# IEC – IEEE CHALLENGE

IEC – IEEE Challenge ID Number: **2012-03-1128**

Title: **100% Green Computing At The Wrong Location?**

Lead-author: **Frank Kienle**

Email: **kienle@eit.uni-kl.de**

Co-authors: **Christian de Schryver**

Institution: **University of Kaiserslautern**

Email: **-**

Date: **28.06.2012**

# 100% Green Computing At The Wrong Location?

Frank Kienle, *Member, IEEE,* and Christian de Schryver, *Member, IEEE*,

**Abstract**

Modern society relies on convenience services and mobile communication. Cloud computing is the current trend to make data and applications available at any time on every device. Data centers concentrate computation and storage at central locations, while they claim themselves green due to their optimized maintenance and increased energy efficiency. The key enabler for this evolution is the microelectronics industry. The trend to power efficient mobile devices has forced this industry to change its design dogma to: "keep data locally and reduce data communication whenever possible". Therefore we ask: is cloud computing repeating the aberrations of its enabling industry?

## I. Introduction

The major goal of every technological evolutionary step is to simplify things and to increase the user convenience, while generating profit for the enabler companies. The recent trend to enable data access at any time, at any place, and from any device is currently changing our daily life in the personal and in the professional domain. We as users can edit documents, communicate and stay informed whenever and wherever we are, without concerning ourselves with the underlying technology: we leave everything to *the cloud*. The permanent availability of web services has already led to higher dynamics and broader markets in business, and to reduced language barriers and to smarter shopping and leisure activities, only to mention a few examples foreseen by Joe Mullich in the Wall Street Journal [1].

But what provides us with high comfort and productivity is not for free: efficient communications networks and data centers are the synapses and nerve cells of the cloud intelligence. And this intelligence needs to be powered.

*Information and communication technology (ICT)* as an umbrella term for any kind of service based on processing digital information has enabled cloud services by providing the necessary computing power and communication bandwidth for the service. Cloud computing is one general expression that we would like in this paper to be seen as three separate technical tasks:

1) The *transport* of the information between the users and the physical processing or storage location,
2) the *processing or computation* of data in data centers,
3) and the *storage* of data.

Obviously, for all three areas efficient infrastructures have to be provided to offer an overall *green* service.

Although the pure ICT industry currently only constitutes 2% of the overall carbon footprint of all industrial sectors [2], it acts as an catalyst for energy reduction in other domains, e.g. with the concepts smart transportation, smart grids, smart buildings, smart engines, environmental information systems, or software for energy optimizations as listed in [3]. Especially cloud computing is a major driver for higher storage and communication requirements.

Cloud computing as a new trend can be analyzed and evaluated looking at three driver categories:

1) The user convenience,
2) economical reasons,
3) and the technical realization.

We absolutely agree that the benefits in the first two categories are obvious, and (in spite of still open legal issues) we strongly believe in cloud computing as a major technology driver. The convenience increase for the user is quite compelling, and it can already be seen that more and more people and companies

F. Kienle and C. de Schryver are with the Microelectronic Systems Design Research Group, University of Kaiserslautern, Kaiserslautern, Germany, e-mail: {kienle,schryver} @eit.uni-kl.de

sacrifice their privacy for the sake of easy-to-use cloud services. For that reason is it hard to imagine that the basic concept of cloud computing will fail.

Furthermore, a lot of fundamental economic arguments for cloud computing exist, for example:

- The utilization rate of data centers can be optimized (for example by using virtual machines), in contrast to underloaded distributed cores (see "The Case for Energy-Proportional Computing" from 2007 [4]).
- Central maintenance and backup saves a lot of work for system administrators.
- Information is permanently present and can be accessed in many different contexts, allowing companies to react much quicker to market changes than in the past.
- By optimizing resource clustering and binding tasks to appropriate hardware architectures, immense speedups and cost savings can be achieved.

Many references for the advantages and drawbacks of cloud computing for users and treasurers can be found in literature.

However, for us the third category of cloud drivers, the technical realization, raises several important questions in its current state. We can observe nowadays that cloud-oriented topologies concentrate on one or a few number of central data storage and processing centers. The justification is very often the alleged higher energy efficiency of compute clusters under the *green IT* label (and the associated energy cost reduction) [3].

