








Energy vs. QoX Network- and Cloud Services Management

Bego Blanco¹, Fidel Liberal¹, Pasi Lassila², Samuli Aalto²,
Javier Sainz³, Marco Gribaudo⁴, and Barbara Pernici⁴

¹ University of the Basque Country, Leioa, Spain

{`begona.blanco,fidel.liberal`}@ehu.eus

² Aalto University, Espoo, Finland

{`Pasi.Lassila,samuli.aalto`}@aalto.fi

³ Innovati Group, Madrid, Spain

`jsg@grupoinnovati.com`

⁴ Politecnico di Milano - DEIB, Milan, Italy

{`marco.gribaudo,barbara.pernici`}@polimi.it

Abstract. Network Performance (NP)- and more recently Quality of Service/Experience/anything (QoS/QoE/QoX)-based network management techniques focus on the maximization of associated Key Performance Indicators (KPIs). Such mechanisms are usually constrained by certain thresholds of other system design parameters. e.g., typically, cost. When applied to the current competitive heterogeneous Cloud Services scenario, this approach may have become obsolete due to its static nature. In fact, energy awareness and the capability of modern technologies to deliver multimedia content at different possible combinations of quality (and prize) demand a complex optimization framework.

It is therefore necessary to define more flexible paradigms that make it possible to consider cost, energy and even other currently unforeseen design parameters not as simple constraints, but as tunable variables that play a role in the adaptation mechanisms.

In this chapter we will briefly introduce most commonly used frameworks for multi-criteria optimization and evaluate them in different Energy vs. QoX sample scenarios. Finally, the current status of related network management tools will be described, so as to identify possible application areas.

1 Introduction

Network Performance- and more recently Quality of Service/Experience/X-based network management techniques (where “X” can represent “S” service, “P” perception, “E” experience or “F” flow, just to give a few examples), focus on the maximization of associated KPIs. Such mechanisms are usually constrained by certain thresholds of other system design parameters, e.g., typically, cost. When applied to the current competitive heterogeneous Internet of Services scenario, this approach may have become obsolete due to its static nature. In fact,

energy awareness and the capability of modern technologies to deliver multimedia content at different possible combinations of quality (and prize) demand a complex optimization framework.

It is therefore necessary to define a more flexible paradigm that makes it possible to consider cost, energy and even other currently unforeseen design parameters not as simple constraints, but as tunable variables that play a role in the adaptation mechanisms. As a result, for example, the service supply will then search for the maximum QoE at the minimum cost and/or energy consumption. In consequence, a certain service will not be offered at a single and specific guaranteed price, but will vary with the objective of obtaining the best (QoE, cost, energy, etc.) combination at a given time.

Unfortunately, most considered design parameters are conflicting, and therefore the improvement of one of them entails some deterioration of the others. In these circumstances, it is necessary to find a trade-off solution that optimizes the antagonistic criteria in the most efficient way. Therefore, the resource allocation problem becomes a multi-criteria optimization problem and the relevance of each criteria gains uttermost importance.

This chapter analyzes the existing optimization frameworks and tools and studies the complexity of introducing utility functions into network/management mechanisms, including fairness considerations. Then, we present cost/energy/*-aware network and cloud services management scenarios. Finally, we address the challenge of introducing energy-awareness in network controlling mechanism and provide a general view of current technologies and solutions.

2 Dealing with Multi-criteria Optimization: Frameworks and Optimization Tools

Regardless the mathematical or heuristic tools applied in order to find (near) optimal solutions in the scope of Internet of Services management mechanisms, all of them share common issues due to the extension of the original definition of the problem to a multi-criteria one. This section provides a summarized compilation of those issues, especially those related to how the decision maker (DM) will take into consideration different antagonistic criteria.

2.1 Generic Definition of the Problem

The classical constrained single criteria problem deals with finding the combination of design parameters (normally represented by a vector x^*) in the feasible space (S) that minimize a single function (1).

$$\exists x^* \in S / \min f(x^*) = z \quad (1)$$

Then the multi-criteria or multi-objective optimization problem, defined as an extension of the mono-criteria one, aims at simultaneously minimizing a collection of requirements keeping the equality and inequality constraints of the feasible space (2).

$$\exists x^* \in S / \min f_i(x^*) = z_i \forall i = 1, 2, \dots, k \quad (2)$$

The optimal solution that minimizes simultaneously all the criteria is most of the times hardly achievable, and is known as utopian solution [5]. Therefore, the actual best solution of the problem should be as close as possible to this utopian solution. The optimization problem must then be redefined to extract from the whole feasible space of solutions, those closer to the utopian solution. That set of solutions characterizes the Pareto-optimal front. The goal of a good multi-criteria optimization problem is the search of a set of solutions that properly represents that Pareto front, i.e., uniformly distributed along that Pareto front.

However, due to the trade-offs among different parameters, in most of the cases there will not exist such a solution which minimizes all the criteria simultaneously. So, the nature of the problem is usually re-defined by introducing the concept of Utility Function, responsible for quantifying the relevance and composite articulation of different criteria. Then, the real formulation of the problem can be expressed mathematically as follows (3).

$$\exists x^* \in S / \min U(z_1, z_2, \dots, z_k) \quad (3)$$

2.2 Incorporating Multiple Criteria in General Optimization Methods

Multiple Objective Optimization (MOO) has been a field of intensive research in different engineering areas. This activity has led to the development of a lot of MOO methods ranging from exact methods to meta-heuristics and including several different nature algorithms.

In this section, we propose a comprehensive taxonomy of the optimization problem synthesized from the works in [13, 14, 21, 30, 31, 36]. The presented taxonomy categorizes the optimization problems according to different perspectives where the main goal is to determine how the multiple criteria are considered by the DM. Table 1 summarizes the characterization of the optimization criteria that are defined as follows:

- **Qualitative vs. quantitative criteria:** refers to how the analyzed criteria are measured. If the DM is able to represent the preference degree of one option against the others by a numerical value, then the criteria are quantitative. Otherwise, the criteria are qualitative, meaning that preference can not be numerically measured or compared and, in consequence, a descriptive value is assigned.
- **Preference articulation:** refers to the point in time the DM establishes its preferences:
 - **A priori preference articulation:** the preferences are defined at problem modeling stage, adding supplementary constraints to the problem (i.e., weighted sum and lexicographic methods).
 - **A posteriori preference articulation:** once the optimization problem provides the set of results from the optimization process, DM's preferences are used to refine the final solution (i.e., in evolutive and genetic methods).

- **Progressive preference articulation:** DM's preferences are gradually incorporated in an interactive way during the optimization process.
 - **Without preference articulation:** when there is no preference definition for the problem (i.e., max-min formulation, global criterion method).
- **Continuous vs. discrete:** refers to the variable type used to describe the optimization criteria. When the optimization problem handle discrete variables, such as integers, binary values or other abstract objects, the objective of the problem is to select the optimum solution from a finite, but usually huge, set. On the contrary, continuous optimization problems handle infinite variable values. In consequence, continuous problems are usually easier to solve due to their predictability, because the solution can be achieved with an approximate iterative process. Since cost/energy aware network and services management must deal with both discrete (i.e., number of servers, route lengths, radio bearers, etc.) and continuous design parameters (i.e., coding bit rate, transmission power, etc.) both techniques should be considered.
- **Constrained vs. not constrained:** refers to the possibility of attaching a set of requirements expressed through (in)equality equations to the optimization problem. In this case, besides finding a solution that optimizes a collection of criteria, it must also meet a set of constraints. Non constrained methods can be used to solve constrained methods, replacing restrictions for penalizations on objective functions to prevent possible constraint violations. As aforementioned, classical network management approached involved considering a single criteria only and establishing Cost and Energy constrains. The proposal in ACROSS to move to a multi-criteria optimization analysis does not necessarily imply getting rid of all the possible constraints.

Those classifications do not result into disjoint categories. In fact, multi-criteria optimization problems in the considered heterogeneous network and services management scenario may fall into one or several of the categories listed above.

Summarizing, before beginning with the process of multi-criteria optimization problem there is a crucial previous step: the definition of the criteria to be optimized, i.e., the preferences of the DM about the suitability of the obtained solution.

Regardless the decision maker being the Cloud/Network/SOA designer or service operator the adaptation algorithm must incorporate the impact of different criteria on their perception of the goodness of any solution. A key factor in the analysis for decision making is indeed the fact that the functions that model decision maker's preferences (criteria or objective functions) are not usually known a priori.

