European Commission

Advances in Clouds

Research in Future Cloud Computing

Expert Group Report

Public version 1.0

Editors
Lutz Schubert [USTUTT-HLRS]
Keith Jeffery [STFC]

Information Society and Media

Reproduction is authorised provided the source is acknowledged.

Printed in Belgium

ADVANCES IN CLOUDS

Report from the CLOUD Computing Expert Working Group

Editors. Lutz Schubert, Keith Jeffery

Abstract. "The Future of CLOUD Computing" report January 2011 characterised CLOUD computing and identified opportunities for Europe based on R&D topics. Since then the 'hype' surrounding CLOUDs has continued and according to Gartner's hype-cycle analysis the technology is sliding into the trough of disillusionment. However, there are more offerings, more in-house development of CLOUDs and more experience in how best to use the technology. Developments have closed some gaps identified in January 2011 but more gaps have opened. This report considers the current, fast-moving, situation and identifies research topics which – if supported – could provide Europe with the know-how to be world-leading in CLOUD Computing.

EXECUTIVE SUMMARY

Nowadays we hear that CLOUD computing is ubiquitous, the way to do ICT and being adopted or at least considered by CIOs. In fact most CIOs are considering CLOUD Computing, some have converted their in-house data centres to using CLOUD technology, some have experimented with outsourcing parts (usually not all) of their ICT production to a public CLOUD and most wish for interoperable hybrid CLOUDs providing seamless elastic ICT resource provision. There is no doubt that CLOUDs have the potential for being the next generation model of utility computing, but with that potential comes the risk of failing to realise the potentials and delivering a solution that only covers a smaller scope of business opportunities and is superseded quickly. Whilst a lot of development and progress has already been made in CLOUD technologies, there still remains a wide range of concerns that need to be addressed in future CLOUD iterations in order to reach its full potentials.

This document is primarily focusing on these concerns, non-regarding all the advances made in CLOUDs over recent years, so as to identify the most pressing research issues. There is still a lot of confusion caused by CLOUD computing terminology and claimed advantages, making it difficult to distinguish which features are truly new and related to CLOUD computing and which features are simply re-branded earlier technologies coming from cluster computing, distributed computing, GRIDs and others, and therefore are potentially of less research concern.

Such concerns include guaranteed SLA and QoS including availability, security, privacy and dynamic elasticity. On the other hand the ‘pay as you go’ paradigm is seen as advantageous for ICT management accounting both in-house and using outsourced ICT systems in public CLOUDs.

To reach the full promises of CLOUD computing, major aspects have not yet been developed and realised and in some cases not even researched. Prominent among these are open interoperation across (proprietary) CLOUD solutions at IaaS, PaaS and SaaS levels. A second issue is managing multitenancy at large scale and in heterogeneous environments. A third is dynamic and seamless elasticity from in-house CLOUD to public CLOUDs for unusual (scale, complexity) and/or infrequent requirements. A fourth is data management in a CLOUD environment: bandwidth may not permit shipping data to the CLOUD environment and there are many associated legal problems concerning security and privacy. All these challenges are opportunities towards a more powerful CLOUD ecosystem.

It must be noted that European ICT industry is not characterised by some large suppliers – who may compete in public CLOUD provision with the major US suppliers – but mainly by innovative SMEs with particular skills especially in provision of software services. A major opportunity for Europe involves finding a SaaS interoperable solution across multiple CLOUD platforms. Another lies in migrating legacy applications without losing the benefits of the CLOUD, i.e. exploiting the main characteristics, such as elasticity etc.

Finally many long-known ICT challenges continue and may be enhanced in a CLOUD environment. These include large data transmission due to inadequate bandwidth; proprietarity of services and programming interfaces causing lock-in; severe problems with trust, security and privacy (which has legal as well as technical aspects); varying capabilities in elasticity and scaling; lack of interoperation interfaces between CLOUD (resources and services) offerings and between CLOUDs and other infrastructures and many more. These problems are in fact research challenges and overcoming them provides Europe with an opportunity to exploit the market.

Time is thereby of the essence: the market continues to evolve and industrial development strives for short-term solutions that satisfy the immediate customer needs. As opposed to this, research projects tend towards development and uptake cycles that exceed this customer drive – accordingly, research must focus on the research challenges that create long-term impact and contribute to shaping the future CLOUD ecosystem.

Major Findings Since The Last Report

Whereas the “Future of Cloud Computing” report [SCH10] focused on reflecting the current position of CLOUDs in the economy, this report tries to analyse the progress made over the last two years and draws conclusion for a European Cloud Research Agenda. The experts noted thereby in particular key economic differences between the USA and Europe that impact on the key challenges to be solved, and thereby differentiates CLOUD research, development, and potentially usage.

First of all it must thereby be noted that the European market is differently structured, with much more small(er) and diversified players, and a general focus towards B2B - the European telecommunication industry is a major exception to these cases though. This setup demands for much more work on integration, federation and interoperation to build up a European wide Cloud Ecosystem that incorporates and exploits this diversification for richer service provisioning.

Problems already challenging enough within more homogeneous environments thus reach a higher level of complexity, such as programming models, software engineering and multi-tenancy that need to be tackled accordingly. At the same time, this offers additional long-term opportunities, tackling the IT development to be expected anyway at an early point.

A major concern raised in this document thereby relates to the different timelines and their impact on research and development, thus highlighting the relevance for *long-term* research strategies, in order to ensure relevance and contribution of the research results to industrial developments and needs, and thereby to increase

Europe's competitiveness and position on the global cloud market. The experts have identified explicit topics and approaches outlined in this document.

These aspects also have a major impact on how we understand cloud computing: not only should this highlight the economic differentiator, but also their long-term relevance, rather than just reflecting our current model. The definition has accordingly been updated to reflect different stakeholders' view with the goal to settle the long-term understanding of clouds.

TABLE OF CONTENTS

INTRODUCTION & BACKGROUND

CLOUD concepts (as opposed to implementations) promise a cost-effective realisation of the utility computing principle, allowing users and providers easy access to resources in a self-service, pay-as-you-go fashion, thus decreasing cost for system administration and improving resource utilisation and accounting. Over the last few years, the CLOUD has generated a major impact on the global IT ecosystem, giving rise to new markets and new user communities.

The concept of CLOUDs is thereby not a novelty in itself – in fact, the principles arose from a direct industrial need to improve resource utilisation without impacting on consumer requirements, i.e. use the available resources more efficiently. Initial data centres and server farms employed load management mechanisms not unlike the base CLOUD principles, to ensure high availability according to current usage. Over time, these principles have evolved into a managed, dynamic server infrastructure, which finally led to the sales success of Amazon's EC2[1]. Multiple providers started to rebrand their infrastructures to "CLOUDs" thereby implicitly shaping the concept and understanding of the term.

Marketing and increased diversification of CLOUD offers created a large hype around the CLOUD that led to strong user expectation pressure, that partially could not be realistically be fulfilled – this is generally the case for any promising technologies or concepts. Marketing tends to promise characteristics that are easily confused with characteristics with different meanings in other domains, potentially leading to the misconception that CLOUDs

- are always available ("anywhere, anytime");
- offer infinite resources over seamless and prompt elasticity;
- expose supercomputer-like high performance;
- lead to minimal costs;
- always provide environmental ('green') benefits.

And even though CLOUDs can offer substantial support in all of these directions, they nonetheless must be limited by the system constraints, which signify the boundaries between these domains, such as mobile computing, high performance computing, internet of things etc. These boundaries are however not clear to all (potential) customers, and sometimes not even to the providers themselves, and it is one of the major research tasks for future CLOUD systems to identify the realistic and feasible boundaries, i.e. the explicit position of CLOUDS within these domains.

Many users and consumers feel compelled to migrate to the CLOUD in order to stay competitive and to benefit from the assumed improvements offered by the according infrastructure. Without the clear boundaries, i.e. without a clear knowledge about which use cases are well-suited for CLOUD environments, there is a high risk that an

[1] http://aws.amazon.com/ec2/

inexperienced user estimates the capabilities and hence the usability of CLOUDs wrongly, leaving him unsatisfied with the actual capabilities he gets from the CLOUD. Few in-depth analyses exist as yet that would provide users with enough technical insight into what capabilities they can actually expect given different use cases.

As users get more experienced in using CLOUD infrastructures, their capabilities, strengths and deficiencies become more and more apparent. Modern providers are therefore under growing pressure to deliver on these expectations, partially arising from their own promises, to satisfy the user needs, lest the interest and uptake in CLOUDs will wane again in the future. According to the Gartner analysis [FEN11], CLOUD computing reached a peak of inflated expectations in 2009 / 2010 and is now in the phase of increasing disillusionment where users become more and more aware of the deficiencies of the system and have to start assessing whether and what is worth moving to the CLOUD.

A. GOAL OF THIS REPORT

In 2010 the European Commission issued a report on the relevance of CLOUD computing for Europe [SCH10]. Within this report, the main developments, the remaining gaps and in particular the strengths, weaknesses and opportunities for CLOUD computing in the European context have been identified and elaborated.

Over the last few years, there has been substantial progress in the domain of CLOUD computing, but at the same time the expectations towards its characteristics have grown immensely. It is the task of this document to capture this progress and assess the impact onto the position of Europe in the global CLOUD landscape, and to identify the concrete research work to be performed in order to strengthen Europe's position in this competitive field.

The CLOUD market is currently a highly dynamic business field with new providers and business models arising effectively overnight, therefore creating a lot of questions about what "a CLOUD" actually is, and implicitly what to expect from it in the short- and long-term future. This makes this a vital time for **deciding the future of CLOUD computing**:

> ! *With the CLOUDs going through the "trough of disillusionment" [FEN08], more and more research questions arise, the results of which will shape what will still be considered "CLOUDs" tomorrow and what their actual capabilities will be. This is a major opportunity for Europe to shape the long-term future development as technology and market matures*

As noted, CLOUDs emerged more from a commercial need, rather than from research and development and thus pursue a more short-term evolution path to satisfy immediate customer needs. To make CLOUDs a long-term sustainable, valuable market domain, Europe must therefore focus in particular on the research and development aspects in CLOUD computing that are of long-term relevance, rather than participating in the short-term fluctuations which not only create a large scope of competition, but also may be of little impact on the global development. In this way Europe can intercept the business opportunity arising from increasing user demand after the ‘trough of disillusionment’.

It must thereby be recognised, that time plays a crucial role in this effort, as the CLOUD domain evolves quickly due to the market drive. The fast development on the market leads to a growing diversification of solutions that make steering increasingly difficult, thus risking that the results bear less long-term relevance. This must also take into account that the research cycle is typically longer than the development cycle in industry.

It is particularly relevant in this context to make sure that the essential aspects of CLOUD computing (and subsequent technologies) are addressed: CLOUDs create interest from a wide range of communities utilising IT resources. Accordingly, there are conflicting expectations and demands towards CLOUD systems (providers, infrastructures, frameworks etc.), to which different markets react in various ways. This leads to a disjointed set of research and development efforts that will create isolated solutions much like the current setup of platforms on the market.

Due to their nature and origin, CLOUD concepts are closely related to other internet based concepts and technologies, in particular Grid, Web Services, Service-Oriented Architectures etc. This is a major opportunity, as existing results, experience and expertise can be reused to build up and to enhance the features and capabilities of CLOUDs. This application of research areas of European strength is highly desirable as a form of continuation of this work, providing easier portability and migration, as well as strengthening of the technological backbone. If this overlap however is not carefully evaluated and exploited, there is a high risk of repeating work already performed in these related domains, thus wasting time and resources.

! *In order to reduce these risks and ensure that relevant results can be achieved that prove to be of* long-term relevance for European *industry, it is therefore urgent and important to identify the core features of CLOUD systems according to their relevance for the future IT ecosystem, and distinguish them clearly from on-going work and effort in related areas.*

This report tries to identify the key aspects of CLOUD systems and their relevance for the future IT ecosystem, including equally industry, academia, government and private persons. It refines the findings of the 2010 report [SCH10] with respect to the progress made over the last 1.5 years in terms of user and provider requirements. It particularly elaborates on the relevance of CLOUD provisioning for Europe and tries to identify the remaining gaps and research issues to be addressed in order to bring CLOUD related technologies towards these goals. Along this line, it tries to establish a set of common definitions that clearly identify the intrinsic aspects of CLOUDs and separates them from other, related areas.

B. About this Document

All content of this document was gathered through a series of working group meetings between invited experts from industry and academia. The meetings focused on discussing the long-term future of CLOUD computing in terms of relevance (and risks) for the European ecosystem, as well as use cases, benefits and technological gaps to enable this development.

1. Document Structure

The document is structured into 6 main sections, addressing the major concerns with respect to ensuring that CLOUD computing will prevail as a relevant asset for future European industry:

Section I provides a set of **common definitions** that captures the intrinsic characteristics and features of CLOUDs. This section elaborates the main features from different perspectives (user, provider, developer) to ensure common understanding.

Section II analyses the **current uptake and usage** scenarios for CLOUDs, and in particular which features are considered most relevant, as well as how advanced the available technologies are with this respect. The section therefore essentially captures the state of the art from both usage and provisioning perspective.

Section III compares the **strengths and characteristics** of CLOUDs against the current usage in order to identify the major potential for advanced uses of CLOUD systems. In other words, this section analyses the difference between current provisioning and the actual potential of CLOUD systems.

Section IV provides a set of exemplary **use cases** on basis of the full potential of CLOUD systems according to the definitions provided above. It elaborates the relevance of such CLOUD systems for the future IT ecosystem in Europe.

Section V extracts the **specific issues** that require more work in order to make the CLOUD vision happen, according to the characteristics, potentials and future use cases elaborated in section IV. The section elaborates the major research issues and assesses their complexity and relevance for realising the long-term CLOUD vision.

Section VI concludes the report with a set of **general considerations** to be taken into account when trying to address the research issues elaborated above. These considerations and recommendations equally concern researcher and providers, as well as project sponsoring groups, such as the European Commission.

2. ACKNOWLEDGEMENTS

This report was made possible by the European Commission; particular thanks go to Maria Tsakali, Rainer Zimmermann and David Callahan for initiating and supporting the effort.

The report could not have been realized without the excellent contributions from all workshop participants.

3. LIST OF CONTRIBUTING EXPERTS

Maximilian Ahrens, Zimory, Germany; Prashant Barot, Oracle, Germany; Francis Behr, Liberty Surf, France; Dr. Ivona Brandic, TU Wien, Austria; Dr. Marcus Brunner, NEC Europe Ltd., Germany; James Clarke, Waterford Institute of Technology – TSSG, Ireland; Dr. Stuart Clayman, University College London, UK; Dr. Thierry Coupaye, France Telecom Orange Labs, France; Dr. Peter Dickman, Google, Switzerland; Dr. Margot Dor, ETSI, France; Dr. Ake Edlund, KTH, Sweden; Prof. Dr. Erik Elmroth, Umeå University, Sweden; Andreas Ebert, Microsoft Corporation , Austria; Jürgen Falkner, Fraunhofer IAO, Germany; Ana Maria Juan Ferrer, Atos Research and Innovation, Spain; Mike Fisher, British Telecom, UK; Dr. habil. Alfred Geiger, T-Systems; Andrés Gómez Tato, Fundación Centro Tecnológico de Supercomputación de Galicia, CESGA, Spain; Prof. Dr. Yi-Ke Guo, Department of Computing, Imperial College London, UK; Dr. Giles Hogben, ENISA, Greece; Prof. Dr. Keith Jeffery, Science & Technology Facilities Council, UK; Dr. Ricardo Jimenez-Peris, Universidad Politecnica de Madrid, Spain; Prof. Dr. Frank Leymann, IAAS, University of Stuttgart, Germany; Prof. Dr. Ignacio M. Llorente, C12G Labs, Spain; Prof. Dr. Beniamino di Martino, University of Naples, Italy; Burkhard Neidecker-Lutz, SAP Research, Germany; Karsten Oberle, Alcatel-Lucent Bell Labs, Germany; Prof. Dr. Dana Petcu, West University of Timisoara, Romania; Thomas Piwek, Deutsche Lufthansa AG, Germany; Dr. Harald Schoening, Software AG, Germany;Lutz Schubert, HLRS, Germany; Prof. Dr. Theodora Varvarigou, SCES NTUA, Greece; Dr. Yaron Wolfsthal, IBM, Israel;

I. CLOUD DEFINITION

With new CLOUD providers popping up everywhere on the market promising various capabilities and the concepts having reached a peak of “maximum expectations” towards its features, it is not surprising that there exist multiple understandings of what makes a CLOUD in the first instance. Accordingly, multiple different definitions exist that show only marginal overlap and leave users, providers and developers equally confused about what can actually be expected from a CLOUD, respectively what kind of capabilities should be provided.

In addition to this, the terminology used and the conceptual context of CLOUDs have a high overlap with other domains, such as Data Centre Clusters, Grids, High Performance Computing, Web Services etc., leading to strong misunderstandings across these domains. With knowledge about the actual capabilities and best use recommendations lacking, this further intensifies the impact from CLOUD migration when inexperienced providers force porting of a specific application to a CLOUD system. Terms such as

- high performance
- accessibility
- scalability (elasticity)
- availability
- etc

thereby tend to specifically lead to misconceptions as all of these terms must actually be interpreted relatively, i.e. compared to other mechanisms for internet-based resource provisioning – in most cases even compared to simple dedicated servers.

It is therefore highly beneficial to agree on common definitions that clearly separate the domains and their characteristics from each other. This also contributes to a better understanding of the potential usage scope of CLOUDs and thus in the long run to best use recommendations in different application cases. Current definitions reflect in particular the current status and capabilities of CLOUDs, thereby not fully capturing the essential long term characteristics, nor the full potential of the usage scope of such systems. Thus, such definitions do not help to clarify the application cases.

Within this section, we will elaborate a (set of) definition(s) of CLOUD systems that equally allows users, providers and developers to identify the usability of the systems in different application contexts. To this end, we will elaborate (a) which characteristics are associated with CLOUD offerings, (b) which of these characteristics are intrinsic to CLOUDs and which belong to other domains, and (c) how and whether CLOUDs adopt or adapt the characteristics from other domains.

As with any essentially “conceptual” domain, there is a wide range of diverse opinions and views on the core characteristics of the CLOUD. The main point of the definition below is thereby not to satisfy all potential interest groups, but to identify core characteristics that are of long-term relevance for Europe.

A. CHARACTERISTICS OF CLOUDS

CLOUDs are not an isolated concept that would allow for a clear set of attributes without any overlap to other domains or interpretations. In other words, there are multiple characteristics associated with CLOUDs, partially due to the relationship and applicability in other domains, partially due to the background and intentions of the providers and finally partially due to the actual CLOUD-specific features added to these other domains.

Among these aspects attributed to CLOUDs are also economic issues that are associated with employing the CLOUD, such as turning CAPEX[2] into OPEX[3], outsourcing IT infrastructure or improving time to market. Though these characteristics are essential motivators for taking up CLOUD technologies, they do not impact on the concepts or technology behind these systems. Accordingly, they play a secondary role for defining CLOUD specifics and therefore will be ignored in the context of this section – they will partially re-arise in the context of the use cases below (see section IV).

1. THE MODERN IT ENVIRONMENT

The modern environment of IT infrastructures is defined by a number of changes in resources, usage and the general economic approach towards service provisioning and consumption. In particular, the shift is denoted by a stronger than ever concentration on the web as a means to interact, work and even to socialise and live. Whilst some of these capabilities have only been made possible by a provisioning model such as the CLOUD, it is nonetheless generally agreed that modern IT providers, users and developers will have to face the following main environmental challenges:

- ***Scale:*** the number of resources on all levels of the IT infrastructures are constantly increasing, ranging from network connects over cores per processor to devices on the internet. These numbers are increasing faster than current administrators or management support tools can cater for. At the same time, the number of users and requests is increasing, too, putting a growing pressure on providers to cope with the scale. Mechanisms for automated management of large scale infrastructures with a tendency towards homogeneous usage are therefore in growing demand. At the same time, outsourcing (in advance and on demand) to compensate the management deficiencies is a desirable means for many IT service providers.

