Modeling Performance and Energy Efficiency of Virtualized Flexible Networks

Raffaele Bolla[1,2], Roberto Bruschi[2], Franco Davoli[1,2(✉)], and Jane Frances Pajo[1,2]

[1] Department of Electrical, Electronic and Telecommunications Engineering, and Naval Architecture (DITEN), University of Genoa, Genoa, Italy
{raffaele.bolla,franco.davoli}@unige.it,
jane.pajo@tnt-lab.unige.it
[2] CNIT (National Inter-University Consortium for Telecommunications) National Laboratory of Smart, Sustainable and Secure Internet Technologies and Infrastructures (S3ITI), Genoa, Italy
roberto.bruschi@unige.it

Abstract. We examine some aspects of modelling and control in modern telecommunication networks, in the light of their evolution toward a completely virtualised paradigm on top of a flexible physical infrastructure. The trade-off between performance indicators related to user satisfaction of services (e.g., in terms of perceived quality, delay and ease of the interaction) and the energy consumption induced on the physical infrastructure is considered with some attention. In this respect, we provide a discussion of potential problems and ways to face them, along with a short description of the approaches taken in some European project activities.

1 Introduction

In recent years, a significant shift in data networking paradigms and in resource allocation mechanisms in networking has been gaining increasing momentum. Whereas in the past bandwidth, among other resources, used to be considered a potential bottleneck to be administered carefully, especially in the user access area (and still is, to some extent, in wireless access), with the increase in available transmission and processing speed, paralleled by an unprecedented increase in user-generated traffic, other factors that were previously concealed have become evident: the legacy networking infrastructure makes use of a large variety of hardware appliances, dedicated to specific tasks, which typically are inflexible, energy-inefficient, unsuitable to sustain reduced Time to Market of new services.

In this context, the search for ways of making resource allocation in telecommunication networks more *dynamic*, *performance-optimized* and *cost-effective* has brought forth the characterizing features of *flexibility*, *programmability* and *energy-efficiency*. The first two aspects are addressed by Software Defined Networking (SDN) [1–4] and Network Functions Virtualisation [5, 6]. In particular, the latter leverages "…standard IT virtualisation technology to consolidate many network equipment types onto industry standard high volume servers, switches and storage, which could be located in

Datacentres, Network Nodes and in the end user premises" [7]. The expected benefits are improved equipment consolidation, reduced time-to-market, single platform – multiple applications, users, and tenants, improved scalability, multiple open eco-systems and, last but not least, exploitation of the economy of scale of the Information Technology (IT) industry: according to [8], the 2016 market for data-centre servers has reached \$32 billion worldwide, with a growth rate of 6%, against \$27 billion world-wide for routers and switches, with growth rate of 1%; in any case, "the main disruption to the market is being provided by the growth of cloud and hosted solutions, which are redefining markets and enabling new competitors to emerge" [8].

SDN and NFV, along with Cloud and Fog Computing (or, more generally Multi-access Edge Computing – MEC [9]) paradigms create the basis for the "softwarisation" phenomenon that is going to find its full development in the 5G ecosystem [10, 11]. The creation of network slices in this context [12] provides the mechanisms to hierarchically abstract and orchestrate resources (both real and virtual) to eventually offer a complete, flexible, isolated and manageable networking environment to vertical industries for the deployment and dynamic instantiation of their applications.

However, it should be kept in mind that the certainly meritorious and notable effort behind the development of architectural concepts, abstractions, and standardised interfaces, as well as of the (most often open source) software constructs enabling their implementation, which has characterised the evolution of such innovative networking ecosystem, does not provide by itself the control and management mechanisms necessary for its proper operational functionality. The intelligence to perform dynamic resource allocation in such complex multi-actor environment must come from data- or model-based control strategies that operate at multiple levels, interact in non-mutually-obstructive fashion, and concur to the accomplishment of common as well as conflicting goals, within well-defined constraints. This consideration brings forward the other two aspects that we mentioned earlier, regarding *performance* and *energy-efficiency*. In this framework, the purpose of the paper is to highlight some of the issues concerning the trade-off between these two aspects in the virtualised networking framework.

The paper is organized as follows. We recall some of the energy-related issues in the next section. In Sect. 3, we summarise the results of some approaches to the problem of joint performance-energy optimisation in softwarised networks, and in Sect. 4 we briefly introduce two recent European projects that address some specific aspects in this general perspective. Section 5 contains the conclusions.

2   Energy-Efficiency Modelling and Control Aspects

Information and Communication Technology (ICT) has been historically and fairly considered as a key objective to reduce and monitor "third-party" energy wastes and achieve higher levels of efficiency. A classic example in this respect has been the use of video-conferencing services; more recent ones are Intelligent Transport Systems (ITS) and, directly affecting the energy sector, the Smart Grid. However, until relatively recently, ICT had not applied the same efficiency concepts to itself; consideration of energy consumption issues started first with wireless networks (see, e.g., [13]) and

datacentres ([14, 15], among others), and later extended to fixed networks and the Internet in general ([16–18], among others).

There are two main motivations that drive the quest for "green" ICT: the environmental one, which is related to the reduction of wastes, in order to impact on the carbon footprint; the economic one, which stems from the reduction of operational costs (OpEx) of ICT services. Indeed, according to the Global e-Sustainability Initiative (GeSI) [19], global ICT emissions (including datacentres, voice and data networks, and end-user devices) of greenhouse gases (GHG) are bound to reach about 1.3 GtCO2e/y (Gtons of CO2 equivalent per year), amounting to 2.3% of overall GHG emissions. On the other hand, it is interesting to observe that ICT's abatement potential is estimated to be 7 times higher (16.1%).

Today's (and future) network infrastructures are characterized by a design capable to deal with strong requests and constraints in terms of resources and performance (large loads, very low delay, high availability, …), and by services that exhibit high variability of load and resource requests along time (burstiness, rush hours, …). The current feasible approach to cope with energy consumption is centred on smart power management (energy consumption should follow the dynamics of the service requests) and on flexibility in resource usage (virtualization to obtain an aggressive sharing of physical resources).

