Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 2 (2010) 99-113

Energy-Efficient Networking: An Overview

László SZILÁGYI, Tibor CINKLER, Zoltán CSERNÁTONY

Budapest University of Technology and Economics,
Dept. of Telecommunications and Media Informatics,
e-mail: szilagyi@tmit.bme.hu; cinkler@tmit.bme.hu; csernatony@tmit.bme.hu

Manuscript received June 30, 2010; revised October 30, 2010.

Abstract: Recently – for economical and ecological reasons –, energy-efficiency has been receiving an emerging attention from both industrial and fundamental researchers. Given the large-scale growth and resource overprovision of communication networks, the issues related to energy consumption get more significant than ever. In this paper, we provide an overview of some of the latest contributions and most important trends on energy-efficient networking. Since today's networks are heterogeneous to a large extent, different concepts and aspects of networking are discussed in separate sections.

Following the introduction of the fundamental discipline of energy-proportional computing (and also its relationship with resource over-provisioning), we give an overview on the power-saving opportunities of core networks with focusing on two of the most-widely applied techniques for energy management. A comparison of circuit and packet switched networking approaches is also made, along with some considerations taken for the cases in which electrical and optical switching are employed. In case of access networks, in addition to considering the commonly used landline connections, we give a brief description about the popular handoff mechanisms for wireless networks. Finally, we consider energy-saving opportunities on data centers and take a short overlook on cloud computing. Despite all the evident differences, being between distinct segments of networking, some of the methods applied at distinct networking technologies show considerable similarities with each other due to the fact that some of the occurring problems exhibit similar patterns in terms of modeling and problem formulation.

Keywords: energy-efficiency, energy consumption, green networking.

1. Introduction

With the explosive growth of communication networks, energy consumption has risen to a major economical (operational expenditures) and ecological (CO_2 emission) concern in the past few years. About 2 percent of the total CO_2 emission is produced by the Information Technology and Communication sector (ICT) which is more than the contribution of the whole aviation industry [1]. A recent study puts more emphasis on this issue by showing that the rise of energy consumption of large communication systems corresponds to Moore's law [2]. Therefore, power consumption has become a critical factor of communication networks, IT facilities, data centers and high performance network elements. Energy-efficient designing helps cutting the Operating Expenses (OPEX) as well [3], [4], and in addition to that, it also might result in more reliable network elements (due to the decrease of heat dissipation).

In order to save energy in communication networks, first, we have to reveal the reasons of energy wastage in existing systems. Energy inefficiency might come from architectural (SW related) and physical design (HW related). From the energy-efficiency point of view, the most important feature of networking is underutilization. While networks are generally designed to handle peak-time traffic, most of the time, their capacity remains (heavily) unexploited. This is called *over-provisioning*. According to [1], the magnitude of network utilization is 33% for switched voice, 15% for internet backbones, 3~5% for private line networks and 1% for LANs, while the energy consumption of network equipments remains substantial even when the network is idle. A rather physical related issue is that the energy consumption of network elements is not proportional to their utilization, i.e. energy cost is a function of capacity, not throughput. These facts result in high energy wastage.

Today's networks are mostly *configured statically*, running at full performance all the time which is not necessary. In order to achieve higher energy-efficiency, network management methods need to be able to dynamically adopt network characteristics to the actual traffic demands during operation. Switching off underutilized (or idle) parts of the network and dynamically adapting transmission rates (with satisfying certain QoS constraints) are ultimately important approaches of designing a greener network. In order to make energy-aware management possible, network elements also should support these features. On-demand frequency-scaling of CPUs and data storage modules and network interfaces with adjustable transmission rates (rate-adaptation support) are all mandatory for attaining greener network elements.

Finding efficient ways for *cooling* network equipments is also a big challenge. In case of data centers, roughly 50% of electricity is consumed by the cooling infrastructure, the other 50% used for computing [5], [6]. Increasing cooling efficiency and using alternative cooling methods have a huge contribution to the electric bill of equipment rooms. Employing alternative energy sources (e.g. solar, wind) for supplying network nodes (base stations) is also a matter of interest nowadays.

In this paper, we provide an overview of the latest results concerning energy-efficient networking, discussing the different functional parts of the ICT infrastructure separately. In *Fig. 1*, the estimated share in energy consumption by different areas of ITC can be seen [7]. Energy-efficiency is examined from an operator's point of view with focusing on networking infrastructure and data centers, but leaving PCs, monitors, printers, and other user equipments out of consideration.