[2] CAPEX: Capital Expenditure
[3] OPEX: Operational expenditure

- ***Heterogeneity:*** together with the number of devices, the range of *types* of resources currently increases significantly. This does not only mean mobile devices and different PCs, but more importantly, different processors with different specifics, capabilities, performance characteristics etc. Just like scale, this divergence is almost impossible to handle for the average IT infrastructure provider, let alone for the developer. Outsourcing to more homogeneous infrastructures can thereby only be considered a temporary solution, as long as the required resource types exist in sufficient amount – with the growing heterogeneity of resources, this however becomes less and less likely in the future (also see discussion sections IV, VI.A.4).
- ***Economic concerns:*** the business models around the IT service and resource provisioning have changed drastically over recent years, moving away from onetime license fees and licensing models tied to individual users or machines, to flat rates and free or low cost online applications. Also, with the growing relevance of the "Prosumer", consumer provided content gets an increasing share of the market and thus reaches a level of real competition to large company provided content.
- ***Mobility:*** in the modern, globalised economy and with modern smartphones and powerful mobile devices, a growing demand for online availability, accessibility to data, mobile office like working environments etc. is notable. With the growth in heterogeneity, scale and usage, this means though that the providers and developers have to cope more than ever with different form factors, interoperability, portability and even compensate connection losses etc.
- ***Energy constraints:*** along with the demand for mobility, comes an implicit demand for less energy consumption to ensure long operation times. But the energy constraints are not restricted to mobile devices, but in fact apply to all modern IT infrastructures with the growing awareness of the impact on the environment, and its implicit energy consumption costs.

2. Intention Analysis

The CLOUD's primary purpose consists in providing services (including resources, applications, tools etc.) under the conditions of the modern IT environments (see above). In the CLOUD case this means in particular a focus on increasing availability of services, data, or even just the infrastructure – with the main intention to reduce cost for resource utilisation. Historically, CLOUD systems arise from the need to ensure that this availability and related quality criteria can be met, even if the load is highly dynamic, i.e. the number of requests, the data size, the work load of the task etc. all vary over time. The classical approach towards this problem consists in employing sufficient resources to ensure that the respective criteria are met under worst case conditions – thus needing to provision more resources than utilized under typical load conditions ("overprovisioning"). Out of the need to reduce the resource utilisation to the current load, the CLOUD management systems were born, which can thus be considered an elaborated form of dynamic load management.

We can therefore state that the (intended) primary characteristics of CLOUD systems are:

- ***Utility computing*:** CLOUDs are the new form of utility computing, following closely up on the (technological) principles of Grids (cf. section I.B.2). They allow easy outsourcing of in particular the IT resources infrastructure, by moving the in-house applications to a dedicated (public) CLOUD provider, thus reducing the cost for administration and management. CLOUDs extend the capabilities of Grids in particular by improving the utilisation of resources, mostly through elastic provisioning (see below).
- ***Elasticity*:** is one of the primary distinguishing factors of CLOUDs from other domains. It allows the environment to – ideally automatically – assign a dynamic number of resources to a task. In other words, its goal consists in ensuring that the amount of resources actually needed is actually available to the respective task or service. This is typically employed to ensure that availability (and similar quality) criteria of a service or resource is always granted with best resource utilisation.
  Most current CLOUD offerings realise elasticity through dynamic replication of the service instance, respectively an according image within the infrastructure – this is often referred to as "horizontal" scalability. This can accommodate for multiple users accessing the same service, but also for hosting multiple tenants on the same resource.
  As opposed to that, only few systems support if a single instance requires more resources to increase performance (rather than availability) – this is referred to as "vertical" scalability. Currently, vertical scalability is almost only supported for storage, which can grow with the requirements of the user, but not for parallelised applications.
- ***Availability and Reliability***: Closely related to availability, CLOUD systems build up on the internet of service principle to expose the services in a highly accessible fashion, i.e. with minimal configuration and device requirements (generally through a browser). CLOUDs enhance this aspect further by virtualising the service / resource access, basically allowing access "anywhere, anytime". The actual resource dedicated to the user may thereby be completely unknown and constantly vary according to the elasticity principle. The principle can in particular be employed to realise self-service capabilities in IT based businesses.
- ***Ease of use*:** There is a lot of debate about whether "ease of use" can be considered an essential characteristic of CLOUDs and what it actually implies. The fact is that CLOUDs can reduce the overhead for managing and administering resources through automation and outsourcing, and should reduce the overhead for creating highly available and reliable services. They enable users to easily provision their own services in a fashion that meets the characteristics (availability, elasticity etc.) above.

3. Implicit Challenges and Requirements

Whilst the CLOUD is often considered a technological solution to the issues listed above, it is essentially nothing more but a provisioning and management concept that can be implemented in various fashions, though with key economic differentiators (cf. [SCH10]). Nonetheless, the conditions and concerns listed above lead to a series of technological problems and requirements that have to be addressed by current and future CLOUD environments in order to ensure proper fulfilment of the CLOUD characteristics.

In particular we can note the following key concerns:

- ***Resource Management*** **(including data & network management):** in order to handle the growing amount of resources, users, but also scope of devices, use cases etc. some form of automated resource management is required that reduces the overhead for the administrator to cater for this diversity and scale. This relates not only to the machines (respectively processing units) themselves, but also to the communication network and to handling the data. The amount of data consumed, produced and distributed increases constantly – way faster than the underlying network, thus leading to increasing challenges with respect to replicating and distributing data, localising it, routing to it, compensating for bandwidth and latency limitations etc.
  Resource management must thereby respect multiple positions, i.e. not only does the own infrastructure grow and become more complex in management, but also the connection to the user, the user devices, the usage context etc. all vary increasingly and have to be handled by the providers. As discussed in more detail below, this in turn relates closely to aspects of developing applications in a fashion that they can actually deal and use the divergence and scale. This obviously also relates to protection of data, separation of environments and dealing with multiple tenants (see below).
  We can therefore identify the following main sub-concerns without going into technical details:
    - Efficient handling of an increasing amount of more and more heterogeneous resources
    - Efficient handling of an increasing amount of information and communication
    - Efficient handling & processing of large quantities of data
    - Interoperability and portability between the resources employed

  To achieve this, it is necessary that resources, their usage etc. can be monitored and (dynamically) reconfigured – ideally autonomously.
- ***Scalability of code and data*****, horizontally and ideally vertically:** is necessary to ensure service quality according to the customer / provider requirements, in particular to allow for dynamic elasticity, i.e. distribution and replication of the instances across the environment. This ensures in particular to meet availability requirements. However, replication and distribution must

carefully cater for aspects of multi-tenancy, consistency, concurrency, workload etc., or else performance may degrade significantly.

CLOUD aware applications therefore need to be significantly differently developed and handled, including equally the structure of code, data and their relationship to each other. This should respect the need for availability (horizontal scale) but also for more performance / bigger resource need of one and the same application. Specialised scalable data management systems for CLOUD environments, such as elastic caches, NOSQL databases etc., already exist and scale fairly well vertically, i.e. change size according to need – however, even in this case, fragmentation etc. can lead to considerable excessive resource consumption.

Obviously this does not only require that the application is organised accordingly, but also that an according CLOUD execution engine / model can exploit this structural information. As an implicit concern, interoperability must be guaranteed at least within the scope of deployment and usage.

- ***Fault tolerance and reliability*:** must thereby be guaranteed to compensate the increasing number of problems arising with the scale and scope of the infrastructure and the according deployment. If the pure hardware mean time between failures (MTBF) of a modern server is even as low as a week, distributing an application and its data across 500 nodes decreases the MTBF to an average of three times per hour. Without any additional means to ensure reliability (and thus execution stability), a typical future server provided application would thus fail too frequent for proper usage. This obviously applies to any large data centre, independent of CLOUDs.

  Reliability can thereby not only be achieved through replication of the respective instances, as with a large degree of horizontal (or vertical) scale replication would become extremely costly. It is therefore necessary to provide means that not only identify the relevant execution & data points, but that also allow for fast rollback, loss-less reset etc.
- ***Multi-Tenancy*:** it should thereby not be forgotten that the applications hosted in a CLOUD environment are principally accessible to multiple users at the same time. This is not restricted to multiple applications / services being hosted on the same physical server, but could also involve actual sharing of code and data, or at least of parts of this code and data. This however is hardly properly exploited by current models, as it implies that the code (and data) have been structured accordingly (see also above). Some PaaS models cater for that right away on middleware level but restrict the developer to make use of the according functionalities.

  Multi-tenancy raises multiple concerns that implicitly impact on the quality of the CLOUD systems and in how far the respective characteristics can be fulfilled – in particular we can note that
  - Consistency of shared data becomes highly complicated with the number of concurrent users and the degree of interactivity expected. Location, degree of replication, concurrency of usage etc. thereby all

play a role. So far, most providers only support "eventual" consistency, i.e. where the order of updates is not (necessarily) maintained, or use other trade-offs to handle their specific consistency needs

- The data (or computation, or even just part of data) of an individual tenant needs to be kept secure and private, even though code, resource, data line etc. are all shared. One typical approach at the moment (at least for IaaS and in some cases also for SaaS and PaaS) consists in virtualising the individual instances.
- Not only data and code need to be isolated and shared, depending on accessibility, but also the usage information needs to be isolated from one another, so as to ensure maintenance of the quality parameters offered for the respective user(s), and to allow accurate accounting.

A typical approach to realising at least parts of these aspects consists in virtualising the environment, i.e. employing virtualisation technologies for hosting the service(s), respective image(s) to be exposed. This allows treating each user's environment separately and thus isolating the code and data assigned to this user. Since virtualisation techniques allow for the dynamic management of the respective image, i.e. start, stop, relocation and replication, this provides an easy fashion to handle multiple instances and thus elasticity. It also enables the host to expose the environment in a common fashion, according to the virtualisation technology used.

Virtualisation has become such a standard approach to solving these issues at least at infrastructure (IaaS) level, that there is a frequent confusion between CLOUDs and virtualisation technologies. However, it must be stressed that virtualisation is just one potential approach to solving these issues and may lead to other issues and concerns not meeting the overall objectives, such as unnecessary overhead for small, lightweight applications with large numbers of users (cf. section III below).

4. General Concerns

Due to the nature of CLOUD computing, in particular related to the internet-based service provisioning / outsourcing context, CLOUD systems have to address a series of additional concerns raised mostly by customers and users, but also by the providers themselves. As will be elaborated in more detail in section I.B.2, these are aspects that typically relate strongly to other internet domains:

To these belong first of all the concerns related to *trust, security and privacy*, that are frequently raised in the context of internet based service provisioning and outsourcing. There is a general distrust regarding outsourcing of data – in particular if the handling of data is out of control and the data itself may be unprotected. This makes new security mechanisms relevant, including e.g. homomorphic encryption to ensure that even the data handler cannot misuse the data. Virtualisation is thereby no solution to this problem and in fact can give raise to additional security concerns (such as stealing images).

Related to the trust aspects it must be noted that many of the *legislative* issues in this context are as yet unsolved and as long as these issues are unclear, other means need to be introduced to compensate for this deficiency, such as *data provenance, location control, restricting contracts* etc. Revisions of the legal framework are currently under way, but security and legislation must ideally reach a level where a user can blindly rely on them without requiring advice from a lawyer.

Since CLOUDs are an economic model for resource provisioning and usage, any according technology must cater for means to support *accounting and billing* of the consumed resources. On the other hand, the user must be able to assess the quality of the provided services, uptime of resources, availability etc. in accordance with the contractual agreements, as well as to get insight about the provider's internal processes to a degree, in other words some form of *auditing* must be supported. Ideally, in order to maintain competitiveness, this also allows the user to *compare* different offerings qualitatively so as to make educated choices with respect to provider selection.

However, competitiveness is limited as long as *portability and interoperability* of systems still are insufficiently supported – the current divergence of offerings in particular tends towards locking a user into a provisioning model with little chance to moving his services into another environment. The implicit cost of this *vendor lock-in* (switching costs etc.) still causes many potential users to refrain from using CLOUD systems in the first instance.

In general it can be said that a central concern for all cases consists in usability of the environment. This particularly affects the CLOUD users who will either consume the offered services or who will develop services for a CLOUD infrastructure, respectively even enhance the CLOUD environment themselves. Accordingly, they require *tool support* for integration of the services, respectively a *programming model* that caters for the types of scalability, distribution, composition etc. to generate efficient new services. Most of all, however, the relevant *expertise* as to how to best exploit the CLOUD, which use cases apply, which cost models apply, which QoS can be maintained etc. is generally lacking.

B. Classification of Characteristics

The characteristics elaborated in the previous section can be differentiated according to multiple criteria. Classically, we distinguish between technological, business / economic and social / legal aspects, as these dimensions provide an insight into the actual work to be performed. Table 1 provides an overview over these characteristics and their classification.

Technical	Business / Economic	Social / Legal	Other
Elasticity / Scalability	Outsourcing	Security	Multi-Tenancy
Virtualisation	Pay per use	Provenance	Ease of Use
Agility & Adaptability	Resource utilisation	Privacy	
Availability	Energy efficiency		
Data Management	Cost efficiency		
Reliability	Metering		
Programmability			

TABLE 1: OVERVIEW OVER THE MAIN CHARACTERISTICS AND THEIR CLASSIFICATION

1. Relevance of the Characteristics

As can be easily seen, not all of the listed characteristics have equal relevance for user or providers. In fact, their relevance is highly dependent on the use case in the first instance, e.g. a purely academic research use case will probably put less relevance on security, auditing and related aspects, whereas an in-house developer may put less emphasis on portability and interoperability etc.

a) Intrinsic & Extrinsic Characteristics

Obviously within the context of this report, we must assume that the use case in question has principal CLOUD requirements, i.e. is actually suitable to exploit the specific characteristics. Without elaborating here which use cases belong to this group, we can generally say that the according application must have specific requirements along the primary capabilities of CLOUD systems. According to the preceding section, this means in particular increased availability under dynamic scaling conditions, i.e.

- Availability
- Elasticity
- Resource Utilisation
- Multi-Tenancy

> ! *We denote availability, elasticity, (improved) resource utilisation and support for multiple tenants as **intrinsic** capabilities of CLOUD systems, i.e. any CLOUD provisioning must be able to adhere to them in some form.*

As opposed to that ***extrinsic*** characteristics are typically defined outside the mere CLOUD domain, but inherited by means of conceptual overlap. To these conceptually related domains count utility computing, internet of services etc., as elaborated in more details below. It should be noted that this does not imply that these aspects are not *extended* by the CLOUD system, e.g. by offering additional capabilities. We can therefore distinguish between *extrinsic extended* and *extrinsic inherited* (i.e. unchanged) characteristics. Extrinsic extended characteristics (such as data management and reliability) are implicitly still CLOUD-specific and of considerable relevance, whereas extrinsic inherited (such as virtualisation and energy efficiency) ones are effectively not CLOUD-specific, though they may be relevant enablers for CLOUDs. See Table 2 for a complete overview.

b) Relevance Observations

The extrinsic / intrinsic categorization does not necessarily reflect the characteristics' relevance associated by the provider or user according to his use case. Instead, it only implies whether the system can be considered a CLOUD, or whether it is actually more related to one of the underlying domains. Vice versa, if an application user or provider does consider the intrinsic characteristics as irrelevant for the given use case, he may want to reconsider the choice of a CLOUD infrastructure in the first instance.

With less technical interest, it is also obvious that in particular commercial providers will consider economic aspects as primary concerns.

The average user will primarily perceive clouds as a means to outsource IT on all levels, whilst maintaining a high level of availability and quality of service. They will implicitly (if not even explicitly) consider non-functional aspects (such as elasticity, reliability, availability, performance, privacy (security) and ease of use) of higher importance than technological ones, as the user will not have to cater for the technological problems behind the provisioning. As opposed to that, a provider will typically have higher concerns regarding the technological problems that actually need to be solved in order to offer the non-functional capabilities to their customers.

2. Relationship to other domains

CLOUDs extend the capabilities of other domains with the specific goal to achieve scalability/ elasticity, availability with optimal resource utilisation, which is as such only partially addressed in other domains. Specifically, CLOUDs belong to the wider areas of Internet of Services (including Web Services, Web3.0, SOA etc.) and Utility Computing (including Grid, Virtual Organisations etc.) and implicitly inherit multiple aspects from these domains, such as virtualisation and outsourcing. Depending on usage, they may extend these characteristics, in particular by adding a new business model.

What is important to stress again in this context is that not all CLOUD characteristics exclusively belong to the CLOUD domain – extrinsic features that generally belong to other domains enable service and utility computing are naturally taken over in CLOUDs.

Specifically, we can distinguish between (cf. Figure 1)

- Characteristics exclusive to the CLOUD (intrinsic features)
- Characteristics belonging to other domains but having to be adapted in order to meet the CLOUD relevant specifics (extrinsic extended)
- Characteristics that belong to other domains and just act as enablers to CLOUD systems, i.e. do not have to be extended (extrinsic inherited)

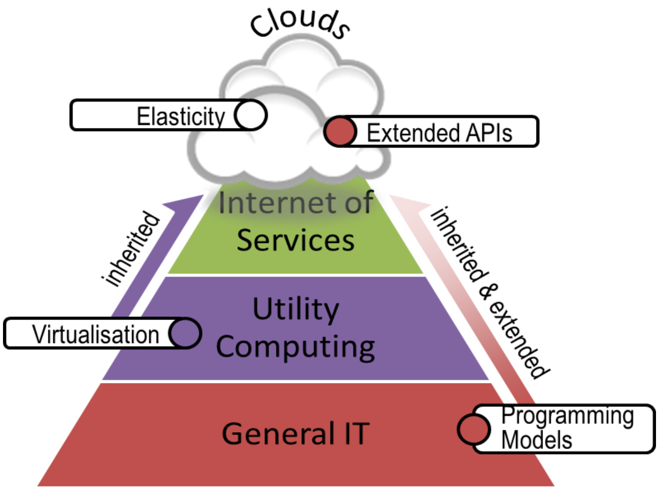

FIGURE 1: INHERITANCE AND EXTENSION OF CHARACTERISTICS ACROSS THE RELATED DOMAINS

! *Any research or development work must be aware of this distinction in order to not reproduce work, as this would not only duplicate effort, but increases the risk of deviation and hence interoperability and usability issues. Instead, there is a high potential to uptake and extend existing work, thus improving quality, portability etc. Any work should therefore first assess the relationship of the addressed aspects with respect to other domains.*

We refer here in particular to three related domains, though the list can be easily extended and refined:

- Internet of Service: covering areas such as Web Services, Web 3.0, Service-oriented architectures etc. In other words, individual service provisioning without specific means for dealing with availability, i.e. just network load balancing
- Utility Computing: including Grids, Virtual Organisations, and also HPC to some degree, as they were originally conceived and realised.
- General IT or more correctly "non-web" IT, including all computer science aspects concerning isolated machines, i.e. without making explicit use of the web. This includes theory of computation, hardware architectures, operating systems etc.

More specifically, the characteristics identified above can be classified with respect to the domains they were originally conceived in and to the ones which take up / extend them (see Table 2).

Columns thereby denote the domain to which the concepts apply, so that any entry within an according column implies that the respective characteristic is specifically adjusted / extended to meet the domain's requirements. In other words, single-column entries imply that all domains to the left inherit the characteristics without signification adaptations, whilst domains to the right do not support this characteristic. If an entry spans multiple columns, it means that the base concept is conceived in the right-most domain and that all domains to the left adapt the concept to their respective needs.