In [16] we have introduced a taxonomy of approaches to energy efficiency in fixed networks, where two broad families of techniques are identified to adapt the consumption to load variations, acting on different time scales: *dynamic adaptation* and *smart standby*. The first one can be further divided into Adaptive Rate (AR) and Low Power Idle (LPI), which aim at adjusting the processor's speed (by adjusting frequency, voltage, or both) according to the load, and at putting part of the hardware into lower-consumption states during idle periods, respectively. The second family of techniques is usually referred to in conjunction with longer "sleeping" periods, and can be used effectively in virtualised environments (e.g., to consolidate functionalities to execute onto a smaller group of servers and to shut down the unused physical machines). Such techniques have been long used in computing devices, where the Advanced Configuration and Power Interface (ACPI, maintained since 2013 by the Unified Extensible Firmware Interface Forum – UEFI) [20] provides a standardized interface between the hardware and the software layers; however, only relatively recently they have found application in networking devices.

The ACPI introduces two power saving mechanisms, which can be individually employed and tuned for each core: Power States (C-states, where C0 is the active power state, and C1 through Cn are processor sleeping or idle states, where the processor consumes less power and dissipates less heat), and Performance States (P-states; while in the C0 state, ACPI allows the performance of the core to be tuned through P-state transitions, by altering the working frequency and/or voltage, or throttling the clock, to perform AR). The adoption of similar concepts in the framework of networking devices (e.g., switches and routers), spawned by the ECONET project [21], among others, has led to the development of the Green Abstraction Layer (GAL) [22, 23], later adopted as ETSI standard 203 237 [24]. The GAL allows power-aware devices, or parts thereof, to communicate their power-performance adaptation capabilities to network control and management entities, and to receive parameter settings

and commands from them, effectively enabling power-performance trade-off, on the basis of suitable optimisation techniques (see, e.g., [25–28]).

In this framework, the application of control and optimisation methods to manage the above-mentioned trade-off has been considered both at the device-level, with the application of Local Control Policies (LCPs), and the network-level, concerning Network Control Policies (NCPs). In many cases, the two can be applied in a hierarchical fashion, where NCPs perform a kind of periodic or event-driven parametric optimisation, in order to adaptively set LCP model parameters (e.g., in terms of the choice of C- and P-states). The various techniques adopted may differ according to the type of model used to represent the physical processes to be dealt with, which basically entail queue and flow dynamics. A very general scheme to highlight the main components and their interaction is represented in Fig. 1.
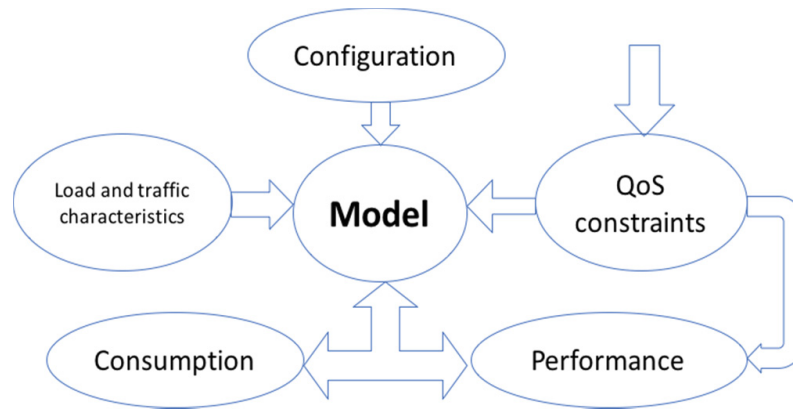
Fig. 1. Identification and interaction of models, inputs, goals and constraints.

In our opinion, two broad categories of modeling techniques are of particular interest here:

- Models based on classical queueing theory, possibly augmented with the explicit consideration of setup times (which stem from the different wakeup periods associated with different C-states of the processors), and taking into account the bursty nature of traffic at the packet level (e.g., the M^X/G/1/SET adopted in [25, 26]) lend themselves to performance analysis or parametric optimization for adaptive control and management policies over longer (with respect to queueing dynamics) time scales;
- Fluid models suitable for real-time control can stem from the first category, as was shown in the pioneering work of J. Filipiak in [29], or even from simpler, measurement-based, stochastic continuous fluid approximations (see, e.g., [30]); very interesting models and techniques for the dynamic control of queues can be found in the Lyapunov optimization approach proposed by Neely [31].

Moreover, control techniques in this setting may be quasi-centralized (owing to the presence of SDN controllers, which can supervise a certain number of underlying switches) or hierarchical (considering LCP-NCP interactions [25–28]), or even based

on more sophisticated completely distributed control techniques stemming from game [32] or team theory [33]. On this basis, and also taking into account the new capabilities offered by network softwarisation in terms of flexibility and programmability, which were mentioned in the Introduction, one would be led to conclude that the premises are there for a – technically and operationally – easier way to apply complex control and management strategies (with the latter operating on longer time scales, but often tightly integrated with the former and autonomic) for truly dynamic Traffic Engineering, accounting for both energy and Quality of Service/Quality of Experience (QoS/QoE) Key Performance Indicators (KPIs).

However, the introduction of NFV changes the perspective quite a bit with respect to "legacy" networking equipment:

- The hardware (HW) that consumes energy belongs to the Infrastructure Provider (InP), which in general may not coincide with the Network Service Providers (NSPs) in a multi-tenant environment;
- The HW is shared by multiple Virtual Machines (VMs) or by Network Slices, through a virtualization environment;
- Queueing models can be identified and used to assess the performance of VMs as function of the virtual resources assigned to them (as well as to control their assignment), but the relation between the performance of the VMs and their energy consumption is not straightforward, involving the virtualization layer, Infrastructure SDN Controllers (IC), Virtual Infrastructure Managers (VIMs), Wide-area Infrastructure Managers (WIM), Resource Orchestrators (ROs), Network Service Orchestrators (NSOs), and Tenant SDN Controllers (TCs) in the overall resource allocation process [12].

In the next Section, we will mention some initial approaches to the problem.