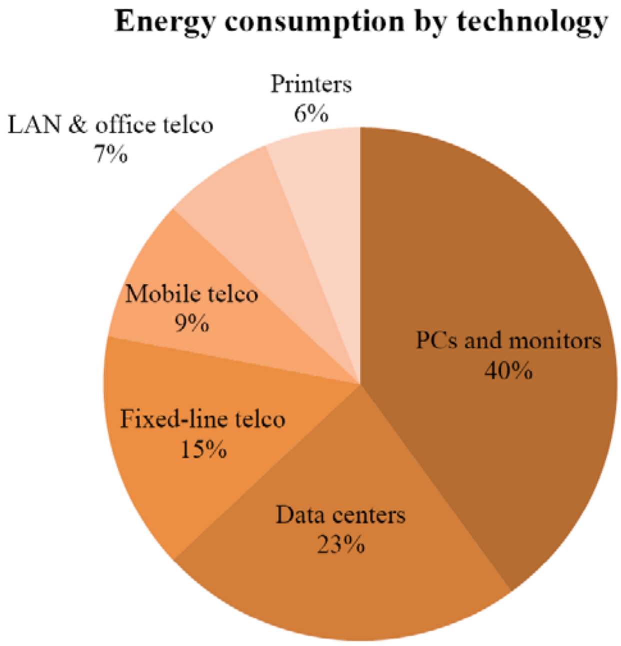

Figure 1: The estimated energy consumption shares of different functional parts of ITC [7].

The paper is organized as follows. Section 2 explains the importance of power consumption dynamic range. In the remaining sections, we provide a closer look at the issues with different areas of communication networks: Section 3 discusses some of the most important features of core networks, while Section 4 and 5 do the same for access networks and data centers, respectively. Finally, in Section 6, we conclude our experiences and reveal a few important directions for future solutions.

2. The dynamic range of power consumption: Rate adaptation versus switching off components

Rate-adaptation is highly important in energy-aware data transmission and data processing for the following reason. For transmission power-rate function is a relationship that gives the required amount of transmission power for a certain rate. Keeping the bit-error probability fixed, the required power is a convex function of the rate for most encoding schemes. This results from Jensen’s inequality which states that transmitting data at lower rates and over longer time intervals has less energy cost compared to faster rate transmissions [8]. For data processing variable adaptive rate control means adjusting CPU frequency for the required performance in order to save power.

The *dynamic range* of power consumption is one of the most important attributes of network elements: it is the relative difference of energy consumption of an element between 0% and 100% of relative utilization. For example, study [6] shows that the utilization–power usage characteristics of a system (consisting of servers) shown in *Fig. 2* has a smaller (50%) dynamic range compared to the one on *Fig. 3* (90%). As suggested by the results, given the typical utilization region (10-50%), a 50% save can be achieved in terms of energy consumption when having a dynamic range of 90%. Consequently, network component vendors should increase the dynamic range of their equipments as much as possible. The dynamic range of 100% addresses the case of *energy-proportional computing*.

However – even when applying rate-adaptation –, network components with low dynamic range still tend to consume a considerable amount of energy when being not utilized at all (if the offset of power usage is too high). Even so, in many cases (apart from rate-adaptation), power consumption can be further reduced by *switching off* certain underutilized (or idle) components. The price to pay here is the potential occurrence of additional delay, however – due to the typically high power on consumption – transition between active and passive operation mode cannot be done arbitrarily often.

A real challenge for these solutions is finding a method to decide whether a network element should operate or not, and if so, then determining the operational rate in specific moments.

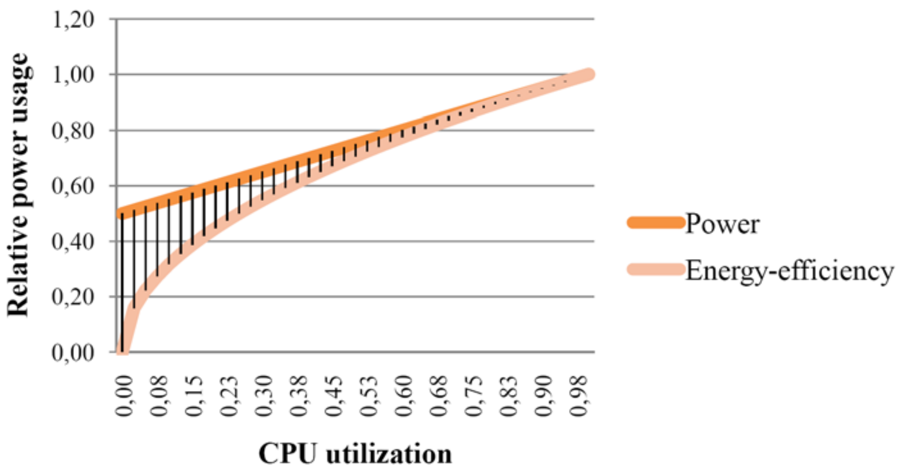

Figure 2: Relative power usage and energy-efficiency for a typical data center scenario.