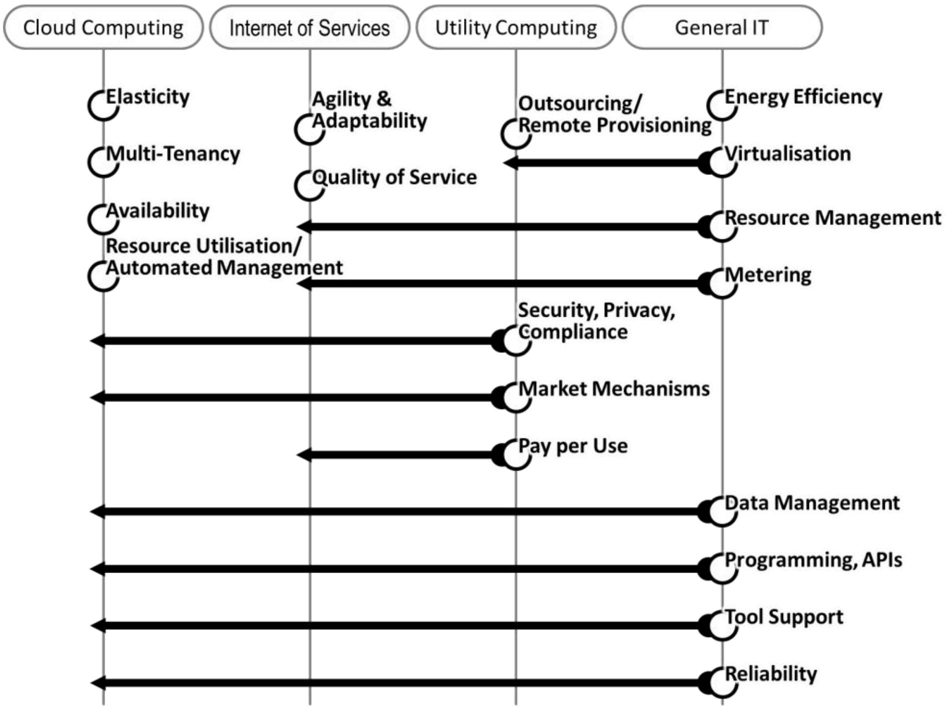

TABLE 2: OVERVIEW OVER THE CLOUD CHARACTERISTICS AND THEIR RELATIONSHIP TO OTHER DOMAINS. INHERITANCE FROM RIGHT TO LEFT. THE FURTHER LEFT A CHARACTERISTIC ADVANCES, THE MORE IT NEEDS TO BE ADAPTED FOR THE RESPECTIVE DOMAIN(S)

It will be noted that this classification is subject to many discussions due to the overlap in terminologies across domains. The terminology here bases on the definitions provided in section I.A. It is worth mentioning in this context that the classification is oriented towards the left, i.e. "which domain provides capabilities that can be exploited on higher-level domains", but there is a noticeable right-orientation, too, where for example utility computing is improved by availability methodologies of the CLOUD, without necessarily turning the respective domain into a CLOUD system – this however exceeds the scope of this report and is only of secondary relevance here.

C. Definition: What is a "CLOUD"

If CLOUDs are to be developed further in the future, it is essential to extract the core characteristics of CLOUD systems, so as to make sure that a common understanding of the work to be performed can be achieved. It can be noted thereby that various definitions exist that partially contradict each other.

It can be noted that most of these definitions boil down to describing CLOUD computing as another means of offering services / resources / servers over the internet[4]. Gartner at least extends this with the aspects of scalability / elasticity and

[4] For example:

serving multiple concurrent users (which does not necessarily imply multi-tenancy)[5]. Whilst these definitions reflect the current usage behaviour, they do not manage to distinguish CLOUDs from other areas of internet-based service provisioning.

The most popular definition is probably the one provided by the National Institute of Standards and Technology (NIST) in 2009, respectively its updated version in 2011. According to this definition

> “ CLOUD computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This CLOUD model is composed of five essential characteristics, three service models (Software / Platform / Infrastructure as a Service), and four deployment models, whereas the five characteristics are: on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service. The deployment models include private, community, public and hybrid CLOUD. [NIS11]

This definition reflects very well the current understanding of CLOUD computing, in particular as long the technological and economic challenges, such as an appropriate model for cost calculation, metering etc. are still not properly addressed. In other words, current definitions reflect the status, but neither the intention behind CLOUDs, nor the direction into which they will (or should) develop. Therefore these definitions cannot serve the purpose of steering research and development and new offerings will constantly readjust our understanding of CLOUD computing.

> “ We’ve redefined CLOUD computing to include everything that we already do [Larry Ellison][6]

1. Towards a Long-Term Definition

In order to represent the core values in the definition and to serve as a principle steering models that implicitly allows developers, users and providers to identify which technologies and aspects fall into the CLOUD domain, and which do not, the “Future of CLOUD Computing” report promoted a basic definition:

PCMag: CLOUD computing refers to services offered on the public Internet or to a private network that uses the same protocols. [http://www.pcmag.com/encyclopedia_term/0,2542,t=CLOUD&i=39847,00.asp]
MacMillan dictionary: a type of computing in which computing resources are shared via the Internet, rather than the use of local servers or personal computing devices.
[http://www.macmillandictionary.com/buzzword/entries/CLOUD-computing.html]
Oxford dictionary: the practice of using a network of remote servers hosted on the Internet to store, manage, and process data, rather than a local server or a personal computer.
[http://oxforddictionaries.com/definition/CLOUD+computing]

[5] Gartner defines CLOUD computing as a style of computing in which scalable and elastic IT-enabled capabilities are delivered as a service to external customers using Internet technologies.
[http://www.gartner.com/it/page.jsp?id=1035013]

[6] http://news.cnet.com/8301-13953_3-10052188-80.html

> “ A 'CLOUD' is an elastic execution environment of resources involving multiple stakeholders and providing a metered service at multiple granularities for a specified level of quality (of service) [SCH10]

As it turned out, however, this definition was considered too restrictive and too imprecise to actually be of value for the intention of specifying and steering the development of CLOUDs. One major aspect was particularly neglected, namely that different stakeholders have different perspectives and hence understandings, goals and intentions with CLOUDs. The following sections try to provide a definition of what we will understand as “CLOUDs” in the future considering these different roles:

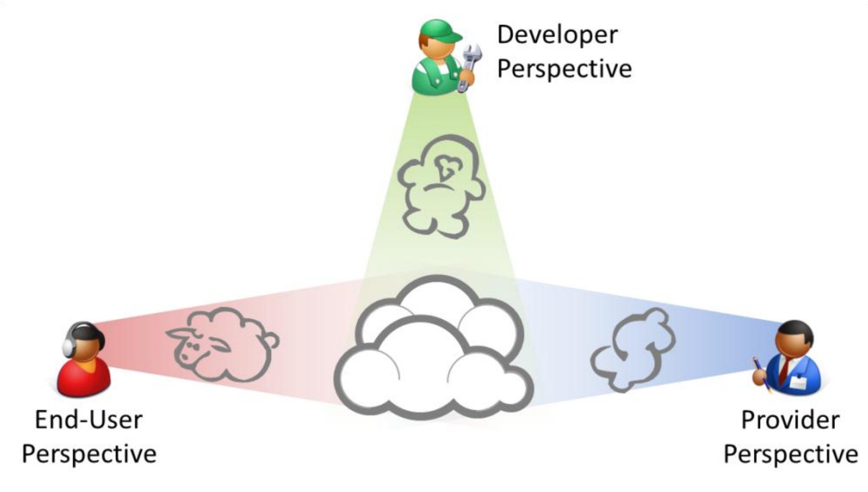

FIGURE 2: SCHEMATIC OF THREE DIFFERENT PERSPECTIVES ON CLOUDS

a) The (Non-Technical) User Perspective

The core stakeholder in CLOUD computing is obviously the customer or end-user who makes actual use of the resources / services offered. As CLOUDs are economy driven, the main incentives for their uptake (and thus provisioning) are cost aspects, expressed through a wide range of factors, such as outsourcing of resources and management, higher availability and thus better service provisioning etc. It does not matter thereby whether the user is the CLOUD provider (or developer) himself, i.e. whether the CLOUD is actually privately owned and used.

From the perspective of this stakeholder, the CLOUD should extend the general internet-based service provisioning model with aspects of high availability, reduced cost (through improved resource usage) and ease-of-use. The typical end-user is thereby not interested in the technical details that enable this behaviour and is arguable to what degree this is restricted to CLOUDs, or whether not future other models will achieve the same behaviour through different approaches which will not be considered “CLOUD” anymore.

! *__CLOUDs__ are environments which provide resources and services to the user in a highly available and quality-assured fashion, thereby keeping the total cost for usage and administration minimal and adjusted to the actual level of consumption[7]. The resources and services should be accessible for a principally unlimited number of customers from different locations and with different devices with minimal effort and minimal impact on quality. The environment should thereby adhere to security and privacy regulations of the end-user, in so far as they can be met by the internet of services.*

b) THE PROVIDER PERSPECTIVE

In our context, a provider hosts the resources that the CLOUD offering (P/S/IaaS) is executed on. This is not necessarily true for e.g. resellers or service providers that outsource the resources, yet the actual CLOUD provider is responsible for managing and providing the resources in a way that economic incentives are met.

From this perspective, the actual resources involved in provisioning become a major issue. The applications, services and framework should aim at making best use of the hosting resources, so as to reduce administrative costs, yet still meet the requirements of the consumer in terms of availability, reliability etc. To meet these goals the environment must be highly dynamic and adaptive, in particular allowing for scale-out and -in on the fly.

! *__CLOUDs__ are dynamic (resource) environment that guarantee availability, reliability and related quality aspects through automated, elastic management of the hosted services – the services can thereby consist in a platform, a service, or the infrastructure itself (P/S/IaaS). The automated management thereby aims at optimising the overall resource utilisation whilst maintaining the quality constraints.*

c) THE DEVELOPER PERSPECTIVE

In order to realise the capacities for both the customer and the provider, the developers (and researchers) have to incorporate according features in their services, platforms and frameworks. Accordingly, the developer in focus here is not only the user of a (CLOUD) platform as a service, but also the developers that aim at turning a resource infrastructure into a CLOUD, respectively that want to adapt existing applications to exploit CLOUD features.

For this type of stakeholders, it is of particular interest, what a CLOUD environment can offer (technically), respectively what the developer needs to respect in his code in order to address CLOUD capabilities. CLOUDs, due to their nature, do not prescribe a specific technical approach to solving these issues, but many technical issues arise implicitly from trying to address them, such as adaptability from the fact that users and devices differ strongly, elasticity from the expected degree of availability, means for failure compensation due to the large scale and heterogeneous usage etc.

[7] Note that this does not necessarily imply optimal nor cheap.

! *__CLOUDs__ are environments which expose services, platforms or resources in a manner that multiple users can use them from different locations and with different devices at the same time without affecting the quality aspects of the offered capabilities (service, platform, resource) - this means in particular availability, reliability and cost-effectiveness. This is realised through automated, elastic management of the services and their environment.*

2. SCOPING CLOUDS

In light of the above discussions and characteristics identified, we can therefore attempt a minimal definition of a CLOUD environment, i.e. the conditions a system has to fulfil in order to rightfully claim being a “CLOUD”:

! *An environment can be called “CLOUDified”, if it enables a large dynamic number of users to access and share the same resource types, respectively service, whereby maintaining resource utilisation and costs by dynamically reacting to changes in environmental conditions, such as load, number of users, size of data etc.*

It must be noted thereby that many additional criteria exist that are generally associated with CLOUDs, but that are not essential in themselves, such as “pay per usage” which for example in a private environment is just a load balancing criteria, or “outsourcing” which again is not effectively true for in-house, private CLOUDs.

II. CURRENT LANDSCAPE

Since publication of the “Future of CLOUD Computing” report in 2010 [SCH10], many new providers and users of CLOUD systems have arisen, thanks to the progress made in research and development both in industry and academia. With respect to this report’s objectives, the primary interest rests on the potential chances and usage for CLOUD computing, i.e. whether and how the “CLOUD” will prevail as an advanced, long-term means for extended utility computing or whether it will not be able to satisfy the according needs.

This section examines which use cases have taken up on CLOUD computing and how mature the current technological development is considering the main characteristics relevant for the respective use cases. It will not repeat an in-depth evaluation of the existing platforms.

A. UPTAKE

It is noticeable that whilst a great number of new providers have arisen, comparatively few new (classes of) users have started to exploit the *full* potentials of the CLOUD. Though this can be interpreted as a lack of interest in CLOUD provisioning, the point is closer to a lack of trust in light of the switching costs, but with a high general interest in exploiting the potentials. The technology is still considered immature despite the progress made and with little expertise existing about the actual benefits for different use cases, most users belong to the group of individual end-users, rather than to organisational industry or enterprises use. On the other hand, the rise of new providers also shows that multiple enterprises and industries (started to or already) employ CLOUD-like concepts in their infrastructure and partially start to opening these features up to a wider customer base. It is this interface between private CLOUDs (hosted within the organisation) and public CLOUDs available for periodic high-resource requirements (without requiring according internal resources) that exposes the research opportunities.

1. USAGE LANDSCAPE

We can note in particular the following business cases currently available on the market that exploit CLOUD systems:

a) SOFTWARE AS A SERVICE DOMAIN

Most uptake is definitely notable in the software as a service domain. This is mainly due to the fact that this domain has always dealt with similar concerns (availability, reachability) and thus generated the first CLOUD concepts. We can note that over the last years, next to the classical web type services, such as office related services (such as project management, customer-relationship management, word processing etc.), tourism related ones (travel booking, translation etc.), auctioning and public administration (document preservation, registries), also some extended services have arisen that put forward more complex demands towards the running system, such as real-time (streaming, media), aggregation (picture search) and communication control (telecommunication).

b) Platform as a Service Domain

Next to SaaS, Platform as a Service is a growing domain for CLOUD systems, essentially enabling the user to generate his own applications / programs / services with CLOUD enabled capabilities. Next to website hosting and collaboration environments, there is a growing tendency for specialised platforms for providing domain specific support, for example in eScience or eEngineering. There is a potential community of ~50.000 independent software vendors in the EU which require new ways to offer solutions and services.

c) Infrastructure as a Service Domain

Most well-known for CLOUD data storage, Infrastructure as a Service provisioning has the least scope of domains, though probably most usage potential. Accordingly, the typical usage domains are still web storage, resource outsourcing and, with growing interest, resource extensions through CLOUD bursting.

d) Other Domain(s)

Next to (direct or indirect) service provisioning, another growing business case related to CLOUDs consists in all areas related to offering expertise, or to reusing knowledge in other domains. This covers in particular consultation, enhanced utility computing and touches upon aspects such as internet of things.

2. Technology Usage

Non-regarding the growth of interest in CLOUDs, there are still very few technology platforms that actually find wider uptake for realising the above mentioned domains. Obviously, some CLOUD systems find wider usage, such as OpenNebula, OpenStack, Amazon and Google, but in general most CLOUD offerings are still realised through proprietary solutions. It is therefore worth noting that current middlewares seem as yet not to have reached the level of usability making new CLOUD systems easy to create.

a) OpenStack and OpenNebula

In the context of European development and open-source development, eyes will invariably fall onto two of these systems / technologies: OpenNebula and OpenStack.

OpenStack[8] is a multi-vendor consortium coordinating several open-source software projects to deploy AWS-like infrastructure, mostly to serve the needs of service providers. The project was started by NASA and Rackspace, almost two years ago, and now counts with the support of other vendors like HP, Dell and Cisco. Currently more than 150 companies have joined the project among which are AMD, Intel,

[8] http://openstack.org/

Canonical, SUSE Linux, Red Hat, Cisco, Dell, and HP. It is free open source software released under the terms of the Apache License.

OpenNebula[9] on the other hand was initiated in the context of EU-funded projects in CLOUD computing. It is an open-source project that started five years ago to develop the industry standard solution for building and managing virtualized data centers and private CLOUD infrastructures. OpenNebula is a very active open source project with a steadily growing community and very large user base, including telecom companies like RIM, system integrators like Logica, supercomputing centers like SARA, or research centers like FermiLab. The project also has collaborations in CLOUD computing innovation and interoperability with big vendors like Microsoft.

OpenNebula provides fully open-source, interoperable software solutions and is being used as an open platform for innovation and interoperability in leading research and infrastructure projects, and as a reference implementation of CLOUD standard specifications.

Both, OpenNebula and OpenStack, deliver fully open-source software to build IaaS CLOUDs, released under Apache license, and developed with an open and transparent process over the Internet. While OpenStack mainly focuses on AWS-like public CLOUD features, OpenNebula offers a comprehensive solution for data center virtualization management, enabling the users to easily build their own private CLOUDs.

OpenNebula is delivered as a single integrated package comprising key functionalities for CLOUD computing, whereas OpenStack delivers a set of products for individual functionalities and capabilities. OpenNebula is committed to implementing major de-facto and de-jure standards, such as Amazon APIs, or the specifications by OGF, DMTF and SNIA. OpenStack builds loosely up on AWS, but primarily incorporates its own standardisation working group, trying to incorporate the requirements from the participating companies.

B. Maturity / Advances Made

Many of the so-called CLOUD offerings only realise partial aspects of the characteristics detailed in section I.A. Often enough, infrastructures are sold as CLOUD for pure marketing reasons, without addressing the essential characteristics. In most cases, the infrastructure only offers some characteristics of CLOUDs, respectively only parts are exploited.

Not employing *all* characteristics does not necessarily imply that the respective system cannot be considered a CLOUD though. As detailed above, we must in this context distinguish between primary (or intrinsic) and secondary (or extrinsic) characteristics – whereas the primary ones are essential for defining an

[9] http://www.opennebula.org/

infrastructure as a CLOUD. Looking at Table 2, it becomes obvious that most extrinsic criteria contribute little directly to forming a CLOUD, but implicitly to forming other domains. Exceptions are obviously those characteristics that are adapted for specific CLOUD requirements, such as a CLOUD specific API. However, such adaptations are only sensible if the underlying system does provide the intrinsic characteristics – otherwise, the extended programming model would either not be applicable or implicitly address a lower-level domain.

It can also be noted from the list of characteristics above, that only few (technological) characteristics are exclusive to the CLOUD – accordingly, modern CLOUD technologies can (and do) exploit a great amount of technologies and concepts developed in other domains. Not surprisingly, it can therefore be generally stated that many technologies in CLOUD computing have reached an overall high maturity over the last few years.

However, often the specific technological advance addresses either isolated use cases or isolated characteristics – in particular available platforms often exhibit great maturity in one specific aspect, but lag behind in other characteristics. Implicitly, the full applicability and maturity of the CLOUD systems depend highly on the use case and the provisioning type intended.

III. POTENTIALS & MISSED OPPORTUNITIES

Despite the maturity, uptake and technological advances made, it is nonetheless notable that CLOUD systems still fall far behind their potential. There are multiple reasons for this deficiency, partially arising from the wrong expectations from marketing – in particular where terminology may easily be misinterpreted in different domains (such as e.g. “high performance”). Further to this, the CLOUD inherits deficiencies from its related internet domains (cf. section I.B.2), so that it is generally unfair to judge it by standards that are not even fulfilled on lower levels – nonetheless CLOUD providers promise high reliability and availability, which still poses major issues in IT. Thus CLOUD providers often find themselves in the cross fire of media attention, due to the large amount of users affected (such as e.g. the Amazon CLOUD outage[10] in 2011 or the Google GMail reset[11] etc.), even though these failures are technologically hardly more drastic than any other web outage.