The label *green* can be justified by two different approaches:

1) The use of (reputed) clean renewable energy to power a system,
2) and strong attempts to reduce the energy consumption and therefore to save energy.

In this paper we will mainly focus on the second point, since we believe that energy is most *green* if it is not used at all.

One famous report from Greenpeace named 'How dirty is your data' [5] focuses on the primary power aspect of data centers themselves. An important statement in this report is that data centers are often powered by dirty energy, although it is obvious that the need for transparency and the source of primary energy is mandatory to call a data center *green*.

The Greenpeace breakdown of the carbon forecast for different ICT device categories shown in Figure 1 clearly states that data centers themselves only account for around 25% of the future $CO_2$ emissions, whereas mobile devices and communications will be the major polluters. This assumption is extensively confirmed by the ICT EU-25 forecast [6].

Another claim of this report is that *'Energy efficiency (at data centers) alone will, at best, slow the growth of the sectors carbon footprint'*. The reason for this is that the energy footprint of the end-user service not only includes the pure computation and storage done in data centers, but also the communication and the user devices. All these technical aspects are strongly interlocked and should not be simply separated, since they as a whole represent the final service.

We fully agree to the messages given in the Greenpeace report that focuses on the data centers, and would like to go one step further by asking: even if a super data center is *green* itself, the current macroscopic architecture of a cloud service seems not to be *green* at all. It seems more to be driven by 'local cost efficiency' and business models instead of environmental aspects.

The Greenpeace report and also the SMART2020 report [2] forecast an amount of 60% of the overall carbon footprint for telecommunication infrastructure and devices. Thus the communication itself (that is not in the main focus up to now) seems to have the biggest impact on the overall carbon footprint.

The huge impact of the mobile communication and thus generating and transporting information is supported by the mobile data traffic forecast from Cisco as illustrated in Figure 2. The highest traffic is generated by videos streamed over the Internet, that are again hosted in central data storage centers.

In summary we can see that the overall carbon footprint is mainly composed of the telecoms and devices plus the data centers. Instead of concentrating on bringing down the carbon footprint of individual parts of the cloud system we should better try to find an optimal global minimum of pollution, including all parts of the system.

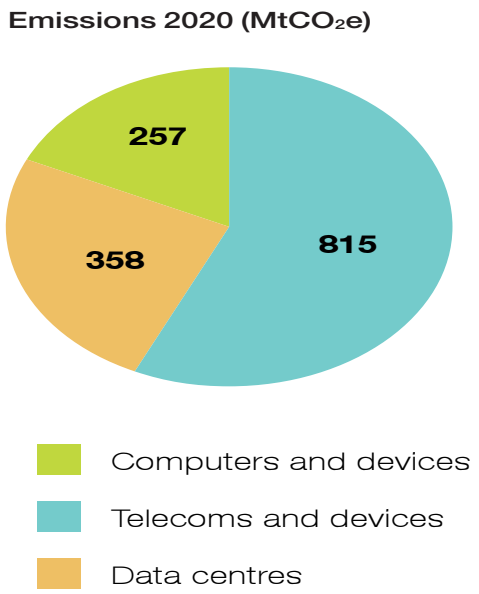


Fig. 1: Carbon forecast 2020 for different devises, source Greenpeace [5] or SMART2020 report [2]

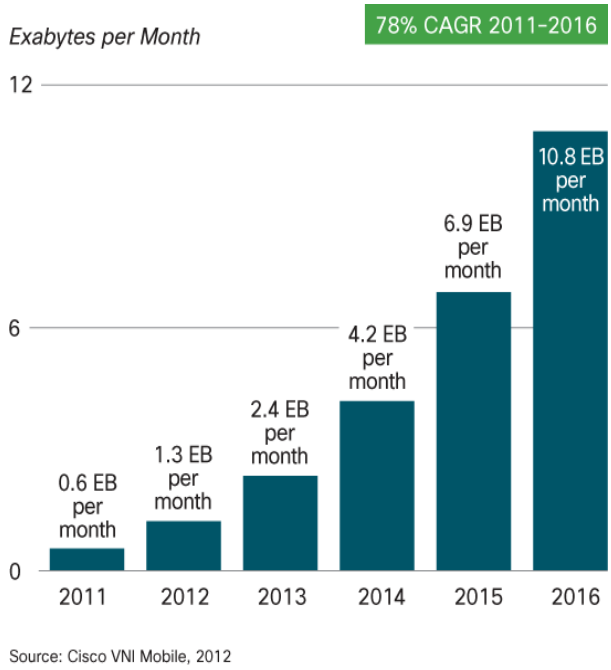


Fig. 2: Mobile data traffic forcast from Cisco [7]