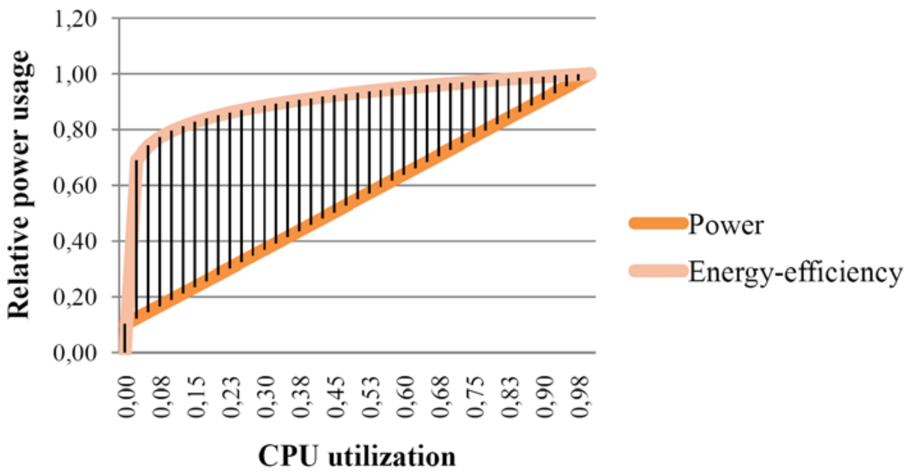

Figure 3: Relative power usage and energy-efficiency for a data center scenario with larger dynamic range [6].

3. Core networks, packet versus circuit-switching

The backbone infrastructure of a communication network transferring aggregated data flows is called a *core network*. Given that the typical rate of utilization shows a considerable variation within a single day, and that (as stated previously) networks are provisioned to handle peak traffic, networks tend to be *underutilized* for most of the time.

An opportunity for energy conservation method is to shut down idle network elements. Since the transition between active and sleeping states requires additional energy and time, inactive periods along with the routing has to be scheduled properly in order to preserve power without exceeding a given value of delay. In [9], a frame-based periodic scheduling is introduced. The network elements work either at a full or at a zero rate. The lengths of active periods are proportional to the traffic loads of network elements while periods cannot be arbitrarily small since transitions consume extra power, however too long periods result in increased delays. In [2], the network is controlled on the basis of traffic statistics: the day is sliced into one-hour long “network configuration periods”, and the state of the system is formalized as a “linear programming problem” with the daily operation scheme being determined to satisfy the specific QoS constraints.

An important feature of core networks is that multiple paths might exist between distinct nodes of them. This fact allows us to distinguish between nodes to determine the ones that can be switched off during low backbone utilization periods. Many high capacity links within a core network actually are compositions of smaller capacity ones. In [10], a method is demonstrated for shutting off cables in bundled links for lower traffic periods. The problem of determining the sublinks to shut down is formulated as an Integer Linear Program (ILP) that is proven to be NP-complete. Given this fact, for tackling this intractability, efficient heuristics have to be proposed. In [10], a fast greedy heuristic solution suggested for a dedicated multicore server is proven to be feasible. Optimization cycles becoming 8 seconds to 50 minutes long (depending on the actual topology) mean a dynamic power-saving solution.

Underutilization means that a link or a network element does not need to operate at its full capacity to satisfy the corresponding QoS constraints. Upon this fact, rate adaptive routers, switches and links can be employed in order to save energy [2], [11], [12]. A suitable hybrid application of element switching-off and rate-adaptation is a promising solution [11].When multiple transmission lines form a single large-capacity link, *rate-adaptation* can be the combination of operations: switching off transmission lines, and tuning the transmission rate of individual lines (width control) [13].

In [14], core networks are examined for their power-efficiency from a different perspective. Today's backbone network infrastructures are composed by optical links between large distance nodes with the processing and switching of traffic being performed mostly in the electrical domain. While the energy consumption of electrical components is getting lower by every year, large reduction could be achieved in terms of power consumption if the whole processing and switching is executed in the optical domain. Although, optical switching has become a reality (MEMS, CMOS), IP, nowadays' dominant packet switched transmission technology, requires random access memory for buffering which is yet to be implemented purely optically. Large fiber delay lines are not practical as they require power-consuming signal regenerators, not to mention the impractical size of them.