A. DEFICIENCIES

There is therefore an increased expectation towards CLOUD offerings, systems and technologies that currently cannot be fulfilled to entire satisfaction. This does not only mean that the expectations exceed the technical feasibility - more importantly it implies that the concepts behind CLOUDs invite for more complex scenarios than can currently be fulfilled. Along that way, it implies that a current interest and need from the business side exists. If CLOUDs are to prevail in future developments, it needs to be examined, how well the concepts (and technologies) can actually meet the expectations and fulfil the according business needs.

> ! *A sustainable CLOUD model is only viable if it meets the long-term requirements from the surrounding ecosystem.*

1. DEVELOPMENT OF THE ECOSYSTEM

First of all it must be noted that the Ecosystem into which CLOUDs are embedded, i.e. their environment both in terms of the IT infrastructure, as well as in terms of the surrounding business models and use cases.

A primary, on-going change is thereby the notable **increase in scale and heterogeneity** of this ecosystem. Not only in terms of the growing numbers of users, but also in terms of use cases, programming models, operating systems, devices, and capabilities of these devices. Whatever service or infrastructure offered, the environment must be able to handle a massive amount of requests from different locations with different connectivity and from different resource types.

[10] http://www.businessinsider.com/amazon-outage-enters-its-second-day-lots-of-sites-still-down-2011-4
[11] http://mashable.com/2011/03/01/google-sorry-gmail-reset/

The growing amount of requests necessitates a large amount of servers / resources, which have to be administrated with the least amount of effort. The common current approach consists in setting up highly homogeneous data centres that can be managed uniformly (i.e. all servers with the same configuration, thus allowing uniform management / administration). This approach is however costly, as it requires that all resources are equally updated, respectively that a sufficient amount of backup devices exists to compensate for hardware failures. Redundancy (and thus reliability) is thereby achieved at the cost of hardware expenditure, rather than through according (currently missing) software management mechanisms. Also, it does not allow exploiting different resource capabilities for different requirements – at least not beyond the range of capabilities supported by the respective server platforms.

At the same time, the growth in usage and in computer capabilities leads to an increasing amount of data being handled, and communicated over the web, whereas however all communication is physically limited in terms of latency and bandwidth – this can only be compensated to a certain degree by the software stack. For example, modern multi-core processors already suffer from the so-called memory wall [MOO08] and the same affect will become more and more notable across the whole network. Similarly, the sheer volume of information exceeds capabilities of current processing approaches, making data clumsy and unmanageable.

The degree of mobility required and expected by the users is also increasing vastly. The internet crosses all countries and borders, and so do its users. More and more people are on the move, taking their working environment, but also their social network, as part of their computing devices and the according internet connectivity with them everywhere. That means that their devices have to cater for the performance requirements with minimal energy consumption, leading to remote computing models, which in turn however increase communication again.

Business processes behave similar and just like the users, modern business interactions more and more cross the boundaries of enterprises, countries and thus, implicitly of regional jurisdictions and legislations. More and more tasks are off-loaded into other countries and thus partially leave the direct control of the supervisor.

> ! *CLOUDs offer a potential to deal with the increasing demand for availability, diversity and mobility implicit to the current growth of the internet by providing means to dynamically react to the changes in requirements, usage, location etc.*

However, current CLOUD systems still fall short in many of these respects:

2. Technological Challenges

Given the current adoption of CLOUDs and IT in general, it may come as a surprise that not all deficiencies have been properly addressed - frequently complex tricks need to be exploited to give the impression of fulfilling any of these aspects, and often enough at the expense of any other criteria. And the pressure of the environment will only increase. Within this chapter we list therefore the main technological concerns and challenges:

a) Scalability

Scalability, or in the context of CLOUDs more precisely: elasticity, even though being one of the key characteristics of CLOUDs, is still a major concern on all levels: code (application logic), data (application and storage), and communication (network). In the CLOUD context, elasticity is generally constrained to horizontal scale, i.e. the replication of instances with the number of requests (respectively similar quality concern) – only with storage, vertical scalability is supported at least to a certain degree, namely the increase / decrease of the available / used storage space.

However, efficient horizontal scalability is not achieved just by replicating the respective code and data instances – in fact in most real use cases, the instances need to share at least some data and ideally should not replicate computations that have higher performance than communication requirements. What is more, spawning (scaling) of objects – no matter whether for the purpose of horizontal or vertical scale – is thereby still slow in modern CLOUD environments and therefore also suboptimal, as it has to take a degree of lag (and hence variance) into account.

Just like in parallel computing, the actual application logic has to be carefully assessed with respect to its internal dependencies and (implicit) requirements towards the middleware, including communication, consistency management, scheduling etc. Vertical scale (parallelism) is thereby hardly addressed in any CLOUD infrastructure, thus reducing its performance significantly, given current processor manufacturing trends [SUT12].

Software engineering and programming models therefore need to be strongly re-assessed in these domains, necessitating more joint expertise from the domains in parallel and distributed computing.

Non-regarding the growth of IT infrastructures (in terms of numbers of processors, connectivity etc.), there are still hard limitations, which to overcome requires a high degree of intelligent program structuring / developing, adaptability etc. Beyond communication limitations, we can also list aspects such as lacking quality of service control on network level, limitations of storage, consistency management etc.

b) Resource Utilisation

As also implied by the limitations of scalability / elasticity above, resources are not utilised as optimally as they could, which implies still that more resources are consumed than necessary. Even though CLOUD systems already reduce the number of resources consumed, they could fare considerably better. This does not only reduce the scope of utilisation but also increases the costs unnecessarily. Notably, higher costs are principally acceptable, as long as the benefit (flexibility, agility etc.) outweighs this cost, respectively as long as the cost for further improvement would be higher than the benefit. This means that the effective utilisation of resources must be simple enough for provider, developer and user.

However, there is generally insufficient experience and expertise about the relationship between pricing, effort and benefit: most users cannot assess the impact of moving to the CLOUD, and frequently enough that information is even missing for internal hosting in the first instance. Not only are there no appropriate cost models as yet, but also the benefit for a customer cannot easily be estimated, as it depends very much on the use case and the application, and hence the expected switching cost, dynamicity in usage etc.

In addition to this, many current technologies, such as most virtualisation approaches, create additional overhead that cannot be sustained with the growing scale anymore.

In order to improve resource utilisation, it is therefore necessary to develop more efficiently scalable technologies, provide effective means to structure code and data so as to exploit the CLOUD characteristics properly, better business models, but also the necessary support to monitor, supervise and control instances. This applies equally to resources and service instances.

Many of these aspects require that the CLOUD system provides capabilities to the user / developer that are currently not fully supported. In particular, it requires that the instances and resources in the CLOUD *can* be controlled in terms of their location, relation (communication linkage and dependency), instantiation (and destruction), replication, consistency etc.

In order to overcome the data & communication problems, this will also require dynamic movement “with” the customer to ensure best communication linkage etc.

c) MULTI-TENANCY ISSUES

The impact of multi-tenancy is easily underestimated, but can raise major technological challenges. Whilst maintenance of consistency across multiple tenants is an obvious concern, isolation creates actually more difficulties. Depending on what is shared and to what degree, it is currently difficult to impossible to distinguish which part of the resource consumption is caused by which user - for example shared communication over a given network line. This however is necessary for accurate usage / cost assessment per individual user.

What is more, this assessment of the individual user in relation to usage, resource access etc. is highly relevant if a per-user quality of service is supposed to be maintained. If the QoS is only observable per resource instance, instead of per user, some users will not get the quality they subscribed to. Furthermore, the quality maintenance measurements, and even just the usage may create interferences, across users, so that regulated environments with a theoretically even quality distribution across all tenants can lead to irregularities [KOH07][PU10]. This obviously makes optimisation accordingly difficult, too. QoS parameters thereby do not necessarily apply equally to all users – ideally users will want to be able to specify their own, application / usage specific parameters that the CLOUD caters for.

Similarly, not only the performance / quality needs to be isolated per user, but also their environment and data. Sharing the same resource can principally imply that each user can access the data and code of the other users – whilst in some cases this may be a design purpose, many use cases will actually require private and secure data / code spaces. This is solved e.g. with virtualised environments, but this approach is not always applicable. Security concerns are therefore one further issue arising from multi-tenancy.

d) Lock-In

Due to the lacking generality in the approaches to realise CLOUDs, there is implicitly a strong divergence between the CLOUD offerings and thus the interfaces they expose and the way they are programmed / controlled. Few CLOUDs allow execution of something other than their own images / services / platform APIs, even though there exist standardisation efforts, such as the Open Virtualization Format[12] for virtual images, or the efforts by OpenStack[13] to agree on standard interfaces (cf. section II.A.2). Standardisation only addresses part of the problem in the first instance, if functionalities and capabilities diverge (portability issues).

Applications developed for one CLOUD often enough have to be redeveloped for other CLOUD providers, if the execution environment is supposed to be shifted. Due to this switching cost, the average user is quickly locked-in into the environment that he started to host his services on. Portability therefore needs to be improved to increase competitiveness. Stacked models, plugins and similar approaches stall this problem but cannot solve it in the long run, as they would require constant updating.

Opening the market and offering new mechanisms enabling a broader participation can similarly to open source strategies reduce the risks of lock-in, by encouraging broad uptake of de-facto standards. This however, too, relies on minimal divergence between offerings (and markets)

This divergence, on the other hand, is often justified, as they reflect specialisation of the respective CLOUD environment, and allow for dedicated innovation according to different speeds of progress in the respective domains. Specialisation can ensure that users in the respective domain get the best services for their respective use case. Not all domains are appropriately supported as yet, though – many of the usage areas that would benefit greatly from CLOUD systems are not yet sufficiently covered.

[12] http://www.dmtf.org/standards/ovf
[13] http://www.openstack.org/

Interoperability and portability across CLOUD systems is however highly complicated and as the web service history as shown, it is dubious whether pure standardisation is sufficient to address this problem. Not only interfaces are affected with the interoperability problem, but also the code itself and its according data structure. For full exploitation of the services, it is not only necessary that the interfaces can be invoked correctly, but also that they can be interpreted unambiguously. This is a well-documented concern, already raised in the Grid and Web Service domains, even if the economic model behind these domains differ (see e.g. [NGG06]).

e) MOVING TO THE CLOUD

Switching cost does not only arise from the portability / interoperability restrictions of the different CLOUD environments, but also simply from the fact that most applications are not CLOUD ready. Not only are the applications not programmed in a fashion that they can exploit the CLOUD characteristics, but they pose implicit requirements not sufficiently addressed by current CLOUD APIs and programming models. It must thereby always be kept in mind that not all use cases are equally fit for CLOUD exploitation – even though this is a common current misconception given by the generic marketing of CLOUDs (cf. section I).

Most applications exhibit properties of multiple modalities of scale (horizontal *and* vertical) which are however difficult to identify, let alone to exploit. Generally, this requires reorganisation of the code, for which a deeper understanding of the algorithmic structure is important. The methods of restructuring need to be improved and their impact in such new domains as CLOUDs better understood (see e.g. [BRO10]). Distributed applications behave differently from sequential ones and raise additional concerns for which programming model, compiler or middleware have to cater for, such as consistency management, message routing etc. The architectural choices of the infrastructure thereby influence immensely what kind of qualities can be expected for the different applications – a relationship that is not yet fully understood though.

In general there is a lack of support for porting applications (source code) with respect to all aspects involved in the process:

- Understanding of the relevant / appropriate use cases and their requirements towards the infrastructure
- Which algorithmic cores have to be restructured how (software architecture)
- How to program for the CLOUD, respectively appropriate distributed programming models
- Interoperability between code and data, respectively means to support the conversion

Notably, moving *to* the CLOUD should not hinder the user / developer to move *out* of the CLOUD again and back onto own resources.

f) DATA CHALLENGES : STREAMING, MULTIMEDIA, BIG DATA

CLOUDs, following the web service development, still have a very classical client-server like organisation, i.e. the services and applications on the CLOUD are invoked for a specific request, which is processed remotely and the result returned to the requesting agent. A growing amount of use cases requires interactivity with the service, even with multiple users at the same time and potentially involves a large degree of different multimedia streams, including e.g. voice, video chat, live virtualisation etc.

Data and communication in general is still one of the main hurdles in the internet though, as the communication demand grows faster than the technological support (see also above). With the latency inherent to the internet and the dynamicity inherent to the CLOUD, meeting the real-time requirements of both interactive and streaming applications with a constant quality is a highly complex task. This must thereby consider all aspects of future ecosystem development, ranging from heterogeneity of resources over number of users to mobility of the devices.

Related to this, the applications need to be enabled to deal with a different representation of data, that is not associated with a specific location anymore and that may not even be constant in the first instance. As the data gets fragmented and dispersed over the network, some support is needed to ensure that relevant data is available when needed and that it is maintained appropriately without complicated programming. This is strongly related to the problem of big data management, where data exceeds the manageability of software and system [LON11][ECO10]

g) PERFORMANCE

One implicit factor related to most of the deficiencies listed above, consists in the performance of the system, respectively of the software / services running on top of it. Due to their high amount of available resources, CLOUDs are often compared to high performance computing systems, i.e. it is assumed that they offer high performance. The effective performance however depends highly on the degree of scalability (a), the utilisation of resources (b) and the communication strength (f). Accordingly, it depends on how the software was written, how the infrastructure is set up and how it is maintained.

The necessity for high data throughput constantly increases, not only due to the amount of users, but also to satisfy the needs of data / computation intense use cases. Even though the number of cores in modern day processors increases, too, the resource need of some application will exceed the capabilities of such processors. Future CLOUD systems must be able to support a wide range of different devices and use cases (cf. section IV) and therefore be able to understand the various strengths and capabilities and relate them to the actual requirements.

h) OTHER

In addition to the research challenges listed above, a large number of additional issues can be listed, most prominently the security and privacy concerns, covering the full range from legal aspects over regulatory issue down to the organisational and technical level. Data Processing Agreements even beyond Europe are already in discussion though[14] and their technological impact will have to be assessed over the next years.

Many of these issues apply generically across the full range of Internet of Services & Things, due to the nature of CLOUDs (see also section I.B.2).

3. ECONOMIC OPPORTUNITIES & EXPERTISE GAPS

It has been noted multiple times in the preceding section that relevant expertise for supporting CLOUD uptake is needed in various contexts. This lack of knowledge hinders uptake to the extent that would be possible, if new users could build up on an existing pool of expertise and in particular experience from a longer time of CLOUD employment. Such a period of knowledge gathering would also have given rise to new cost and business concepts to effectively deal with the CLOUD and thus help new uptakers to assess the value / benefit of the CLOUD for their purposes.

That such knowledge and experience is missing is not only due to the fact that CLOUDs are a comparatively new market phenomenon, but also due to missing environments to conduct experiments on, and to missing knowledge about appropriate testing parameters for different use cases. Platforms for testing, in particular when in-house, are typically way below the scale necessary for executing experiments under realistic conditions and out-house testing can quickly exceed the dedicated budget for testing purposes. What is more, in order to execute appropriate tests, the usage behaviour to be expected needs to be parameterised and simulated by the respective environment to provide realistic results.

In general there is still very little experience about what behaviour to expect from platforms and usage of such scale and dynamicity, and implicitly also little is known about the impact from such environments onto the code, the quality etc. Modern providers such as Amazon therefore tend to run their environment first in a test phase during which they try to assess the effective quality of service they can maintain and offer. The user behaviour is thereby not only generally a variable, but mostly unknown – even though some observations from non-CLOUDified usage could principally be taken over to the CLOUD environment.

At setup phase it is therefore difficult to judge what requirements towards the environment and its applications will actually be posed and thus whether e.g. the QoS management capabilities will not be exceeded at run-time.

[14] See e.g. http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm

Without the knowledge of impact, requirements and user tendencies, it is however also very difficult to assess the actual costs involved in moving to the CLOUD (“switching cost”) and to assess the potential benefits from this move, thus allowing effective economical comparison between CLOUD and classical environments. Individual calculators per provider do not consider the full aspects involved in migration, let alone allow for comprehensive comparison between providers - some form of mapping between the regular application and usage behaviour characteristics to CLOUD based ones are required to perform this calculation. It also requires that user patterns and CLOUD characteristics can be categorized and reproduced.

Given enough fine-granulated information, this also allows economical comparison between different CLOUD offerings in terms of the intended application use cases. Next to behaviour characteristics, such comparison also needs to be able to assess and represent the value of supporting features, such as degree of isolation supported, auditing details etc. Again, however, little knowledge exists how to actually assess and compare the value of these characteristics.

More generally formulated, it is of high relevance to identify

a) Which use cases are most suited for (which) CLOUD and under what circumstances and
b) The best practice recommendations for moving to the CLOUD.

Non regarding the high interest in CLOUDs, there is still a high fear of CLOUDs as a cost / risk factor: CIOs are still sceptical because they have to manage the implicit risks, and IT departments fear for their jobs. The risks are thereby not only associated with costs and effort, but also arise from additional aspects, such as legalistic concerns, business and cost models etc. Much of the current distrust in CLOUDs arises e.g. from the fact that users currently have very little insight into where their data is actually hosted. This is a major obstacle to uptake of CLOUD systems in particular in the commercial environment. As opposed to this, academic and private users may not necessarily care about the actual data location, as long as the expected quality of service is maintained and cost is low. This group of users therefore forms one of the major initial uptakers. Small enterprises may form a major secondary line of uptakers, where some conditions are relaxed for the benefit of reducing infrastructure costs.

The biggest problem thereby consists however still in the fact that no appropriate business model exists as yet that satisfies all roles and long-term requirements. Not only is it unclear how much payment to ask for which feature / service, it is also unclear how providers, first of all telecommunication industry, may actually earn money with providing CLOUDs. These concerns may become major obstacles to the uptake of new technologies.

B. ROOM FOR IMPROVEMENT

In addition to the actual limitations and problems of commercial CLOUD offerings, there is a wide scope of additional potential that is currently just ignored due to these deficiencies. Not only because they limit the uptake and thus the experience in usage, but more importantly, because CLOUDs could offer more capabilities, if these obstacles would be overcome. There are clear indicators that the potential of the CLOUDs have not been exploited to the degree actually possible just by maintaining the according concept(s). These include:

1. LARGER INFRASTRUCTURES IN TERMS OF NUMBERS OF RESOURCES (FEDERATION & INTEROPERABILITY)

CLOUD providers nowadays offer their infrastructures typically as "isolated" platforms, where a user cannot easily switch between providers without significant additional costs. This not only restricts the resource scope, but also the usability, as well as the potential for users to combine capabilities according to need. CLOUDs need to be open on as many levels as sustainable possible to achieve economies of scale in terms of users, devices and applications without being commoditized.

In particular in Europe, where the average provider will host less resources than the major players in the US or China, federating multiple infrastructures will provide an opportunity to serve resource needs beyond the local capacities. This also allows selecting the infrastructure according to the security, privacy and legislation needs. For example, execution of distributed business processes (Business-Process-as-a-Service), i.e. composite services above the SaaS-layer may lead to significantly faster adoption[15]. Notably, federation may impact on the pricing / business model (in terms of competition etc.)

2. BETTER USER SUPPORT (SPECIALISATION VS. GENERALISATION)

Though even generic platforms imply specialisation to specific needs, they mostly try to address as many use cases as possible thus not offering the performance and capabilities needed for specific users. Given the growing interest in services comprising multiple domains, it is thereby also not clear to which degree CLOUD offerings should be specialised, respectively generic.

Specialisation could thereby by a specific opportunity for European Independent Software Vendors (ISV) and similar specialised service providers which have strong niche or local market positions, even though they only have limited positions on the software market worldwide [AUM10].