2.3 Complexity of Defining Multi-criteria Utility Functions to be Incorporated in Network/Management Mechanisms

Considering the relevance of the choice of a multi-criteria utility function, different tools aiding at this task will be reviewed in this section.

Table 1. Characterization of the optimization criteria for DM.

Description type	Qualitative
	Quantitative
Preference articulation	A priori
	A posteriori
	Progressive
	None
Type of variables	Continuous
	Discrete
Constraint definition	Constrained
	Not constrained

- Goal attainment
- MAUT (Multi-Attribute Utility Theory)
- Preference relations
- Fuzzy logic
- Valuation scale

Goal Attainment. This basic format restricts the feasible space with the most relevant set of alternatives according to the DM’s preferences (Fig. 1). Such preferences if represented mathematically usually result in a n-dimensional shape or contour in the decision space limiting those solutions acceptable by the DM (similar to that imposed by the constraints in the design space). It is a simple and direct format, that just splits alternatives into relevant/non-relevant groups. However, it only offers little information about preferences, not providing any hint about the predilections of the DM.

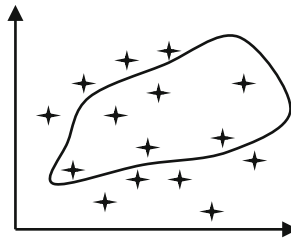


Fig. 1. Selection set within feasible space.

The work in [8,9] describe the use of goal attainment preference modeling in multi-criteria algorithms.

Multi-Attribute Utility Function. In this case, the utility function is build by describing the repercussion of an action regarding a specific criterion. Each action is assigned a numerical value, so that the higher the value, the more preferable the action. Then, the assessment of an action becomes the weighted sum of the numerical values related to each considered criterion. This representation format is capable of modeling DM's preference more precisely than the Selection Set. Nonetheless, it also means that the DM must evaluate its inclinations globally, comparing each criterion against all the others, which is not always possible. Therefore, this type of utility function is suitable just for the cases in which a perfect global rationality can be assumed [46]. For example, this classic model is commonly employed in economics and welfare field.

Besides, utility has an ordinal nature, in the sense that the preference relation between the possible choices is more significant than the specific numerical values [4]. So, this leaves the door open to discarding the numerical value of the utility, as it is shown next.

Preference Relations. This representation format models the inclination over a set of possible choices using a binary relation that describes the qualitative preference among alternatives. Then, a numerical value is linked to that relation, defining the preference degree of alternative x_i against alternative x_k in a quantitative way [37].

This format of preference modeling provides an alternative to the assignment of a numerical value to different utility levels, allowing the comparison of alternatives pairwise, providing the DM higher expressiveness to enunciate his preferences (i.e., similar to Analytic Hierarchy Process – AHP [39]). Outranking methods employ this format of preference representation.

Fuzzy Logic. This format allows the introduction of uncertainty over the preferences under analysis. In order to avoid ambiguity in the definition process of the preferences, each “ x_i is not worse than x_k ” is attached a credibility index. In this sense, fuzzy logic becomes a useful tool [46], as a general framework for preference modeling where certain sentences are a particular case.

The obstacle of using fuzzy logic with credibility indexes is the weakening of the concept of truth. The infinite possible values of truth between absolute truth and falseness have an intuitive meaning that does not correspond to their formal semantics. In addition, there are other problems, such as the formulation of the credibility index itself.

Valuation Scale. This preference formulation defines a formal representation of the comparison between possible choices that expresses both the structure of the described situation and the variety of manipulations that can be made on it [37]. This type of sentences are appropriately expressed in logical language. But classical logic can be too inflexible to acceptably define expressive models. In consequence, other formalisms must be taken into account to provide the model with the required flexibility.

Conclusion. Preference or criteria description format plays a crucial role in the definition of the nature and structure of the information the DM employs to set his predilections up towards the different possibilities. The selection of the best representation format will rely on the characteristics of the specific area of expertise. Sometimes, inclinations will be better expressed using numerical values, and in other cases using more natural descriptions, such as words or linguistic terms.

The final goal is to contrast the impact of the potential actions with the purpose of making a decision. Therefore, it is necessary to establish a scale for every considered criterion. The elements of the scale are denoted degrees, levels or ranks.

Table 2. Summary of methods to define multi-criteria utility functions ordered by complexity.

Goal attainment	Simple, just relevant/non-relevant categorization of preferences
Multi-Attribute Utility Theory (MAUT)	Utility function as a weighted sum of numeric values assigned to criteria, needs perfect global rationality
Preference relations	Modeled through binary relations to define preferences pairwise
Fuzzy logic	Introduces uncertainty through a credibility index
Valuation scale	Establishes a formal representation of the preference between alternatives

Table 2 summarizes the aforementioned methods to define multi-criteria utility functions represent the preferences of the decision maker related to the multiple criteria to be optimized. This table also orders them according to its complexity, starting with the simpler Goal Attainment method and ending with the completer Valuation Scale.

2.4 Multi-criteria Problems Solving Mechanisms

Once the optimization problem is modeled or formulated, the solution is found after the application of an optimization method.

Most optimization algorithms frequently imply an iterative searching process. Beginning with an initial approach to the solution, the algorithm performs consecutive steps towards the termination point. The search strategy states the difference among the diverse methods and there is no universal method applicable to any kind of problem. Table 3 shows a classification of the main optimization solving families.

Table 3. Classification of optimization solving methods.

Weighted sum [21]	The multiple objective functions are aggregated in a single function by the assignment of weights
Random search [13]	Generate random numbers to explore the search (feasible) space
Tabu search [14]	Iteratively make movements around the current solution constrained by a group of forbidden or tabu movements
Physical programming [31]	Incorporate preferences without the need of weight assignment. Address both design metrics and constraints in the same way, integrating them into the utility function
Lexicographic [6]	Objective functions are processed in a hierarchical basis
Genetic and evolutionary [12, 15]	Imitate the optimization process of the natural selection. Employ techniques such as heredity, mutation, natural selection or factor recombination to explore the feasible space and select the current solution
Simulated annealing [14]	Imitate the iterative process of cold and heat application for metal annealing by increasing or decreasing the difference between the ideal solution and the current approach
Ant colony optimization (ACO) and swarm optimization [40]	Imitate animal behavior related to their intra-group communication or their search for the optimal ways towards the food
Outranking methods [46]	Build an ordered relation of the feasible alternatives based on the defined preferences over a set of criteria to eventually complete a recommendation

2.5 Fairness Consideration

Traditionally, the goal of any optimization problem has been the search for the optimum solution for a given situation among all the possible ones in the feasible solution space. This optimality meaning has often been understood as a Pareto Optimum, i.e., the result of the maximization/minimization of the objective functions (or criteria), where the result of none of the objective functions can be improved, but at the expense of worsening another one. Finding a Pareto-optimal solution means finding the technically most efficient solution. And applying this concept to the field of networking, this optimality results on the optimum distribution of resources among the flows traveling through the network.

Obviously, an optimal distribution of resources not always implies an equitable use of them. Indeed, in some cases it may lead to absolutely unfair situations that entail the exhaustion of some resources. In that sense, the efficient assignment of resources derived from the direct application of optimization

algorithms may leave without service some customers or final users, due to the provision of all the benefit to others (see examples in [9,38]). Obviously, the global utility of the system is the maximum, but the result is clearly unfair, and the situation worsens as the heterogeneity of the final users increases.

The conflict between the maximization of the benefit, the optimal resource allocation and the fairness of the distribution is a field that has been widely analyzed in Economy, as part of microeconomics or public finances. The conclusion is that the incompatibility between fairness and efficiency is not a design problem of the optimization algorithms, but of the formulation of the problem to be optimized, where the fairness concept must be included. The difficulty rises up since efficiency is an objective or technical goal that, in consequence, can be measured and assessed quantitatively. This has nothing to do with the concept of fairness, a subjective concept whose assessment is not trivial.

Although fairness may initially seem to be easy to define, it has a variety of aspects that complicate its proper delimitation. Taking the sense of equanimity, an equitable distribution of resources could be defined as an evenly split available resource assignment among the flows competing for them. The disadvantage of this distribution is that it does not take into account the specific necessities of each flow. If all the flows obtain the same portion of resources, those with lower requirements benefit from a proportionally higher resource quantity.