3 Managing the QoS-Energy Trade-off in Virtualised Networks

We have already noted that a significant reduction in Capital Expenditure (CapEx) should be realised by the economy of scale achievable with the adoption of general-purpose HW; which are then the main OpEx sources that can be reduced by technological advancement? They appear to be the ones related to energy consumption and network management, which roughly account for a figure equivalent to the entire infrastructure CapEx. Whereas it is true that local virtualisation of Base Stations can provide significant energy saving [34], it is not so straightforward to determine whether virtualisation in the backhaul network would reach the same result, unless specific energy-aware solutions are included in future 5G technologies. In [35] we have considered a use case corresponding to the virtualisation of a Serving Gateway (SGW) in the Evolved Packet Core (EPC), sketched in Fig. 2.

A number of aspects have been taken into account to cope with energy consumption: (i) appropriate downsizing of the parts of SGW that cannot be completely virtualized; (ii) dynamic activation of the minimum number of VMs to support the current traffic and their consolidation to the minimum number of servers;
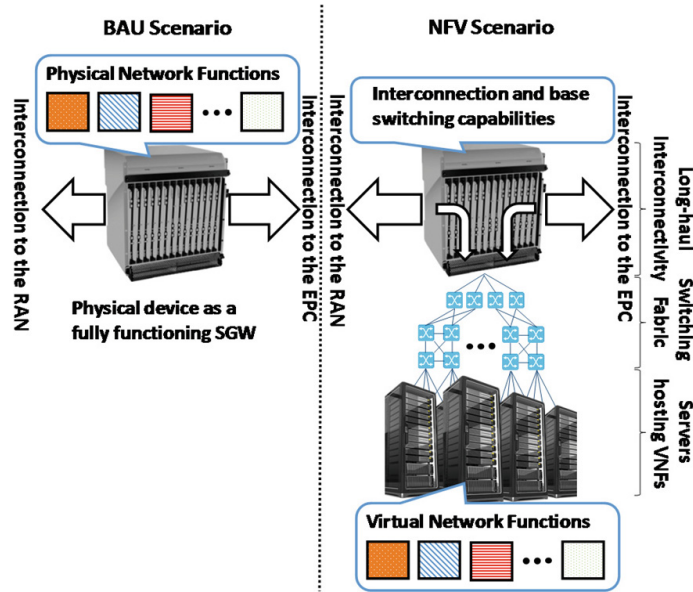
Fig. 2. SGW in the Business-as-Usual (BAU) and in the virtualization scenario [35].

(iii) optimisation of the interconnection switches' topology and enablement of energy-awareness capabilities; (iv) scaling of the throughput of a server through Amdahl's law [36], to account for parallelization. Even under these favourable assumptions, the power consumption of the virtualized Service Router (vSR), at the same target delay, results to be at least twice that of the "traditional" SR of the BAU solution.

Indeed, once fixed the silicon technology, energy consumption largely depends on the number of gates in the network device/chip hardware. The number of gates is generally directly proportional to the flexibility and programmability levels of HW engines. If we fix a target number of gates by using General Purpose CPUs, we obtain maximum flexibility, but reduced performance/power ratio; on the other hand, by using very specialized Application-Specific Integrated Circuits (ASICs), one would obtain minimum flexibility, but greatly enhanced performance/power ratio. Other technologies (e.g., network/packet processors) provide performance between these boundaries.

Essentially, there are three basic enablers at chip/system level: (i) dynamically programmable resources able to perform multi-purpose services; (ii) specialized HW for offloading to speed-up basic functionalities; (iii) standby capabilities to save energy if a resource is unused. The presence of general-purpose HW offers the possibility of moving services among the components of a node, or among nodes in a network. When the workload is low, many services can aggressively share single general-purpose HW resources. Thus, even if a general-purpose/programmable resource consumes more energy than ASIC-based solutions, a smaller number of HW elements can be left active, in order to effectively handle the current workload. Then, if programmability *for* energy efficiency is sought, two main issues need to be considered:

- which basic (sub-)functionalities need to be moved (and "frozen") to the offloaded specialized engines (best performance in terms of bps/W)?

- which ones have to remain in the programmability space (lower performance but stronger sharing and more evolution opportunities)?

The solutions need to be identified by considering and effectively supporting the newest trends in Internet technology evolutions. With these considerations in mind, we can turn back to the modelling aspects, and to the related control strategies that can be devised to jointly optimise performance and energy consumption. In this respect, as a possible example, we briefly summarise here the approach that has been taken in [37]. The scenario addressed is represented in Fig. 3. We consider a set of VMs dedicated to perform certain (virtualized) network functions (VNFs) on incoming traffic streams of various nature. For the sake of simplicity, a one-to-one correspondence is assumed between VNFs and VMs; the rationale behind this is that for a VNF consisting of multiple VMs the overall VNF performance can be derived from the individual VMs' performances, according to the chaining defined by the VNF provider. In any case, the VNF consolidation reduces to a VM consolidation problem. The VMs are initially placed among a given set of multicore servers through a First-Fit Decreasing (FFD) bin-packing algorithm [38] based on the workloads specified in the Service Level Agreement (SLA). Since such specifications are generally derived from peak workloads, the main goal here is to dynamically manage VM consolidation in each server according to actual workload variations, by jointly tuning the ACPI configuration and minimizing the number of active cores.
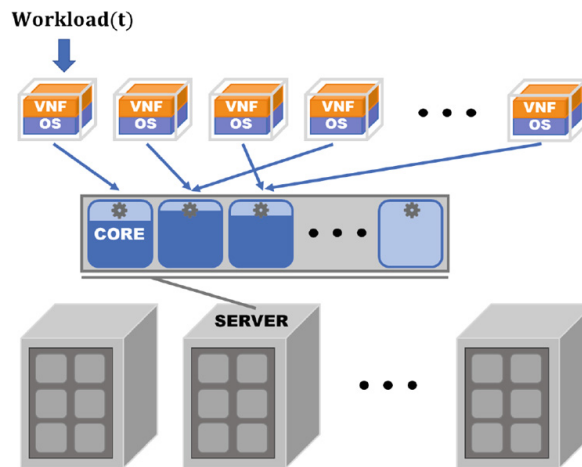
**Fig. 3.** Reference framework for VNF consolidation [37].

An $M^X/G/1/SET$ queueing model can be effectively used to represent the energy consumption, when considering the *aggregate workload* (see Fig. 4) produced by all VMs insisting on the core.