We claim that the current structure of super data centers is inefficient in terms of the overall energy consumption and that distributed data centers of intermediate size will be way more efficient. In our opinion, this is a fundamental structural problem and we could learn some lessons from the cloud enabling industry, the microelectronics sector that is facing the energy problem for a long time now and already went through some paradigm changes affecting the design process of efficient architectures. Therefore we claim:

**Green computing at the wrong place is not green in total!**

## II. The Cost of Data Transfer, a Microelectronics Perspective

Together with the evolving changes in modern social life and the increasing demand of permanent network access, the recent progress in microelectronics industry is one of the key enablers of the modern ICT sector. This trend is massively boosted by the business models of big cloud service providers like Google, Apple, Microsoft and may more. However, from today's microelectronics point of view, cloud services with centralized data centers seem like a step backwards from the system architectural concepts that have evolved over the last years.

We will illustrate our thesis by shortly discussion two problems: the limited communication bandwidth and the costs of data movements.

### *A. Inter-Device Communication Bandwidth is Limited*

Up to now the technological advance in microelectronics design and manufacturing has managed to keep pace with Moore's Law that states 'Every 1.5 years the number of transistors will double per given area' [8]. The European Network of Excellence on High Performance and Embedded Architecture and Compilation (HiPEAC) claims in its latest road map that *data deluge* is already outperforming Moore's Law nowadays [9]. As a further example they point out that the on-chip complexity of a modern 4G modem is 500x higher than the complexity of a 2G modem, whereas a 4G modem only provides a 100x of throughput. Although the first explorations for future technologies in the 5G communication standard

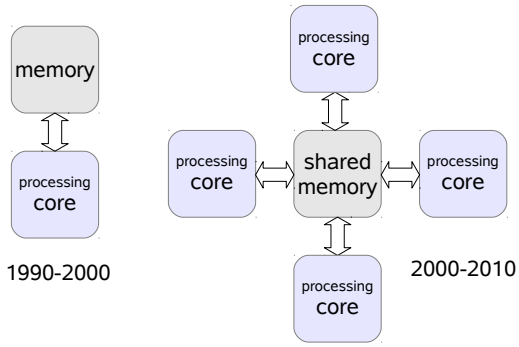


Fig. 3: Processor architecture trends within decades, from centralized storage and the first step towards distributed computing and storage

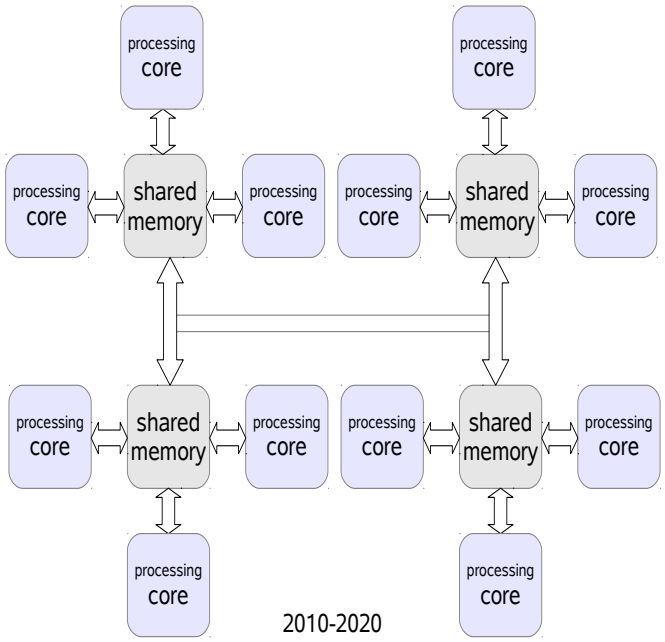


Fig. 4: Current heterogeneous distributed storage architecture with the main design rule: store data as close as possible to the processing units, see e.g. [10].