Changing the definition of fair distribution to that assigning the resources proportionally on the basis of flow requirements is neither the ideal solution. In this case, the most consuming items are benefited, i.e., those which contribute more to the network congestion, to the detriment of lighter transmissions and consequently, of the global performance of the network.

In addition, other aspects such as cooperation must also be considered. There may be some nodes in the network not willing to give up their resources to other transmissions, and so, this kind of behavior should be punished. But, what happens when a node doesn't give up resources to the network due to the lack of them? It would be the case of a node with low battery or low capacity links. Would these be reason enough to reduce the transmission resources that have been assigned? In this case, would the distribution be fair? This conflict remains unsolved, although some approaches have been formulated and are discussed next.

The work in [8] presents several interpretations of the concept of fairness. In one hand, there is the widely accepted *max-min fairness* definition [38], usually employed in social science. It is based in the search for consecutive approaches to the optimum solution in a way that no individual or criterion can improve its state or utility if it means a loss for a weaker individual or criterion.

Translating this concept to communications, the distribution of network resources is considered max-min fair when all the minimum transmission rates of the data flows are maximized and all the maximum transmission rates are minimized. It is proven that this fairness interpretation is Pareto-efficient.

Another interpretation of fairness that also searches for the trade-off between efficiency and equity is the *proportional fairness* [26]. A resource distribution

among the network flows is considered proportional when the planned priority of a flow is inversely proportional to the estimated resource consumption of this flow. It can also be proven that the proportional fairness is Pareto-efficient.

Both aforementioned interpretations the bandwidth is shared to maximize some utility function for instantaneous flows. This means that the optimality of the resource assignment is measured for a static combination of flows. Taking into account the real random nature of the network traffic, it is necessary to define the utility in terms of the performance of individual flows with finite duration. And in this case, it is not so clear that the max-min or proportional fairness concepts reach an optimum result. With random traffic, the performance and, in consequence, the utility depend on precise statistics of the offered traffic and are hard, if not impossible, to be analytically assessed.

Sharing flows under a *balanced fairness* criterion [9], the performance becomes indifferent to the specific traffic characteristics, simplifying its formulation. The term balanced fairness comes from the necessary and sufficient relations that must be fulfilled to guarantee the insensitiveness in stochastic networks. This insensitiveness entails that the distribution of the active flow number and, in consequence, the estimated throughput, depends just on the main traffic offered in each route.

Balanced fairness makes it possible to approach the behavior of the elastic traffic over the network and, in addition to the insensitiveness property, it also makes it possible to find the exact probability of the distribution of concurrent flows in different routes and then evaluate the performance metrics.

The balanced fairness is not always Pareto-efficient, but in the case that existing one, it will be one of a kind.

3 Cost/Energy/*-Aware Network and Cloud Services Management Scenarios

Once most well known multi-criteria optimization techniques are introduced, the next step is to analyze the application scenarios. This section overviews several research scenarios where energy-aware control of different systems has been considered as part of the ACROSS project. The scenarios include the following: modeling and analysis of performance-energy trade-off in data centers, characterization and energy-efficiency of applications in cloud computing, energy-aware load balancing in 5G HetNets and finally incorporating energy and cost to opportunistic QoE-aware scheduling.

3.1 Modeling and Analysis of Performance-Energy Trade-Off in Data Centers

An increasing demand for green ICT has inspired the queueing community to consider energy-aware queueing systems. In many cases, it is no longer enough to optimize just the performance costs, but one should also take into account the energy costs. An idle server (waiting for an arriving job to be processed) in the

server farm of a typical data center may consume as much as 60% of the peak power. From the energy point of view, such an idle server should be switched off until a new job arrives. However, from the performance point of view, this is suboptimal since it typically takes a rather long time to wake the server up. Thus, there is a clear trade-off between the performance and energy aspects.

The two main metrics used in the literature to analyze the performance-energy trade-off in energy-aware queueing systems are ERWS and ERP. Both of them are based on the expected response time, $E[T]$, and the expected power consumption per time unit, $E[P]$. The former one, ERWS, is defined as their weighted sum, $w_1 E[T] + w_2 E[P]$ and the latter one, ERP, as their product, $E[T] \cdot E[P]$. Also, generalized versions of these can be easily derived.

Here we model data centers as queueing systems and develop policies for the optimal control of the performance-energy trade-off. For a single machine the system is modeled as an M/G/1 queue. When considering a whole data-center, then a natural abstraction of the problem is provided by the dispatching problem in a system of parallel queues.

Optimal Sleep State Control in M/G/1 Queue: Modern processors support many sleep states to enable energy saving and the deeper the sleep state the longer is the setup delay to wake up from the sleep state. An additional feature in the control is to consider if it helps to wait for a random time (idling time) after busy period before going to sleep. Possible approaches for the sleep state selection policy include: randomized policy, where processor selects the sleep state from a given (optimized) distribution, or sequential policy, where sleep states are traversed sequentially starting from the lightest sleep state to the deepest one. Analysis of such a queueing system resembles that of classical vacation models.

Gandhi et al. see [17], considered the M/M/1 FIFO queue with deterministic setup delay and randomized sleep state selection policy but without the possibility of the idle timer, i.e., the timer is either zero or infinite, and they showed for the ERP metric that the optimal sleep state selection policy is deterministic, i.e., after busy period the system goes to some sleep state with probability 1 (which depends on the parameters). Maccio and Down [29] added the possibility of an exponential idle timer in the server before going to sleep, and showed for the ERWS cost metrics and for exponential setup delays that the optimal idle timer control still sets the idle timer equal to zero or infinite, i.e., the idle timer control remains the same. Gebrehiwot et al. considered the more general M/G/1 model with generally distributed service times, idle timer distributions and setup delays, both ERP and ERWS cost metrics (and even slightly more generalized ones) and randomized/sequential sleep state selection policies. Assuming the FIFO service discipline, it was shown in [20] that even after all the generalizations the optimal control finally remains the same: the optimal policy (a) either never uses any sleep states or (b) it will directly go to some deterministic sleep state and wake up from there. This result was shown to hold for the Processor Sharing (PS) discipline in [19] and for the Shortest Remaining Processing Time (SRPT) discipline in [18]. Thus, it is plausible that the result holds for any work-conserving discipline.

Energy-Aware Dispatching with Parallel Queues: The data center can be modeled as a system of parallel single-server queues with setup delays. The system receives randomly arriving jobs with random service requirements. The problem is then to identify for each arrival where to dispatch arriving new jobs based on state information available about the system, e.g., the number of jobs in the other queues. Another modeling approach is to consider a centralized queue with multiple servers, i.e., the models are then variants of the multiserver M/M/n model.

In the parallel queue setting and without any energy-aware considerations, the optimality of the JSQ policy for minimizing the mean delay with homogeneous servers is one classical result, see [48]. However, in an energy-aware setting the task is to find a balance for using enough servers to provide reasonably low job delay while taking into account the additional setup delay costs, and to let other servers sleep to save energy. Achieving this is not at all clear. For the centralized queue approach, Gandhi et al. proposed the delayed-off scheme, where servers upon a job completion use an idle timer, wait in the idle state for this time before going to sleep, and new jobs are sent to idle servers if one is available or otherwise some sleeping server is activated. An exact analysis under Markovian assumptions was done in [16], and it was shown that by appropriately selecting the mean idle timer value, the system keeps a sufficient number of servers in busy/idle state and allows the rest to sleep. An important result has been only recently obtained by Mukherjee et al. in [33], which considers the delayed-off scheme in a distributed parallel queue setting; it was shown that asymptotically delayed-off can achieve the same delay scaling as JSQ, i.e., is asymptotically delay optimal, and at the same time leaves a certain fraction of servers in a sleep state, independent of the value of the idle timer and the setup delay. This result holds asymptotically when the server farm is large with thousands of servers.

However, in a small/moderate sized data center there is still scope for optimization. In this setting the use of MDP (Markov Decision Process) and Policy Iteration has been recently considered by Gebrehiwot et al. in [28], where the data center is assumed to consist of two kinds of servers: normal always-on servers and instant-off servers, which go to sleep immediately after queue empties, i.e., there are no idle timers, and an explicit near optimal policy is obtained for minimizing the ERWS metric that uses as state the number of jobs in the queues and the busy/sleep status. Also, size-aware approaches with MDP have been recently applied by Hyytiä et al. in [24, 25].