Then, two ways can be considered to enforce performance constraints: (i) coarsely, by imposing a limit on the maximum utilization for each core; (ii) more precisely, by computing the average system latency on the basis of the model. This requires the knowledge of the second moments of the batch size (obtainable from measurements of the second moment of the busy period) and of the packet service time (directly
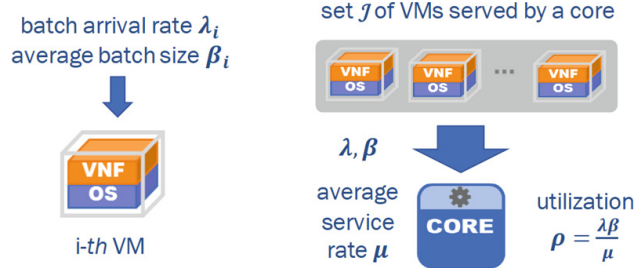
Fig. 4. Single VNF/VM pair and aggregate core workload.

measurable). Based on the model, decision rules can be defined to design an energy- and performance-aware consolidation policy in the space $(\lambda, \lambda\beta)$ of all pairs of aggregate batch arrival rate $\lambda = \sum_{i=1}^{|\mathcal{J}|} \lambda_i$ and aggregate workload $\lambda\beta$, where $\beta = (1/\lambda) \sum_{i=1}^{|\mathcal{J}|} \lambda_i \beta_i$ (see Fig. 4). The resulting policy is sketched in Fig. 5, where C_x and P_y are the C- and P-state values, respectively. The scheme includes energy- and performance-aware workload classification rules that define the most energy efficient configuration to be applied to the serving core.
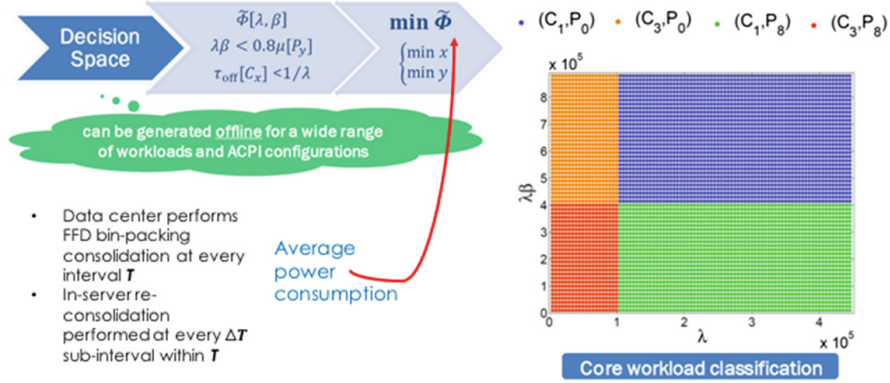
Fig. 5. Dynamic and long-term decision strategies.

Evaluations performed in [37] show that, despite VMs' workload variations, the total system workload is kept relatively stable, at approximately 18% below SLA specifications, and the policy applying both in-server consolidation and power scaling can gain about 10% with respect to baseline scenarios without the same capabilities. In a scaled-down datacenter example with 500 servers (with 2 octa-core processors each) and 10,000 VMs, at the average European Union electricity prices for industrial consumers of 0.12 €/kWh during the second half of 2014, this can turn to an annual saving of approximately 19,000 €.

We conclude the discussion in this Section with some further remarks about the evolution of the GAL. As we have already noted, a main challenge in the virtualised environment is that the correspondence between the HW that consumes energy (and

belongs to the Infrastructure Provider) and the virtualized objects (VMs, containers, …) that execute Network Functions (and belong to the Network Service Provider) is not so straightforward as in the legacy networking infrastructure: the execution mediated by the hypervisor and its scheduling policies, the resource allocation performed through multiple functional modules, the presence of multiple tenants, among other factors, do not allow establishing a "direct" relation between virtual resources (e.g., vCPU) and Energy Aware States of the HW. As we noted in [39] in a more specific context, possible lines of action may comprise: (i) the use of queueing models for aggregated traffic only (per server/core); (ii) the adoption of simpler aggregated models for HW energy consumption (e.g., Generalized Amdhal's Law [40]), and of more detailed queueing models for execution machines; (iii) the introduction of "virtualized" Energy Aware States as backpressure from the Infrastructure Provider to create incentives toward tenants to become energy-aware (currently under investigation in ETSI for a second version of the GAL).

## 4   The Experience of Two H2020 European Projects

In conjunction with our previous discussion, two projects that we are currently coordinating, funded by the European Commission under the Horizon 2020 program, touch some of the issues we have raised and are attempting to provide some answers.

4.1   INPUT – In-Network Programmability for Next-Generation Personal Cloud Service Support

The INPUT project [41], started in January 2015, aims at designing a novel infrastructure and paradigm to support Future Internet personal cloud services in a more scalable and sustainable way and with innovative added-value capabilities. The INPUT technologies will enable next-generation cloud applications to go beyond classical service models, and even to replace physical Smart Devices SDs), usually placed in users' homes (e.g., network-attached storage servers, set-top-boxes, video recorders, home automation control units, etc.) or deployed around for monitoring purposes (e.g., sensors), with their virtual images, providing them to users "as a Service" (SD as a Service – SDaaS; see Fig. 6).

The INPUT Project defines a *virtual image* as a software instance that dematerializes a physical network-connected device, by providing its virtual presence in the network and all its functionalities. Virtual images are meant to realize smarter, always and everywhere accessible, performance-unlimited virtual devices into the cloud. They can be applied both to provide all the functionalities of fully dematerialized physical devices through the cloud, and to add potentially unlimited smartness and capacity to devices with performance- and functionality-constrained hardware platforms.

