The LEAP Technology Landscape

Lower Energy Acceleration Program (LEAP) Solutions, Adoption Factors, Impediments, Open Problems, and Scenarios

Roberto Verdecchia
Vrije Universiteit Amsterdam
The Netherlands
r.verdecchia@vu.nl

Patricia Lago
Vrije Universiteit Amsterdam
The Netherlands
p.lago@vu.nl

Carol de Vries
PhotonDelta
Eindhoven, The Netherlands
carol@photondelta.com

TARGET AUDIENCE

This technology landscape is intended for all stakeholders that aim at contributing to building a future-proof energy efficient digital infrastructure, from business organizations like data centers, software development companies, telecommunication service providers, to business customers, NGOs and end users; but also governmental organizations, decision makers and funding agencies.

Preferred citation: Verdecchia, R., Lago, P., & de Vries, C. (2021). *The LEAP Technology Landscape – Lower Energy Acceleration Program (LEAP) Solutions, Adoption Factors, Impediments, Open Problems, and Scenarios.* LEAP Initiative, Amsterdam Economic Board.

1 BACKGROUND

With the introduction of high bandwidth data transfers, affordable data plans, the generalized migration to the cloud of software applications and data management, and the popularization of streaming services, digital infrastructures are experiencing an ever-growing demand of data consumption. As expected, the related energy consumption is steadily increasing over time. This motivated sector leaders like Microsoft, Google and Amazon, to increasingly adopt in recent years renewable energy resources, *e.g.,* solar and wind farms, as a means to lower the environmental impact of their hyperscale data centers. Nevertheless, adopting renewable energy can be considered only as part of the solution, as (i) such adoption does not tackle the need to optimize the use of cloud resources, and (ii) the production of renewable energy will not meet its demands already in the near future. With the global transition toward the adoption of renewable energy resources, the whole society will need them. Therefore, exploiting renewables for the future data infrastructures will not, as such, make them sustainable, but rather they need to become energy efficient, too, not to compete with the other industrial sectors. This is especially true when considering the Netherlands, which constitutes a prominent European "data hub" distributed over a relatively small geographic area[1].

For the last decades the digital infrastructure industry has been able to maintain a relentless pace of introducing new generations of faster and more energy efficient computing hardware approximately every two years. Nevertheless data consumption is rising faster than the improvement in energy efficiency and now also the so called "Dennard scaling" [4], that allowed lower power consumption with each new semiconductor generation, is irremediably coming to an end. In addition, with the ever-growing increase of data transport speeds, the power consumed in wiring and communication is rising more than linearly. To maintain the increase in data processing power, new solutions are needed.

In this context, the Lower Energy Acceleration Program[2] (LEAP) was launched to explore alternative solutions towards a sustainable growth of the data center industry. The aim of LEAP is to accelerate the transition to a sustainable digital infrastructure in which we integrate innovative developments at the heart of our energy system and provide a solution for spatial planning with circular use of materials. One of the topics in the LEAP program is the development of a technology landscape for energy efficient digital infrastructures. This landscape focuses on three different temporal horizons, namely:

- Horizon 1 (H1): State of the art (today)
- Horizon 2 (H2): Within the next 4-6 years (near future)
- Horizon 3 (H3): Beyond 6 years (future)

In this report, we provide insights into the research conducted by the Software and Sustainability (S2) research group[3] at Vrije Universiteit Amsterdam focusing on H2 and the role of software technology for energy efficient digital infrastructures; and PhotonDelta focusing on H3 and the role of hardware technology.

In this research, we have carried out a series of interviews with various stakeholders of the data center industry, on the future directions that will lead to more energy efficient digital infrastructures. The collected results were then refined and validated *via* focus groups involving additional participants. In total, 45 participants took part to this study.

It is important to note that the presented landscape focuses on the *energy efficiency* of digital infrastructures. Hence, other sustainability aspects of digital infrastructures, such as life cycle assessment, carbon footprint, circularity and waste management, etc. fall outside the scope of this landscape.

The report is structured as follows. In Section 3 we provide a broad overview of the energy efficient digital infrastructure solutions prospected in the 3 horizons. Section 4 presents the key adoption factors of the solutions, while Section 5 presents the most prominent impediments for their adoption. In Section 6, we present open problems regarding energy efficient digital infrastructure solutions. Finally, Section 7 sketches the four future scenarios emerging from the study, followed by an outlook to the next steps in Section 8.

[1] https://www.pbl.nl/publicaties/grote-opgaven-in-een-beperkte-ruimte

[2] https://amsterdameconomicboard.com/en/initiatief/leap-lower-energy-acceleration-program

[3] https://s2group.cs.vu.nl

2 HOW TO READ THIS LANDSCAPE

Open for inspiration. In this report we describe the concepts that we uncovered in the study, as well as the time horizons in which the participants placed them. This suggests a strategy about when each concept is expected to be available on the market provided we invest in research and development to create them – hence a landscape. However, this landscape can (and should) be read in various ways depending on the reader's perspective. In general, it should be an inspiration to think out-of-the-box about solutions and innovations needed to develop a future-proof and energy efficient digital infrastructure. In particular it can act as a strategy to achieve a goal over time, *e.g.,* to position technological solutions over time, so that they incrementally build upon one another (see the sketch in Figure 1.(a)); or to incrementally realize target scenarios that help creating socio-technical solutions which contribute to a systemic way of thinking (see Figure 1.(b)). The aim of the landscape is to inspire to take action. LEAP will assess which innovations to take forward in its next phase of the LEAP initiative, to further accelerate the transition. Note: this landscape focuses on energy efficient solutions; the energy use for production, materials scarcity, toxicity or waste will need to be a topic for additional research.

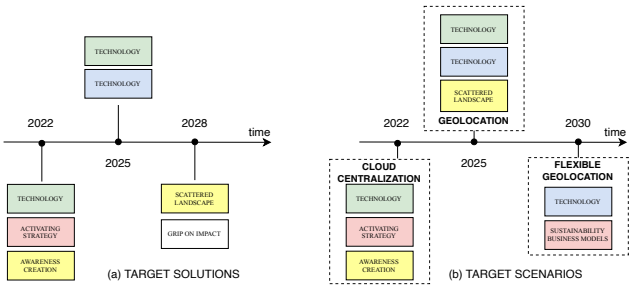

Figure 1: Landscape illustrations (a) per target solutions and (b) per target scenarios

Key of reading. Throughout the report, concepts such as solutions and open problems, are categorized *via* acronyms. The general structure of the acronyms is "[sustainability dimension]-[concept type]". *Sustainability dimensions* can be technical (T), social (S), environmental (E), economic (Ec); paradigm shifts (PS); or general (G) when not tight to any particular one. *Concept types* can be solutions (S), adoption factors (AF), impediments (I), or open problems (OP). For example, "Domain-specific Hardware (T-S)" is a technological solution. It must be noted that our main focus is on novel technologies/solutions; as such, existing technologies (*e.g., software virtualization* like virtual machines and containerization) are left implicit even if they will undergo continuous improvement and optimization, over time, to contribute to energy efficiency.

The solutions are temporally ordered throughout the landscape horizons, showcasing the technology readiness in terms of widespread-adoption and full impact, as perceived by the participants of this study.

