughput for TCP using the public Internet is never more than 148 Mbps due to the protocol overhead and lack of network resources isolation.

Regarding the latencies, the public Internet offers lower values because the paths may be routed more directly (stitched paths bounce out of the core in Ghent and Amsterdam). End-to-end latencies using the stitched links are 144 ms and 118 ms for TACC and UC respectively. In Section IV we analyze the impact of such latency in the task allocation decision process.

IV. PERFORMANCE EXPERIMENTS

A. Processing Power

Before deciding what is the best location to execute a given task, it is necessary first to assess the available processing power of each testbed. In order to benchmark the nodes, we have used High-Performance Linpack (HPL)\(^5\), a library to solve numerical linear algebra commonly used to perform benchmarks on high performance computers. We have selected

\(^5\)http://www.netlib.org/benchmark/hpl/
Table II: End-to-End Link characterization.

<table>
<thead>
<tr>
<th>Link</th>
<th>Type</th>
<th>Bandwidth (UDP)</th>
<th>Bandwidth (TCP)</th>
<th>Latency (round-trip)</th>
<th>Jitter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citylab - Chameleon-TACC</td>
<td>End-to-End</td>
<td>861 Mbps</td>
<td>724 Mbps</td>
<td>144 ms</td>
<td>19.9 us</td>
</tr>
<tr>
<td>Citylab - Chameleon-UC</td>
<td>End-to-End</td>
<td>861 Mbps</td>
<td>725 Mbps</td>
<td>118 ms</td>
<td>22 us</td>
</tr>
<tr>
<td>Citylab - Chameleon-TACC</td>
<td>Public Internet</td>
<td>N/A</td>
<td>125 Mbps</td>
<td>128 ms</td>
<td>200 us</td>
</tr>
<tr>
<td>Citylab - Chameleon-UC</td>
<td>Public Internet</td>
<td>N/A</td>
<td>148 Mbps</td>
<td>103 ms</td>
<td>100 us</td>
</tr>
</tbody>
</table>

Table III: Comparison of HPL results.

<table>
<thead>
<tr>
<th>Testbed</th>
<th>Node Type</th>
<th>Execution time (s)</th>
<th>Gflops</th>
<th>CPU</th>
<th>Cores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citylab</td>
<td>APU2c4</td>
<td>77.79</td>
<td>1.098</td>
<td>AMD GX-412TC</td>
<td>4</td>
</tr>
<tr>
<td>Virtual Wall</td>
<td>Pcgen03</td>
<td>10.43</td>
<td>8.188</td>
<td>Intel Xeon E5620</td>
<td>8</td>
</tr>
<tr>
<td>Chameleon</td>
<td>Haswell</td>
<td>7.79</td>
<td>10.96</td>
<td>Intel Xeon E52670</td>
<td>12</td>
</tr>
</tbody>
</table>

Figure 3: Wattage consumption of running CPU-intensive task.

some of the most common nodes available in each testbeds: Haswell in Chameleon (more than 500 nodes), Pcgen03 in Virtual Wall (more than 100 nodes) and APU2c4 in Citylab.

Computing performance of each node type is presented in Table III in terms of Gflops and execution time. Since Haswell nodes are equipped with more processing power, the same HPL problem (input matrix size $N = 5040$ and block size $NB = 128$) it is solved 10 times faster (and can perform 10 times more float operations per second) than in the APU2c4 node at Citylab. Performance of Pcgen03 nodes lay in between, but closer to the figures of Haswell nodes.

Besides computing capacity, task offloading decisions can be taken also considering power consumption. CPU intensive tasks could be executed not only faster in the cloud but also more efficiently. In order to illustrate this, we show the power consumption of each node type in Figure 3. These figures have been calculated by using the PowerTop utility.

While APU2c4 and Pcgen03 nodes behave similarly, Haswell nodes have slightly lower wattage. This could result in a significant difference when integrating the overall consumption over extended periods of time, i.e., when consumption of data transmission can be neglected. Although this difference is partially influenced by the nodes’ CPU, the major contribution on the power consumption reduction is in the lower average CPU % used time required in Haswell nodes.

B. Task Offloading

In order to assess the task offloading possibilities, we have first run experiments to test the scalability of a general purpose CPU-intensive task that do not require large amounts of data. This has been tested by running a simple counter-loop Python script with $O(n^2)$ that computes a triangular matrix of $n = 2000$ rows [29]. Figure 4a shows the average computing time of each node type independently when scaling the task from one single instance to 30 simultaneous instances.