Virtual and physical SDs can be made available to users at any time and at any place by means of virtual cloud-powered *Personal Networks* (PNs), which constitute an underlying secure and trusted service model (Personal Network as a Service – PNaaS). PNs provide users with the perception of always being in their home Local Area Network with their own (virtual and physical) SDs, independently of their location.
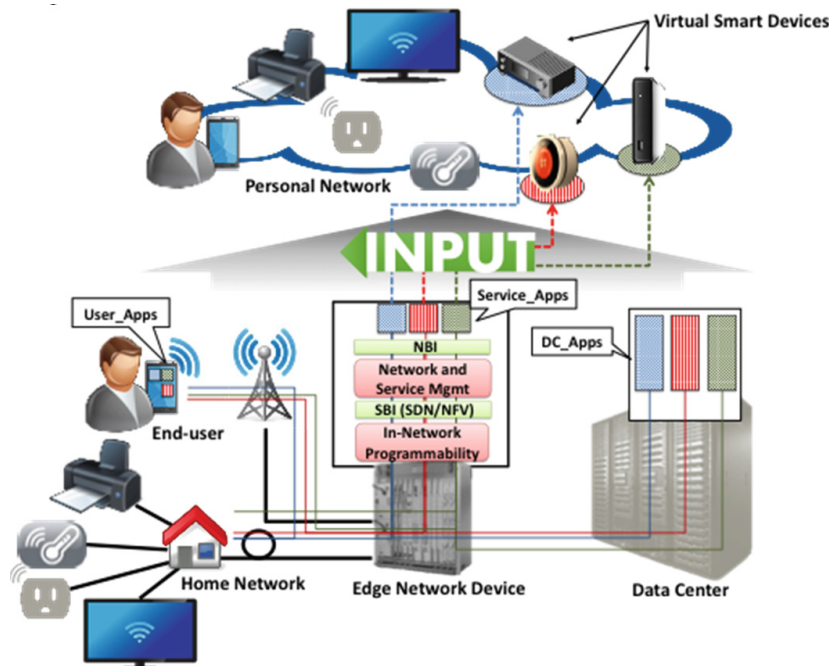
Fig. 6. The INPUT project general concept and structuring.

To achieve these ultimate objectives, the INPUT Project overcomes current limitations on the cloud service design due to the underlying obsolete network paradigms and technologies, by:

- introducing computing and storage capabilities to edge network devices ("in-network" programmability) in order to allow users/telecom operators to create/manage private clouds "in the network";
- moving cloud services closer to end-users and smart devices, in order both to avoid pointless network infrastructure and datacentre overloading, and to provide lower latency reactiveness to services;
- enabling personal and federated cloud services to natively and directly integrate themselves with the networking technologies close to end-user SDs, in order to provide new service models (e.g., the PN concept);
- assessing the validity of the proposed in-network cloud computing model through appropriately designed use cases and related proof-of-concept implementations.

As a side effect, the INPUT Project aims at fostering future-proof Internet infrastructures that will be "smarter," fully virtualized, power vs. performance optimised, and vertically integrated with cloud computing, with a clear impact on Telecom Operators', Service Providers' and end-users' CapEx and OpEx. In this respect, the INPUT Project is extending the programmability of network devices, to make them able to host cloud service applications, which cooperate with those in users' terminals and datacentres to realize the aforementioned cloud services. The INPUT approach and its infrastructural impact can contribute to the top line growth of European Telecom Operators and help increasing their revenue opportunities, enabling them to offer their infrastructure in support of novel value-added personal cloud services with reduced investments and

operating expenses. To this purpose, "in-network" programmable network devices have been designed on top of state-of-the-art off-the-shelf hardware with advanced power management capabilities, and suitable consolidation and orchestration mechanisms have been developed to optimize energy consumption and user-perceived QoE.

Central to the INPUT architecture are the concepts, illustrated in Fig. 6, of cloud applications (*Service_Apps*) hosted in network edge devices, and of their capability of cooperating with and of offloading corresponding applications residing in the users' smart objects (*User_Apps*) and in datacentres (*DC_Apps*), to realize innovative personal cloud services.

The presence of such Service_Apps allows user requests to be manipulated before crossing the network and arriving at datacentres in ways that enhance performance. Such manipulations can include pre-processing, decomposition and proxying. Moreover, Service_Apps take advantage of a vertical integration in the network environment, where applications can benefit from network-cognitive capabilities to intercept traffic or to directly deal with network setup configurations and parameters. The integration of Service_Apps at the network edge level is a fundamental aspect, since this level is the one where the Telecom Operator terminates the user network access, and a direct trusting/control on user accounts and services is performed. Therefore, this level is the best candidate to host personal Service_Apps, and to provide novel network-integrated capabilities to the cloud environment in a secure and trusted fashion. To achieve this purpose, the INPUT Project has been also focused on the evolution of network devices acting at this level beyond the latest state-of-the-art SDN and NFV technologies, and on how to interface them with the "in-network" programmability. This approach enables the reduction of reaction times of cloud applications, by exploiting the ability to directly access network primitives, and by providing improved scalability in the interactions of the network with users and datacentres.

As shown in Fig. 7, the INPUT Project is centred on a multi-layered framework that allows, on one hand, multiple Personal Cloud Providers to request IT (e.g., in terms of computing, storage, caching, etc.) and network resources of the Telecom Infrastructure Provider via an extended Service Layer Agreement. On the other hand, in order to minimize the OpEx and increase the sustainability of its programmable network infrastructure, the Telecom Infrastructure Provider can make use of advanced Consolidation criteria that allow Service_Apps to be dynamically allocated and seamlessly migrated/split/joined on a subset of the available hardware resources. The unused hardware components can enter low-power standby states. The presence of these power management criteria and schemes is a key aspect for the maximisation of the Return on Investment (RoI) of the INPUT technology to Telecom Infrastructure Providers.