3 LANDSCAPE OVERVIEW

Figure 2 gives an overview of the solutions described in this section. It is important to note that, while in Figure 3 software-centric solutions are reported, it is in general hard to distinguish between hardware- and software-centric solutions for most concepts presented in this section (*e.g., non-Von Neumann architectures* are interlaced with *novel software architectures*). Hence, in the reminder of this section, solutions are presented without making a distinction between hardware- and software-centric ones for the sake of clarity.

3.1 H1: Solutions for Today

Solutions belonging to the H1 are characterized as being readily available for adoption. While the focus of this study is on horizons H2 and H3, we captured also the solutions in H1 that have been mentioned by the participants, or well-known solutions that are being adopted in H1 but are expected to reach further maturity or full potential in H2. Examples may include software virtualization solutions (*e.g.,* virtualization and containerization also mentioned in the H1 report [9]) which are instrumental to maximize energy efficiency in the H1 scenario of cloud centralization (see Section 7.1), and implement more innovative scenarios like energy-driven dynamic consolidation, and the flexible geolocation highlighted in Section 7.2. Although well known, there is (still) significant room for optimization to maximise the energy saving potential [1].

Moving to the Cloud (PS): During H1 we see a first paradigm shift, already occurring in present time, namely *moving to the cloud*. This paradigm shift entails moving data, computational, and software capabilities from on premise to the cloud. In other words, rather than owning resources locally, resources are accessed on-demand, as provided by renowned cloud computing services (*e.g.,* Amazon Web Services, Microsoft Azure, and the Google Cloud Platform). This paradigm shift influences the rise in popularity of software applications specifically designed to be deployed on the cloud, such as *cloud-native and serverless applications*. For instance, a study from Eclipse about cloud computing growth between 2008 and 2014 [5], found that 86% of companies were already using between 1 and 4 different types of cloud computing services, and predicted that 50% of all IT would be cloud-based between 2019 and 2025. Figures from 2020 report 90% of companies being cloud based [6], hence far exceeding the expectations. As another example, while cloud-stored data witnessed a yearly growth of 20%, about 99% is waste as hardly ever used [3].

Heuristics for Hyperscale Hardware Management (T-S): Moving to the cloud entails a growing centralization of software and hardware resources in hyperscale digital infrastructures. Therefore solutions relative to this paradigm shift are prominently characterized by the energy optimization of this type of digital infrastructures, *i.e., heuristics for hyperscale hardware management*. A frequent adopted solution regards *heat management*, such as *efficient cooling* strategies (*e.g., immersion cooling*) and the *reuse of dissipated thermal heat*. In addition, energy consumption of hyperscale digital

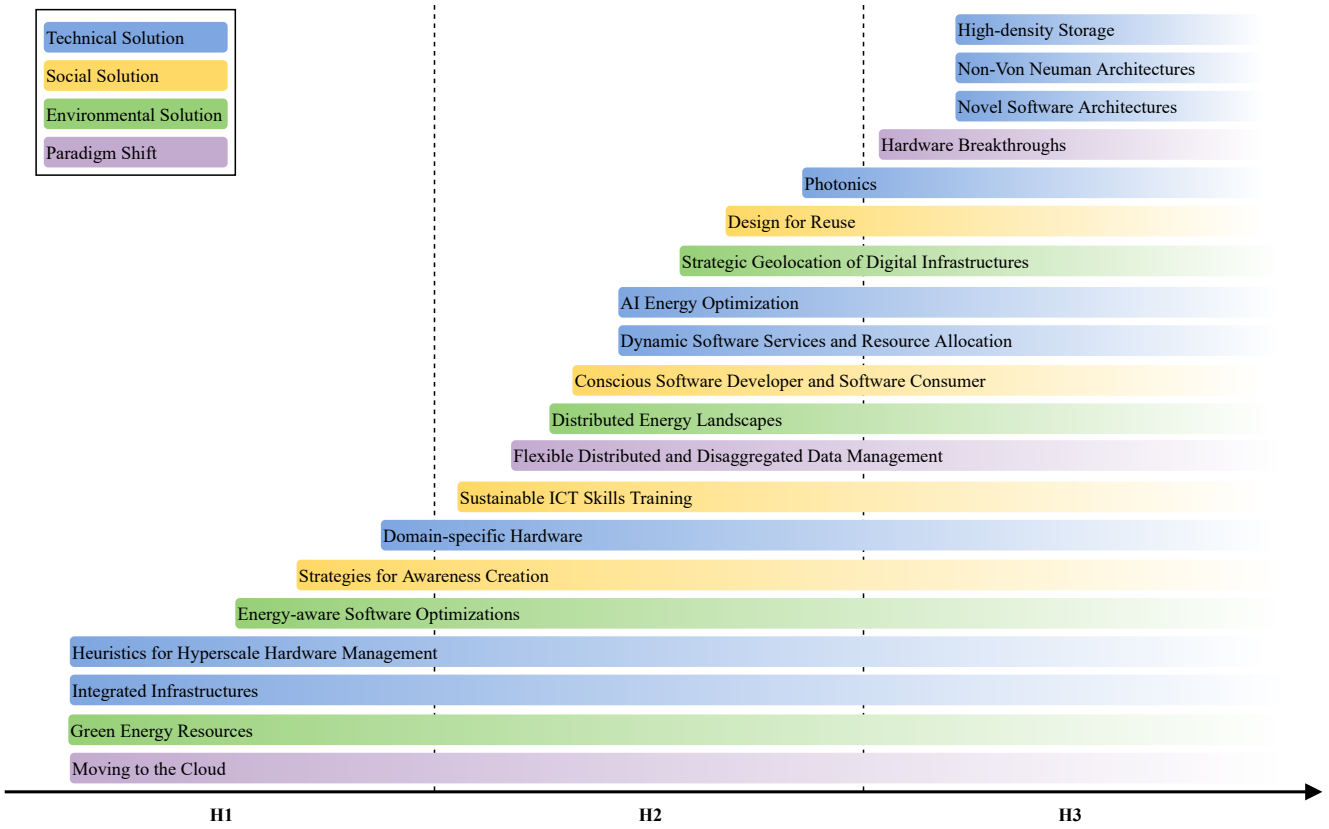

Figure 2: Overview of the temporal distribution of solutions (H1-H3)

infrastructures can be lowered by adopting *energy-aware storage optimization*. This solution often entails moving the data that requires high transfer speeds to solid-state dive storages (SSDs), while archiving less-frequently accessed data *via* long-term backup storage solutions, *e.g.,* Amazon Glacier[4], that are far less performant, but also more energy efficient.

Green Energy Resources (E-S): Another prominent category of solutions regards the adoption of *green energy resources*, *e.g.,* solar and wind farms, which often envisions the proximity of the future hyperscale digital infrastructures to green energy resources.

Energy-aware Software Optimizations (E-S, H1): *Energy-aware software optimizations* of the applications running on the digital infrastructures is another category of solutions that start to arise in H1. An overview of this type of solutions, distributed over the three horizons, is reported in Figure 3. Related to this horizon, a cloud-centric specific solution is what is referred to as *kill zombie systems*, *i.e.,* the detection and shut down of idle servers to ensure that no energy is wasted to keep unused hardware resources running. In addition, the transition to the cloud encourages the *use on demand* of resources, enabled by *event-based software engineering*, and allowing to timely consume cloud resources only when certain triggers appear in the event stream. Related to the *moving to the cloud* paradigm shift, H1 sees the rise of *cloud-native* and *serverless applications*.