---

[15] Early promoters of service composition through workflows already appear on the market, such as the European SME RunMyProcess (http://www.runmyprocess.com/), or OutSystems http://www.outsystems.com/). But also the Amazon Simple Workflow Service (http://aws.amazon.com/swf/) and the Microsoft Windows Workflow Foundation (http://msdn.microsoft.com/en-us/netframework/aa663328) already offer cloud support.

3. EASE OF USE (PROGRAMMABILITY)

CLOUDs are generally not easy to use to program for, let alone allow easy conversion of existing applications into the respective environment. In other words, an existing / proprietary application cannot make easy use of CLOUD capabilities and development of a new application that fully exploits the CLOUD features is difficult, as the scalability concepts are not inherent to current programming models.

4. MORE TRUST (LEGISLATION, POLICY, SECURITY, PRIVACY ETC.)

A CLOUD provides "global" data hosting, possibly across multiple legal jurisdictions, raising compliance issues for both users and providers. The additional lack of proper data encryption modes that support remote computing without decryption makes many users distrust the CLOUD – without additional means, such as homomorphic encryption, data will always be accessible to the provider himself, at least. But it is not only the lack of security support that is cause for distrust, but also the lack understanding of CLOUD behaviour, quality of service to be realistically expected, performance rating etc. (see above). Legislative issues concern the whole internet economy and need to be solved at political level (see also [EC11][FIA10]).

5. BETTER RESOURCE UTILISATION, LESS COST (MANAGEABILITY & EFFICIENCY)

Though the dynamic management of resources is the main asset of CLOUDs, this capability is as yet far from optimal, being effectively an NP-complete problem. Resources are still over- and underutilized, and code and data are not effectively distributed, let alone adapted to the infrastructure. Even the potential benefits for energy efficient computing are not exploited to their best yet – in fact, there is hardly any quantifiable data that allows for measuring, let alone for exploiting energy efficiency aspects to their minute details. For example, the impact of the relationship between code behaviour and hardware layout are even unknown to processor manufacturers as yet.

The key point to improvement hereby consists in being able to dynamically spread out applications running on local (self-owned) infrastructures (which may be a CLOUD) beyond the own infrastructure into an outsourced environment, thereby fully respecting the associated QoS, including privacy, security etc.

IV. FUTURE USAGE & RELEVANCE FOR EUROPE

Modern infrastructures host tens of thousands of servers and hundreds of thousands of computing units (if cores are included), yet it must be expected that this number will increase by orders of magnitude in the future. In addition to this, modern infrastructures incorporate and have to interact with more and more diverse resources, ranging from mobile devices over large scale platforms to web-enabled “things”, such as washing machines. Implicitly, not only the scale of infrastructures is increasing immensely, but also their heterogeneity and complexity, thus making effective management of the environment highly problematic. At the moment, one administrator can roughly deal with 1.000 servers under best circumstances, i.e. assuming strict homogeneity in its setup [MIL09] – new methods are therefore needed to improve the manageability of such infrastructures even further, on both user and provider side. This equally affects all levels of provisioning as the underlying environment needs to be maintained in order to offer the respective services / capabilities.

The need for servers and hosting infrastructures will only increase further in the future with the large scope of mobility of end-users. Not only human users increase the load on the network though, also the web-enabled things are expected to be effectively “always connected”, to enable updates of firmware or software at any time. CLOUDs have thus become a necessity to deal with the modern demands for internet provided services / applications / resources and the implicit requirement for large hosting infrastructures. CLOUDs are thus part of the next “mainstream computing”.

A. RELEVANCE OF CLOUD SYSTEMS FOR EUROPEAN INDUSTRY & RESEARCH

CLOUDs are important to Europe for three major reasons (cf. [ASH12]):

1, they provide a means for industry, especially SMEs, to utilise more cost-effective IT thus gaining commercial benefits;

2. they provide a means for industry, especially SMEs, to access more advanced ICT than through usual architectures allowing more ‘adventurous’ use of computing, and to move faster into new markets with less financial risk, thus gaining commercial benefits;

3. they provide an opportunity for IT providing industry – especially SMEs – to offer their services in an open marketplace and gain consequent commercial benefits;

European industry is highly heterogeneous and builds on a strong basis formed by small to medium enterprises. It is exactly these enterprises that benefit the most from the enhanced management and hosting capabilities that the CLOUD offers, by outsourcing resources, their administration etc. In addition to this, small to medium enterprises typically show a high innovative flexibility, but lack the capability to quickly, develop, test and host new services, for which again the CLOUD could contribute essentially, if developed further accordingly.

European industry – especially SMEs – can be advanced in concepts requiring high performance or high throughput computing. Unable to afford the capital required for in-house provisioning, CLOUDs provide cost-effective mechanisms for industry to procure the required ICT facilities. They also have the potential to lower the entry cost into new solutions and reduce the cost for failure.

Finally, Europe has always been strong in offering and providing services [DGT08] for end-users and businesses. Over recent years, with the financial crisis and similar factors, this trend has however decreased slightly in the global economy [MON09]. With the fast development of service provisioning over IT, there is a high risk for Europe to fall more and more behind the US and potentially China, as these countries quickly adopt the internet based service provisioning principles. The limitations in particular of SMEs, however, make it difficult for them to establish and go along with these technologies (cf. section III.A), as long as no new adoption and adaptation platform is made available, such as an enhanced CLOUD infrastructure.

Europe must therefore play a more active role in shaping the development on the IT and in particular the CLOUD environment. There are multiple opportunities for the European ICT industry which will be elaborated in this chapter.

1. General Future Developments & Requirements

It has been mentioned multiple times within this report that the future IT infrastructure is growing extremely in size and heterogeneity. An increasing number of users make use of online services of all kinds and more and more applications exploit the benefits of data and code outsourcing for improved maintenance and availability. This leads to an increasing load on networks and servers, which in turn are constrained in terms of bandwidth, latency, performance etc. Service and resource providers all over the world need to ensure that they can deal with this increasing amount of usage in a most cost-effective way and respecting the given restrictions and limitations. Though automated resource management capabilities have increased considerably over time, they are still insufficient to deal with the expected scope and heterogeneity.

Moreover, the need for resources will exceed the availability of individual providers, in particular if they cannot afford to host large scale infrastructures such as major US providers do. In particular for business processes and services spanning multiple providers, a simple CLOUD provisioning model is insufficient. Similarly, maintaining availability under mobile and widely dispersed constraints requires exploitation of CLOUD providers from other countries to extend the local infrastructure. The globalisation of the internet therefore can only lead to a globalisation of the CLOUD ecosystem, too – on all levels. This equally implies technical, economical, legal and political concerns. Current federation and interoperability support is still too weak to realise this, though.

Diversity is thereby not restricted to the types and qualities of the CLOUD offerings that users may want to integrate into a single system, but is also affected by the growing heterogeneity of device and client platforms, data structures, applications and even operating systems (Windows, Mac, Linux, Android, iOS etc.). In order to reach a large customer base, providers must therefore host environments that are highly adaptable and offer a wide scope of interoperability and portability.

Even though the number of processing units and thus the implicit performance at least for multi-task execution is constantly increasing, mobile devices nonetheless will outsource most of their processing towards the internet. The clients thus effectively resemble browsers more and more, where comparatively little computation and data handling is executed locally, so that the hosts need to cater for the main processing, but implicitly also for the communication (see e.g. Google's Chromebook[16]). The average user typically also owns multiple devices between which he wants to share his data and ideally also his applications and working environment. With this comes the demand for constant online availability – non-regarding technical deficiencies.

The high degree of mobility, together with the growing awareness for carbon emission and energy consumption also lead to an increasing demand for "green-ness" of computing. The average customer will thereby only be aware of the local consumption (i.e. power consumption of his device), but the demands on global level create growing pressure on providers to also host green infrastructures.

2. Economics & Economic Interests

To realise these future requirements is particularly of economic interest: providers aim at providing the user needs and interests in order to ensure that their respective offerings sell. But next to the direct economic exploitation of CLOUDs, comes the indirect one from the user side: CLOUDs offer a unique model to outsource tasks, in particular management and administration of resources, thus allowing enterprises to concentrate on their core business, without having to deal with additional administrative tasks. With the CLOUD concepts of elasticity, it is furthermore guaranteed that payment reflects interest and usage from the customers – at least to a certain degree.

In general, economic interest focuses in particular on aspects such as scale, elasticity, availability and outsourcing to reduce capital expenditure and to realise payment in relationship to consumption (pay-per-use). Within this section, we will list some of the key economic interests. We will thereby not distinguish between the perspectives from academia, industry and end-user, as this would exceed the scope of this discussion:

[16] http://www.google.com/intl/de/chromebook/features.html

a) Provider Interest

The primary interest of providers consists obviously in providing better services to their customers, in the sense that they should have increased interest in making use of the offered services. This includes not only higher availability of the application and (its) data, but also all other aspects of quality of interest to the customer. Moreover, however, this includes specialisation and extension of services to provide additional capabilities to the specific users' domain. The attraction for the user thereby consists in easier development of applications in his domain, better performance, or new capabilities.

In addition, providers will want to reduce the cost incurring from maintaining and using their resource infrastructure. This means for one that the management overhead of the large scale of resources should be reduced to a minimum, but second and more importantly, that the utilisation of resources is optimised, so that the minimum amount of resources is required for provisioning of the services in the desired quality.

b) Consumer Interest

Similarly, the consumers' primary interest consists in easy usage, but in particular in reducing the cost and overhead for maintaining a resource infrastructure. Infrastructure as a Service allows enterprise users to outsource the infrastructure they need for consuming or offering their own services, thus reducing exactly this overhead in management. But it also offers them a low cost entry point for offering new services, respectively for maintaining an infrastructure, of which the required size is not known a priori. For example, a new service offered to the end-user behind the CLOUD consumer, will at initialisation time create mixed interest in the community, so that the actual long term requirement of resources is not known.

This equally applies to services and resources consumed within the enterprise environment itself. Through CLOUD usage, the number of resources required within the enterprise can be kept to a minimum, extended dynamically with the resources offered by the CLOUD host. Thus, the consumer can realise a large resource scope than otherwise possible. Ideally, he will thereby want to realise mixed private public CLOUDs, where essential data and code is retained at his private site according to their sensitivity.

Next to extending the resource scope, CLOUD obviously also offers an extension of the consumer's capabilities to build up new and enhance existing business opportunities.

c) Developer / Researcher Interest

Finally, developers' and researchers' main interest consists primarily in overcoming the technical obstacles towards realising the requirements that would in the long run improve the capabilities of CLOUD providers. This should thereby respect the specific circumstances of European CLOUD provisioning and consumption, as these may differ strongly from the US and China. As such, it can be said that the developers' and researchers' *economic* interest in CLOUDs is mainly indirect in nature, namely to realise technologies that are of long-term validity for building up a CLOUD ecosystem. Europe has a significant community of independent software vendors who exhibit great interest in transitioning their solutions and applications to the CLOUD. To ensure such long-term impact, researchers must work together to steer the actual development and progress towards realising these requirements, so as to create a stronger knowledge base, but also to ensure commonalities and interoperability of the results.

Additionally, CLOUDs offer unique testing environments for researchers and testers without having to acquire and manage an according resource infrastructure locally, thus saving costs and overhead for tests of code systems that show a high degree of scale and dynamicity. With the according support from the infrastructure, this could also cater for adaptability and dynamicity requirements towards the tested service.

B. The Future Landscape of European CLOUDs

It is obvious that Europe would benefit greatly from a framework that can offer the full CLOUD characteristics. It is however less clear from the preceding sections which use cases may actually be able to fully exploit these characteristics and therefore contribute to the European interests as listed above. Whilst it is obvious that specific use case profit from individual capabilities (as described in more detail in section II), it needs yet to be examined, how industry, and in particular how Europe can benefit from the full potential of CLOUDs. Within this section we elaborate potential use cases and their relevance for (European) industry. Note that these differ from the NIST use cases [NIS11] insofar, as the latter describe the essential low-level steps involved in complex application provisioning, whereas this document describes use cases on the level of business models, and therefore concrete opportunities for future cloud applications in Europe.

1. Global CLOUD Ecosystem

Europe is part of the global IT infrastructure, of which CLOUDs form a subset that actively spans national, legislative and enterprise boundaries. In the future, global CLOUD environments that can host encompassing and new innovative business models – ecosystems – will become increasingly important. This also follows the logical development towards higher mobility and mobile offices across all such boundaries. Such ecosystems will make a wide variety of use cases possible, such as more efficient logistics and transportation, global disaster warning and management systems etc.

2. Tools and Service Market

Foremost, European CLOUDs would offer extended capabilities (such as availability and efficiency) to the tools and service market, i.e. realise improved services that ensure availability under highly dynamic conditions, that provide personalised capabilities and enable the fast creation of / support for new business models. This may equally span areas of public interest (improved medical services, eGovernment & municipal services, "digital cities"), domain specific and industrial use cases (specialised CLOUDs, business service support, administration outsourcing), and of individual interest (personalised services, mobile support).

3. Utility Computing

Utility computing is a growing field of IT for which CLOUDs are perfectly suited: infrastructure as a service allows easy outsourcing of resources and their management, and efficient PaaS models can allow quick development and hosting of complex services. Effectively this means outsourcing any capabilities / computation that cannot (or should not) be executed locally. This is by no means restricted to commodity computing, but can also include data handling on demand, or high performance computing of explicit computing tasks in a domain. This covers a wide scope of potential domains, such as health & medical cases, engineering, eScience, collaborative design etc. This also includes support for startup networks through European wide CLOUD resources.

4. Consultation

Another highly relevant business case relates to counselling expertise for CLOUD provisioning / hosting, as detailed in section III.A.3. Such counselling is particularly relevant as long as no common business expertise about CLOUD systems in the different usage domains exists, i.e. as long as CLOUDs are not a fully recognized commodity. Relevant expertise thereby covers both the technical and the economical side, enabling to give educated information about specific use cases, their requirements, porting of applications, best use recommendations, usage profiles etc.

C. Analysis: Additional Characteristics

It will be noted that the future use cases exceed the technological scope of the characteristics identified in section I.A. In other words, they require additional capabilities that are as such not directly supported by CLOUDs. Accordingly, it could be argued, that they are not intrinsic to the CLOUD concepts – however, since they partially capture environmental development and requirements that will generally arise in the global IT ecosystem and affect multiple domains, one may equally argue that they are essential for any future CLOUD system developments.

Instead however, we extend the hierarchy depicted in Figure 1 by an additional domain, namely the global CLOUD ecosystem (cf. Figure 3) which inherits features from all lower levels and extends them to a managed, federated global resource mesh.

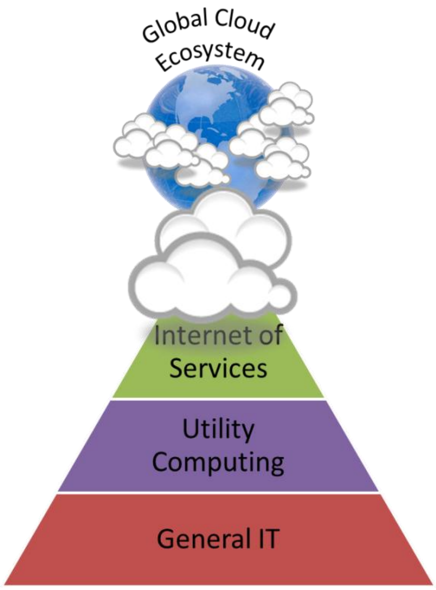

FIGURE 3: EXTENDED DOMAIN HIERARCHY

We can therefore generate a relationship diagram similar to Table 2 with the "Global CLOUD Ecosystem" as an additional domain. It will be noted that even though all three capabilities represent extended characteristics of the global CLOUD ecosystem, similar aspects have already been addressed specifically in the context of Virtual Organisations and global service markets (i.e. ecosystem).

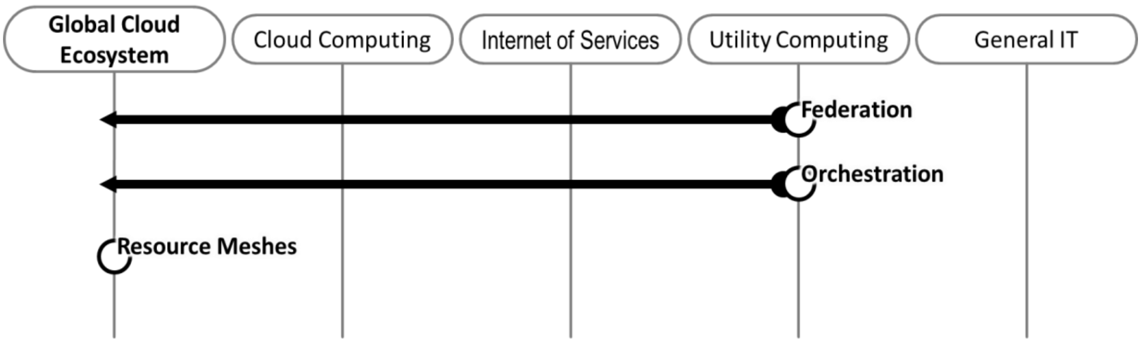

TABLE 3: EXTENDED LIST OF DOMAINS AND CHARACTERISTICS

We thereby introduce the new term of "Resource Meshes" or "Open World Infrastructure", denoting an environment of resources that expose services on all stacks or tiers, i.e. ranging from different flavours of Network as a Service, over compute resources (IaaS) up to software and services (PaaS & SaaS) across boundaries and instances as a single integrated environment.

V. RESEARCH ISSUES

Despite the seeming advances made in CLOUD provisioning, many major obstacles remain as yet basically unsolved. Whilst minor advances have been made, aspects such as federated CLOUDs are far from completion yet – communication restrictions play thereby a major development wall. This is clearly due to the fact that these issues have been addressed only insufficiently by industry as yet, due to their complexity and challenges, as well as their limited short-term business interest. This does by no means limit the value of these aspects, as detailed above – rather, the incremental approaches commonly pursued by industry still struggle with the actual adoption and exploitation of CLOUDs.

The preceding sections have identified most of the according obstacles and discussed them in some detail. For the purpose of reference, we will quickly summarise them in the following:

- **Economic models & expertise:** the market structure of CLOUDs is as yet generally unclear – when and how CLOUDs are profitable is still mostly subject to guesswork and experimentation. Most existing data is constrained to specific use cases and platforms. New cost and pricing models are needed and more knowledge needs to be gathered as to when it is sensible to move to CLOUDs, and how much cost & effort this implies.
- **Scale & heterogeneity of modern IT environments:** the scope of user devices, as well as infrastructure resources has grown beyond easy manageability both in terms of size / scale, and diversity / heterogeneity. Optimising the resource utilisation becomes increasing complex and is hardly supported by either code or data hosted. Automation is thereby crucial.
- **Communication limitations:** the increased scale leads to increased communication and data traffic that exceeds the physical connection and processing limitations, and also the current network management methodologies.
- **Modalities of scale:** code and data are not prepared for exploiting the CLOUD environment and all its characteristics. Algorithms exhibit thereby both modalities of scale, horizontal and vertical, and the environment has to deal with the according impact on execution.
- **Programming models:** are still oriented towards single-core sequential execution. In order to fully exploit the CLOUD characteristics, new approaches to distributed programming and execution are needed that cater for horizontal and vertical scale. Similar restrictions apply to composed and aggregated applications and standard communication between instances.
- **Resource Management / Usage specific behaviour:** different use cases exhibit different code behaviour and requirements. Generally, CLOUDs do not explicitly exploit these differences, in particular in terms of location control, multi-tenancy etc. so as to optimise the execution performance in terms of all

quality parameters. For example, moving data with the user to ensure availability not only through replication.