3.2 Characterization and Energy-Efficiency of Applications in Cloud Computing

Modeling Applications. With the goal of improving energy efficiency in cloud computing, several authors have studied the different factors that are causing energy loss and energy waste in data centers. In [32], the different aspects are discussed in detail, and idle runs are discussed as one of the causes for energy waste, as already mentioned earlier in this chapter. Low power modes have been proposed in the literature both for servers and storage components, however

their benefits are often limited due to their transition costs and inefficiencies. To improve energy efficiency and reduce the environmental impact of federated clouds, in the EU project ECO₂Clouds [47] an adaptive approach to resource allocation is proposed, based on monitoring the use and energy consumption of resources, and associating it to running applications. The demand for resources can therefore be associated to applications requesting resources, rather than only to the scheduling of resources and tasks in the underlying cloud environment.

Along this line, we have studied within ACROSS how different types of applications make use of resources, with the goal of improving energy efficiency.

As mentioned in Sect. 3.1, to compare different solutions in terms of response time and power consumption, the two main approaches are ERWS and ERP. An alternative, which allows evaluating energy efficiency at application level, is the *energy per job* indicator. This indicator allows comparing different solutions in terms of work performed, rather than on performance parameters, and to discuss ways of improving energy efficiency of applications in terms of application-level parameters.

Another aspect which has been considered is that increasing resources is not always beneficial in terms of performances, as the systems may present bottlenecks in their execution which may cause inefficiencies in the system: in some cases, the additional resources will worsen energy efficiency, as the new resources are not solving the problem and are themselves underutilized. As a consequence, in considering energy efficiency in applications in clouds, some aspects can better characterize the use of resources:

- *Shared access to resources*: during their execution application can request access to shared resources with an impact on energy consumption due to synchronization and waiting times.
- *The characterization of the application execution patterns*: batch applications and transactional applications present different execution patterns: in batch applications the execution times are usually longer with larger use of resources, but response time constraints are not critical; in transactional applications, response times are often subject to constraints and the allocated resources must guarantee they are satisfied.

These application-level aspects have an impact on the resource allocation criteria in different cases. In the following, we discuss how to model batch and transactional applications considering these aspects with the goal of choosing the number of resources to be associated to an application in terms of VMs with the goal of minimizing the energy-per-job parameter.

Batch Applications. Batch applications have been studied in detail in [22] to consider the following aspects: number of VMs allocated for executing a batch of similar applications, shared resources (in particular shared storage access and access synchronization), heterogeneous deployments environments for VMs, with servers with different capacity.

While for the details we refer to [22], we summarize here the main characteristics of the approach. The general goal is to minimize idle time to improve

energy efficiency, while avoiding to increase execution time for each application in the batch, which would result in an increase of the total energy. We assume that in computing the energy per job, idle time is distributed to all applications being run on the system in an equal basis. Queuing models have been developed to represent applications, in terms of computing nodes to execute the application and storage nodes for data access, which is assumed to be shared, with the possibility of choosing between asynchronous access and synchronous access (with synchronization points). In both cases the critical point is represented by the ratio between the service time for computing nodes and the service time for storage access: going beyond this point the energy per job is increasing without significant benefit in execution times.

An example is shown in Fig. 2, where it is clear that increasing the number of VMs for an application after the critical point is mainly resulting in a loss of energy efficiency, both with synchronous and asynchronous storage access.

Transactional Workloads. For transactional workloads, the main application-level parameter affecting energy consumption is the arrival rate. In fact, assuming an exponential distribution of arrivals, if the arrival rate λ is much lower than the service time, the idle times will be significant. On the other hand, getting closer to service time, the response time will increase, as shown in Fig. 3. The details of the computations can be found in [23]. The paper also describes how different load distribution policies for VMs can influence energy-per-job. Assuming again that idle power is uniformly distributed to all VMs running on the same host, three policies have been evaluated: (1) distributing the load equally; (2) allocating larger loads to VMs with lower idle power; (3) allocating larger loads to VMs with higher idles power. Initial simulation results result in Policy 2 being the worst, while Policy 1 and 3 are almost equivalent, with Policy 1 resulting in better energy-per-job and Policy 3 in better response times [23].

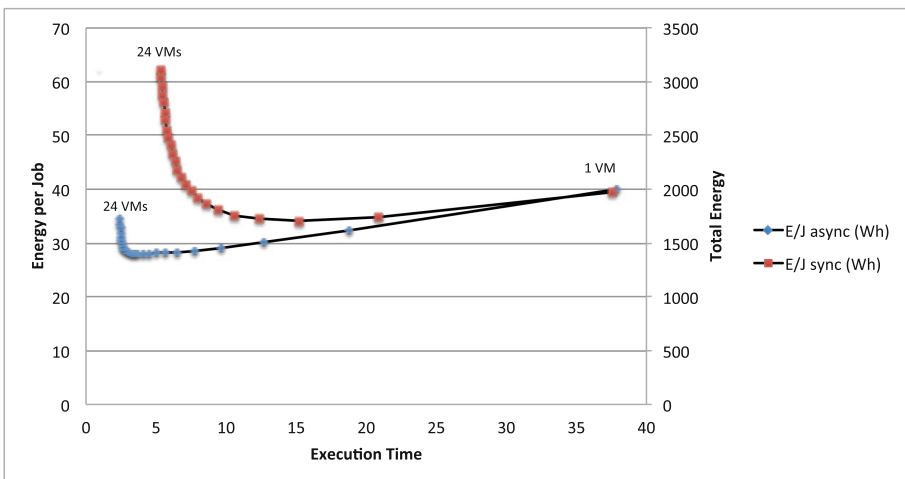


Fig. 2. Energy per job in batch applications

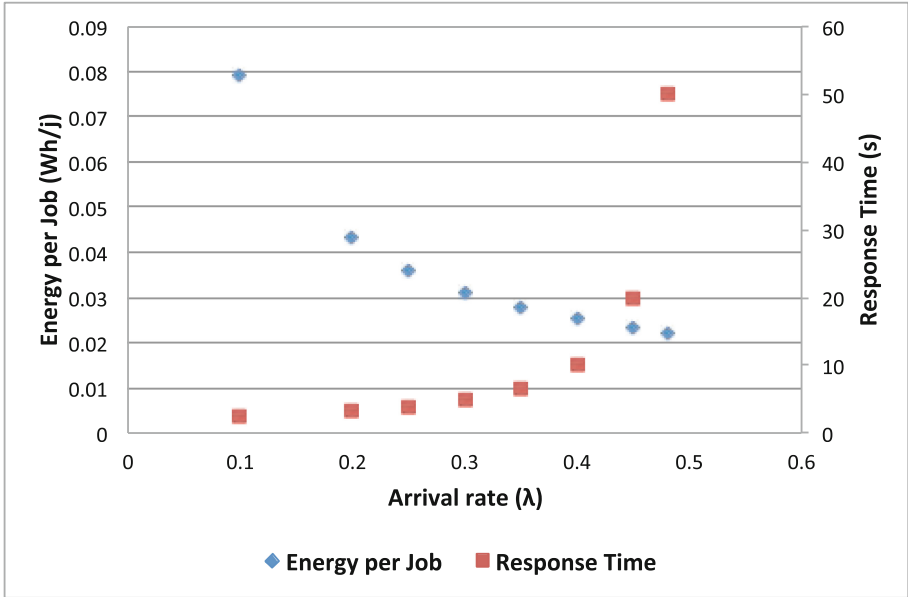


Fig. 3. Energy-per-job in transactional applications [23]

3.3 Energy-Aware Load Balancing in 5G HetNets

The exponential growth of mobile data still continues and heterogeneous networks have been introduced as a vital part of the network architecture of future 5G networks. Heterogeneous networks (HetNet) especially alleviate the problem that the user data intensity may have spatially large variations. These are network architectures with small cells (e.g., pico and femtocells) overlaying the macrocell network. The macrocells are high power base stations providing the basic coverage to the whole cell area, while the small cells are low power base stations used for data traffic hotspot areas within a macrocell to improve spectral efficiency per unit area or for areas that the macrocell cannot cover efficiently.