Integrated Infrastructures (T-S): Another solution, which spans also across H2 and H3, regards the creation of *integrated infrastructures* born from the tight collaboration between software and hardware companies, that develop dedicated hardware built to satisfy the needs of software companies in a sustainable fashion.

Domain-specific Hardware (T-S): A solution specific to hardware regards instead the appearance during H1 of *domain-specific hardware*. Such hardware components are designed to efficiently solve specific problems, *e.g.,* the appearance of new graphics processing unit (GPU) types developed and optimized specifically for deep learning.

Strategies for Awareness Creation (S-S): A solution addressing the social dimension of the sector and much broader in scope, regards *strategies for awareness creation*, *e.g.,* monitoring and communicating about the environmental impact of data production, manipulation, and usage, hence striving towards a conscious use of energy, and a behavioural change in the data consumption patterns. *Design for reuse* implies a possible implicit tradeoff between energy consumption optimization and hardware waste, as utilizing longer lived technologies can imply a delay in using the newer more efficient technologies.

[4]https://aws.amazon.com/glacier/

3.2 H2: Solutions for the near future

Flexible Distributed and Disaggregated Data Management (PS): Occurring in H2, this paradigm shift foresees a transition from hyperscale data centers to *flexible distributed and disaggregated data management*. With the steady advancements in communication technologies, and the growing affordability of computational power, *edge computing* is expected to gain a widespread popularity in the near future. At the same time the "edge" will take different shapes from what we first thought, varying from static mini-clouds on premise to flexible "follow-the-need services". This will imply to elastically move a vast number of computational tasks as near as possible to both the consumer premises and the increasingly decentralized energy production, *e.g.,* via *onboard computing*, and *distributed networks of computational nodes*.

Strategic Geolocation of Digital Infrastructures (E-S): Transitioning towards a flexible distributed and disaggregated data management will allow for the *strategic geolocation of digital infrastructures*. With this solution, digital infrastructures can be strategically positioned close to their end-users, in order to ensure high bandwidth, and keeping low the energy consumption of data flow and related communication.

Dynamic Software Services and Resource Allocation (T-S): In addition, distribution and disaggregation supports better profiling of energy consumption patterns, allowing for *dynamic software services and resource allocation*. At a coarser level of distribution and disaggregation, *mini-clouds* can be seen as intermediate steps supporting the transition towards a completely distributed paradigm. This solution enables to profile data usage patterns, and dynamically allocate services and resources by considering also the specific action performed on the data, *e.g.,* data transport, storage, or manipulation. For example, the energy efficiency of data staging can be optimized *via* profiling by analyzing its frequency of use, and subsequently allocating the best fitted resources to reduce data traffic and improving performance at the same time.

Distributed Energy Landscapes (E-S): In addition to mitigating the energy waste of both, or either, hyperscale and co-location data centers (*e.g.,* due to idle times or suboptimal virtualization practices), the shift towards a distributed paradigm enables also the use of *distributed energy landscapes*, supported by *smart energy grids*, to locally produce and consume energy, avoiding the inevitable energy dissipation characteristic of a centralized monolithic system.

AI Energy Optimization (T-S/E-S): Another characteristic aspect of H2 is the prominent role that artificial intelligence (AI) will play. As for other computational tasks, AI is expected in the near future to shift further towards distribution and disaggregation, enabled *via* (i) novel *federated learning algorithms*, supported by the appearance of *edge AI*, (ii) *data optimization/compression strategies* allowing to transfer high volumes of curated information rather than raw data, and (iii) *approximate computing*, *i.e.,* the provisioning of results of acceptable quality, rather than optimal, in order to reduce energy consumption. As AI training/serving are known as particularly energy greedy computational tasks [12], future developments of AI require applying energy efficiency software engineering to AI-based systems, a field which is currently rapidly gaining traction [13]. In addition, promising prototypes showcase how energy consumption of AI-based systems could be reduced by utilizing *AI-dedicated hardware components*, *i.e., AI on chip*.

Energy-aware Software Optimizations (E-S, H2): While in H1 *energy-aware software optimizations* apply software engineering practices for energy efficiency, in H2 it will be instrumental to create innovation as *energy efficient distributed software*, *e.g.,* flexible distribution and disaggregation, smart virtualization. While current advancements towards stable and reliable edge computing are promising, more significant research will be required to systematically shift towards a distributed adaptation paradigm. This will soon require to consolidate aspects such as *serverless architectures*, and *optimized service orchestration strategies*. In addition, the shift towards integrated infrastructures, spawn from tight collaborations between software and hardware manufacturers, will require further advancements in fields such as *infrastructure partitioning*, *i.e.,* the partition of hyperscale data centers to optimize and sustain different tasks and workloads. Current trends predict the widespread popularization of innovative software optimizations, such as *fine-grained dynamic load balancing*, and *AI-enabled optimization of software energy consumption* (*e.g.,* to manage virtualization and scheduling tasks). Other solutions specific to energy-aware software optimizations (cf. Figure 3) regard (i) the *software virtualization of hardware resources*, virtualizing pools of hardware resources in order to ensure the seamless allocation and use of heterogeneous hardware components available on the cloud, (ii) *workload optimizations*, carried out to dynamically tune hardware and software resources to best fit the task at hand, and further advancements in the field of *energy-driven software engineering*, allowing to refactor software applications to make them more energy efficient, while maintaining unvaried their delivered functionality.

Sustainable ICT Skills Training (S-S): H2 is characterized by the popularization of **sustainable ICT skills training**, carried out both in academic and industrial settings. Such educational paths are necessary to systematically create profiles able to reason about and address the energy sustainability of digital infrastructures in all types of industrial sectors.

Conscious Software Developers and Consumers (S-S): In addition to educational training of sustainable ICT skills, advancements in energy efficiency measurement and monitoring allow for the rise during H2 of *conscious software developers and software consumers*, *i.e.,* people who develop a renewed sense of responsibility regarding the sustainability of the ICT solutions they implement and use. The rise of responsible software developers and software consumers on the one hand leads to more sustainable software solutions "by design", and on the other hand increases the (quality of) sustainability requirements of digital infrastructures demanded by consumers.