From an energy-efficiency point of view, it is important to note that – despite the connectionless nature of IP – above 90% of the traffic within backbone networks is transported via the connection oriented Transmission Control Protocol (TCP) [14] (consider applications such as IP television, voice over IP, video conferencing or online gaming services, all those by which a very high quality of service is required).

As being shown in *Fig. 4*, within a node of an IP network, more than half of the energy is consumed by the traffic processing and forwarding engine (TP/FE).

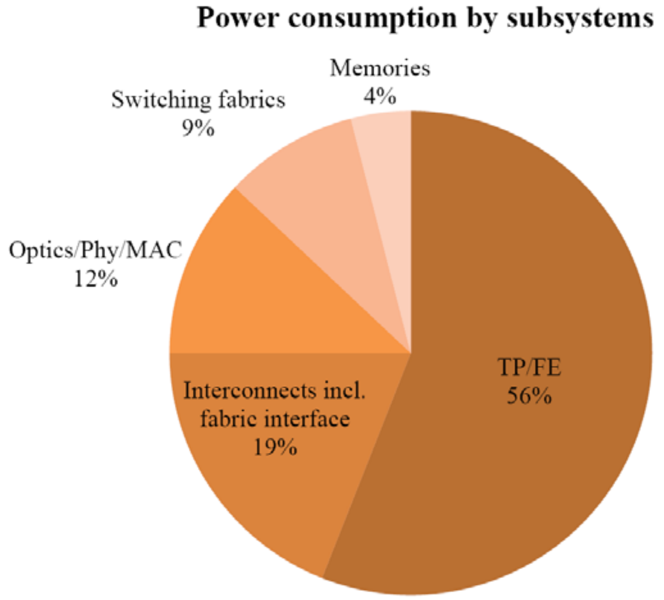

Figure 4: An estimation of power consumption by subsystems (the efficiency of power conversion is not considered) [14].

On the contrary, the implementation of all-optical, circuit-switched nodes does not require complex traffic processing and large memory units (within the electrical domain).

In contrast, all-optical circuit-switched nodes whose implementation doesn't require complex traffic processing, large memories in the electrical domain and optical to electrical to optical conversions consume less than 10% of the power consumed by ordinary optical packet-switched architectures using electronic traffic processing. Similarly, electronic circuit-switched designs consume 43% less power than packet-switched one. Taking all these into account, future energy-efficient core networks nodes might use circuit switching in the core, along with packet switching edge nodes in order to reduce complexity, thus power consumption.

4. Access networks

Since most of the physical elements are located in the segment of access networks, the energy saving by each type of elements is multiplied by a large factor. This makes an important contribution to the reduction of total consumption [15].

Nowadays, the most widely deployed technology for broadband *landline connections* still employs copper lines for bearers. With the continuous increase of bandwidth requirements, broadband penetration and bandwidth demands, more energy is required than ever. Although new transmission technologies, such as VDSL2, allow higher speeds, they induce increased complexity and power consumption [16]. By today’s networking technologies, serving high-bandwidth demands together with sustainable energy consumption can only be achieved through progressive optical fiber deployment in FTTCab, FTTB (and also in the longer-term FTTH) architectures which are expected to shorten the copper access network and to boost the overall performance of xDSL systems. The deployment of such systems however requires certain capital expenses to be involved which makes this technological shift a gradual one. Dynamic Spectrum Management [17] and energy-aware solutions of such kind can make copper technologies more sustainable, but they only yield the industry little additional time for making the required technological shift towards optical networks.

Mobile operators with radio access networks (2G, 3G etc.) have to provide services for very large physical areas and for several subscribers as well [2]. Given the necessity for several base stations, operating them requires a large amount of energy. As mobile broadband is expanding rapidly, the energy-efficiency of radio access networks is expected to receive even more significant attendance in the future.

As presented in *Fig. 5*, base stations take most of the energy requirements of cellular networks [18]. For that reason, in the remaining parts of this section, we focus on the issues of radio base stations.