In HetNets, when a user arrives in the coverage area of a small cell it can typically connect to either the local small cell or to the macrocell, as illustrated in Fig. 4. Typically, the small cells offer in its coverage area a possibility for achieving high transmission rates. However, depending on the congestion level at the small cell it may be better from the system point of view to utilize the resources of the macrocell instead. This raises the need to design dynamic load balancing algorithms. In 5G networks the energy consumption of the system will also be an important factor. Thus, the load balancing algorithms must be designed so that they take into account both the performance of the system, as well as the energy used by the whole system.

Consider a single macrocell with several small cells inside its coverage area. The small cells are assumed to have a wired backhaul connection to the Internet.

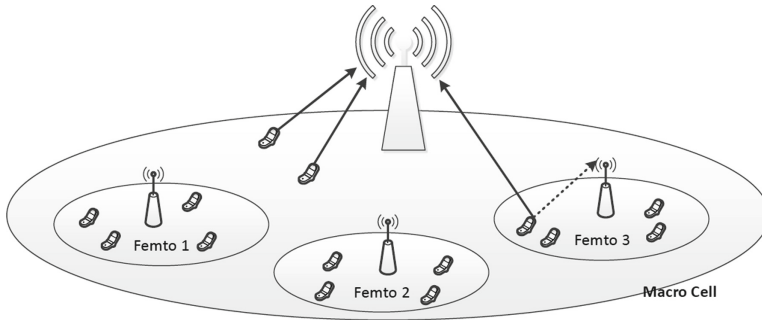


Fig. 4. User inside a femtocell may connect either to the local small cell (femto) or the macrocell to achieve better load balancing.

They typically also operate on a different frequency than the macrocell and hence do not interfere with the transmissions of the macrocell. From the traffic point of view, each cell can be considered, whether it is the macrocell or a small cell, as a server with its own queue each having its own characteristics. The traffic itself may consist, for example of elastic data flows. The load balancing problem then corresponds to a problem of assigning arriving jobs or users to parallel queues. The difference to a classical dispatching problem, where an arrival can be routed to any queue, is that in this case the arrival can only select between two queues: its own local queue or the queue representing the macrocell.

In order to include the energy aspects in the model, the macrocell must be assumed to be operating at full power continuously. This is because the macrocell provides the control infrastructure and the basic coverage in the whole macrocell region and it can not be switched off. However, depending on the traffic situation it may be reasonable to switch off a low power small cell since the small cells typically have power consumption at least an order of magnitude lower than the macrocell. The cost of switching off a base station is that there may be a significant delay, the so-called set up delay, when turning the base station back on again. The queuing models used for the small cells must then be generalized to take this into account.

The resulting load balancing problem that optimizes for example the overall weighted sum of the performance and the energy parts of the whole system is difficult. However, it can be approached under certain assumptions by using the theory of Markov Decision Processes. This has been done recently by Taboada et al. in [42], where the results indicate that a dynamic policy that knows the sleep state of the small cells and the number of flows when compared with an optimized randomized routing policy is better able to keep the small cells sleeping and it thus avoids the harmful effect of setup delays leading to gains for both the performance and energy parts, while at high loads the energy gain vanishes but the dynamic policy still gives a good improvement in the performance.

3.4 Incorporating Energy and Cost to Opportunistic QoE-Aware Scheduling

One of the fundamental challenges that network providers nowadays face is the management for sharing network resources among users' traffic flows so that most of traditional scheduling strategies for resource allocation have been oriented to the maximization of objective quality parameters. Nevertheless, considering the importance and the necessity of network resource allocation for maximizing subjective quality, scheduling algorithms aimed at maximizing users' perception of quality become essential.

Thus, to overcome the lacks found in the field of traffic flow scheduling optimization, during the last years we have analyzed the following three stochastic and dynamic resource allocation problems:

1. Subjective quality maximization when channel capacity is constant [44],
2. Subjective quality maximization in channels with time-varying capacity [43],
3. Mean delay minimization for general size distributions in channels with time-varying capacity [41, 45].

Since these problems are analytically and computationally unfeasible for finding an optimal solution, we focus on designing simple, tractable and implementable well-performing heuristic priority scheduling rules.

For this aim, our research is focused on the Markovian Decision Processes (MDP) framework and on Gittins and Whittle methods [41, 43–45] to obtain scheduling index rule solutions. In this way, first of all, the above scheduling problems are modeled in the framework of MDPs. Later, using methodologies based on Gittins or/and Whittle approaches for their resolution, we have proposed scheduling index rules with closed-form expression.

The idea of Gittins consists in allocating resources to jobs with the current highest productivity of using the resource. The Gittins index is the value of the charge that provides that the expected serving-cost to the scheduler is in balance with the expected reward obtained when serving a job in r consecutive time slots, which results in the ratio between the expected total reward earned and the expected time spent in the system when serving a job in r consecutive time slots.

On the other hand, the Whittle approach consists in obtaining a function that measures the dynamic service priority. For that purpose, the optimization problem formulated as a Markov Decision Process (MDP) can be relaxed by requiring to serve a job per slot on average, which may allow introducing the constraint inside the objective function. Then, it is further approached by Lagrangian methods and can be decomposed into a single-job price-based parametrized optimization problem. Since the Whittle index is the break-even value of the Lagrangian parameter, it can be interpreted as the per cost of serving. In such a way, the Whittle index represents the rate between marginal reward and marginal work, where marginal reward (work) is the difference between the expected total reward earned (work done) by serving and not serving at an initial state and then employing a certain optimal policy.

As a first step towards ACROSS targeted multi-criteria optimization, it is worth mentioning the utility-based MDP employed in [43, 44] for QoE maximization. This function depended on delay only but we plan to extend it to a generic problem aimed at maximizing a multivariate objective function. Considering the meaning of work and reward in Whittle related modeling, such extension could demand the modification of the structure of the problem itself (i.e., alternative MDP) or just considering different criteria in the work/reward assignments.

Although we carried out some very preliminary tests with LP and AHP based articulation of preferences for QoE vs. energy optimization in [27] we plan to further analyze index rules techniques in the multi-criteria problem.

4 Current Technologies and Solutions

Research on energy-aware control has been actively pursued in the academia already for a long time, and Sect. 3 introduced several scenarios that have analyzed and given valuable insights to the fundamental tradeoff between energy efficiency and QoS/QoE. Due to the rising costs of energy, the industry is also actively developing solutions that would enable more energy efficient networks. Next we review industry efforts towards such architectures and finally we introduce a framework for energy-aware network management systems.

4.1 Industry Efforts for Integrating Energy Consumption in Network Controlling Mechanisms

New network technologies have been recently started to consider cost/energy issues in the early stages of the design and deployment process. Besides the infrastructure upgrade, the incorporation of such technologies requires the network managers must handle a number of real-time parameters to optimize Network energy /cost profile. These parameters include, among others, the sleep status of networks elements or the activation of mobile resources to provide extra coverage or change in performance status of some of the processors in the network.

The fact is that energy consumption in networks is rising. Therefore, network equipment requires more power and greater amounts of cooling. According to [27]. By 2017 more than 5 zettabytes of data will pass through the network every year. The period 2010–2020 will see an important increase in ICT equipment to provide and serve this traffic. Smartphones and tablets will drive the mobile traffic to grow up to 89 times by 2020, causing energy use to grow exponentially. For example, mobile video traffic is expected to grow 870%, M2M (IoT) 990% and Applications 129%. As a consequence, ICT will consume 6% of Total of Global Energy consumption: in 2013 it was 109,1 GW according to the energy use models at different network levels shown in Table 4.

Table 4. Energy use models.

Devices	Networks
PC's 36,9 GW	Home & Enterprise 9,5 GW
Printer 0,9 GW	Access 21,2 GW
Smartphones 0,6 GW	Metro 0,6 GW Aggregation and transport
Mobile 0,6 GW	Edge 0,7 GW
Tablets 0,2 GW	Core 0,3 GW
	Service Provider & Data Center 37,1 GW

One of the challenges the industry faces is how to support that growth in a sustainable and economically viable way. However, there is an opportunity for important reductions in the energy consumption because the networks are dimensioned in excess of current demand and even when the network is low in traffic the power used is very important and most of it is wasted [34]. The introduction of new technologies will provide a solution to improve the energy efficiency at the different scenarios (see Table 5).

Table 5. Scenarios for energy efficiency increase.