Design for Reuse (S-S): H1 is characterized by an average lifecycle expectancy of digital infrastructure hardware components equal to approximately two years. Differently, H2 is expected to witness a growing trend of *design for reuse* and *hardware lifecycle management* practices. The adoption of these circular economy strategies allows to produce longer lasting and maintainable hardware components, amortizing the financial cost and environmental impact of hardware

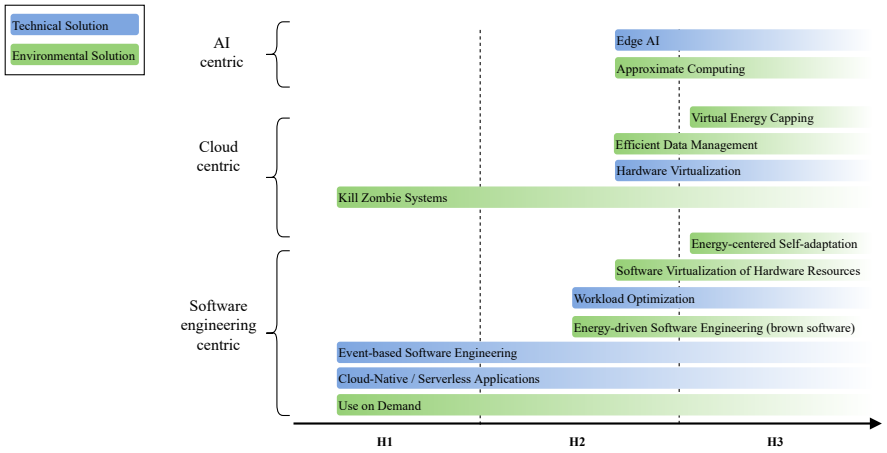

Figure 3: Zoom-in into "Energy-Aware Software Optimization" solutions

production over a longer time span. Obviously, this corresponds to a trade-off between energy-efficiency and use of critical material which is not a one-size-fits-all solution.

3.3 H3: Solutions further away

Hardware Breakthroughs (PS): H3 regards solutions that will appear in the longer term, *i.e.,* beyond 6 years. From this study, H3 appears to mainly consider novel *hardware breakthroughs*. Such hardware breakthroughs, described in the reminder of this section, are expected to drastically change how portions of digital infrastructures are designed and operate. It is important to bear in mind that in the foreseeable future, such solutions will not substitute the hardware technologies currently adopted in digital infrastructures, but will rather co-exist with them.

Photonics (T-S): Prominently, H3 is marked by a widespread use of *photonics*. The demand of low latency communication will be steadily growing while the shift towards distributed paradigms, starting in H2, will gain momentum. While photonics is already in use during H1, it is expected to become much faster and ubiquitous during H2, and will be of widespread and consolidated use in H3. In addition to inter-server/client communication (*copackaged optics*), which is already showcasing promising results, the widespread transition towards optic communication is expected to occur also *within* data centers, with the replacement of micro-electronic hardware with optics-based one by adopting *integrated photonics*.

Non-Von Neumann Architectures (T-S): In parallel, current advancements in *non-von Neumann* research showcase how, during H3, complex computational tasks will be executed at a fraction of the energy consumed by the hardware of today. Prominently, the evolution of **neuromorphic computing** and similar solutions can lead to groundbreaking energy savings required to carry out computational-intensive tasks. This notable change of hardware technologies can lead to drastic changes in the underlying hardware structures of data centers. In addition, current progress in *high-density storage solutions* showcase promising digital infrastructure storage optimizations, *e.g.,* by making use of electron spins or relations between protons and electrons, and enabling to store high volumes of data at a negligible energy cost. A consideration can also be made regarding *quantum computing*: the successful implementation of quantum computers can find applications in a dedicated class of computing, fundamentally out of reach for conventional computers, *e.g.,* quantum physics, molecular chemistry of logistics. Nevertheless, such computational tasks will with high probability fall outside the domain of conventional digital infrastructures. In addition, the low temperature at which quantum computers may have to operate (-272°C), can pose a serious concern regarding their energy efficiency.

Novel Software Architectures (T-S): As hardware is expected to present some considerable breakthroughs in H3, software will be needed to evolve and adapt to the new underlying hardware. This will require the creation of *novel software architectures*, in order to evolve software systems to best fit the drastic technology changes implied by the hardware breakthroughs of H3.

4 ADOPTION FACTORS

In this section, we present the elicited key adoption factors that, if present, would facilitate and even accelerate the adoption of the solutions presented in Section 3. An overview of the adoption factors (as well as impediments and open problems described further on) are depicted in Figure 4.

Technology Readiness (T-AF): The most-frequently mentioned adoption factor is *technology readiness*, especially relevant for H3. The technology readiness results to be a key adoption factor, as developing and on-boarding a preliminary solution with not clearly understood benefits and drawbacks implies a great risk. While solutions presented in H1 are production ready, the technology readiness of solutions appearing in H2 and H3 is hard to define. While the positioning of solutions throughout the landscape depicted in Figure 2 orders the solutions temporally, it is important to bear in mind that such positioning is not definitive, and may change as future developments of the solutions take place.

Ease of Integration (T-AF): Another prominent adoption factor, specific to the software solutions, entails the *ease of integration*, allowing to integrate solutions without any drastic change in the normal functioning of data centers. This results to be another key adoption factor, as the normal functioning of digital infrastructures cannot be interrupted while integrating a new solution, and the return of investment of adopting a certain solutions cannot be hindered by the cost of integrating the solution.

Digitization and Digitalization (T-AF): The adoption of solutions also highly depends on the future maturity of *digitization* (*i.e.,* the conversion of information into digital form) and digitalization (*i.e.,* the adoption of digital technologies in business processes). Digitization and digitalization processes can either pave the way, or inhibit, the development of sustainable digital infrastructures. This depends on the progress that digitization and digitalization will make in the future, and the extent to which their advancements will consider sustainability aspects.

Support for Trade-off Decision Making (S-AF): General to all solutions is a clear understanding of potential tradeoffs and hence the *support for trade-off decision making*, as energy savings should not deteriorate the quality of provided services. This entails also a systematic analysis of implementation and deployment costs involved, in order to understand the economic implications of proposed solutions, *i.e.,* other business cases.

Holistic Paradigm Shift (Ec-AF): Specific to the movement towards a distributed paradigm is instead the requirement of a **holistic paradigm shift**, allowing to distribute the cost of research and development across a wide range of stakeholders, instead of burdening with the implied risk a single party. Additionally, the widespread paradigm shift allows stakeholders ensuring that the undertaken change will be used and supported by other parties, hence mitigating the potential risk of developing silos technologies, *i.e.,* technologies that are hard to interface with others present on the market.

5 IMPEDIMENTS

Unclear Impact (G-I): Related to the key adoption factors, are a list of impediments, which might hinder the adoption of the solutions reported in Section 3. The first and most important impediment is the **unclear impact** of the solutions on service provision quality (*e.g.,* performance), energy savings, and evolution of technology ecosystems. This impediment is related to the *technology readiness* adoption factor, and can be mitigated only by conducting systematic experimentation to evaluate/measure the impact of the landscape solutions.

Adversity to Change (G-I): The *unclear impact* impediment can lead to *adversity to change* of certain parties (*e.g.,* telecommunication and cloud providers), as reshaping currently consolidated technologies may lead to uncharted situations, that have to be clearly analyzed and understood before undertaking major investments.

Lack of Leading Champions (G-I): The study uncovered a relation between the impediment *resilience to change* and the *lack of leading champions*, *i.e.,* leading figures in organisations, or influential organisations themselves, that take initiative and steer the change towards the next generation of sustainable technologies. This impediment is further discussed in the related open problem *lack of guidance* reported in Section 6.

Unclear Use Cases and Business Cases (G-I): An impediment characteristic to research oriented solutions (*e.g.,* quantum computing) is the yet *unclear use cases and business cases*, which have to be understood before they can be successfully adopted in industrial contexts.

6 OPEN PROBLEMS

In this section, we report the encompassing open problems of the next generation of energy efficient digital infrastructure solutions.