have started[1], it is currently unclear how the communication can keep track with the fast performance increase of data processing within mobile devices and data centers. In summary we can state:

**The growth of communication bandwidth is far behind Moore's Law.**

### *B. Moving Data is Costly*

In former times, a programmer's view on a desktop or server computer was very pleasant: one *central* processing unit handling all the data coming from a more or less sophisticated memory and IO system, as illustrated in Figure 3. An idyllic world also for the big players of CPU manufacturing IBM, Intel, AMD and Sun Microsystems: their economical ecosystem was nicely manageable, and the computational performance was the limiting factor. The business models of those companies strongly relied on increasing the of single CPUs by architectural and technological improvements, and power consumption was not in the main focus for many years.

However, this has rapidly changed with the increasing computational demand of the mobile (and battery powered) device market and the point when desktop CPUs started to hit the power wall [11]. The power wall means the trend of consuming exponentially increasing power with each factorial increase of operating frequency and thus performance increase.

To overcome this problem, the maxim of single CPUs in a computer had to be amended, and the CPU design companies had to change their view on processor architectures and business models. Already in 1999, the Intel Fellow F.J.Pollack claimed that the architectural change is to use heterogeneous architectures, i.e. to utilize different cores for different jobs [12]. The statement of Pollack can be summarized in the so called *law of diminishing returns of silicon area* which is the trigger of two major (current) trends to solve the power problem:

1) Smaller dedicated cores are more energy efficient than bigger general purpose cores, since the memories are closely connected to the processing units (as shown in Figure 4).

[1]http://www.eetimes.com/electronics-news/4373902/Dresden-sets-up-5G-communications-research-lab

2) Communication between clusters has to be reduced as much as possible.

Today, especially the second point (the communication on chip and thus the data movement) is a major hot topic in the high-performance computing domain. For example, Nvidia stated at the super computing conference SC10 that nowadays moving bits around within a chip is more costly than the actual computing or storage [13]. The high cost of data movement, that means in technical terms fetching the operands and data costs more than executing the operation, drives all processor designers and manufacturers to re-think about their architectures and their organization of storage and thus the mandatory communications with a distributed memory system, as shown in Figure 4.

Within the HiPEAC report [9] for future challenges, the following statement of Nvidia Chief Scientist is explicitly highlighted:
*It's not about the FLOPs any longer, it's about data movement. And further, it's not simply a matter of power efficiency as we traditionally think about, it's about locality. . . . Algorithms should be designed to perform more work per unit data movement . . . programming systems should further optimize this data movement.* [14]

In summary we see that state-of-the-art power-optimized embedded devices show a strong heterogeneous system architecture, with dedicated cores for dedicated types of jobs. They strictly follow the rule:

**Keep data locality and reduce communication.**

### C. Heterogeneity in the Cloud

Cloud computing is a hype nowadays and many cloud services are already available for private and professional use. They offer an absolutely new convenient user experience with increasingly sophisticated data collection, storage, mining and evaluation strategies. To enable these services, user data has to become an open good for the service providers in many cases, and the integration with current privacy and security regulations poses many challenging questions. For instance, on March 1, 2012 Google revised its privacy policy and merged all the formerly distinct services together. Facebook had to cope with several breaches of their privacy policies in the past, and also Apple is keeping more and more track of the user's data and activities with iCloud and iAd. In the public domain, besides secret and police services also governments have shown a growing interest in data collection and controlling pretensions in the past (the "glass citizen"). All these services in the end rely on centralized data management and are based on the paradigm:

*Reduce data locality and maximize communication.*

It is important to note that **this approach is the exact opposite of the approved low-power design methodology** that has emerged in the microelectronics industry.

This development shows more the orientation towards new business models or supposed security on the cost of the customer's privacy than real effort towards *green* services. It is obvious that *energy efficiency* and *green IT* are the well-appreciated vindications, but definitively no stable justifications for the trend towards the cloud. The opposite is the case: more and more data has to be transferred to central physical locations (and out of the user's control), and the energy for the required communication effort is growing exponentially.