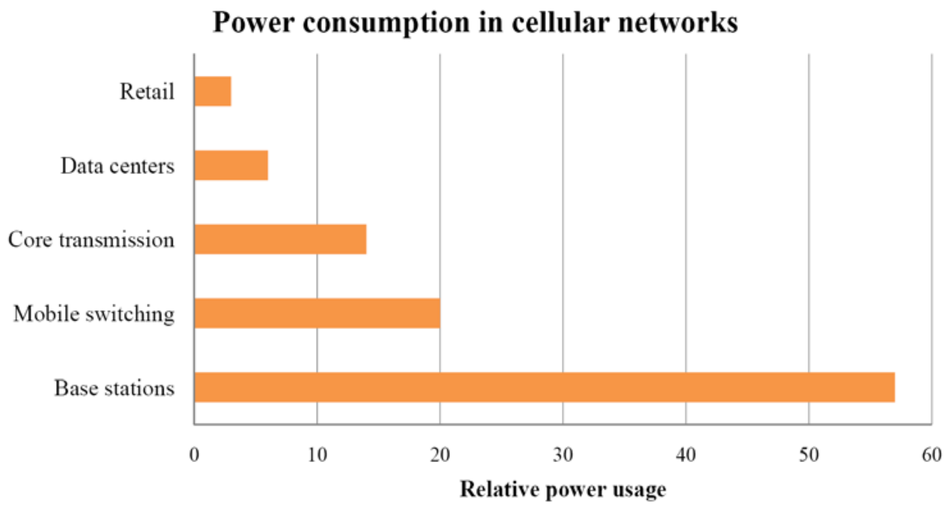

Figure 5: Power consumption in cellular networks [18].

Since radio access networks are also dimensioned for peak traffic loads, a large amount of energy is being wasted due to underutilization. The reduction of traffic level in some portions of a cellular network comes from the combination of two effects: first, the characteristics of the typical day-night behavior of users, and second, the daily swarming of users carrying their mobile terminals around different kinds of areas (residential, office districts, etc.). On the one hand, this induces the need for a large capacity in all areas at peak usage times, on the other hand, it reduces resource requirements during lightly occupied time intervals (day for residential and night for office districts) [15].

Energy can be saved by switching off certain underutilized cells. When some cells are switched off, we assume that radio coverage and service provisioning can be taken care of by the cells which remain active, possibly with a smaller increase in the emitted power, in order to guarantee the availability of service over the whole area. This energy-saving approach assumes that the original network dimensioning is essentially driven by traffic demands, as it normally happens in metropolitan areas, comprising a large number of small cells [15]. In regular cellular topologies, around 25-30% of the energy is possible to save by dynamically powering off active cells during under-utilized (low-traffic) periods [15], [19]. *Fig. 6* illustrates the method itself.

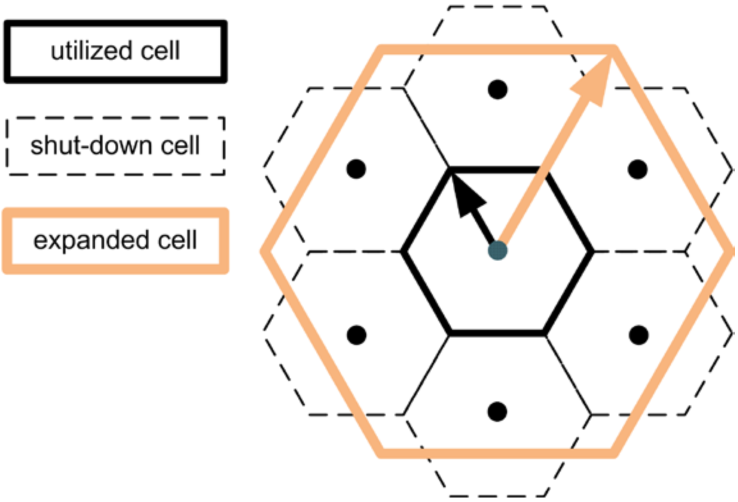

Figure 6: Dynamically powering off cells in cellular networks.

Rate-adaptation is an applicable procedure for wireless links too. Modern wireless devices are equipped with rate-adaptive capabilities which allow the transmitter to adjust its transmission rate over time. This can be achieved in various ways that include adjusting the power level, symbol rate, coding scheme, constellation size and various combinations of these approaches. In order to satisfy QoS requirements, existing systems can change transmission power/coding scheme dynamically depending on the noise floor. If large bandwidth is not required, "low power" transmission over longer time periods might also fit for the actual application [8].

In many areas, multiple access technologies are available. Matching bandwidth requirements with available access technologies, while taking energy-efficiency into consideration can save energy. There are two types of handoffs considered: vertical and horizontal handoffs. *Vertical handoff* is an interworking technique between different networking technologies whose aim originally is to assure the best connectivity to applications for mobile terminals by providing a transparent and seamless roaming between systems of different technologies. *Horizontal handoff* means the switching within the same technology (or layer). Initially, handoff decisions were aimed to be made on the basis of a lot of different factors such as QoS, cost-of-service, etc. [20]. Lately, there has been revealed that by suitably involving energy consumption into the objective (function) of the handoff, energy-efficiency aspects can also be taken into consideration [21], [22], [23], [24]. In *Fig. 7*, an illustration of horizontal and vertical handoff of two base station cells and a WLAN cell can be seen.