Need for a Coordinated Change & Scattered Landscape (S-OP): Related to the *lack of leading champions* is the perceived *need for a coordinated change*, enabling stakeholders to jointly progress, while sharing costs/risks involved, and avoid a *scattered landscape*, characterized by compartmentalized technology silos adopted only by few companies. In addition, real progress needs shared responsibility and shared accountability throughout the whole value chain, which makes all parties responsible for their actions towards the sustainability of digital infrastructures, and empowers them.

Lack of Activating Taxation Strategy (Ec-OP): A key driver towards a communal paradigm shift, and a current open legislation problem related to energy usage, is the **lack of activating taxation strategy** that should both activate all stakeholders, and target other factors than just electricity or carbon emissions. Energy bills are among the highest costs of data centers, nevertheless the current taxation strategies do not drive disruptive changes in energy consumption patterns of data centers[5]. In a foreseeable future, smart taxation strategies may be formulated, such as higher taxation of fossil energy sources, and dynamic pricing based on real-time energy demands. Similarly, e-waste production is currently only marginally supervised, but more stringent legislation could lead in the future to the popularization of "design for reuse" and other optimizations of data center hardware life cycle strategies.

Lack of Policies and KPIs (Ec-OP): An open problem related to missing activating taxation strategies is the *lack of policies and KPIs*[6]. Specifically, this open problems regards the current lack of regulations, standards, and policies regarding sustainability requirements of digital infrastructures. The complexity of this problem is also due to the current absence of KPIs, such as sustainability labels for software and hardware components, that can track and guide the progress of stakeholders developing sustainable digital infrastructures. The introduction of KPIs throughout value chains can also support an enhanced monitoring, regulation, and resolution of sustainability concerns of complex ICT business cases. The

[5] A quite successful activation strategy of the Dutch Government concluded in 2020, was the MJA (*Meerjarenafspraken*, or in English multi-annual agreements), a nationwide initiative intended to improve energy efficiency of ICT products, services and processes by granting tax exemptions when the efficiency targets were met. Results and details are online at www.rvo.nl/onderwerpen/duurzaam-ondernemen/energie-besparen/mja3-mee. Some of the resulting practices are reported in [8].

[6] Key Performance Indicators.

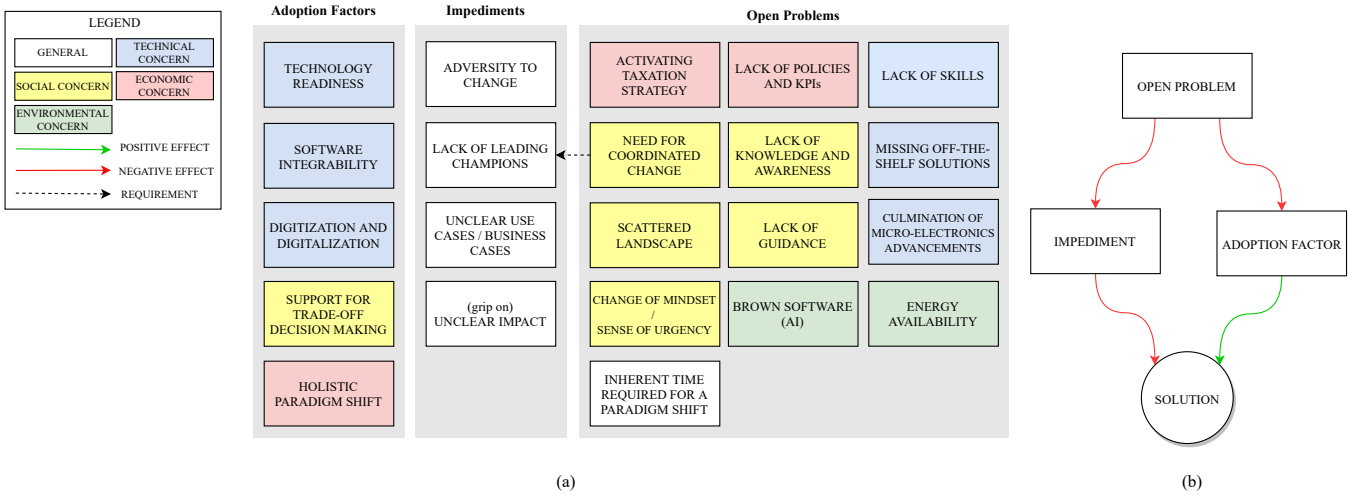

Figure 4: (a) Summary of adoption factors, impediments, and open problems. (b) Network of dependencies.

lack of policies should be addressed at both the national- and the international level, as numerous hardware and software companies are nowadays characterized by an international nature, both in terms of multi-national companies, and companies operating at a global scale.

Lack of Guidance (S-OP): Another open problem of social nature is the perceived *lack of guidance*, *i.e.,* guidance supporting companies in becoming more environmental sustainable, and supporting them in the systematic adoption of energy efficient solutions. This guidance can from either from a governmental institution, a research consortium, or even a private company, that champions and supervises the common endeavour of parties towards more sustainable digital infrastructures.

Change of Mindset / Sense of Urgency (S-OP): An open problem that can inhibit a shift towards energy efficient solutions is the need of a *change of mindset*, in order to give a higher priority to sustainability of digital infrastructures, which often is neglected in favour of other goals such as business targets an customer satisfaction.

Brown Software (T-OP): Another open problem regards applying energy-efficiency practices to the software for and in data centers which is energy inefficient, so-called *brown software*. This problem is becoming more prominent due to the ever-growing adoption of AI, which is at the moment characterized by severe software energy inefficiencies. As AI-based systems become more and more adopted, this open problem may lead to serious energy supply shortcomings if left unaddressed.

Missing off-the-shelf-solutions (T-OP): As a general trend, the current stall to move towards the next generation of energy efficient digital infrastructures, implies a current perception of *missing off-the-shelf-solutions* to optimize the energy consumption of data centers. This translates in the current need of creating or adapting *ad-hoc* solutions, instead of having the possibility to efficiently and effectively applying solutions that are available.

Energy Availability: During H1, we observe a reliable and satisfactory energy supply. Nevertheless current energy consumption trends display the rapid transitions towards electric of applications that used to run on fossil fuel. A prominent example is transportation, *e.g.,* electric cars. To support this transition, it is estimated that the electric grid in the Netherlands will need to at least double in capacity during H2. This combined with the abandonment of brown energy resources, poses two urgent and still open energy-related problems: (i) new challenges in the capacity planning of the energy infrastructure, and (ii) the impending scarce supply of renewable energy, due to the usage from other sectors and the need for space and materials. In addition, this transition will also lead to the inclusion of new notable consumers of electric energy, such as the chemical industry and hydrogen producers, who will require the implementation of dedicated energy lines and energy buffers. Nevertheless transitioning towards electric energy entails some slow processes (from both a technical and regulatory point of view), *e.g.,* installing new cables and transformers. This slow installation processes may require years in order to set up a large power connection in a geographical region, putting practical limits on the growth of certain areas. In addition, as digital infrastructures are moving towards the adoption of green energy resources, they will start to compete with industries of other nature, and even the public sector. This poses both a practical and political problem, that is currently still open.