## III. Economic and Social Impacts on Green Computing

Big market players like Google, Apple or Microsoft are not moving into cloud services without cause. In general, a company's care about environment is not a pure philanthropic gesture, but usually substantiated by strong economic reasons. The business models behind cloud services therefore have a big impact on how *green* the service can be in the end. In fact, if a company can increase its profit, where's the intention to make it in a green way? This is especially interesting due to the fact that according to the Cisco "Global Cloud Index" from 2011, the estimated traffic generated by cloud services will triple from 2012 to 2015 [15].

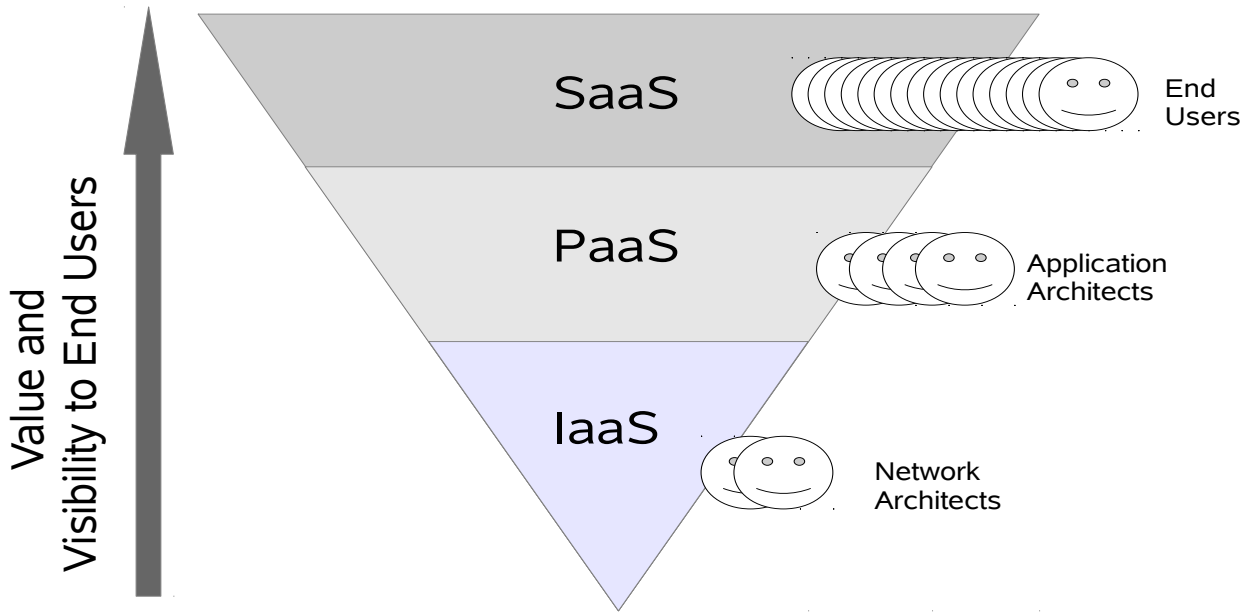


Fig. 5: Visibility of the different layers with respect to costumers

When tackling this question, we have to look a little bit closer on the business models of cloud service providers, that in our opinion are quite orthogonal to green infrastructures. The current trend for every software or hardware company is the so-called *XaaS* model, what stands for 'everything as a service'. In the specific field of cloud computing, three main models can be observed:

- *IaaS*: infrastructure as a service
- *PaaS*: platform as a service
- *SaaS*: software as as service

The abstraction level from the underlying technology and at the same time the awareness by the user increases from the IaaS to the SaaS model, as shown in Figure 5. At the same time, the technical correlation with energy costs decreases from IaaS to SaaS. The companies providing the infrastructure for data processing, communication and storage therefore seem to have the greatest interest for energy saving, since this directly reduces their expenses.