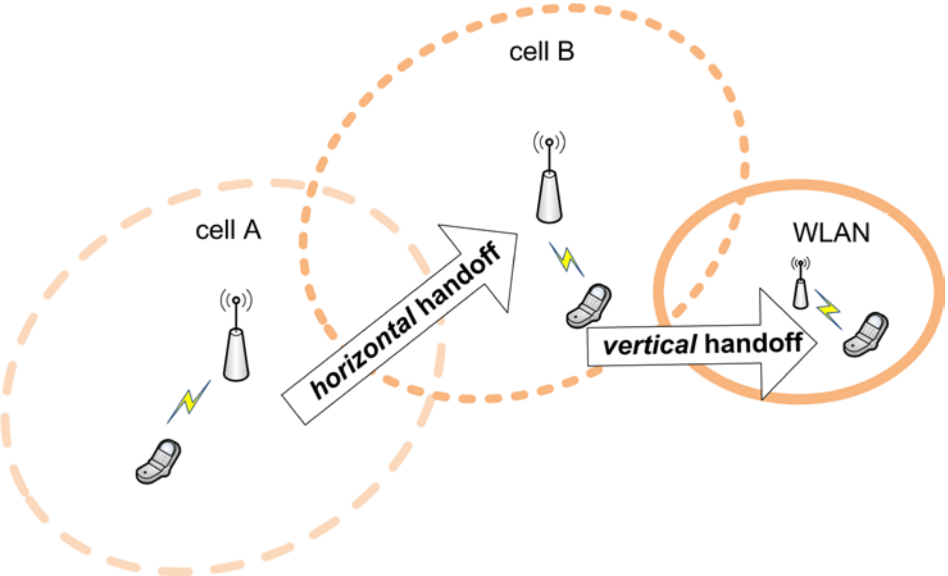

Figure 7: An illustration of vertical and "horizontal" handoff mechanisms between cellular networks and WLAN.

Base stations, exploiting alternative energy resources, already exist [25]. Employing alternative energy resources such as sun and wind for supplying base stations make mobile communication "greener", and in the same time these more sustainable sources ease the deployment of BSs far from the electric grid.

5. Data centers and cloud computing

The majority of networked services are being hosted in data centers. With the rapid growing of the demand for such data-intensive services (e.g. cloud computing, video broadcasting and online social networking), the employment of large *data centers* is of emerging significance. By now, they have evolved to networks consisting of tens of thousands of high data-rate servers. Such systems require special design and management approaches to provide the demanded robustness, performance and reliability.

As indicated in *Fig. 8*, according to [6], servers in a typical large system are being under-utilized in most of the time. As presented in *Fig. 5*, the power consumption is still significant in case of data centers. The problem is with the power consumption dynamic range, as stated above. CPUs are easily tunable, but they are not the major consumers. HDDs and even memory modules are much harder to manage from the aspect of power consumption. Switching off underutilized servers is not feasible in a large scale for servers, as applications and data are usually distributed over numerous machines. Switching off components has too large a penalty, for HDD-s latency becomes in order of 1000 times bigger, and the spin up consumption consumes energy which can be saved in minutes of being off. In order to increase the efficiency of servers, their utilization should be maximized, while machines should be created with as wide dynamic power range as possible (energy proportional computing). Provided the dynamic range indicated in *Fig. 2* could be achieved, 50% of the energy could be saved compared to servers with characteristics shown in *Fig. 1*.

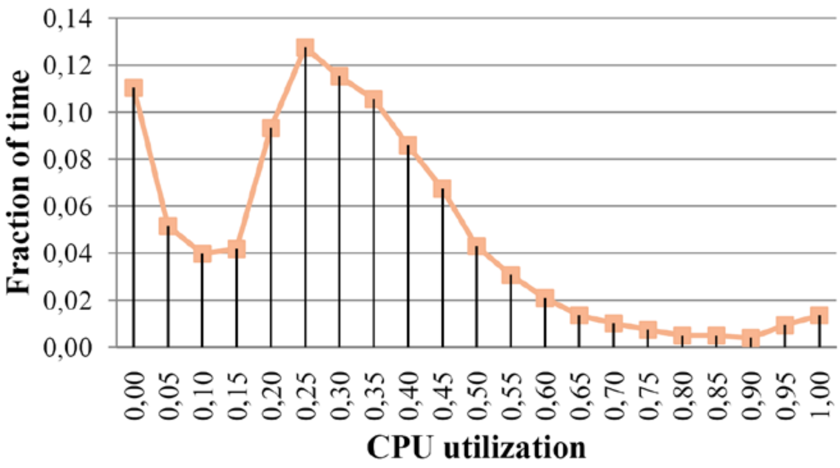

Figure 8: CPU utilization in a typical large-scale system [20].