Lack of Knowledge and Awareness (S-OP): Related to this problem, industrial contexts often suffer of a **lack of knowledge and awareness** regarding what ICT sustainability really means, which could be mitigated with dedicated education and training programs. This is also reflected in a current *lack of skills* regarding energy efficiency of digital infrastructures, that often leads to a lack of knowledge on how to address sustainability in practice, and possibly the adoption of suboptimal solutions.

In the following, we report two open problems that, given their different and encompassing nature, are reported separately.

Inherent Time Required for a Paradigm Shift (G-OP): The first general problem regarding any new technology, which is present also in the context of energy efficient digital infrastructure solutions, is the *inherent time required for a paradigm shift*. From historical data, we know that technology leaps often take between 10-20 years to gain traction, as time is required in order to clearly understand its pros and cons of the technology, and to be adopted by a wider audience. While the information and communication technology domain is characterized by an extremely fast innovation cycle, such consideration holds to a large extent also for this domain.

Culmination of Micro-electronics Computational Advancements (G-OP): The last open problem, laying at the root of the growing concerns of data center energy consumption, is the **culmination of micro-electronics computational advancements**, notably displayed by the culmination of Moore's Law [11]. As reducing further power consumption of micro-electronics is becoming physically impossible, there is no degree of freedom left to further optimize the energy consumption of such technology. Hence, in order to improve the energy efficiency of digital infrastructures, two different possibilities are available, paving the landscape for H2 and H3, namely (i) transforming the current data processing paradigms and data volumes, and (ii) systematically transitioning towards the next generation of hardware technologies.

7 SCENARIOS AND INVOLVED STAKEHOLDERS

In this section, we sketch four scenarios that emerged from this study to address the energy efficiency of the future digital infrastructures. These scenarios exploit the solutions described in the previous sections, and are ordered temporally from the one that is currently taking place, to the one that can only be achieved in a long time horizon.

7.1 Scenario 1: Cloud Centralization

This first scenario entails the migration of software and hardware resources from on-premise to a centralized remote cloud. This scenario is connected to the first paradigm shift (*moving to the cloud*) presented in Section 3.1, that occurs during H1. An overview of this scenario is depicted in Figure 5.

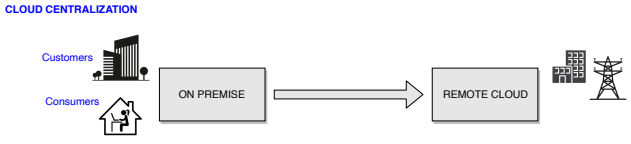

Figure 5: Cloud Centralization (Scenario 1)

Intuitively, this scenario strives towards more energy efficient digital infrastructures by delegating sustainability concerns to prominent cloud providers (*e.g.,* Amazon, Microsoft, and Google). The sustainability of this solution is based on the assumption that hyperscale digital infrastructures deploy already some energy efficient solutions, *e.g., green energy resources* and *energy-aware software optimizations* (see Section 3.1), that might be currently hard to be achieved by customers concerned with managing only a fraction of their software and hardware resources. In addition, the sustainability of this scenario is supported by a high level of *use on demand*, as the consumption of cloud resources is directly associated to a economic expense by the consumers.

As presented in Section 3.1, this scenario has to be regarded only as a temporary scenario adopted as a transition towards more energy efficient scenarios (presented in the reminder of this section). In fact, solutions adopted by hyperscale digital infrastructures, *e.g., hyperscale hardware management*, should be regarded as heuristics to mitigate the environmental impact of digital infrastructures, rather than making them sustainable in the long term. In addition, geographical space limitations, and the current trends of green energy resources development, pose serious concerns regarding the growth of hyperscale digital infrastructures, as their centralized paradigm can constitute a challenge for grid operators to ensure that the required infrastructure facilitates all consumers in any specific geographical area. In addition, the increasing user mobility (reflected in the pervasive use of mobile devices) is turning remote-cloud data storage and -traffic into important bottlenecks [14].

7.1.1 Stakeholders of Scenario 1. The most prominent stakeholders involved in this scenario are:

- *Cloud providers*, who manage their hardware and software resources, and make them available to customers in form of services;
- *Customers and Consumers*, who make use of cloud services, and migrate to various extents their hardware and software capabilities to the cloud; they may include cloud-based applications accessed by both office-workers and home-workers;
- *Hardware Producers*, who supply hardware components to cloud providers, or support them in implementing their own hardware solutions (*e.g.,* via *integrated infrastructures*);
- *Telecommunication providers*, who provide communications services to connect the different entities of the scenario;
- *Governments*, who supervise the energy taxation and regulation of customers and cloud providers.

7.2 Scenario 2: Flexible Geolocation

The scenario is characterized by a hybrid nature, in which remote clouds and micro-clouds coexist. An overview of this scenario is depicted in Figure 6.

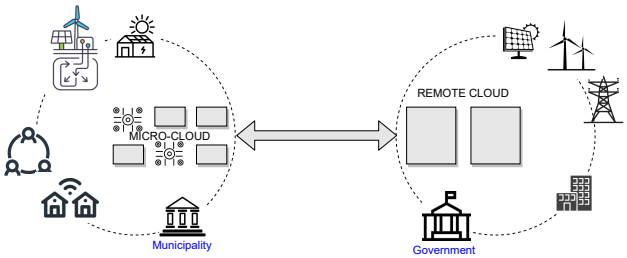

Figure 6: Flexible Geolocation (Scenario 2)

The "flexible geolocation" scenario is supported by the paradigm shift occurring in H2 *flexible distributed and disaggregated data management* (see Section 3.2). Specifically, as *edge computing*, *distributed energy landscapes*, and *dynamic software services and resource allocation* gain traction, it is possible to exploit for the sake of sustainability the energy- and computational resources available at the edge of ICT networks. This allows to distribute both the computational and energy consumption load between different geographical areas, hence mitigating the energy consumption centralized in specific geographical areas characteristic of Scenario 1. In addition, this scenario enables the appearance of *hardware, software, and energy prosumers*, *i.e.,* consumers of hardware or software resources that can not only make use of their local or personal hardware capabilities when possible, but can also make use of local energy smart grids in order to consume energy in more self-sustainable fashion. Flexible geolocation brings a systemic vision of the energy sector and the ICT sector working together, hence enabling both novel economies of scale and stability in terms of energy needs and quality of service, for both sectors. For example, the geolocation of heat production from *e.g.,* data centers and heat consumption from *e.g.,* greenhouses creates mutually-beneficial ecosystems (see also Section 8).

7.2.1 Stakeholders of Scenario 2. The most prominent stakeholders involved in this scenario are:

- *Cloud providers*, see Section 7.1.1;
- *Customers*, see Section 7.1.1;
- *Hardware Producers*, see Section 7.1.1;
- *Prosumers*, who make use of their hardware, software, and energy resources;
- *Smart Energy Grid Providers*, who supervise and manage access to smart grid services;
- *Telecommunication providers*: see Section 7.1.1;
- *Governments*, see Section 7.1.1;
- *Municipalities*, who make urban decisions about *e.g.,* spatial planning for data centers, and the support of urban solutions influencing the production and consumption of energy.

This scenario also uncovers the need for centralized government (*e.g.,* ministries) and decentralized government (*e.g.,* municipalities) to synchronize their decisions, and strategies.