- **Migrating applications:** The full impact and consequences of moving applications to the CLOUDs is as yet not clear. As long as the switching cost is high (lack of programming model, lacking interoperability etc.), the benefits and the scope of impact must be clear for the use cases that actually benefit from CLOUDs.
- **All time issues:** the CLOUD is another step on the evolution of the IT environments, and thus subject to the same concerns as all the others. The specific nature of CLOUDs intensifies many of these problems even further. To these belong first of all security and data protection, but also classical optimisation and management problems.

A. Research Types & Their Relevance

In order to satisfy the goals and requirements outlined in this document and to overcome the major arising obstacles, we can identify a set of research issues that need to be addressed. We can thereby distinguish between three types of relevance:

(1) **essential (fundamental) research** that is essentially required to realise the respective functionalities at all. As CLOUDs extend the internet based service / resource provisioning most according research is closer related to the Internet of Services in general and can be applied to other provisioning domains, too

(2) **extending research** that builds up on existing research & development work – in other words that extends existing base line features to optimise them, adapt them to specific conditions etc.

(3) **application / use case specific research** that serves the purpose(s) of a specific use case and thus is limited in usability for other cases. This is only sensible once the underlying capabilities fulfil at least the minimum quality relevant for this use case.

It was already noted in the preceding section that most of the work relates to the second category, as CLOUDs leverage grids and web services etc., thus exploiting a solid conceptual basis, as most IT in fact. This however does not imply that grids and web services either have addressed the respective capabilities necessary (depending on the feature in question).

A major issue to be kept in mind in this context concerns the rapid progress and changes in the wider CLOUD context, its concept, associated technologies etc. Whilst this document tries to address these issues by identifying the long-term relevant aspects, the development nonetheless needs to be captured and a governance process introduced to ensure that (a) the roadmap is maintained (i.e. updated) accordingly and that (b) deviations are kept to a minimum. It must be stressed hereby again how relevant it is to maintain the long-term aspect of CLOUD research and development.

Bearing this in mind, we can note in particular that the following issues still pose research questions:

1. Essential Research Issues

CLOUDs essentially build up on well-known IT domains, in particular distributed computing, web-based utility computing and load management, which all have a sufficiently covered foundational basis. This means that there are no substantial foundations to be covered to make CLOUDs happen at all – after all, CLOUDs are already a commercial reality. Nonetheless, the unique combination of these aspects give raise to concerns that none of these domains have to address individually. To realise the according functionalities, most existing results either apply workarounds, or simply ignore these aspects, thereby limiting the capability scope of CLOUDs.

We can therefore classify these research issues as *essential* to realise the global CLOUD ecosystem vision:

a) Business and Cost Models:

CLOUDs open up a large scope of new business models, but also implicitly generate a lot of problems to make these models viable. Self-hosting and –provisioning becomes easily affordably this way but may not generate any revenue directly itself, unless other cost and payment models have been found. So far, services on the CLOUD are expected to be free or at least at low cost, and the cost for maintaining the underlying network is frequently completely neglected, so that the revenue stream for telecommunication industry is low. Also the growing competition between CLOUD (service) providers makes it more and more difficult for users to identify and compare the value of offerings. The same effect has already been observed with the “dotcom bubble” and should be avoided a second time. This asks for complete new models to generate revenue, to calculate and compare costs etc.

This relates also to identification of appropriate market mechanisms that support trading of CLOUD services (including applications, technologies) in a fashion that allows easy identification, comparison and provisioning, i.e. registration of new services. Services should thereby ideally be “fungible” to allow for easy provider switching. Obviously such markets must be developed in close adherence to the cost and payment models.

b) Data Management and Handling:

Though plenty of work has been invested into distributed data bases etc., the special concerns of CLOUDs with respect to dynamicity, distribution, concurrent access, locality etc. pose issues that have as such hardly been addressed. Data production and consumption is constantly growing, with the environment (both soft- and hardware) not being ready to handle this scope: applications do not encode the relationship between code and data properly, let alone that CLOUDs can exploit this relationship efficiently. Most data is still structured completely monolithically without respecting the environment and the need for communication, let alone for distribution and consistency. Data could and should move with the need put forward by the environment (i.e. location of user and compute unit).

Data management therefore is not restricted to maintenance of data, but its efficient computing, communication and distribution so as to ensure fulfilment of the overall quality parameters. This must thereby respect the physical limitations, the privacy requirements (such as laid out by the upcoming ISO 27018 standard) and the expected speed of growth in order to be of long term relevance. In other words, radically new means to process data on all levels are required, with a long-term perspective on scalability and growth.

c) Resource Awareness

The relationship between resources and their impact on execution of different applications / use cases is still widely unknown. Even modern day processor manufacturers mostly guess what kind of capabilities are most suited for future processing requirements, leading to a wide diversification of partially specialised and partially general purpose compute units. In CLOUD computing, this diversification becomes even more relevant, not only for executing the services and applications, but also for serving the right form factors of the user devices. To provide well-adapted, personalised services, and to address the restrictions given by the environment (cf. data management), the CLOUD thereby should not only be aware of the resources itself, but ideally also of the context and environment of usage, including aspects such as trajectory to predict connectivity / availability concerns etc.

The scope of context is thereby unclear and given the concepts of CLOUDs as such, the degree of information and the distribution / organisation of handling it over the environment needs to be assessed further.

d) Multi-tenancy Impact

Multi-tenancy goes beyond the fact that multiple tenants reside on the same server with their own isolated domain, but also implies that these tenants may furthermore share code and data on different levels with one another. Already the isolated environments cause performance side effects that have hardly been addressed, let alone solved. One major issue thereby consists in extrapolating the actual usage information for monitoring and accounting, which requires isolation of the individual tenant’s environment and context, even if it overlaps with other tenants’ contexts. The relationship should thereby be bi-directional, i.e. not only allowing the provider to extrapolate performance / usage information, but also to control it and therefore understand the impact and relationship between the contributing factors, in order to assure quality of service.

e) PROGRAMMABILITY

In order to cater for data management, resource awareness, performance and related issues, it is imminent that the program (and its relationship to data) is structured accordingly and exploits the specific features of the distributed, dynamic environment. Programming models are however still generally oriented towards single-core, local, sequential execution, following the principles laid out by Turing and von Neuman. Thus it is not only difficult to develop an efficient CLOUD application / service, because the according knowledge is missing, but more importantly because the programming model does not cater for it in the first instance and demands a high amount of expertise and effort to compensate for these deficiencies. New programming models must therefore address aspects of distribution, parallelisation and replication from base up.

Much work exists from the domains of distributed and parallel computing, but these approaches do not properly address the code-data relationship that needs to be exploited for aspects of concurrency, consistency, replication etc.

f) NETWORK MANAGEMENT

Network management for CLOUD computing and IT+Network altogether are important and currently not well covered areas. CLOUDs require the seamless provisioning and management of resources on all levels, in order to fulfil the base characteristics of CLOUDs, in particular regarding availability. As such, wrong network management can cause e.g. starvation of individual services, due to the problems of isolation and lacking management / control over the connectivity. Such management therefore needs to take the full scope of relationships between software, hardware and network into consideration, as they influence each other strongly, and cater for the increasing scale of usage and infrastructures.

The main goal consists thereby in improving the service capabilities and quality – therefore it is not clear, whether future services should be able to control the network, or whether to continue the classical line of high quality over a best effort network. This will depend strongly on the type of service offered.

One approach may consist in turning the network connectivity itself into a service, a.k.a. Network as a Service (NaaS). The network must thereby become as elastic as the CLOUD, maintaining the right quality of service in terms of bandwidth and latency, as required by the application / user. This might become a CLOUD market in its own right for CLOUD providers to get the right network resources, but also for internet service providers in general. Software-Defined Networks[17] ("SDN") may prove to be a valuable concept in this context as it will allow developers and administrators to define the data flow across the routed network.

[17] http://www.technologyreview.com/biotech/22120/

g) Legislation and Policies

Legislative and policy issues in the internet have been a concern for multiple years now, but yet still remain widely unsolved. With the introduction of CLOUD systems and its high dynamicity and sheer borderless usability, even more legislative concerns arise and even seemingly clear regulations can lead to contradictions that are difficult to solve.

It should be noted in this context though, that legislation cannot be regarded a research issue as such, but foremost qualifies as a political concern. Even though it necessitates legal research to promote solutions, it is outside the scope of ICT. Nonetheless, the decisions taken on political levels, respectively also the according concerns expressed by users and providers has impact on the technical solutions and the policies for e.g. management enactment, which have to be observed closely in future research and development iterations.

2. Extending Research Aspects

As has been noted in the preceding sections, the major amount of research and development work in the context of CLOUDs consists in adapting, adjusting and extending results from related domains, that form a basis for CLOUDs (see also Figure 1).This does not mean that some of these adaptations do not require complete new approaches for example on an algorithmic level to address the specific requirements put forward by the CLOUD.

a) Mixed CLOUD Provisioning Modalities

Mixed, hybrid and even cross-provider spanning federated CLOUDs are as yet essentially academic, i.e. have been mostly subject to research with little to no commercial support available. Research is thereby generally pursuing the Grid established model of “Virtual Organisations” where providers (CLOUDs) are identified according to requirements and execution is orchestrated across providers. There are also some concepts for “CLOUDs-of-CLOUDs” but they so far hardly apply to realistic usage settings, not only due to interoperability issues across proprietary CLOUDs, but also due to lack of manageability means across boundaries.

As CLOUDs will not completely replace local infrastructures, support for hybrid models need to be elaborated further anyhow, so as to enable providers to make use both of local and remote resources. This allows for increased security, but also to improve energy and execution performance, given that locality of the resources can be properly exploited and aligned to the execution / data distribution model.

Research can build up strongly from the according work performed in the domains of web services and Grids that both addressed cross-organisational work distribution and resource assignment, as well as means to integrate different frameworks and middlewares. Implicitly, standardisation plays an important role in this context.

b) Autonomicity and Adaptability

The scale and complexity of systems increase faster than any current management system (and administrator) can cope with efficiently and effectively. New systems must be enabled to manage and adapt themselves autonomically on a large scale. This includes aspects of dealing with reliability, in particular to handle and compensate faults to make the system resilient against the failures inherent to any IT system.

Moreover, in order to support future usage scenarios, which consider mobile devices and “Things” contributing with their storage and processing capabilities, these autonomic management systems need to evolve from considering dedicated resources only, to dynamic resource management systems, where resources will be heterogeneous, discoverable and dynamically attached/detached

Autonomic and adaptive systems have been a research issue for quite some time, with some impact already notable on modern computer architectures, but with most results being academic in nature so far (see e.g. the areas of Multi-Agent-Systems). The relationship and applicability of these results with / for CLOUDs still needs to be assessed.

c) (CLOUD) Software Engineering

Along similar lines, the vast resources and their capabilities are currently not programmable: not only has code become less portable with the divergence of resource types, but more importantly, the degree of scale is not met by any current programming model: the additional constraints of communication, concurrency, multi-tenancy are not catered for. What is more, there is no way of representing and planning it (large scale hybrid software engineering) in the first instance. It is necessary to restructure the way not only data is structured, but also code and its relationship to data and with one another is maintained in order to exploit the features of the CLOUD. It must thereby also be considered that the CLOUD behaves more non-deterministic than other IT infrastructures, which poses issues of reproducibility. It must thereby be kept in mind that not all applications benefit from scale, in particular not in their current implementation, so that new expertise and a new “programming base” needs to be elaborated. This must be able to identify and exploit (horizontal & vertical) scalability aspects throughout the whole application lifecycle and across all stacks. Future developers will thereby require a good understanding of how to map organisational & business problems into innovative CLOUD solutions.

Appropriate engineering principles must however not only consider the way software and data is organised, but also how software and hardware, and the management of the IT infrastructures relate to each other. Operation of an IT infrastructure is in strong relationship to what kind of services / software runs best and how it needs to be developed. Paradigms such as DevOps [JEN11] may thereby play an important role in the future.

There have been multiple attempts already to extend software engineering and also programming principles (cf. above) with a notion of distribution. The Grid / web services community focused in particular on distribution, whereas the High Performance Computing domain is working on expressing and exploiting parallelism (vertical scale) and concurrency (related to horizontal scale) for decades now.

d) SECURITY & CONFIDENTIALITY

Dynamicity, heterogeneity, elasticity, multi-tenancy etc. all pose additional security constraints beyond currently available solutions and beyond the concerns of legislation and policy. A particular confidentiality concern in the outsourcing model thereby consists in the fact that the resource host always has full access to the data, even in encrypted cases as the computation requires (local) decryption; as well as lacking control over the location of data with unclear legislation between countries / providers. This demands for enhanced programmability support, but in particular for efficient encrypted computation execution.

Whilst security, encryption, authentication etc. have all been concerns of IT for decades now (and will always remain concerns, as not only the security mechanisms, but also the means to circumvent them advance), some aspects such as e.g. homomorphic encryption may offer relatively new approaches.

e) COMPLIANCE

Security is also related to aspects of ensuring compliance CLOUD resource behaviour, i.e. not only enabling the gathering / provisioning of monitoring data, but also ensuring, respectively validating their correctness. This does not only mean extracting the respective information from the source of mixed data that includes performance information etc. for all tenants simultaneously, but, more importantly, to verify their correctness against the promised quality of service. In other words, the user needs to be able to ensure that the provided monitoring information is the right data and to interpret the terms with respect to the promised quality.

Quality of Service monitoring and management are old concerns in terms of IT research and development. Even though they will required extension and adaptation to CLOUDs, the main changes relate to the new terms that need to be catered for and that potentially cannot easily be monitored / extracted. Compliance particularly needs to be adapted to CLOUD concerns.

f) INTERACTIVITY, REAL-TIME, STREAMING

One particular case of quality criteria to be met by the CLOUD in order to address the promised in terms of application hosting, remote office etc. consists in meeting real-time constraints. In the context of application usage via the CLOUD, this primarily concerns interactivity constraints. However, with the growing demand for multimedia integrated applications, streaming concerns put forward similar requirements towards CLOUD management.

Approaches must thereby be able to cater for the side effects of multi-tenancy usage on networks, resources etc. but must also keep in mind the impact of dynamicity of the CLOUDs that may lead to replication, relocation etc. of the providing services at run-time, thereby affecting the performance and in particular timing behaviour.

Real-time provisioning over the internet has been an on-going research concern for multiple years now. At the latest with the introduction of mobile end-users, dynamicity has thereby been a specific research aspect, from which CLOUDs can derive essential approaches and mechanisms.

3. Use Case Specific Work

This area of work obviously very much depends on the specific industrial application context addressed. Implicitly, the major research aspect relates to realising specialised CLOUDs that adhere to the requirements of the respective use case. However, a set of more general research recommendations can be identified that are common to serve specific use cases:

- Commonalities between use cases need to be identified and the according framework developed in a fashion that allows for maximum reusability
- More support for easy and quick development of services is needed to enable small companies to quickly produce and provide their respective services. This also includes aspects of enhancing and extending existing services by quickly integrating and building up on them.
- Gather knowledge about the implications from using a CLOUD (switching cost, effort, which use cases benefit etc.) and how to make the transition happen, which still poses one of the biggest hindrance for CLOUDs in specific usage domains.
- Combination of code and data from different sources to create enriched services needs to be examined (and supported) more
- For this also, testing environments should be available to allow quick and easy development and testing the service, and what is more benchmarking and comparing the results. For this however, also the means and interpretation of such comparison is still missing.

Note that specialised CLOUDs pose higher emphasis on user friendly provisioning and usage, such as for brokerage, aggregation and personalisation. This also requires more flexible licensing models from provider side to offer their ISV codes accordingly.

B. Realisation Effort Analysis

Not only relevance of the obstacles faced differs strongly, but also the complexity of reaching them. In other words, the degree of effort that needs to be invested in order to overcome a specific obstacle, respectively to realise the according issue to a certain degree. It will be noticed thereby that most of the research issues / obstacles listed have no clear degree from which on they can be considered solved, or – more importantly – they move to another research type, e.g. from being essential to extending research.

1. COMPLEXITY

The complexity to realise these issues thereby seems to depend directly on the degree of "newness", i.e. in how far the work can base on existing results and in how far paradigm changes in themselves are required to achieve the results. However, this is frequently confused with the seeming complexity of a completely new task that arises from the lack of expertise of how to handle the given situation – the solution may thereby however be simple, just unknown. As opposed to that, solutions that have been in development and use for a long time are sometimes extremely difficult to advance further or move into a complete new area, whereas completely new approaches can have more flexibility. Therefore, extending research aspects can have a higher complexity than essential and foundational research.

In light of the obstacles identified above, we can generally say that a major part of the research complexity arises from the more or less recent paradigm change in computing, i.e. the move from the classical single-instance computer and the homogeneous infrastructure to the widely dispersed and connected mesh of heterogeneous resources which are parallel in themselves. In other words, in modern IT and internet, we face multiple paradigm changes that affect the research, development and usage equally – in particular we can note the following main movements:

- from single-processor, isolated machines to highly distributed and parallel computers
- from homogeneous infrastructure setups to complete heterogeneity on all levels
- from applications that scale better than the computer to infrastructures that offer more resources than the application can use
- from (computational) performance as the main obstacle to communication as the restricting factor
- from large, centralistic providers with fixed prices to a diverse set of small(er) providers with pay per use models (and even prosumers)

All these changes affect how computing needs to be handled and therefore constitute the difficulty to solve the problem. In effect, such paradigm changes bring in complete new aspects to IT that have not or only rarely been addressed before. As noted above, this does not necessarily imply that the solution is complex in itself, but that it is highly complex to overcome all issues involved. We can thereby generally state that complexity is composed by the following factors

- the research effort needed to create a solution (without assessing its quality)
- the degree of expertise already existent in this field

- the acceptance of existing approaches to be overcome
- the uptake needed in order to create impact

2. Timescale

Next to complexity of developing a solution for the issues above, the expected time to its realisation is a key factor that plays an important role in particular in uptake of the solution and competition with other developments on the market. Generally, the longer it takes to generate a useable solution, the less like its uptake, unless it is a completely disruptive solution even at the time of its “market” (respectively community) injection. Accordingly, it is generally advisable to generate results in an iterative fashion by prioritising the aspects addressed.

Along the same line, the duration for which a product / a solution is expected to stay relevant in the field of application is a relevant co-factor with respect to its timescale. In general we can note that the following aspects influence the timescale of a research results

- The time for its development
- The longevity of the solution
- The market development (in terms of when / for how long the solution is required)
- The competition (in terms of when other results make the solution obsolete)

The combination of these aspects can be regarded as an indicator for whether it is worth addressing the problem in the first instance. For example a solution that is only relevant for a short time, but takes a long time to develop is probably not worth investigating in the first instance. It can be noted thereby that the time of development is related to the complexity of the respective research issue, i.e. that solving complex problems is generally more time-consuming than addressing simpler ones – at the same time it can be expected that the relevance, and hence longevity of a solution to a complex problem tends to be higher than for simple issues.

In this context we can roughly classify solutions according to the impact and their relevance they most likely will have on the market:

- Disruptive, new aspects: have not been addressed before and thus pose completely new challenges. They can give rise to completely new opportunities.
- Continuation: concerns all aspects that have been addressed but not fully concluded before – this means, that either some issues remain un- or not satisfyingly solved, respectively that adaptations to new requirements have to take place.

- Closed issues: are of no real research concern anymore, even if pure development tasks may still remain. They typically represent well-established approaches to common problems.
- Ever-Greens: effectively belong to the domain of "Continuation", with the specific extension that they are unlikely to be ever solved, as they create new issues with the advances in IT (e.g. security, optimisation, programmability)

There is a notable relationship between the impact criteria and the type of relevance, i.e. essential research issues tend to be disruptive, and extensions may fall under continuation of work. However, this does not imply that continuation of work may not lead to disruptive results, which is of particular relevance in the CLOUD context:

Due to the nature of the CLOUD, many aspects fall either in Continuation or Ever-Greens, since the technologies employed in CLOUD systems are neither disruptive nor completely new to the IT domain. More surprisingly, though, hardly any fall into the class of closed aspects – mostly due to the fact that the nature of the primary characteristics has some impact on how to handle some of the aspects that would generally be considered more or less closed.