More detailed presentations of the INPUT outcomes can be found in the project-related publications [42]. Instrumental to the realisation of the project demonstrators and prototypes has been the development of OpenVolcano (Open Virtualization Operating Layer for Cloud/fog Advanced NetwOrks) [43, 44], a comprehensive open-source platform for Fog and Mobile Edge Computing.
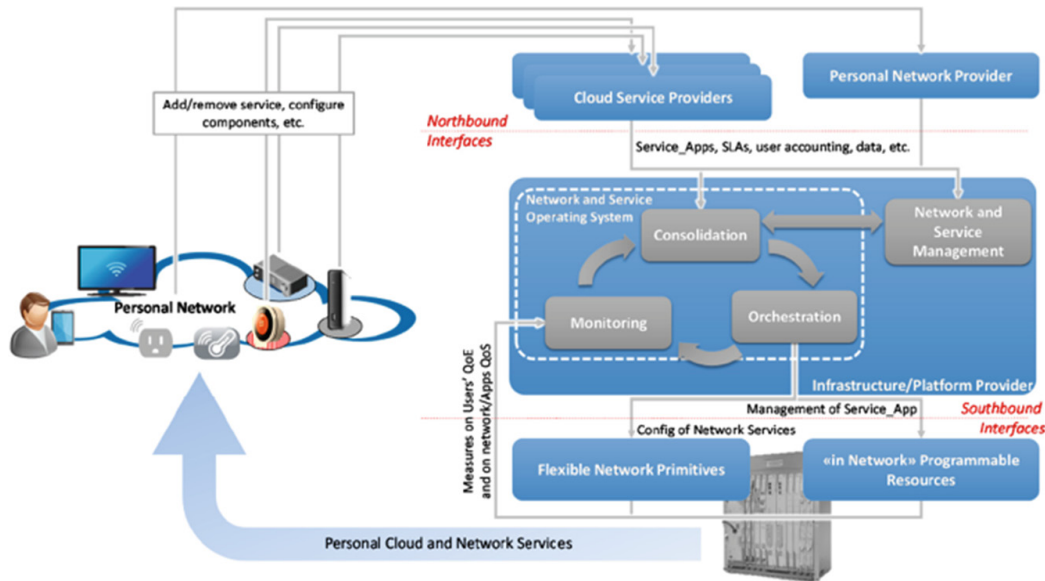
Fig. 7. INPUT architectural elements and their interaction.

4.2   MATILDA – A Holistic, Innovative Framework for Design, Development and Orchestration of 5G-Ready Applications and Network Services Over Sliced Programmable Infrastructure

MATILDA [45, 46] is an H2020 5G PPP (Public Private Partnership) project started in July 2017. Its vision is to design and implement a holistic 5G end-to-end services operational framework tackling the lifecycle of design, development and orchestration of 5G-ready applications and 5G network services over programmable infrastructure, following a unified programmability model and a set of control abstractions. MATILDA aims to devise and realize a radical shift in the development of software for 5G-ready applications, as well as virtual and physical network functions and network services, through the adoption of a unified programmability model, the definition of proper abstractions and the creation of an open development environment that may be used by application and network functions developers.

Intelligent and unified orchestration mechanisms will be applied for the automated placement of the 5G-ready applications and the creation and maintenance of the required network slices. Deployment and runtime policies enforcement is provided through a set of optimisation mechanisms realising deployment plans based on high-level objectives and a set of mechanisms supporting runtime adaptation of the application components and/or network functions based on policies defined on behalf of a services provider.

Multi-site management of the cloud/edge computing and IoT (Internet of Things) resources is supported by a multi-site virtualized infrastructure manager, while the lifecycle management of the supported VNF Forwarding Graphs (VNF-FGs), as well as a set of network management activities, are provided by a multi-site NFV Orchestrator (NFVO). Network and application-oriented analytics and profiling mechanisms are supported based on both real-time and a posteriori processing of the collected data from a set of monitoring streams. The developed 5G-ready application components,

applications, virtual network functions and application-aware network services are made available for open-source or commercial purposes, re-use and extension through a 5G marketplace.

The MATILDA project envisions the design and development of a holistic framework that supports the tight interconnection among the creation of 5G-ready applications and of the on-demand required networking and computational infrastructure, in the form of an application-aware network slice, and the activation of the appropriate networking mechanisms for the support of industry-vertical applications.

The MATILDA layers along with the main artefacts and key technological concepts comprising the MATILDA framework per layer are depicted in Fig. 8.
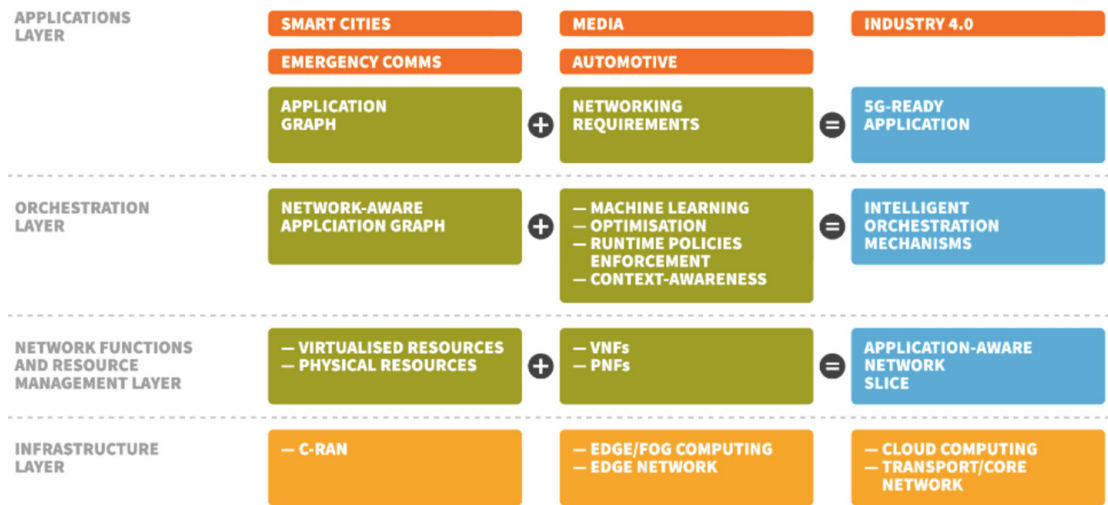
Fig. 8. The MATILDA general architectural framework.

The Applications layer corresponds to the Business Service and Business Function layer and regards the design and development of the 5G-ready applications per industry-vertical, along with the specification of the associated networking requirements. The Orchestration Layer regards the support of deployment and optimisation mechanisms of the 5G-ready applications over the available multi-site programmable infrastructure. Orchestration refers to both the application components and the attached virtual network functions and includes a set of intelligent mechanisms for optimal deployment, runtime policies enforcement, data mining and analysis, and context awareness support. The Network Functions and Resource Management Layer regards the implementation of the resource management functionalities over the available programmable infrastructure, as well as the lifecycle management of the activated virtual network functions. The Infrastructure Layer consists of the data communication network spanning a set of cloud computing and storage resources.