Home: Sleep mode	
Office: Cloud	
Access: VDL2, Vectoring, VoIP	Wireless Access: LTE Femto, Small, HetNet IP: MPLS Backhaul Fixed Wireless: Microwave Backhaul for Wireless 2G 3G, Fiber Copper: VDL2, Vectoring, VoIP, PON
Metro: IP/MPLS Transport, Packet Optical	
Edge: IP Edge	
IP Core: Next Gen IP Router and Transport (10 Gb)	
Service Provider & Data Center	

Current forecasts estimate that the trend will be to manage energy consumption and efficiency policies based on different types of traffic. Two organizations pursuing this goal are introduced next.

GeSI Global e-Sustainability Initiative (GeSI) [2]. Building a sustainable world In collaboration with members from major Information and Communication Technology (ICT) companies and organisations around the globe, the Global e-Sustainability Initiative (GeSI) is a leading source of impartial information, resources and best practices for achieving integrated social and environmental sustainability through ICT.

In a rapidly growing information society, technology presents both challenges and opportunities. GeSI facilitates real world solutions to real world issues both within the ICT industry and the greater sustainability community. We contribute to a sustainable future, communicate the industry's corporate responsibility efforts, and increasingly drive the sustainability agenda.

Members and Partners: ATT, Telecom Italia, Ericsson, KPN, Microsoft, Nokia, Nokia Siemens.

Green Touch [3]. GreenTouch is a consortium of leading Information and Communications Technology (ICT) industry, academic and non-governmental research experts dedicated to fundamentally transforming communications and data networks, including the Internet, and significantly reducing the carbon footprint of ICT devices, platforms and networks.

4.2 C-RAN: Access Network Architecture of Future 5G Networks

Cloud computing represents a paradigm shift in the evolution of ICT and has quickly become a key technology for offering new and improved services to consumers and businesses. Massive data centers, consisting of thousands of connected servers, are fundamental functional building blocks in the implementation of cloud services. With the rapidly increasing adoption of cloud computing, the technology has faced many new challenges related to scalability, high capacity/reliability demands and energy efficiency. At the same time, the huge increase in the processing capacity enables the use of more accurate information that the control decision may be based on. This justifies the development of much more advanced control methods and algorithms, which is the objective of the work as described earlier in Sect. 3.

To address the growing challenges, the research community has proposed several architectures for data centers, including FatTree, DCell, FiConn, Scafida and JellyFish [7]. On the other hand, vendors, such as, Google, Amazon, Apple, Google etc., have been developing their own proprietary solutions for the data centers which has created interoperability problems between service providers. To push forward the development of architectures addressing the challenges and to enable better interoperability between cloud service providers, IEEE has launched the IEEE Cloud Computing Initiative which is developing presently two standards in the area: IEEE P2301 Draft Guide for Cloud Portability and Interoperability Profiles and IEEE P2302 Draft Standard for Intercloud Interoperability and Federation.

Cloud-based approaches are also considered as part of the development of the future 5G networks. Namely, in the C-RAN (Cloud-Radio Access Network) architecture [35] the radio access network functionality is moved to the cloud. This means that all the radio resource management and cell coordination related functionality requiring complex computations are implemented in the cloud. This makes the functionality of the base stations simpler and hence also cheaper to manufacture. However, this places tough requirements on the computing capacity and efficiency of the centralized processing unit, essentially a data center, and

the interconnection network between the base stations and the data center. Several projects based on the C-RAN architecture have been initiated in the Next Generation Mobile Networks (NGMN) consortium and EU FP7 [10], and the C-RAN architecture will most likely be considered also in the standardization by 3GPP.

4.3 A Framework for Energy-Aware Network Management Systems

Considering the problem modeling and the existing optimization frameworks described in the previous sections, the challenge now is the integration of energy consumption in network controlling mechanisms. The networks in the data centers and in the operators world are showing a fast evolution with growing size and complexity that should be tackled by increased flexibility with softwarization techniques.

Emerging 5G Networks now exhibit extensive softwarization of all network elements: IoT, Mobile, and fiber optics-based transport core. This functions should be integrated in a network management environment with autonomous or semi-autonomous control response capabilities based on defined SLA's and applying policies and using simulated scenarios and past history learning.

By monitoring the energy parameters of radio access networks, fixed networks, front haul and backhaul elements, with the VNFs supporting the internal network processes, and by estimating energy consumption and triggering reactions, the energy footprint of the network (especially backhaul and fronthaul) can

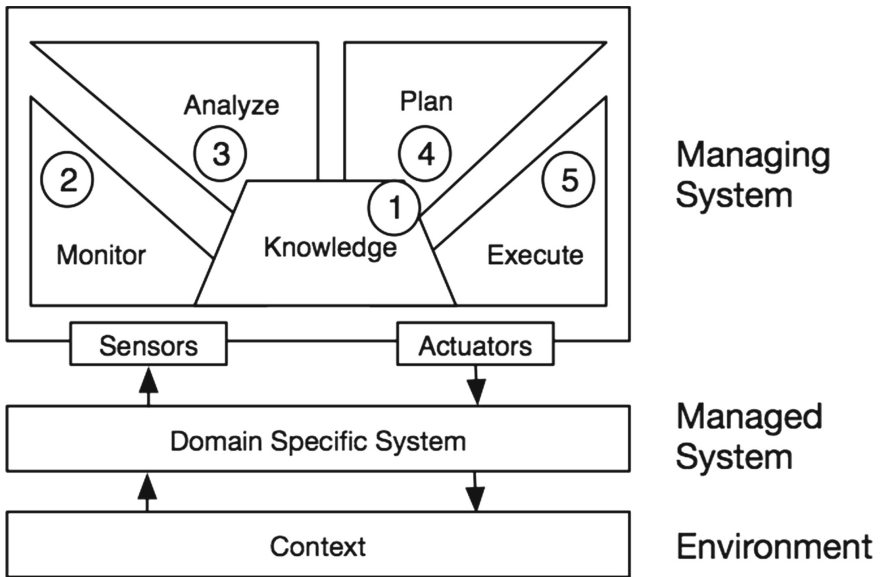


Fig. 5. MAPE-K diagram.

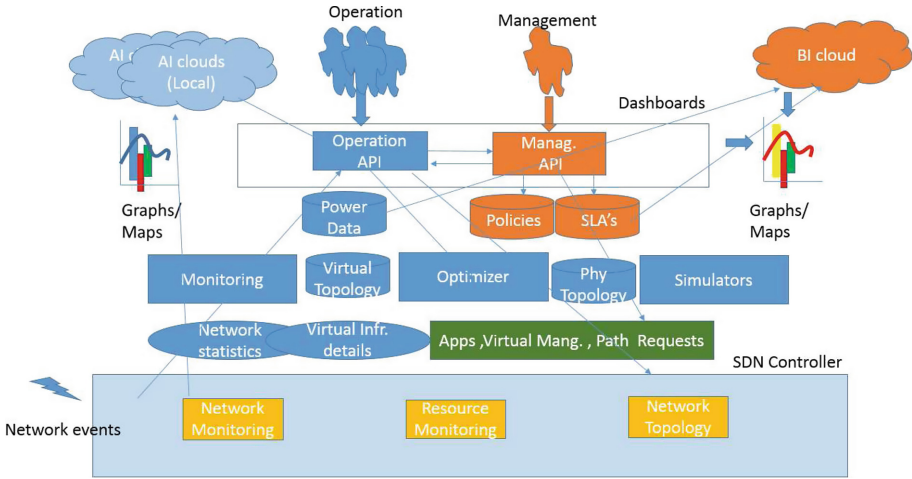


Fig. 6. Functional description of an energy management and monitoring application.

be reduced while maintaining QoS for each VNO or end user. An Energy Management and Monitoring Application can be conveniently deployed along a standard ETSI MANO and collect energy-specific parameters like power consumption and CPU loads (see Figs. 5 and 6). Such an Energy Management and Monitoring Application can also collect information about several network aspects such as traffic routing paths, traffic load levels, user throughput and number of sessions, radio coverage, interference of radio resources, and equipment activation intervals. All these data can be used to compute a virtual infrastructure energy budget to be used for subsequent analyses and reactions using machine learning and optimization techniques [11].