On the other hand, the *concepts* of CLOUDs involve disruptive aspects that may cause long-term impacts on both future economic models, as well as technological progress. It can in particular be noticed that CLOUDs extend the relevance of distributed and parallel computing, mimicking development on processor technologies. This in turn may mean that either continued work may have disruptive impact in the CLOUD domain, as well as that continued work may spawn of concepts for disruptive technologies, if examined more closely in the CLOUD.

C. Research Issues and their Classification

According to the discussion in the preceding sections we can analyse the relevance, as well as complexity and constitution of the research issues identified in more detail:

1. Data management:

We are in the era of 'Big Data'. The EC report 'Riding the Wave' [HLG11] emphasised the volumes of data. Furthermore the data becomes ever more complex with multimedia and syntax (structure) and ever increasingly complex semantics (meaning). The requirements and expectations of end-users in being able to utilise the data – as datasets or in 'mash-ups' – increase exponentially. Since CLOUD facilities are usually centralised, and most data is collected from distributed sensors and over networks there is a paradoxical problem. We need new approaches to resolve this paradox; to gain the elastic advantages of CLOUDs but coupled to the distributed nature of data collection.

a) Research tasks

- Handling of big data across large scales;
- Dealing with real-time requirements – particularly streamed multimedia;

- Distribution of a huge amount of data from sensors to CLOUD centres;
- Relationship to code – there is a case for complete independence and mobile code – move the code to the (bulky) data;
- Types of storage & types of data – there is a need for appropriate storage for the access pattern (and digital preservation) pattern required. Different kinds of data may optimally utilise different kinds of storage technology. Issues of security and privacy are also factors.
- Data structuring & integrity – the problem is to have the representation of the real world encoded appropriately inside the computer – and to validate the stored representation against the real world. This takes time (constraint handling) and requires elastic scalable solutions for distributed transactions across multiple nodes;
- Scalability & elasticity are needed in all aspects of data handling to deal with 'bursty' data, highly variable demand for access for control and analysis and for simulation work including comparing analytical and simulated representations;

b) COMPLEXITY & TIMELINE:

This topic has been a long-term concern, especially since the capacity of and speed of networks has failed to keep up with the capacity and speed of storage. Incremental solutions from existing technology have low complexity and short timescales; a really novel (set of) solution(s) is of high complexity and long timescale.

2. COMMUNICATION & NETWORK

Like data management the user expectations are much greater than installed capacity. The lack of adoption of IPv6 and the missing capabilities to allow bandwidth reservation and optimised sharing is impeding progress for certain time-critical applications. It is clear that mobile device access is increasing very rapidly with a growth of >2 times per year. This will put demands on mobile networks which are now averaging 200Kb/sec and need to rise to Mb and higher speeds.

a) RESEARCH TASKS:

- Guaranteeing bandwidth / latency performance, but also adjusting it on demand for individual tenants (elastic bandwidth / latency): this is a real issue for an increasing number of applications. It is necessary for the network to exhibit some elasticity to match that of the CLOUD centres. This may require network slices with adaptive QoS for virtualising the communication paths;
- Compensating for off-line time / maintain mobile connectivity (internationally): intermittent mobile connectivity threatens integrity in computer systems (and also allows for potential security breaches). This relates to better mechanisms for maintaining sessions / restarting sessions from a checkpoint;
- Isolating performance, connectivity etc.: there is a requirement for the path from end-user to CLOUD to be virtualised but maintaining the QoS and any

SLA. This leads to intelligent diagnostics to discover any problems in connectivity or performance and measures to activate autonomic processes to restore elastically the required service.

b) COMPLEXITY & TIMELINE:

- There are complexities in service agreements and contracts with the providers of low-medium complexity and short term timescale. However, rapidly changing technical options can make this a medium complexity medium term problem.
- The provision of autonomic elastic connectivity is highly complex and with a long timescale.

3. RESOURCE DESCRIPTION & USAGE:

This is an emerging area commonly under the term 'metadata'. The need is to describe the resources in such a way that autonomic software can utilise the descriptions to make appropriate decisions. GRIDs provide some background design and experience. However, the roles of users and software against particular data resources is not yet well defined and the relationship of such a (customised) service via SLA and QoS to the underlying CLOUD and communications resources is distinctly unclear.

a) RESEARCH TASKS:

- Generic ways to define characteristics: there is a need for an architecture of metadata to a common framework (with internal standards) to describe all the components of a system from end-user to CLOUD centre;
- Way to exploit these characteristics (programmatically, resource management level): the way in which software (dominantly middleware but also, for example, user interface management) interacts with and utilises the metadata is the key to elasticity, interoperation, federation and other aspects;
- Relates to programmability & resource management: there are issues with the systems development environment such that the software generated has appropriate interfaces to the metadata;
- Depending on the usage, "resources" may incorporate other services
- Virtualisation – by metadata descriptions utilised by middleware -
    - Of all types of devices
    - Of network
    - Of distributed infrastructures
    - Of distributed data / files / storage

b) COMPLEXITY & TIMELINE:

The basic requirements can be achieved with medium complexity in a medium timescale. However, a full elastic autonomic solution is of high complexity and requires a longer timescale.

4. Resource management (Provider Support)

This links to the topic above. However while above we discuss how the system utilises descriptions of the virtualised resource, this section discusses the optimisation of resource usage. It requires system knowledge of user intent in order to optimise the resources available to the conflicting demands in an elastic and virtualised manner. Clearly this topic relates to metadata, software architecture, programmability and systems development as well as code execution management within a framework of fault tolerance and managed quality.

a) Research tasks:

- Deal with scale and heterogeneity: the metadata has to have rich enough semantics;
- Multidimensional, dynamic and large scale scheduling respecting timing and QoS;
- Efficient scale up & down: this requires dynamic rescheduling based on predicted demand;
- Allow portable programmability: this is critical to move the software to the appropriate resource;
- Exploit specifics on all levels: high performance and high throughput applications tend to have specific requirements which must be captured by the metadata;
- Energy efficient management of resources: in the 'green environment' the cost of energy is not only financial and so good management practices – another factor in the scheduling and optimisation of resources – have to be factored in;
- Resource consumption management : clearly managing the resources used contributes to the expected cost savings in an elastic CLOUD environment;
- Advanced reservation: this is important for time or business critical tasks and a mechanism is required;
- Fault tolerance, resilience, adaptability: it is of key importance to maintain the SLA/QoS

b) Complexity & Timeline:

The complexity is directly dependent on degree of divergence in the future – which is expected:

- Current complexity: medium and medium timescales required
- Long term complexity: very high (if divergence increases) with much longer timescales

This will be a continuing issue as doubtless divergence will occur and complexity increase.

5. Multiple Tenants

From the supplier point of view the management of multi-tenancy is very important and techniques to achieve this with appropriate resource management, security and other aspects are required urgently. At the moment multi-tenancy is not used optimally and there is a lack of control (lack of tools and management information). This relates to the lack of appropriate metadata described above.

a) Research Tasks:

- Isolate performance, isolate network slices: this is needed to manage resources and security;
- No appropriate programming mechanism: this requires research and development to find an appropriate systems development method, probably utilising service-oriented techniques;
- Co-design of management and programming model: since the execution of the computation requires management of the resources co-design is an important aspect requiring the programmer to have extensive knowledge of the tools available in the environment;

b) Complexity & Timeline:

- Lack of programmability: medium to high complexity with medium to long timescales;
- Lack of control: medium complexity moving to high complexity - and long timescales – with divergence;
- Optimisation: high complexity and long timescales;

6. Federation (& interoperability, portability)

With increasing numbers of CLOUD service providers at IaaS and PaaS levels, federation to provide (by increasingly linked suppliers) greater resilience and elasticity – as well as to incorporate diverse data sources and code bases – becomes a winning strategy. However, the currently leading CLOUD suppliers have steadfastly refused to agree to a common interface (API). There are issues of portability as well as complex issues of legality of access rights. Communication constraints thereby pose additional technological challenges.

a) Research Tasks:

- Portability, orchestration, composition: this is a huge and important topic requiring research into semi-automated systems development methods allowing execute time dynamic behaviour;
- Merged CLOUDs: virtualisation such that the end-user does not realise the application is running on multiple CLOUD providers' offerings;
- Management: management of an application in a federated environment requires solutions from the topics listed above but with even higher complexity;

- Brokering algorithms: are needed to find the best services given the user requirements and the resource provision;
- Sharing of resources between CLOUD providers: this mechanism would allow CLOUD providers to take on user demands greater than their own capacity by expanding elastically (with appropriate agreements) to utilise the resources of other CLOUD suppliers;
- Networking in the deployment of services across multiple CLOUD providers: this relates to the above and also to the Networking topic earlier;
- SLA negotiation and management between CLOUD providers: this is complex with technical, economic and legal aspects;
- Support for context-aware services: is necessary for portability of (fragments of) an application across multiple CLOUD service providers;
- Common standards for interfaces and data formats: if this could be achieved then federated CLOUDs could become a reality;
- Federation of virtualized resources (this is not the same as federation of CLOUDs!) is required to allow selected resources from different CLOUD suppliers to be utilised for a particular application or application instance. It has implications for research in
    - Gang-Scheduling
    - End-to-End Virtualisation
- Scalable orchestration of virtualized resources and data: co-orchestration is highly complex and requires earlier research on dynamic re-orchestration/composition of services;
- CLOUD bursting, replication & scale of applications across CLOUDs: this relies on all of the above.

b) COMPLEXITY & TIMELINE:

- Complexity arises mostly from growing divergence as discussed above;
- Essential work could lay the basis for the next steps in moving towards a standard framework and within that particular standards to allow federation of applications (as services) across CLOUDs. The initial work is medium complexity and medium timescale but if the diversity continues to increase becomes highly complex with a much longer timescale;
- Optimisation: medium to high complexity and long timescales – especially for co-optimisation across multiple CLOUD service providers;
- This topic is very important and rapid action could provide significant savings in future CLOUD deployment and significant opportunities for European SMEs to have a stable platform upon which to construct and offer services;

7. PROGRAMMABILITY & USABILITY

The newer hardware resources cause existing systems development and programming techniques to be hopelessly outdated. The weight of legacy code impeded progress and causes inefficient use of CLOUD resources. Completely new systems development / programming methods are required where metadata providing parameters related to scaling (vertical and horizontal), elasticity, security etc are present and where the relationship to metadata allows autonomic management.

a) RESEARCH TASKS:

- Restructure algorithms / identify kernels: in order to place in the new systems development context – this is re-use of old algorithms in a new context;
- Design models (reusability, code portability, etc): to provide a systematic basis for the above;
- Control scaling behaviour (incl. scale down, restrict behaviour etc.): this requires to be incorporated in the parameters of the metadata associated with the code;
- Understand and deal with the interdependency of (different) applications with the management of large scale environments
- Different levels of scale: this is important depending on the application requirements and the characteristics of different scales need to be recorded in the metadata;
- Integrate monitoring information: dynamic re-orchestration and execution time changes to maintain SLA/QoS require the monitoring information to be available to the environment of the executing application;
- Multi-tenancy: as discussed above this raises particular aspects related to systems development and programmability;
- Ease of use: the virtualised experience of the end-user depends on the degree with which the non-functional aspects of the executing application are hidden and managed autonomically;
- Placement optimization algorithms for energy efficiency, load balancing, high availability and QoS: this is the key aspect of scheduling resources for particular executing applications to optimise resource usage within the constraints of SLA and QoS;
- Elasticity, horizontal & vertical: as discussed before this feature is essential to allow optimised resource usage maintaining SLA/QoS;
- Relationship between code and data: the greater the separation of code and data (with the relationships encoded in metadata) the better the optimisation opportunities. Includes aspects of external data representation;
- Consider a wide range of device types and according properties, including energy efficiency etc.; but also wide range of users & use cases (see also business models): this concerns the optimal use of device types for particular applications;
- Personalisation vs. general programming: as programming moves from a 'cottage knitting' industry to a managed engineering discipline the use of general code modules and their dynamic recomposition and parameterisation (by metadata) will increasingly become the standard practice. However this

requires research in systems development methods including requirements capture and matching to available services.

b) COMPLEXITY & TIMELINE:

Work in this domain can base partially on existing results but to date these are relatively primitive so the complexity is medium with medium timeframe. As opposed to that, the development of new models implies a high research & development complexity, mostly due to uptake and long timescales

8. POLITICAL & LEGISLATORY

There is a requirement for research in these areas but from an ICST perspective these issues mainly provide constraints within which the technology has to develop. The issues are important because the reduce trust in CLOUDs and hinder uptake.

a) RESEARCH TASKS:

- Privacy concerns: especially in international data transfers from user to CLOUD;
- Location awareness: required to certify conformity with legislation;
- Self-destructive data; if one-off processing is allowed;

b) COMPLEXITY & TIMELINE:

The negotiations are in international law and therefore of high complexity and long timescales.

9. SECURITY

Once any part of ICT goes 'out of house' there are security concerns. As well as legalistic problems (there may be differential protection by law in different countries / regions) there are a series of technical concerns. These arise from the aspects of multi-tenancy, international data and code distribution and associated human resource issues. In a complex ICT operation involving many people in many countries with varying standards of professionalism and ethics there are increased security risks.

a) RESEARCH TASKS:

- Process applications without disclosing information: Homomorphic security: this offers some chance of preserving security (and privacy);
- Static & dynamic compliance: this requires the requirements for compliance to be available as metadata to be monitored by the running application;
- Interoperability, respectively common standards for service level and security: this relates to standard interfaces since the need is to encode in metadata;
- Security policy management: policies change with the perceived threats and since the CLOUD environment is so dynamic policies will need to also be dynamic.
- Detection of faults and attacks: in order to secure the services, data and resources, threads need to be detected early (relates to reliability)

- Isolation of workloads: particular workloads of high security may require isolation and execution at specific locations with declared security policies that are appropriate;

b) COMPLEXITY & TIMELINE:

- Security is a continuous ‘evergreen’ problem and with CLOUD computing it becomes even more complex so there is medium complexity and timeline rising too high.
- Homomorphic encryption: this has high complexity and a long timescale;
- Improved metadata with security parameters and monitoring information; this is of high complexity and long timescale;

10. BUSINESS & COST MODELS

A real advantage of CLOUDs is ICT cost management by ‘pay as you go’ with associated monitoring, accounting and billing. A second advantage is the encouragement of “prosumers” so that a market develops with traded resources. Energy costs – expected to be reduced - play a part in the cost model. In-depth experience in this area is often lacking and more evidence is needed.

a) RESEARCH TASKS:

- Accounting, billing, auditing: pricing models and appropriate dynamic systems are required including monitoring of resources and charging for them with associated audit functions. This should ideally be supported by integrated quota management for both provider and user, to help keep within budget limits
- Monitoring: common monitoring standards and methods are required to allow user choice over offerings and to match user expectations in billing. There are issues in managing multi-tenancy accounting, real time monitoring and the need for feedback from expectations depending on resource usage and costs.
- Expertise: The lack of expertise requires research to develop best practice. This includes user choices and their effect on costs and other parameters and the impact of CLOUDs on an ICT budget and user experience. Use cases could be a useful tool.

b) COMPLEXITY & TIMELINE:

- New business models have consequences in a wide economic context: high complexity and long timescale
- Monitoring & security: medium complexity but linked with dynamic autonomic behaviour complexity rises to high as does timeframe
- Expertise: time intensive, but low-medium complexity

VI. RECOMMENDATIONS & CONSIDERATIONS

“CLOUDs” are not a sophisticated and established technology as yet – instead, CLOUD offerings generally suffer from serious (long-term) drawbacks. These drawbacks are less well known due to the lack of expertise in the CLOUD domain, and a clear definition of user expectations of the CLOUD offerings due to the high hype factor. There is however a strong potential and need for a long-term stable basis, upon which CLOUD (and future internet) technologies can continuously evolve, as the internet and in general the scale of available resources and their heterogeneity increases [ZIM11].

In order to reach this vision however, it is necessary for researchers, developers and supporters equally to consider environmental circumstances, such as the on-going development and the impact of user demands on the timescales. Within this section we therefore compile a set of general considerations to be kept in mind in this context, as well as a set of recommendations towards European economy and funding bodies regarding how research results in this direction can be improved.

Timing is thereby of utmost criticality, considering the on-going efforts and changes on the market / in industry.

A. CONSIDERATIONS

Europe is not alone in addressing the research and development issues of future CLOUDs, and is therefore in a position somewhat between competition and collaboration – competitive in order to ensure its position in a global environment, collaborative in order to enable the vision of a global CLOUD ecosystem in which Europe plays a key role (cf. [KRO11]). In order to ensure this development, European research and development must take a series of circumstances into consideration:

1. RESEARCH AND DEVELOPMENT IS CONSTANTLY PROGRESSING

Research and development are not standing still. In particular in areas with high industrial drive, such as CLOUD computing, solutions and new technologies are developed quickly according to user demand and market evolution – potentially on cost of quality and interoperability. As opposed to that, (academic) research generally moves slower, even though the quality of the results and their long-term relevance may be higher. In addition to this, research results are generally taken up only slowly, as the users do not have immediate confidence in them and as their maturity is comparatively low at first, i.e. their uptake often requires initial adaptation work, whereas company-own solutions have been adjusted to the according requirements from the beginning. Even if this initial cost is minor, the implicit benefit has to be made absolutely clear to the uptakers.

Due to these different timelines, there is an implicit risk for divergence (see discussion in the Introduction section). Even now, there is industrial development which must / will definitely conflict with the research agenda laid out here. In particular short-term aspects of direct economic interest will already be highly advanced. These cover for example

- Payment models
- Accounting & billing models
- Use case expertise
- Hybrid (part local, part remote) applications

This does not mean however that all aspects of these areas have been fulfilled (cf. section V.B) and that these domains would not benefit from more advanced methods. It does mean though, that in order to provide valuable advances over the industrial efforts, an according long-term perspective must be the main focus of such work.

2. A Solid Technological Basis Exists – From Within and Outside Europe

EU research has to build on top, and go beyond, what exist today (from US, or whatever origin it has). It has to be forward looking more than focussing on providing EU solutions to issues already addressed, in order to achieve proper impact, and taking the most of invested funds

3. The European Economic Landscape Tends towards Small- to Medium Scale Heterogeneous Players

Europe has more generally more diverse, small providers who cannot individually compete with large providers such as Amazon, Microsoft or Google; also, the market is much more fragmented. Accordingly, the European resource infrastructure is much more heterogeneous, as it spreads over more small(er) scale providers. Therefore, in order to offer capabilities of similar scale and scope, Europe is much more in need of means for integration, federation and interoperability.

This heterogeneity on infrastructure level should thereby not be regarded as a hindrance to large scale cloud provisioning, but instead as a specific advantage in terms of combination of specialized environments, knowledge and capabilities to offer enhanced *composed* cloud services.

It must also be noted, that US enterprises tend to be more (end) user-oriented than the (large) European players. This means that the American industry aims for the mass of “small” customers, whereas the European one focuses more on business-2-business interactions. With the primary uptakers in clouds being in particular smaller players, Europe may miss out on the major customer base.

Any economic / business model for the cloud must reflect these differences and account for them.

4. THE ENVIRONMENT AND ECOSYSTEM WILL DEVELOP FURTHER

The IT market and industry has not yet reached its full potential and the scope (of users and devices), as well as the heterogeneity (of use cases and resources) will just continue to grow. This growth will thereby exceedingly push the boundaries of performance, networking, communication etc.