On the other hand, IaaS and PaaS are pure B2B products, and their visibility from a user perspective is small. The user in general is only interested in the final service he or she sees. SaaS products as the interconnection point to the user may take the *green* label of the underlying IaaS and PaaS layers and make it visible to the user in the value added chain, but only for marketing reasons. Users can salve their conscience by using a *green* service. This leads to the strange situation that the primary demand for energy saving is actually far away from the user, at the closest point to technology in the whole service framework.

Furthermore, collecting all the computation in central data centers decreases labor costs and the maintenance effort, and business models relying on user data collection and data mining will prevent any effort to decentralize cloud services. In our opinion it is not clear at first glance why the cloud computing industry itself should be interested in moving towards decentralized data centers, and so to reduce communication and thus overall energy consumption while maybe sacrefying the local large scale advantage (energy consumption). Indeed, we currently observe a turning away from this power saving paradigm towards traditional centralized computing. For example, in the iPhone 4 Apple implemented a voice recognition system that runs locally on the device. With the introduction of the iPhone 4S and the (much more mighty) Siri, the data processing has been shifted to data centers in the cloud. The question is if this has only

happened for economic reasons, or if Apple is exploiting the strong link between Siri and iCloud: if the technological progress of the chip industry would allow to execute Siri on the iPhone locally, would Apple move Siri's voice processing back to the device, or rather enhance it with new features and thus keep the data control?

Besides economic efficiency and pollution control, there are other reasons that might trigger decentralized cloud structures. An important aspect that is directly experienced by the cloud service user is the communication latency. In contrast to the communication bandwidth that determines the maximum throughput of data that can be send or received by a device, the latency is the reaction time between a user initiated event and the arrival of the service's answer. The latency is independent of the amount of data that is transmitted, and can be crucial for a service to be accepted by users: perceived slow reaction times are annoying and may cause the user to change the service,

However, the communication latency is strongly correlated with intermediate stations like media changes (from air to wire) switches, and routers in the communication chain between the device and the data center. More smaller and distributed data centers located closer to the end-user devices will therefore help to bring the latency down. Besides that, security and safety aspects, robustness, or legal reasons can be further reasons to decide for a (partly) decentralized cloud service structure.

Nevertheless, it's the user in the end who makes use of the cloud services, and his or her consumers' behavior actually determines if a service can become *green* in total. The big challenge therefore is to evolve a global view on cloud services, ranging from the data centers at the one end over the communication network until the user as an individual in social life. First thoughts on this aspect in the domain of cyber-physical systems have recently been formulated by the German National Academy of Science and Engineering, acatech [16]. In the end, the cloud service users have the biggest influence on the architectures and the amount of *green* clouds that we will see in a few years. Therefore we think it is mandatory right now to raise the awareness of the underlying relations in public, allowing people to develop a feeling about the environmental costs of cloud services and finally contribute to *greener* computing maybe by a more responsible handling of those.

## IV. Harmonized Education Towards Green ICT

The motivation to take steps into the direction of *greener* life results from a social consensus saying that environmental protection is important for all of us. In order to really make people change their way of living, comprehensive information and education concepts have to be developed, and based on the public legitimation, the government has to enact the appropriate laws that induce companies and users to go for environmental friendly solutions.

The connection between a powered light bulb and its impact on the environment is obvious to everyone, not only due to the latest news coverage and extensive attempts to alert the value of electrical energy to the general public. However, many people still believe that ICT services in general and cloud services in particular are *green* by definition: to recognize that firing 50 Google queries or browsing Facebook for one hour is not for free is challenging, but we believe that this awareness is the key to cleaner cloud services.

Therefore we strongly want to point out the importance of education for two groups of people:

- The end users of cloud services in general,
- and cloud system software and hardware architects in particular.

Only comprehensive education will lead to a enduring change of system architecture paradigms and service consumer behavior.

An example for the success of this approach is the movement towards renewable energies in many countries, as stated in the Greenpeace report [5]. Based on a comprehensive information campaign, the road maps that have been developed over the last years rely on a broad common consent of the population. Together with the supported expertise of engineers on this field, the so-called smart grid initiatives could be developed.