APU2c4 (Citylab) nodes have always higher computing times compared with Pcgen03 (Virtual Wall) and Haswell (Chameleon) nodes. After 4 simultaneous tasks, the computing time needed per tasks increases almost linearly. A detailed (magnified) view of the processing times of Pcgen03 and Haswell nodes is shown in Figure 4b. Here, we also include the measured RTT between Virtual Wall and Chameleon using a stitched link to TACC (i.e., 140ms). This illustrates that after about 10 simultaneous tasks, the computing time plus link latency benefit the task offloading to the Chameleon testbed over Virtual Wall.

These former results evidence that it is always preferable to offload CPU-intensive tasks from the edge to the cloud independently of the number of concurrent instances, as also stated by Yorick et al. [30]. Additionally, for the case of more than 10 such concurrent tasks, the average time it takes to complete each individual task is lower when using the Chameleon testbed, even considering the transatlantic latency.

Finally, we show performance on end-to-end task execution latency. In this case, we consider tasks that require large amount of locally sourced data (e.g., large files or data streams are required for the task). Figure 5 presents the evolution of the time it takes to receive (through MQTT) and parse a JSON object of different sizes (performing a hash of it afterwards).

Since these tasks require to move bulk data, the overall task latency always favors the execution on the edge. Comparing Chameleon (TACC) and Virtual wall, Figure 5 shows that for data larger than ~100 MB, the offloading to Chameleon is preferable despite the transatlantic latency in the stitched link.

Unlike the previous experiments with negligible sourced data, end-to-end results of file parsing tasks evidence that overall task latencies are mainly subjected to data transmission times, and that only for large amounts of data, the processing power plays a major role.

\(^6\)https://01.org/powertop/
V. MAIN TAKEWAYS AND LESSONS LEARNED

• **Importance of multi-tier task offloading.** The presented deployment and its results evidence the convenience of having one or more levels of task offloading. By having the opportunity to offload tasks from the edge to tier-1 or tier-2, not only it is possible to reduce the overall task latency, but also to minimize power consumption. This multi-layered approach allows the network to fulfill both real-time and computationally intensive task in a flexible manner.

• **Importance of stitched links.** Public networks offer limited connectivity which for many experiments is not enough. Some transport protocols like UDP are not available, network segments are shared and available throughput is limited. Therefore, to perform experiments with carrier-grade QoS, dedicated stitched links are needed. However, stitched links are still not fully programmable. Upcoming controlled testbeds like FABRIC will allow users to truly control the end-to-end path.

• **Importance of sourced data.** Another evidence is the importance of the size of the sourced data when making task offloading decisions. When tasks need to parse or filter large amount of data, keeping locally the task or some parts of it seems to be the best option, as for the significant price to pay in latency when moving the data to the cloud. Only when tasks are computationally intensive, the cloud is a better option.

• **Importance of host tuning.** There are also some technical aspects to be taken in account. Currently, Citylab can only be linked to other Fed4FIRE+/Chameleon testbeds through GRE tunnels. This limits the MTU for end-to-end connections to 1368, and higher MTU values will cause the network to underperform. Additionally, the stitched link has a one-hop latency of more than 115 ms. Default TCP parameters are not configured for such high-latency links. We refer the users to TCP host tuning manuals to increase the transmission buffers and to use the TCP congestion algorithm BBR [31].

• **Next logical steps.** In this work we have given an initial characterization and discussion of results of a transatlantic multi-domain joint-experiment deployed in the Citylab, Virtual Wall and Chameleon testbeds, focusing on network and per-node performance. Since a major potential in these cloud testbeds is in their scaling capacity, future works deploying few hundreds of nodes and orchestrating them (e.g., by using Kubernetes and MANO solutions) would unveil the full potential of these global experiments. Additionally, a more thorough study of the trade-offs between power consumption and processing time would be useful for these use cases that aim to optimize power consumption and have less constrained latency requirements (e.g., AI training). Finally, an additional next step would be to study task prioritization for these use cases with different tasks, some with high-priority tasks that need lower latency and some others with lower priority that can be executed slower, but more efficiently in the cloud.

VI. CONCLUSION

In this work we have demonstrated a global multi-domain experiment that uses a multi-tier hierarchical architecture to enable edge/cloud task offloading at different levels. In order to do this, we have integrated three testbeds, Citylab, Virtual Wall and Chameleon, located in Belgium and USA (connected via The Nederlands), describing the deployment and characterizing it in terms of end-to-end network performance and per-node computing capacity. Obtained results confirm the importance of the edge/cloud task offloading to obtain the required latency requirements. We also give some preliminary
figures on when and where is more convenient execute a task, given the number of simultaneous tasks or size of the sourced data required by the task, finding a trade-off between edge and cloud. Finally we discuss our experiences and present some logical future research paths required to unveil the full potential of such multi-domain global experiments.

VII. ACKNOWLEDGEMENT

Funded by the European Union’s Horizon 2020 Fed4FIRE+ project (no.723638) and Celtic+ FlexNet project (2016/3).

REFERENCES