The CLOUD concept offers an enormous potential to deal with this growth in scale and complexity, but it must thereby not be confused with the "final answer" to this problem. As the IT market expands further, more problems will arise that sooner or later will require additional capabilities that the CLOUD concept cannot satisfy, thereby giving rise to a new model.

> ! *CLOUD computing is not the final solution for the internet of services, nor for utility computing – something will come after the CLOUDs.*

B. RECOMMENDATIONS

In order to reach the vision of a European CLOUD ecosystem embedded in a global market and business environment, it is however insufficient to "just" pursue a set of research and development actions that address the major concerns to realising this vision. Rather, it is necessary to enact a series of policies that not only ensure that the right issues are addressed timely, but also that the right actions are taken to enable and encourage at least European-wide uptake in the according business sectors.

To achieve this, the experts make the following recommendations to be put into action by the European Commission, industry and academia. These recommendations thereby go beyond the pure support and implementation of the research topics as identified in the preceding sections of this report.

#1 Ensure Progress in CLOUD Research

CLOUDs are embedded into the wider domain of utility computing and therefore overlap strongly with other related domains that have been subject to multiple years of research and development. There is hence a substantial risk that research and development is repeated, rather than taken up and adapted to the CLOUD.

> ! *Any CLOUD related effort must clarify its contribution & relevance in terms of the essential CLOUD concerns as identified by this report*

This implies that any research and development proposal must position itself with respect to the key CLOUD concepts and obstacles as laid out by this document and thereby clearly elaborate the basis from which it builds up, in particular from the related domains (cf. Figure 1). To this end

! *Strong monitoring (concertation and coordination) actions need to be put in place that can evaluate the progress made in CLOUDs and related domains*

Due to this strong multidisciplinarity of CLOUDs – both in terms of technology and usage –

! *Holistic approaches are needed that encourage the collaboration and reuse of results and knowledge between related domains, including Grids, Web Services and High Performance Computing*

#2 Focus on Concerns of Long Term Relevance

Industry and research pursue different timelines and commercial development is constantly progressing. Any short-term attempt in research therefore runs first of all the risk of becoming obsolete already during the research period and secondly may only contribute to the increasing diversification rather than steering the realisation of essential CLOUD capabilities (cf. section I).

! *Publicly-funded research should address the longer-term horizon intercepting the requirements and providing opportunities for European industry to take up in shorter term timescales*

Long-term topics however run the risk that their relevance for the industrial sector is not clear, or not aligned with the more short-term oriented development (risk of diversification) and therefore hinders uptake. This means that the commercial interest must always be clear, at least in the form of indicating benefit and impact (analyse use cases and uptake likelihood).

! *Encourage industry to participate in shaping the vision of the future CLOUD ecosystem and to provide as much insight into their research and development agenda as possible.*

This is a unique chance for European industry to specify their needs and get support for their long-term plans. Long-term research should thereby clearly focus on

- Helping to create a bigger economic opportunity in ICT and non-ICT sectors
- Helping to increase the competitiveness of the European companies in the global market

#3 Enable the Fast Transition to the CLOUD

Developers (both academic and commercial) get little support for realising and testing CLOUD applications and services: first of all, this implies the generation of the necessary tools, programming models and expertise to support movement to the CLOUD.

! *Research must focus on outcomes that can support the transition towards the CLOUD and that themselves can be quickly and easily transferred to the market*

In addition to this, however, it is also necessary to cater for environments that allow for large-scale tests and that provide innovative startups with a quick entry point into the market.

! *Large scale testbeds with technical, economic and legal aspects supported should be provided to test European solutions in a realistic environment and to showcase demonstrators*

This requires that such environments are easily accessible without any impact from legislation or security concerns, even though IPR issues are fully maintained. This may imply that *national* providers are needed to ensure maximum compliance.

#4 Encourage Large Scale European Providers

To compete in the global CLOUD ecosystem, higher resource needs have to be satisfied than is currently possible with the small scale, heterogeneous infrastructures typical for the European market.

! *Europe must encourage the hosting and provisioning of large scale European CLOUD providers, respectively the uptake and support of federated resource infrastructures by encouraging providers to collaborate and adhere to common standards*

A major obstacle to this consists however still in the unclear payment and cost models for future CLOUD provisioning, which requires that all aspects of provisioning (network, resources) are taken into consideration.

! *Future cost / business models must equally adhere to the requirements of the prosumer, as for the ones of major enterprises*

This must thereby also respect the legalistic and security concerns of the user, in particular in eGovernment related and public service domains.

#5 Encourage SME Providers

The European market tends towards (comparatively) small scale, diversified and heterogeneous providers in terms of their resource infrastructures. However, they produce innovative services and need a large-scale interoperable CLOUD provisioning in order to achieve a market for their offerings.

! *Europe must encourage the development of services for CLOUD users within the context of multiple CLOUD infrastructure and platform suppliers in an integrated manner utilising standards to ensure interoperability and portability.*

As for large-scale provisioning of infrastructures and platforms, a major obstacle to this is the unclear payment and cost models for future CLOUD provisioning, which requires that all aspects of provisioning (network, resources) are taken into consideration.

> **!** *Good economic business models are required to encourage SMEs to provide services for the open, integrated CLOUD environment in Europe.*

#6 Promote Open Source Solutions

Long-term concerns will not be solved within a single project iteration, instead it is to be expected that multiple iterations will be required to close the gaps between state of the art and vision. This alone already requires that results will be easily accessible and open for further work.

Open source solutions should produce the best outcome by harnessing the widest possible range of expertise.

Moreover, open source development can serve as a technology transfer instrument across domains and communities, and encourage wider interoperability between solutions spawning of from the core open technological development. Through community wide support longevity and uptake of the results are further encouraged.

#7 Encourage the Development and Adoption Of Standards

Interoperability will play an increasing role with the growth of CLOUDs across national and technical boundaries, as well as with the growing uptake from different communities and domains. It also expands the potential market for SME-supplied services of general applicability. Along similar lines, portability is a necessity to reduce the vendor lock-in and allow future users to easily switch between providers according to their needs. This all necessitates development and usage of standards in a coordinated fashion.

> **!** *Europe must therefore encourage central standardisation bodies or coordination efforts that ensure commonality between interfaces, protocols, techniques etc. without creating a wide diversity of standards.*

#8 Think Ahead

Some Research Funding should be used to predict ’over the horizon’ requirements and potentially available technologies in order to guide particularly SME service providers and reduce the risk elements of their business plans and to provide a roadmap for large-scale CLOUD infrastructure and platform providers.

C. CONCLUDING REMARKS

Europe can greatly benefit from the CLOUD capabilities, not only in order to reduce management costs, but in specifically by enabling small to medium enterprises to easily offer new, scalable services beyond current limitations, and by establishing a culture in which new types of services can be easily developed, tested and provided. In addition to this, the CLOUD specific capabilities will enable hosting the next generation of platforms appropriate for current (limited) service offerings, and make a scalability possible that is unprecedented.

The research required is not only technological, but also in legal, economic, environmental and standardisation areas.

It should be noted also, that the research results also contribute significantly to related areas of research, including scalable programming (HPC, multicore), future networking, general service provisioning, storage management etc. by providing solutions on the software stack, steering the requirements and compensating deficiencies.

ANNEX A: COMMON TERMINOLOGY

All terms and definitions basing on the “Future of Cloud Computing” report [SCH10] and “the NIST Definition of Cloud Computing” [NIS11].

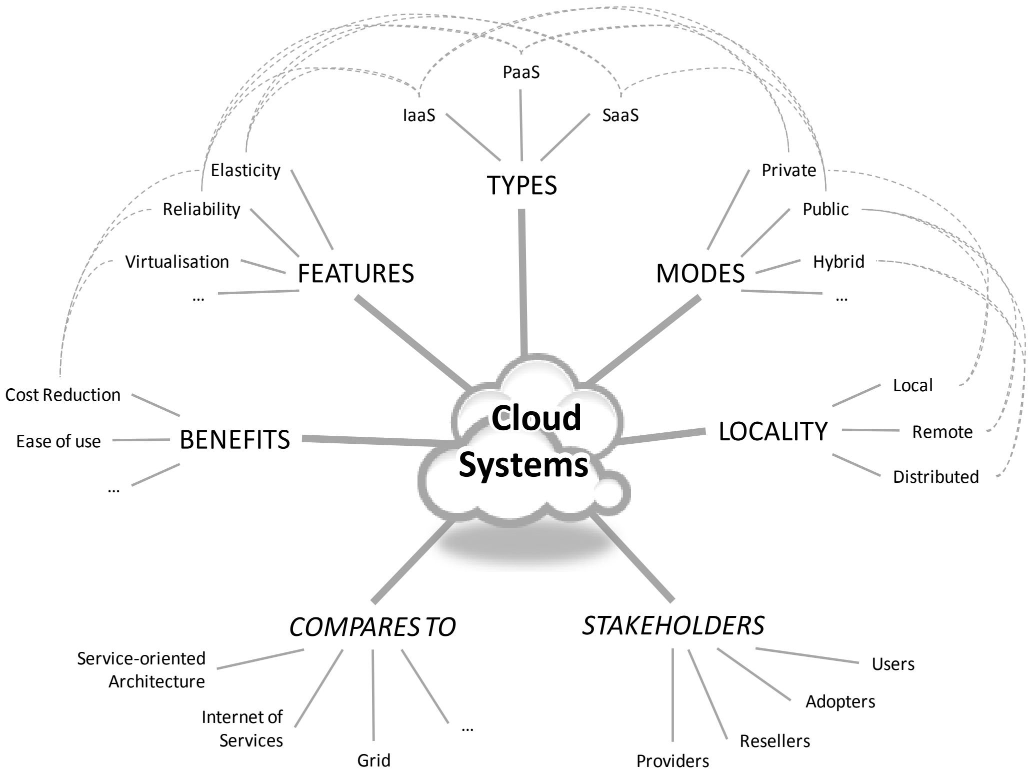

FIGURE 4: NON-EXHAUSTIVE VIEW ON THE MAIN ASPECTS FORMING A CLOUD SYSTEM ACCORDING TO THE 2010 REPORT [SCH10]

A. TYPES OF CLOUDS

(CLOUD) Infrastructure as a Service (IaaS) provide (managed and scalable) resources as services to the user. Accordingly, different resources may be provided via a service interface:

Data & Storage CLOUDs deal with reliable access to data of potentially dynamic size, weighing resource usage with access requirements and / or quality definition.

Compute CLOUDs provide access to computational resources, on which the user can principally host any software. The resources are typically exposed as part of a “virtualized environment” (not to be mixed with PaaS below), in which CLOUDified services and applications can be executed. They offer additional capabilities over a simple compute service.

(CLOUD) Platform as a Service (PaaS), provide computational resources via a *platform* upon which applications and services can be developed and hosted. PaaS typically makes use of dedicated APIs to control the behaviour of a server hosting engine which executes and replicates the execution according to user requests (e.g. access rate).

(CLOUDs) Software as a Service (SaaS), are offering implementations of specific business functions and business processes that are provided with dedicated CLOUD capabilities, i.e. they provide applications / services *using* a CLOUD infrastructure or platform, rather than providing CLOUD features themselves. The user can thereby not control the underlying cloud infrastructure.

B. Deployment Types (CLOUD Usage)

Private CLOUDs are owned by the respective enterprise and / or leased for exclusive use by a single organisation. Functionalities are not directly exposed to the customer, though in some cases services with CLOUD enhanced features may be offered.

Public CLOUDs are provisioned to the general public. Enterprises may use CLOUD functionalities from others, respectively offer their own services to users outside of the company. This also allows other enterprises to outsource their services to such CLOUD providers. The scope of functionalities thereby may differ.

Special Purpose CLOUDs. Most IaaS CLOUDs have a "general purpose" appeal to them, as they can be equally used for a wide scope of use cases and customer types. As opposed to this, PaaS CLOUDs tend to provide functionalities more specialized to specific use cases: specialization implies providing additional, *use case* specific *methods*.

Community CLOUDs are provisioned for the exclusive use of a specific community. These may be, but don't have to be *Special Purpose CLOUDs*, if the CLOUD infrastructure is also adapted in terms of functionalities to the respective community.

Hybrid CLOUDs are any combination of CLOUD deployments as listed above. For example, in some cases, the cloud user wants to retain certain functionalities (services, data) within premises, so that the *Hybrid CLOUD* would consist of a mix of *private* and *public CLOUD* infrastructures so as to achieve a maximum of cost reduction through outsourcing whilst maintaining the desired degree of control.

Meta CLOUDs incorporate multiple cloud infrastructures to allow for the provisioning of meta-services across boundaries and layers. They differ from *Hybrid CLOUDs* in so far, as they aggregate capabilities on all layers to offer enhanced capabilities, independent of the underlying deployment types.

C. CLOUD Environment Roles

(CLOUD) Providers offer *CLOUDs* to the customer. Hosts of CLOUD enhanced services (SaaS) are typically referred to as *Service Providers*.

(CLOUD) Resellers or Aggregators aggregate CLOUD platforms from *CLOUD providers* to provide larger resource infrastructures or to provide enhanced features. This relates to *community CLOUDs* in so far as the CLOUD aggregators may expose a single interface to a merged CLOUD infrastructure.

(CLOUD) Adopters or (Software / Services) Vendors enhance their own services and capabilities by exploiting CLOUD platforms from *CLOUD providers* or *CLOUD resellers*. The CLOUD enhanced services thus typically become *software as a service*.

(CLOUD) Consumers or Users make *direct* use of the CLOUD capabilities (cf. below), not to improve the services and capabilities they offer, but to make use of the direct results.

(CLOUD) Tool Providers do not actually provide CLOUD capabilities, but supporting tools such as programming environments, virtual machine management etc.

Prosumer. Future market developments will enable the user to become provider and consumer at the same time.

REFERENCES

[ASH12] Ashford, W. (2012). Neelie Kroes calls for speedy EU uptake of cloud computing. ComputerWeekly. Available online: http://www.computerweekly.com/news/2240114460/Neelie-Kroes-calls-for-speedy-EU-uptake-of-cloud-computing

[AUM10] Aumasson, A., Bonneau, V., Leimbach, T. & Gödel, M. (2010). Economic and Social Impact of Software & Software-Based Services. Cordis (Online), BE: European Commission. Available online: http://cordis.europa.eu/fp7/ict/ssai/docs/study-sw-report-final.pdf

[BRO10] Brooks, C. (2010). How to build an application for the Cloud. SearchCloudComputing. Available online: http://searchcloudcomputing.techtarget.com/feature/How-to-build-an-application-for-the-cloud

[DGT08] Directorate General for Trade (2008). EU performance in the global economy. European Commission. Available online: http://trade.ec.europa.eu/doclib/docs/2008/october/tradoc_141196.pdf

[EC11] DG INFSO Unit F5 (2011). Workshop on ‘Socio-Economics in Trustworthy ICT’. Cordis (Online), BE: European Commission. Available online: http://cordis.europa.eu/fp7/ict/security/docs/workshop-socio-economics-in-trustworthy-ict.pdf

[ECO10] The Economist (2010). Data, data everywhere. Available online: http://www.economist.com/node/15557443?story_id=15557443

[FIA10] Future Internet Assembly (2010). FIA Ghent Report. Available online: http://www.future-internet.eu/uploads/media/r-fia-ghent.pdf

[FEN08] Fenn, J. & Raskino, M. (2008). Mastering the Hype Cycle: How to Choose the Right Innovation at the Right Time. Harvard Business Press

[FEN11] Fenn, J. & LeHong, H. (2011). Hype Cycle for Emerging Technologies 2011, Gartner. Available online: http://www.gartner.com/DisplayDocument?doc_cd=215650

[JEN11] Jennings, R. (2011). How DevOps brings order to a cloud-oriented world. SearchCloudComputing. Available online: http://searchcloudcomputing.techtarget.com/feature/How-DevOps-brings-order-to-a-cloud-oriented-world

[HLG10] High level Expert Group on Scientific Data (2011). Riding the wave - How Europe can gain from the rising tide of scientific data. Cordis (Online), BE: European Commission. Available online: http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

[KOH07] Koh, Y., Knauerhase, R., Brett, P., Bowman, M., Wen, Z. & Pu, C. (2007). An Analysis of Performance Interference Effects in Virtual Environments. IEEE International Symposium on Performance Analysis of Systems & Software.

[KRO11] Kroes, N. (2011). The clear role of public authorities in cloud computing. Blog Entry. Available online: http://blogs.ec.europa.eu/neelie-kroes/public-authorities-and-CLOUD/

[LON11] Longbottom, C. (2011). Big data: large problems or huge opportunities? ComputerWeekly. Available online:

http://www.computerweekly.com/news/2240105424/Big-data-large-problems-or-huge-opportunities
[MIL09] Miller, R. (2009). How Many Servers Can One Admin Manage? Data Center Knowledge. Available online: http://www.datacenterknowledge.com/archives/2009/12/30/how-many-servers-can-one-admin-manage/
[MON09] Monteagudo, J. & Dierx, A. (2009). Economic performance and competition in services in the euro area: Policy lessons in time of crisis. European Commission, Economic and Financial Affairs, occasional papers 53. Available online: http://ec.europa.eu/economy_finance/publications/publication15841_en.pdf
[MOO08] Moore, S. (2008). Multicore Is Bad News For Supercomputers. IEEE Spectrum. Available online: http://spectrum.ieee.org/computing/hardware/multicore-is-bad-news-for-supercomputers
[NGG06] Next Generation GRIDs Expert Group (2006). Future for European Grids: GRIDs and Service Oriented Knowledge Utilities - Vision and Research Directions 2010 and Beyond. Cordis (Online), BE: European Commission. Available online: ftp://ftp.cordis.europa.eu/pub/ist/docs/grids/ngg3_eg_final.pdf
[NIS10] Badger, L., Bohn, R., Chandramouli, R., Grance, T., Karygiannis, T., Patt-Corner, R. & Voas, J. (2010). Cloud Computing Use Cases. National Institute of Standards and Technology. Available online: http://www.nist.gov/itl/cloud/use-cases.cfm
[NIS11] Mell, P. & Grance, T. (2011). The NIST Definition of Cloud Computing. National Institute of Standards and Technology. Available online: http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf
[PU10] Pu, X., Liu, L., Mei, Y., Sivathanu, S., Koh, Y. & Pu, C. (2010). Understanding Performance Interference of I/O Workload in Virtualized Cloud Environments. IEEE 3rd International Conference on Cloud Computing.
[SCH10] Schubert, L., Jeffery, K. & Neidecker-Lutz, B. (eds). (2010). The Future of Cloud Computing – Report from the first Cloud Computing Expert Working Group Meeting. Cordis (Online), BE: European Commission. Available online: http://cordis.europa.eu/fp7/ict/ssai/docs/Cloud-report-final.pdf
[SUT11] Sutter, H. (2011). Welcome to the Parallel Jungle!. Dr. Dobb's Journal. Available online: http://www.drdobbs.com/parallel/232400273
[WHI11] Whittaker, Z. (2011). Case study: How the USA PATRIOT Act can be used to access EU data. ZDNet. Available online: http://www.zdnet.com/blog/igeneration/case-study-how-the-usa-patriot-act-can-be-used-to-access-eu-data/8805
[ZIM11] Zimmermann, R. (2011). Cloud Computing. FIA Budapest plenary session. Available online: http://www.future-internet.eu/fileadmin/documents/budapest_documens/FIA-Budapest_Consolidated_final_300611.pdf pages 33-36

European Commission

ADVANCES IN CLOUDS – Research in Future Cloud Computing

Luxembourg: Publications Office of the European Union

2012 — 84 pp. — 21 x 29,7 cm