The key technological concepts and artefacts comprising the proposed MATILDA framework and constituting its unique selling points are:

- A conceptual architecture to support the provision of 5G end-to-end services tackling the overall lifecycle of design, development and orchestration of 5G-ready applications and 5G network services over a programmable infrastructure.
- A set of meta-models representing the vertical industry applications' components and graphs, the virtual – and physical – network functions and forwarding graphs.
- An innovative collaborative environment supporting the design and development of 5G-ready applications and VNF-FGs, including a web-based integrated development environment (IDE), verification and graphs' composition mechanisms.
- An orchestrator that is in charge of the optimal deployment and orchestration of the developed applications over the available programmable infrastructure – taking into account a set of objectives and constraints, as well as the defined policies, along with the instantiation of the required network functions for the support of the infrastructural-oriented functionalities. Policies enforcement is going to be supported by a context-awareness engine, able to infer knowledge based on a set of data monitoring, analytics and profiling production streams.
- A multi-site virtual infrastructure manager supporting the multi-site management of the allocated resources per network slice, along with a multi-site NFVO supporting the lifecycle management of the network functions embedded in the deployed application's graph and a set of network monitoring and management mechanisms.
- A novel analytics and unified profiling framework consisting of a set of machine learning mechanisms, of design time profiling and runtime profiling, toward the production of advanced analytics and software runtime profiling.
- A marketplace including an applications' and virtual network functions' repository and a set of mechanisms for the support of diverse 5G stakeholders.

5   Conclusions

The evolution of networks in the light of their "softwarisation" and virtualisation process and of the integration of diverse forms of access and transport paradigms has gained even greater impulse with the advent of 5G, the fifth-generation mobile network. In this framework, flexibility and programmability have become of paramount relevance, and new challenging KPIs have been set. Our attention has been focused on the aspect of network control and management of this complex heterogeneous environment, which appears to be sometimes hidden behind the architectural and operational constructs that form the basis for the virtualisation of resources and the orchestration of the physical and virtual elements. At the same time, we have tried to highlight the interaction of resource allocation policies performed for the purpose of attaining QoE/QoS-related KPIs with energy consumption of the physical network elements. The relation between certain network operations and their impact on energy consumption of the infrastructure is somehow blurred by the numerous mediating software and orchestration levels necessary to achieve the virtualised functionalities that ensure the two desired (and by now non-renounceable) aspects of network and

services flexibility and programmability. In this respect, we have examined some potentially critical aspects, pointed out possible ways of coping with them, and briefly described some current project approaches.

References

1. Software-defined networking: the new norm for networks. In: Open Networking Foundation Whitepaper (2012). https://www.opennetworking.org/images/stories/downloads/sdn-resources/white-papers/wp-sdn-newnorm.pdf
2. Nunes, B.A.A., Mendonça, M., Nguyen, X.-N., Obraczka, K., Turletti, T.: A survey of software-defined networking: past, present, and future of programmable networks. IEEE Commun. Surv. Tut. **16**(3), 1617–1634 (2014)
3. Kreutz, D., Ramos, F.M.V., Veríssimo, P.E., Rothenberg, C.E., Azodolmolky, S., Uhlig, S.: Software-defined networking: a comprehensive survey. Proc. IEEE **103**(1), 14–76 (2015)
4. ONF TR-521, SDN Architecture (2016). https://www.opennetworking.org/images/stories/downloads/sdn-resources/technical-reports/TR-521_SDN_Architecture_issue_1.1.pdf
5. ETSI GS NFV 002 V1.1.1 (2013-10) Network Functions Virtualization (NFV); Architectural Framework. http://www.etsi.org/deliver/etsi_gs/nfv/001_099/002/01.01.01_60/gs_nfv002v010101p.pdf
6. Mijumbi, R., Serrat, J., Gorricho, J.-L., Bouten, N., De Turck, F., Boutaba, R.: Network function virtualization: state-of-the-art and research challenges. IEEE Commun. Surv. Tut. **18**(1), 236–262 (2016)
7. Network functions virtualisation. In: Introductory White Paper, ETSI (2012). https://portal.etsi.org/nfv/nfv_white_paper.pdf
8. Synergy Research Group (2016). https://www.srgresearch.com/articles/enterprise-spending-nudged-downwards-2016-cisco-maintains-big-lead
9. http://www.etsi.org/technologies-clusters/technologies/multi-access-edge-computing
10. Manzalini, A., et al.: Towards 5G software-defined ecosystems – technical challenges, business sustainability and policy issues. In: IEEE SDN Initiative Whitepaper (2016). http://resourcecenter.fd.ieee.org/fd/product/white-papers/FDSDNWP0002
11. Expert Advisory Group of the European Technology platform Networld 2020, Strategic Research and Innovation Agenda (2016). Pervasive Mobile Virtual Services. https://www.networld2020.eu/wp-content/uploads/2014/02/SRIA_final.pdf
12. Ordonez-Lucena, J., Ameigeiras, P., Lopez, D., Ramos-Munoz, J.J., Lorca, J., Folgueira, J.J.: Network slicing for 5G with SDN/NFV: concepts architectures and challenges. IEEE Commun. Mag. **55**(5), 80–87 (2017)
13. Suarez, L., Nuaymi, L., Bonnin, J.-M.: An overview and classification of research approaches in green wireless networks. EURASIP J. Wirel. Commun. Netw. **2012**, 142 (2012)
14. Orgerie, A.C., Dias de Assunção, M., Lefevre, L.: A survey on techniques for improving the energy efficiency of large scale distributed systems. ACM Comp. Surv. **46**(4), 1–35 (2014)
15. Moghaddam, F.A., Lago, P., Grosso, P.: Energy-efficient networking solutions in cloud-based environments: a systematic literature review. ACM Comp. Surv. **47**(4), 64.1–64.32 (2015)
16. Bolla, R., Bruschi, R., Davoli, F., Cucchietti, F.: Energy efficiency in the future internet: a survey of existing approaches and trends in energy-aware fixed network infrastructures. IEEE Commun. Surv. Tut. **13**(2), 223–244 (2011)