The application can optimally schedule the power operational states and the levels of power consumption of network nodes, jointly performing load balancing and frequency bandwidth assignment, in a highly heterogeneous environment. Also the re-allocation of virtual functions across backhaul and front haul will be done as part of the optimization actions, in order to cover virtual network functions to less power-consuming or less-loaded servers, thus reducing the overall energy demand from the network.

Designing software systems that have to deal with dynamic operating conditions, such as changing availability of resources and faults that are difficult to predict, is complex. A promising approach to handle such dynamics is self-adaptation that can be realized by a Monitor-Analyze-Plan-Execute plus Knowledge (MAPE-K) feedback loop. To provide evidence that the system goals are satisfied, regarding the changing conditions, state of the art advocates the use of formal methods.

Research in progress [1] tries to reinforce the approach of consolidating design knowledge of self-adaptive systems with the traditional tools of SLA's and policy modules and in particular with the necessity of defining the decision criteria

using formalized templates and making it understandable for a human operator or manager via the human interfaces and dashboards as shown in Fig. 6. This figure shows the proposed architecture of an advanced Network Monitoring and Management System that includes energy management. At the top are the two agents responsible for the management of the network: On one side those responsible for the negotiating the SLA's with the customers and of establishing the policies of the operation. On the other side those responsible for the detailed technical operation. These roles are supported by a set of applications and reside in the corresponding specialized cloud environments. The Business Intelligent cloud helps the Management API to generate dashboards for the optimization of the operation business results, issuing recommendations to the managers or autonomously implementing decisions. Those decisions will be based dynamically on contractual commitments, market conditions and customer's needs. The operational cloud supports the technical operations with specialized technical AI dashboards using available information from many sources: Network monitoring information including real and historical performance data from the network, power data and network statistics. Simulated data can be used to support the operation by providing hypothetical failure scenarios, possible solutions and the impact of applying those solutions. This helps together with the historical data with the analysis of the consequences of possible decisions when trying to solve specific incidents. As in the Business application the operational cloud will analyse the scenarios and select the optimal configuration autonomously or mediated by the operator interaction via the corresponding dashboards reducing the total energy footprint of the network. At the bottom of Fig. 6 is the SND network Controller with access to Network and Resource Monitoring and Topology that reacts to real Network events implementing the required network solution as directed by the layers above.

5 Conclusions and Foreseen Future Research Lines

5.1 Conclusions

This chapter has addressed the challenges of combining energy and QoS/QoE issues in the management mechanisms of network and cloud services. Unfortunately, these design parameters are usually conflicting and it is necessary to introduce multi-criteria optimization techniques in order to achieve the required trade-off solution.

So, as a first step, the common issues related to multi-objective optimization problems and mechanism have been depicted. These issues include typical preference articulation mechanisms, typical optimization methods and fairness considerations. Then, most well-known optimization methods have been briefly summarized in order to provide Internet of Services research community with a broad set of tools for properly addressing the inherent multi-criteria problems.

Finally, in the multiuser/multiservice environments considered in ACROSS, how resources are distributed and the impact into different kind of users must be carefully tackled. As analyzed, fairness is most of the times considered once

the algorithm has selected the most efficient (i.e., optimal) solution. However, the incompatibility between fairness and efficiency is not a design problem of the optimization algorithms, but of the formulation of the problem to be optimized, where the fairness concept must be included.

The next step is to model and analyze the problem of including the performance/energy trade-off into different scenarios in the scope of the ACROSS project. We start this analysis studying the use case of data centers modeled as queuing systems to develop policies for the optimal control of the QoS/QoE-energy balance. The trend in this area is to focus in small/moderate size data centers.

Then, the second scenario focuses on the way different applications use the resources available in cloud environments and its impact in terms of energetic cost. Considering that increasing resources does not always benefit the performance of the system, we analyze two application-level approaches in order to improve energy efficiency: the characterization of the application execution patterns and the shared access to resources.

Next, we show an example of energy-aware load balancing in 5G HetNets where cells of different sizes are used to adapt the coverage to the variations of user data traffic. We discuss the challenge of designing a load-balancing algorithm that considers both the performance of the system and the energy consumption of the whole system. The discussion suggests a MDP approach for the multi-criteria optimization problem.

The last analyzed scenario presents a network services provider that shares resources among different traffic flows. The goal here is to introduce energy and cost into opportunistic QoE-aware scheduling. The research focuses on the use of MDP framework to model the scheduling problem and the application of Gittins and Whittle methods to obtain scheduling index rule solutions.

Finally, the chapter compiles the current state of emerging technologies and foreseen solutions to the energy/performance trade-off issue in network and cloud management systems addressed in the ACROSS project. Based on the expected huge increase of network traffic and, in consequence, of energy consumption, the design of upcoming network management systems must face the challenge of addressing power efficiency while still meeting the KPIs of the offered services. Industry is already fostering innovative initiatives to integrate energy issues into network controlling mechanisms.

In this direction, we present C-RAN architecture as the cloud-based solution for the future 5G access network. This approach moves all the radio resource management and cell coordination functionality to the cloud. The increasing complexity of the service management and orchestration in the cloud requires advanced network control methods and algorithms. Therefore, as final conclusion, we suggest a framework to include energy awareness in network management systems that implements a MAPE-K feedback loop.

5.2 Future Work

The joint research accomplished in the scope of the COST ACROSS action has allowed the identification of common interests to develop in future collaborations. Remaining under the umbrella of Energy/Cost-aware network management, this future work will strongly rely on the application of multi-criteria optimization techniques in order to cope with conflicting performance objectives.

As previously concluded, the consideration of fairness in a optimization process does not fall to the multi-criteria optimization algorithm. On the contrary, it must be considered in the formulation of the design problem itself. Therefore one of the issues that will be addressed in future work grounded in the result of the COST ACROSS action is the inclusion of fairness among users/services/resource allocation in the definition network and services management optimization.

Besides, analyzing the problem of the introduction of energy-awareness in load balancing processes in 5G HetNets, another of the proposed future research lines is to use MDP and Policy Iteration in order to optimize the dispatching problem focusing in small/moderate size data centers. Similarly, we also found common interests in the development of further analysis of index rules techniques in the multi-criteria problem of opportunistic QoE-aware scheduling.

Finally, research in progress envisages innovative initiatives to integrate energy issues into network controlling mechanisms and interactive management approaches including self-adaption features.

Acknowledgment. The research leading to these results has been supported by the European Commission under the COST ACROSS action, supported by COST (European Cooperation in Science and Technology), and by Spanish MINECO under the project 5RANVIR (no. TEC2016-80090-C2-2-R).

References

1. Decide Project ICT H2020 ID: 731533: ICT-10. Software Technologies. Technical report
2. GeSI home: thought leadership on social and environmental ICT sustainability. <http://gesi.org/>. Accessed 09 Oct 2017
3. GreenTouch. <https://s3-us-west-2.amazonaws.com/belllabs-microsite-greentouch/index.html>. Accessed 09 Oct 2017
4. Aleskerov, F.T.: Threshold utility, choice, and binary relations. *Autom. Remote Control* **64**(3), 350–367 (2003)
5. Andersson, J.: A survey of multiobjective optimization in engineering design. University, Linköping, Sweden, Technical Report No: LiTH-IKP-R-1097 (2000)
6. Belton, V., Stewart, T.J.: *Multiple Criteria Decision Analysis: An Integrated Approach*. Kluwer Academic Publishers, New York (2002)
7. Bilal, K., Malik, S.U.R., Khan, S.U., Zomaya, A.Y.: Trends and challenges in cloud datacenters. *IEEE Cloud Comput.* **1**(1), 10–20 (2014)
8. Bonald, T., Massoulié, L., Proutière, A., Virtamo, J.: A queueing analysis of max-min fairness, proportional fairness and balanced fairness. *Queueing Syst.* **53**(1–2), 65–84 (2006)