The architecture of a data center influences the power consumption. Different architectures for a given number of servers mean different links and switching fabrics, different ways of connecting them to each other. The applied links and switching fabrics determine both the throughput and energy consumption. However, more sophisticated structures might provide bigger throughput, smaller delays, higher access speeds, but also require more power. Architects have to make a compromise between throughput and power consumption [26]. Today's popular symmetric architectures (BCube, DCell, Fat-tree, Balanced tree) are quite rigid and make it hard to apply them for the given constraints (throughput, power). Many systems are unnecessarily over-provisioned due to this factor wasting much power. Future architectures should employ more asymmetricity to provide more fine-grained datacenter designs.

It is important to note that the concept of cloud computing itself has got a large potential to make the whole IT industry greener [27]. Concepts like Software as a Service (SaaS) or Infrastructure as a Service (IaaS) helps companies and governments not only to reduce cost with optimized resource usage, but to keep their business greener. Cloud computing service providers have large scale data centers and server farms and provide computing services over the Internet making companies having neither to invest in their own server parks nor to worry about over-provisioning their systems. The capital expenditure of a startup company can be restricted to investing in thin clients to access the demanded services. Although such infrastructures might be more energy-efficient compared to the legacy approach (i.e. ordinary companies providing a single PC for every employer while operating a more or less over-provisioned server park), large scale data centers still tend to be more energy-efficient.

6. Summary

In this paper, we have revealed a number of possibilities for saving power in different parts of the network. We have started the discussion with presenting the general concept that the over-provisioning of the resources results in suboptimal energy-efficiency by making the energy consumption of the network disproportional to its utilization. In the following, we have given a review on the various power-saving opportunities of underutilized network elements for core networks via sleeping and rate adaptation. Moreover, it has been shown that by deploying circuit-switched all-optical networks, it can be capitalized on that the vast majority of Internet traffic is transported via the connection-oriented TCP. Afterwards, access networking has been discussed: among landline connection technologies, optical-based FTTx solutions seem to be promising, while for wireless networks, horizontal and vertical handoff mechanisms have been presented for adjusting range of radio communication. Finally, energy-efficient ways for operating data centers and the concept of cloud computing have been briefly overviewed.

Based on the state-of-the-art, we suspect that the biggest challenge will be the effective control and cooperation of network components of varying energy-awareness. Today’s large-scale communication networks are of great heterogeneity in terms of the employed networking technologies. Consequently, network management software has to handle different networking equipment types and generations at the same time. For both complexity and heterogeneity reasons, there is a fundamental need for a shift to be made from centralized solutions towards a distributed approach. Designing future’s energy-efficient network components and architecture needs strong inter-technological cooperation to match HW capabilities with management techniques.

References

[1] Nordman, B., “Energy Use and Savings in Communications”, in *Proc. of IEEE ICC*, 2009, Keynote.

[2] Bolla, R. et al., “Energy-Aware Performance Optimization for Next-Generation Green Network Equipment”, in *Proc. of ACM SIGCOMM, Workshop on Programmable routers for extensible services of tomorrow*, 2009, pp. 49-54.

[3] Qureshi, A., Weber, R., Balakrishnan, H., Guttag, J. and Maggs, B., “Cutting the Electric Bill for Internet-Scale Systems”, in *Proc. of ACM SIGCOMM, Conference on Data communication*, 2009, pp. 123-134.

[4] Odlyzko, A. M., “Data networks are lightly utilized, and will stay that way”, *Review of Network Economics*, vol. 2, no. 3, pp. 210-237, 2003.

[5] Fan, X., Weber, W. D., Barroso, L.A., “Power Provisioning for a Warehouse-Sized Computer”, in *Proc. of ACM International Symposium on Computer Architecture*, 2007, pp. 13-23.

[6] Barroso, L. A. and Hölzle, U., “The Case for Energy-Proportional Computing”, in *Proc. of IEEE Computer*, 2007, pp. 33-37.

[7] Kumar, R. and Mieritz, L., “Conceptualizing ‘Green IT’ and data centre power and cooling issues”, *Gartner Research Paper*, No. G00150322, 2007.