7.3 Scenario 3: Seamless Continuum

This scenario is supported, among others, by advancements in *hardware virtualization*, *workload optimization* and *dynamic software services and resource allocation* that take place in H2 (see Section 3.2). This scenario is characterized by a pool of shared hardware and software resources, constituted by the resources made available by both micro-clouds and remote clouds. An overview of this scenario is depicted in Figure 7.

The shared pool of resources constitutes in this scenario a "seamless continuum", *i.e.,* hardware and software resources are allocated at runtime, in order to select from the resource available from the pool the ones which are best fitted to provide a certain service. This scenario is supported by advancements in communication technologies, both from a software and hardware perspective, allowing higher transmission speeds, and automatically avoiding congestions by profiling and rerouting data traffic as needed. For example, by considering the AI domain, the computational intensive training of an AI model can be delegated to AI accelerators available on the remote cloud, while the subsequent classification based on the trained model can be executed by a device as close as possible to the end-user. This scenario allows to progress towards the sustainability of digital infrastructures by seamlessly selecting the hardware and software resources most fitted to the task at hand, while leveraging the hybrid nature presented in Scenario 2.

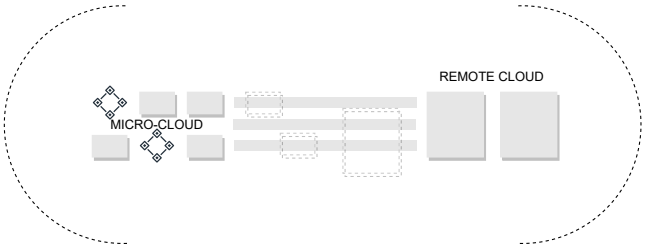

Figure 7: Seamless Continuum (Scenario 3)

7.3.1 Stakeholders of Scenario 3. The most prominent stakeholders involved in this scenario are:

- *Cloud providers*, see Section 7.1.1;
- *Customers*, see Section 7.1.1;
- *Hardware Producers*, see Section 7.1.1;
- *Prosumers*, see Section 7.2.1;
- *Smart Energy Grid Providers*, see Section 7.2.1;
- *Telecommunication providers*: see Section 7.1.1;
- *Governments*, see Section 7.1.1;
- *Municipalities*, see Section 7.2.1;
- *"Seamless Continuum" supervisors*, who supervise the distribution of resources available in the pool. This could be a new type of *aggregator*.

7.4 Scenario 4: Follow Time, Space, and Energy

This scenario builds upon the previous ones, with specific emphasis on the dynamic allocation of resources characteristic of the *seamless continuum* (Scenario 3). More specifically, differently from Scenario 3, in this scenario resources are allocated based on both their software and hardware capabilities, and on the availability of the energy the resources need, the proximity of resources, and the timeliness of the task at hand. An overview of this scenario is depicted in Figure 8, where any digital infrastructure resource can be used seamlessly, hence conceptually linking anything and everything (illustrated by the grey circle in the figure).

As an example for Scenario 4, we can consider a computational intensive task, requiring eventual consistency, that has to be carried out daily. As the execution of such task can be postponed during the day, it is possible to allocate the task to the high computational power available in the remote cloud, and executing it when the energy demand of the remote cloud is low (*e.g.,* at nighttime). If instead the load of the remote cloud never presents a decrease, it is possible to allocate the task to a network of edge computing nodes available in the proximity of the end-user, in order to carry out the task throughout the day by making use of locally available energy.

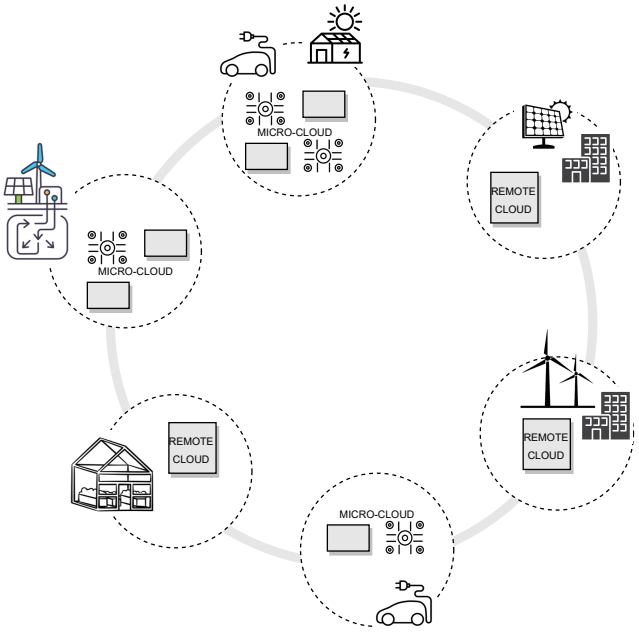

Figure 8: Follow Time, Space, and Energy (Scenario 4)

In this last scenario, the sustainability of digital infrastructures is achieved by making use of information regarding the task at hand, the energy availability, and proximity of resources, in order to achieve a sustainable service provisioning without any apparent degradation of its quality aspects. In addition, this scenario enables the dynamic prioritization of energy resources used based on energy re-use, prosumption, and overall sustainability of the resources.

7.4.1 Stakeholders of Scenario 4. The most prominent stakeholders involved in this scenario are:

- *Cloud providers*, see Section 7.1.1;
- *Customers*, see Section 7.1.1;
- *Hardware Producers*, see Section 7.1.1;
- *Prosumers*, see Section 7.2.1;
- *Smart Energy Grid Providers*, see Section 7.2.1;
- *Telecommunication providers*: see Section 7.1.1;
- *Governments*, see Section 7.1.1;
- *Municipalities*, see Section 7.2.1;
- *"Seamless Continuum" supervisors*, see Section 7.3.1;

As the scenarios are incremental, they can be ordered temporally throughout the three horizons of the landscape. An overview of such ordering is provided in Figure 9.

Interestingly, in line with our study, the Gartner report on the data centers of the future foresees that *"by 2025, 85% of infrastructure strategies will integrate on-premises, colocation, cloud and edge delivery options, compared with 20% in 2020* [2].

8 NEXT STEPS

Due to the continuous increase of data and digital services, the energy required to operate digital infrastructures is steadily increasing, so much so that it is rapidly reaching its feasibility limits

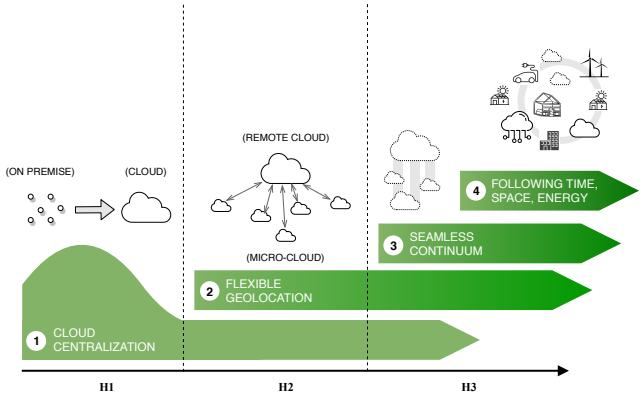

Figure 9: Scenarios across Horizons

unless new technologies are applied. Open problems, impediments, adoption factors, and solutions create a network of dependencies (see Figure 4.(b)[7]) framing the concerns relevant for deciding on the future sustainability of digital infrastructures. To mitigate environmental repercussions, the current adoption of green energy resources, and hardware/software optimizations for hyperscale infrastructures has to be intended exclusively as part of the solution, as it will not be able to scale with the ever growing data consumption demands. In the near future, this can be mitigated by shifting towards a more distributed architectural paradigm, bringing data and its processing closer to the consumer premises, with a mix of strategic data center geolocation and edge computing. Distribution and disaggregation will dynamically promote energy consumption patterns based on renewable energy supply, flexibility of services, and smart transfer and use of information, hence mitigating the environmental impact characteristic to modern hyperscale data centers. This will allow to progress in a sustainable fashion, till novel hardware breakthroughs will occur, setting new standards of low energy data storage, communication, and processing. Rather than seeing digital infrastructures completely replaced due to the introduction of new technology, the most likely progress will entail building heterogeneous digital infrastructures, where old and new technologies co-exist to provide a seamless service provision, but with a sustainable mindset.