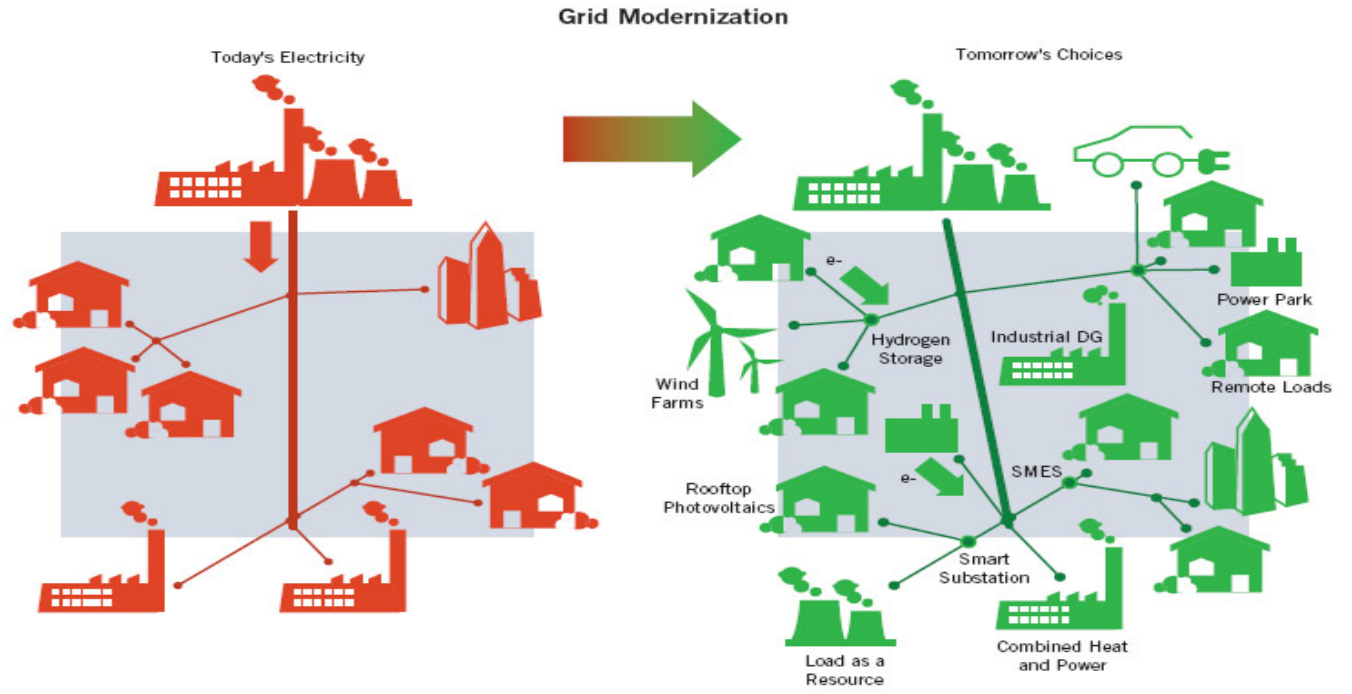


**Fig. 1. The IEEE's version of the Smart Grid involves distributed generation, information networks, and system coordination, a drastic change from the existing utility configurations.**

Fig. 6: Smart grid for electric utilizes, the move from centralized to heterogenious power generation [17]

It is quite interesting to notice that within a smart grid, new power generation sources like wind farms or photovoltaics the originally centralized power generation is moving towards decentralized, heterogeneous power generation. The IEEE vision of a distributed power generation is shown in Figure 6 [17]. Obviously that matches our ideas for the future structure of cloud services.

The acatech study on the future cyber-phsical systems [16] concludes that many people are lacking basic know-how in the fields of Internet and cloud technologies (p.198). Even if a majority is still able to understand and use the service itself, they are very often aware of the impacts on aspects like privacy, security and energy consumption. Since the authors are with the University of Kaiserslautern in Germany, we found out that even graduate students with a strong background in microelectronics (the absolute specialists) have problems to get an overview over this topic.

One reason we could figure out is strongly related to the modern modular teaching system: dependencies between lectures are currently removed as much as possible to make it easier for students to select the time and location where they would like to take the lecture. This has led to a strong concentration of separate core aspects in the lectures, for example optimized processor design is taught in a different lecture than memory system organization, not to talk about environmental or social impacts.