[8] Zafer, M. A., Modiano, E., “A calculus approach to energy-efficient data transmission with quality-of-service constraints”, in *Proc. of IEEE/ACM, Transactions on Networking*, vol. 17, issue 3, pp. 898-911, 2009.

[9] Andrews, M., Anta, A. F., Zhang, L., and Zhao, W., “Routing and scheduling for energy and delay minimization in the powerdown model”, in *Proc. of IEEE INFOCOM*, 2010, pp. 21-25.

[10] Fisher, W., Suchara, M., and Rexford, J., “Greening backbone networks: reducing energy consumption by shutting off cables in bundled links”, in *Proc. of ACM SIGCOMM, Green Networking*, 2010, pp. 29-34.

[11] Nedevschi, S., Popa, L., Iannaccone, G., Ratnasamy, S., and Wetherall, D., “Reducing Network Energy Consumption via Sleeping and Rate-Adaptation”, in *Proc. of 5th USENIX Symposium on Networked Systems Design and Implementation*, 2008, pp. 323-336.

[12] Bolla, R., Bruschi, R., Davoli, F., Ranieri, A., “Performance Constrained Power Consumption Optimization in Distributed Network Equipment”, in *Proc. of IEEE ICC, Workshop on Green Communications*, 2009, pp. 1-6.

[13] Kant, K., “Power Control of High Speed Network Interconnects in Data Centers”, in *Proc. of IEEE ICC*, 2009, pp. 145-150.

[14] Aleksic, S., “Analysis of Power Consumption in Future High-Capacity Network Nodes”, in *Proc. of IEEE/OSA JOCN*, vol. 1, issue 3, 2009, pp. 245-258.

[15] Marsan, M. A., Chiaraviglio, L., Ciullo, D. and Meo, M., “Optimal Energy Savings in Cellular Access Networks”, in *Proc. of IEEE ICC, Workshop on Green Communications*, 2009, pp. 1-5.

[16] Bianco, C., Cucchietti, F., Griffa, G., “Energy consumption trends in the Next Generation Access Network - a Telco perspective”, in *Proc. of INTELEC*, 2007, pp. 737-742.

[17] Cioffi, J. M., Zou, H., Chowdhery, A., Lee, W., and Jagannathan, S., “Greener Copper with Dynamic Spectrum Management”, in *Proc. of IEEE GLOBECOMM*, 2008, pp. 1-5.

[18] Zuckerman, D., “Green Communications – Management Included”, in *Proc. of IEEE ICC*, 2009, Keynote.

[19] Marsan, M. A., Meo, M., “Energy Efficient Management of two Cellular Access Networks”, in *Proc. of ACM SIGMETRICS, Performance Evaluation Review archive*, vol. 37, issue 4, 2010, pp. 69-73.

[20] Hasswa, A., Nasser, N., Hassanein, H., “Generic Vertical Handoff Decision Function for Heterogeneous Wireless Networks”, in *Proc. of IFIP Conference on Wireless and Optical Communications*, 2005, pp. 239–243.
[21] Choi, Y. and Choi, S., “Service Charge and Energy-Aware Vertical Handoff in Integrated IEEE 802.16e/802.11 Networks”, in *Proc. of IEEE INFOCOM*, 2007, pp. 589-597.
[22] Yang, W.-H., Wang, Y.-C., Tseng, Y.-C., and Lin, B.-S. P., “An Energy-Efficient Handover Scheme with Geographic Mobility Awareness in WiMAX-WiFi Integrated Networks”, in *Proc. of IEEE WCNC*, 2009, pp. 2720--2725.
[23] Seo, S. and Song, J., “Energy-Efficient Vertical Handover Mechanism”, in *Proc. of IEICE Transactions on Communications*, vol. E92-B, no. 9, 2009, pp. 2964-2966.
[24] Petander, H., “Energy-aware network selection using traffic estimation”, in *Proc. of ACM MICNET*, 2009, pp. 55-60.
[25] Barth, U., Wong, P., Bourse, D., “Key Challenges for Green Networking”, in *Proc. of Ercim News* 79, 2009, pp. 13.
[26] Gyarmati, L., Trinh, T. A., “How Can Architecture Help to Reduce Energy Consumption in Data Center Networking?”, in *Proc. of ACM SIGCOMM*, 2010, pp. 183-186.
[27] Chang, V., Bacigalupo, D., Wills, G., and De Roure, D., “A Categorization of Cloud Computing Business Models” in *Proc. of IEEE/ACM CCGRID*, 2010, pp. 509-512.