9. Bonald, T., Proutière, A.: On performance bounds for balanced fairness. *Perform. Eval.* **55**(1–2), 25–50 (2004)
10. Chih-Lin, I., Huang, J., Duan, R., Cui, C., Jiang, J.X., Li, L.: Recent progress on C-RAN centralization and cloudification. *IEEE Access* **2**, 1030–1039 (2014)
11. Casetti, C., Costa, L.C., Felix, K., Perez, G.M., Robert, M., Pedro, M., Pérez-Romero, J., Weigold, H., Al-Dulaimi, A., Christos, B.J., Gerry, F., Giovanni, G., Leguay, J., Mascolo, S., Papazois, A., Rodriguez, J.: Cognitive network management for 5G by 5GPPP working group on network management and QoS the path towards the development and deployment of cognitive networking list of contributors. Technical report, 5GPPP Network Management & Quality of Service Working Group (2017). <https://bscw.5g-ppp.eu/pub/bscw.cgi/d154625/NetworkManagement.WhitePaper.1.pdf>
12. Coello, C.: Handling preferences in evolutionary multiobjective optimization: a survey. In: Proceedings of the 2000 Congress on Evolutionary Computation, CEC00 (Cat. No.00TH8512), vol. 1, pp. 30–37. IEEE (2000)
13. Ehrgott, M., Gandibleux, X.: Multiple Criteria Optimization: State of the Art Annotated Bibliographic Surveys. Kluwer Academic Publishers, New York (2002)
14. Farina, M., Deb, K., Amato, P.: Dynamic multiobjective optimization problems: test cases, approximations, and applications. *IEEE Trans. Evol. Comput.* **8**(5), 425–442 (2004)
15. Fonseca, C., Fleming, P.: Genetic algorithms for multiobjective optimization: formulation discussion and generalization. In: International Conference on Genetic Algorithms, vol. 93, pp. 416–423, San Mateo, California (1993)
16. Gandhi, A., Doroudi, S., Harchol-Balter, M., Scheller-Wolf, A.: Exact analysis of the M/M/k/setup class of Markov chains via recursive renewal reward. *Queueing Syst.* **77**(2), 177–209 (2014)
17. Gandhi, A., Gupta, V., Harchol-Balter, M., Kozuch, M.A.: Optimality analysis of energy-performance trade-off for server farm management. *Perform. Eval.* **67**(11), 1155–1171 (2010)
18. Gebrehiwot, M.E., Aalto, S., Lassila, P.: Energy-aware server with SRPT scheduling: analysis and optimization. In: Agha, G., Van Houdt, B. (eds.) QEST 2016. LNCS, vol. 9826, pp. 107–122. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-43425-4_7
19. Gebrehiwot, M.E., Aalto, S., Lassila, P.: Energy-performance trade-off for processor sharing queues with setup delay. *Oper. Res. Lett.* **44**(1), 101–106 (2016)
20. Gebrehiwot, M.E., Aalto, S., Lassila, P.: Optimal energy-aware control policies for FIFO servers. *Perform. Eval.* **103**, 41–59 (2016)
21. Hamacher, H.W., Pedersen, C.R., Ruzika, S.: Multiple objective minimum cost flow problems: a review. *Eur. J. Oper. Res.* **176**(3), 1404–1422 (2007)
22. Ho, T.T.N., Gribaudo, M., Pernici, B.: Characterizing energy per job in cloud applications. *Electronics* **5**(4), 90 (2016)
23. Ho, T.T.N., Gribaudo, M., Pernici, B.: Improving energy efficiency for transactional workloads in cloud environments. In: Proceedings of Energy-Efficiency for Data Centers (E2DC), Hong Kong, May 2017
24. Hyttia, E., Richter, R., Aalto, S.: Energy-aware job assignment in server farms with setup delays under LCFS and PS. In: 2014 26th International Teletraffic Congress (ITC), pp. 1–9. IEEE, September 2014
25. Hyttia, E., Richter, R., Aalto, S.: Task assignment in a heterogeneous server farm with switching delays and general energy-aware cost structure. *Perform. Eval.* **75–76**, 17–35 (2014)

26. Kelly, F.P., Maulloo, A.K., Tan, D.K.H.: Rate control for communication networks: shadow prices, proportional fairness and stability. *J. Oper. Res. Soc.* **49**(3), 237–252 (1998)
27. Liberal, F., Taboada, I., Fajardo, J.O.: Dealing with energy-QoE trade-offs in mobile video. *J. Comput. Netw. Commun.* **2013**, 1–12 (2013)
28. Gebrehiwot, M.E., Aalto, S., Lassila, P.: Near-optimal policies for energy-aware task assignment in server farms. In: 2nd International Workshop on Theoretical Approaches to Performance Evaluation, Modeling and Simulation (2017)
29. Maccio, V., Down, D.: On optimal policies for energy-aware servers. *Perform. Eval.* **90**, 36–52 (2015)
30. Marler, R., Arora, J.: Survey of multi-objective optimization methods for engineering. *Struct. Multi. Optim.* **26**(6), 369–395 (2004)
31. Marler, T.: A study of multi-objective optimization methods: for engineering applications. VDM Publishing, Saarbrücken (2009)
32. Mastelic, T., Oleksiak, A., Claussen, H., Brandic, I., Pierson, J., Vasilakos, A.V.: Cloud computing: survey on energy efficiency. *ACM Comput. Surv.* **47**(2), 33:1–33:36 (2014). <https://doi.org/10.1145/2656204>
33. Mukherjee, D., Dhara, S., Borst, S., van Leeuwen, J.S.H.: Optimal service elasticity in large-scale distributed systems. In: Proceedings of ACM SIGMETRICS (2017)
34. Nedeveschi, S., Popa, L., Iannaccone, G., Ratnasamy, S.: Reducing network energy consumption via sleeping and rate-adaptation. *NsDI* **8**, 323–336 (2008)
35. NGMN Alliance: suggestions on potential solutions to C-RAN. Technical report (2013). http://www.ngmn.org/uploads/media/NGMN_CRAN_Suggestions_on_Potential_Solutions_to_CRAN.pdf
36. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, New York (2006). <https://doi.org/10.1007/978-0-387-40065-5>
37. Öztürk, M., Tsoukiàs, A., Vincke, P.: Preference modelling. In: Figueira, J., Greco, S., Ehrgott, M. (eds.) Multiple Criteria Decision Analysis: State of the Art Surveys. International Series in Operations Research & Management Science, vol. 78, pp. 27–59. Springer, New York (2005). https://doi.org/10.1007/0-387-23081-5_2
38. Pioro, M., Dzida, M., Kubilinskas, E., Ogryczak, W.: Applications of the max-min fairness principle in telecommunication network design. In: Next Generation Internet Networks, pp. 219–225. IEEE (2005)
39. Saaty, T.L.: Decision making with the analytic hierarchy process. *Int. J. Serv. Sci.* **1**(1), 83–98 (2008)
40. Shaw, K.: Including real-life problem preferences in genetic algorithms to improve optimisation of production schedules. In: Second International Conference on Genetic Algorithms in Engineering Systems, vol. 1997, pp. 239–244. IEE (1997)
41. Taboada, I., Jacko, P., Ayestaa, U., Liberal, F.: Opportunistic scheduling of flows with general size distribution in wireless time-varying channels. In: 2014 26th International Teletraffic Congress (ITC), pp. 1–9. IEEE, September 2014
42. Taboada, I., Aalto, S., Lassila, P., Liberal, F.: Delay- and energy-aware load balancing in ultra-dense heterogeneous 5G networks. *Trans. Emerg. Telecommun. Technol.* **28**, e3170 (2017)
43. Taboada, I., Liberal, F.: A novel scheduling index rule proposal for QoE maximization in wireless networks. *Abs. Appl. Anal.* **2014**, 1–14 (2014)
44. Taboada, I., Liberal, F., Fajardo, J.O., Ayesta, U.: QoE-aware optimization of multimedia flow scheduling. *Comput. Commun.* **36**(15–16), 1629–1638 (2013)

45. Taboada, I., Liberal, F., Jacko, P.: An opportunistic and non-anticipating size-aware scheduling proposal for mean holding cost minimization in time-varying channels. *Perform. Eval.* **79**, 90–103 (2014)
46. Tsoukias, A., Vincke, P.: A survey on non conventional preference modelling. *Ric. Operativa* **61**(5–48), 20 (1992)
47. Wajid, U., Cappiello, C., Plebani, P., Pernici, B., Mehandjiev, N., Vitali, M., Gienger, M., Kavoussanakis, K., Margery, D., García-Pérez, D., Sampaio, P.: On achieving energy efficiency and reducing CO₂ footprint in cloud computing. *IEEE Trans. Cloud Comput.* **4**(2), 138–151 (2016). <https://doi.org/10.1109/TCC.2015.2453988>
48. Winston, W.: Optimality of the shortest line discipline. *J. Appl. Probab.* **14**(1), 181–189 (1977)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