17. Bianzino, A.P., Chaudet, C., Rossi, D., Rougier, J.-L.: A survey of green networking research. IEEE Commun. Surv. Tut. **14**(1), 3–20 (2012)
18. Idzikowski, F., Chiaraviglio, F.L., Cianfrani, A., López Vizcaíno, A.J., Polverini, M., Ye, M.Y.: A survey on energy-aware design and operation of core networks. IEEE Commun. Surv. Tut. **18**(2), 1453–1499 (2016)
19. Global e-Sustainability Initiative (GeSI), SMARTer2020: The Role of ICT in Driving a Sustainable Future. http://gesi.org/SMARTer2020
20. http://uefi.org/specifications
21. Bolla, R., Bruschi, R., Davoli, F., Cucchietti, F.: Setting the course for a green internet. Science **342**(6164), 1316 (2013)
22. Bolla, R., et al.: The green abstraction layer: a standard power management interface for next-generation network devices. IEEE Intern. Comp. **17**(2), 82–86 (2013)
23. Bolla, R., et al.: A northbound interface for power management in next generation network devices. IEEE Commun. Mag. **52**(1), 149–157 (2014)
24. Green Abstraction Layer (GAL): Power management capabilities of the future energy telecommunication fixed network nodes (2013) ETSI Std. 203 237 version 1.1.1. http://www.etsi.org/deliver/etsi_es/203200_203299/203237/01.01.01_60/es_203237v010101p.pdf
25. Bolla, R., Bruschi, R., Carrega, A., Davoli, F.: Green networking with packet processing engines: modeling and optimization. IEEE/ACM Trans. Netw. **22**(1), 110–123 (2014)
26. Bolla, R., Bruschi, R., Carrega, A., Davoli, F., Pajo, J.F.: Corrections to: "Green Networking with Packet Processing Engines: Modeling and Optimization". IEEE/ACM Trans. Netw. (2017). https://doi.org/10.1109/TNET.2017.2761892
27. Niewiadomska-Szynkiewicz, E., Sikora, A., Arabas, P., Kołodziej, J.: Control system for reducing energy consumption in backbone computer network. Concurr. Computat. Pract. Exper. **25**, 1738–1754 (2013)
28. Kamola, M., Niewiadomska-Szynkiewicz, E., Arabas, P., Sikora, A.: Energy-saving algorithms for the control of backbone networks: a survey. J. Telecommun. Inform. Technol. **2**, 13–20 (2016)
29. Filipiak, J.: Modelling and Control of Dynamic Flows in Communication Networks. Springer-Verlag, Berlin (1988)
30. Bruschi, R., Davoli, F., Mongelli, M.: Adaptive frequency control of packet processing engines in telecommunication networks. IEEE Commun. Lett. **18**(7), 1135–1138 (2014)
31. Neely, M.J.: Stochastic Network Optimization with Application to Communication and Queueing Systems. Morgan & Claypool, San Rafael (2010)
32. Bruschi, R., Carrega, A., Davoli, F.: A game for energy-aware allocation of virtualized network functions. J. Elec. Comp. Eng. **2016**, 4067186 (2016)
33. Aicardi, M., Bruschi, R., Davoli, F., Lago, P.: A decentralized team routing strategy among telecom operators in an energy-aware network. In: Proceedings of SIAM Conference on Control and its Application, Paris, France, July 2015, pp. 340–347 (2015)
34. Rahman, M.M., Despins, C., Affes, S.: Analysis of CAPEX and OPEX benefits of wireless access virtualization. In: IEEE Internat Conference Communications (ICC), Workshop on Energy Efficiency in Wireless Networks & Wireless Networks for Energy Efficiency (E2Nets), Budapest, Hungary (2013)
35. Bolla, R., Bruschi, R., Davoli, F., Lombardo, C., Pajo, J.F., Sanchez, O.R.: The dark side of network functions virtualization: a perspective on the technological sustainability. In: IEEE Internat Conference on Communications (ICC), Paris, France, 2017, pp. 1–7 (2017)
36. Woo, D., Lee, H.-S.: Extending Amdahl's Law for energy-efficient computing in the many-core era. IEEE Comput. **41**(12), 24–31 (2008)

37. Bruschi, R., Davoli, F., Lago, P., Pajo, J.F.: Joint power scaling of processing resources and consolidation of virtual network functions. In: 5th IEEE International Conference on Cloud Networking (CloudNet), Pisa, Italy, October 2016, pp. 70–75 (2016)
38. Coffman Jr., E.G., Garey, M.R., Johnson, D.S.: Approximation algorithms for bin packing: a survey. In: Hochbaum, D.S. (ed.) Approximation Algorithms for NP-Hard Problems. PWS Publishing, Boston (1996)
39. Bolla, R., Bruschi, R., Davoli, F., Depasquale, E.V.: Energy-efficient management and control in video distribution networks: "Legacy" hardware based solutions and perspectives of virtualized networking environments. In: Popescu, A. (ed.) Guide to Greening Video Distribution Networks - Energy-Efficient Internet Video Delivery, pp. 25–57. Springer (2018)
40. Cassidy, A.S., Andreou, A.G.: Beyond Amdahl's Law: an objective function that links multiprocessor performance gains to delay and energy. IEEE Trans. Comp. **61**(8), 1110–1126 (2012)
41. http://www.input-project.eu/
42. http://www.input-project.eu/index.php/outcomes/publications
43. Bruschi, R., Lago, P., Lombardo, C., Mangialardi, S.: OpenVolcano: an open-source software platform for fog computing. In: 28th International Teletraffic Congress (ITC 28) 1st International Workshop on Programmability for Cloud Networks and Applications (PROCON), Wuerzburg, Germany, September 2016
44. http://openvolcano.org/
45. Bolla, R., et al.: Design, development and orchestration of 5G-ready applications over sliced programmable infrastructure. In: 29th International Teletraffic Congress (ITC 28) 1st International Workshop on Softwarized Infrastructures for 5G and Fog Computing (Soft5 2017), Genoa, Italy, September 2017, pp. 13–18 (2017)
46. http://www.matilda-5g.eu/index.php