Cross-disciplinary education is hard to realize under these circumstances. As a starter, we have come up with a special *Green Computing Seminar* in the last semester in our department, and also other universities are going in that direction. However, only a small part of the whole society will ever attend these offers. Therefore we propose to re-focus on interdisciplinary connections between the more and more complicated subjects on all levels of education, from play to graduate school.

**All-embracing education is the key to greener clouds in the end.**

## ACKNOWLEDGMENT

We would like to thank our boss Prof. Wehn for hot and generative discussions. Special thanks to Dropbox, Google, Apple, Microsoft and a dozen more cloud services that have enabled the synthesis of this work.

# REFERENCES


[1] J. Mullich, "16 Ways The Cloud Will Change Our Lives," *The Wall Street Journal*. [Online]. Available: http://online.wsj.com/ad/article/cloudcomputing-changelives (last access June 2012)

[2] SMART 2020, "Enabling the low carbon economy in the information age," 2012. [Online]. Available: http://www.smart2020.org/_assets/files/02_Smart2020Report.pdf (last access June 2012)

[3] T Systems, "White Paper Nachhaltigkeit und ICT," 2012. [Online]. Available: http://download.sczm.t-systems.de/ContentPool/de/StaticPage/52/85/04/528504_WhitePaper_Green-ICT-ps.pdf (last access June 2012)

[4] L. Barroso and U. Holzle, "The Case for Energy-Proportional Computing," *Computer*, vol. 40, no. 12, pp. 33 –37, dec. 2007.

[5] GREENPEACE, "How dirty is your data," 2011. [Online]. Available: http://www.greenpeace.org/international/en/publications/reports/How-dirty-is-your-data/ (last access June 2012)

[6] European Commission DG INFSO, "Impacts of Informatin and Communication Technologies on Energy Efficiency," 208. [Online]. Available: ftp://ftp.cordis.europa.eu/pub/fp7/ict/docs/sustainable-growth/ict4ee-final-report_en.pdf

[7] Cisco Systems, Inc, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 20112016," 2012. [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html (last access June 2012)

[8] G. E. Moore, "Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp.114 ff." *IEEE Solid-State Circuits Newsletter*, vol. 20, no. 3, pp. 33–35, 2006.

[9] H. E. N. of Excellence on High Performance, E. Architecture, and Compilation, "Computing Systems: Research Challenges Ahead The HiPEAC Vision 2011/ 2012," www.hipeac.net (last access June 2012), 2011.

[10] EETimes, "Getting started with multicore programming," http://www.eetimes.com/design/embedded/4007623/Getting-started-with-multicore-programming-Part-1 (last access June 2012), 2008.

[11] J. M. Rabaey, *Low Power Design Essentials*, 1st ed. Springer, 2009.

[12] F. J. Pollack, "New microarchitecture challenges in the coming generations of CMOS process technologies (keynote address)(abstract only)," in *Proceedings of the 32nd annual ACM/IEEE international symposium on Microarchitecture*, ser. MICRO 32. Washington, DC, USA: IEEE Computer Society, 1999, pp. 2–. [Online]. Available: http://dl.acm.org/citation.cfm?id=320080.320082

[13] NVIDIA Corporation, "GPU Computing," 2010. [Online]. Available: http://www.nvidia.com/content/pdf/sc_2010/theater/Dally_SC10.pdf (last access June 2012)

[14] N. Bill Dally, Chief Scientist, http://www.hpcwire.com/hpcwire/2011-08-18/taking_a_disruptive_approach_to_exascale.html?featured=top (last access June 2012).

[15] Cisco Systems, Inc., "Cisco Global Cloud Index: Forecast and Methodology, 2010-2015," http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns1175/Cloud_Index_White_Paper.pdf (last access June 2012), 2011.

[16] A. N. A. of Science and Engineering, "Acatech study: agendaCPS (cyber physica systems)," http://www.acatech.de/ (last access June 2012), March 2012.

[17] P. E. Technology, "Smart Grid Success Will Rely On System Solutions," http://powerelectronics.com/power_systems/smart-grid-success-rely-system-solutions-20091001/ (last access June 2012), Oct 2009.