To progress towards a energy-aware future of digital infrastructures, it is paramount that people gain awareness of the sustainability of the digital services they develop and use. This allows to hold every party present throughout value chains accountable for the sustainability of their actions, potentially transitioning towards *designing for less*, *i.e.*, ensuring that only what is really needed is produced and consumed.

In addition, the urgency to tackle the sustainability of digital infrastructures requires to promptly activate all stakeholders involved. This is reflected in the current need to revise and innovate the *modus operandi* of funding agencies, as only timely interventions can resolve the current (un-)sustainability trends of digital infrastructures. The adaptation of funding schemes requires also the

[7] The diagrams use the notation defined in [10]

strategic focus on significant national-wide sectors in the Netherlands, such as the flower industry, and the digital infrastructure industry itself.

To build the sustainable future of digital infrastructures that might be, however, all stakeholders must act together: cloud providers, cloud customers, technology providers, consumers, government, and researchers - *we are all decision makers*. Further, we *must* take into account, both qualitatively and quantitatively, possible *rebound* effects [7, 15] of optimized features of current and future infrastructures, *e.g.*, the more data/processing speed is made available, the more data is being consumed, which in turn causes a further need for speed.

ACKNOWLEDGEMENTS

This research received funding from the Netherlands Enterprise Agency Project "Energy Efficient Digital Infrastructures" (project number RVO-TSE2200010) and support from the LEAP Initiative of the Amsterdam Economic Board.

Our sincere gratitude goes to the 45 participants who took part to this study and/or provided feedback on the results, for their time, invaluable insights, interest, and support. Types of companies who participated in the study were hardware manufacturers, cloud providers, software service providers, consultancy firms, academic- and research institutes, funding agencies, and digital infrastructure customers. In addition to the participants that chose to remain anonymous, we would like to thank (in alphabetical order) Joris van den Aker (TNO), Orhan Alici (Senior Solutions Architect at Redhat), Nicola Calabretta (Assistant Professor in Electro-Optical Communication Systems at TU Eindhoven), Jeroen Cox (Strategic Lead Energy & Environment at KPN), Aaron Ding (TU Delft), Sagar Dolas (Adviser at SURF Innovation Lab), Niels Hensen (ITB2 Datacenters), Hans Hilgenkamp (University of Twente), Robbert Hoeffnagel (Adviser at SDIA/Green IT Amsterdam), Judith Inberg (University of Twente), Jos Keurentjes (University of Twente), Sjaak Laan (Director Consulting Expert at CGI), Jan-Willem Lammers (Principal Solutions Architect at VMware), Johan Mentink (Assistant Professor of Physics at Radboud University), Erik Negenman (Application Expert Low Voltage at Global Operations, Schneider Electric), Job Oostveen (TNO), Theo Rasing (Professor of Physics at Radboud University), Patty Stabile (Associate Professor on Neuromorphic Photonics at TU Eindhoven), Jeroen van der Tang (Public Policy Manager Duurzaamheid at NLdigital), Loek Wilden (CDCAP Teamleader Digital Services & Execution at Schneider Electric).

REFERENCES

[1] Barillas, A., Miller, C., Ramesh, S., 2021. Digital transformation and a net zero emissions Europe. Technical Report. Aurora Energy Research.

[2] Cecci, H., Cappuccio, D., 2020. Your Data Center May Not Be Dead, but It's Morphing. Technical Report. Gartner.

[3] Cook, G., Dowdall, T., Pomerantz, D., Wang, Y., 2015. Clicking Clean: A Guide to Building the Green Internet. Greenpeace.

[4] Dennard, R.H., Gaensslen, F.H., Yu, H.N., Rideout, V.L., Bassous, E., LeBlanc, A.R., 1974. Design of ion-implanted mosfet's with very small physical dimensions. IEEE Journal of Solid-State Circuits 9, 256–268.

[5] Desjardins, J., Ang, C., Lu, M., 2014. Cloud Computing Growth - Visual Capitalist. www.visualcapitalist.com/cloud-computing-growth. URL: www.visualcapitalist.com/cloud-computing-growth.

[6] Galov, N., 2020. 25 Cloud Computing Statistics in 2020 - Will AWS Domination Continue? https://hostingtribunal.com/blog/cloud-computing-statistics/. URL: https://hostingtribunal.com/blog/cloud-computing-statistics/. accessed: 2021-5-1.

[7] Gossart, C., 2015. Rebound Effects and ICT: A Review of the Literature, in: ICT Innovations for Sustainability, Springer International Publishing. pp. 435–448. URL: http://dx.doi.org/10.1007/978-3-319-09228-7_26.

[8] Gu, Q., Qing, G., Patricia, L., Henry, M., Simone, P., 2013. A Categorization of Green Practices Used by Dutch Data Centers. Procedia computer science 19, 770–776. URL: http://www.sciencedirect.com/science/article/pii/S1877050913007096.

[9] Harryvan, D., 2021. Savings Options in Datacenters and Server Rooms in 2021 and 2025 (in Dutch). Technical Report. Certios BV.

[10] Lago, P., 2019. Architecture Design Decision Maps for Software Sustainability, in: IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS), pp. 61–64. URL: http://dx.doi.org/10.1109/ICSE-SEIS.2019.00015.

[11] Moore, G.E., 1998. Cramming more components onto integrated circuits. Proceedings of the IEEE 86, 82–85.

[12] Saenko, K., 2020. Why AI is so Power-hungry. Ars Technica URL: https://arstechnica.com/science/2020/12/why-ai-is-so-power-hungry.

[13] Schwartz, R., Dodge, J., Smith, N.A., Etzioni, O., 2020. Green AI. Communications of the ACM 63, 54–63. URL: https://doi.org/10.1145/3381831.

[14] Vellante, D., 2021. A new era of innovation: Moore's Law is not dead and AI is ready to explode. SiliconANGLE URL: https://siliconangle.com/2021/04/10/new-era-innovation-moores-law-not-dead-ai-ready-explode/.

[15] Widdicks, K., Ringenson, T., Pargman, D., Kuppusamy, V., Lago, P., 2018. Undesigning the Internet: An exploratory study of reducing everyday Internet connectivity, in: International Conference on ICT for Sustainability (ICT4S), pp. 384–397. URL: https://easychair.org/publications/paper/VjQ1.