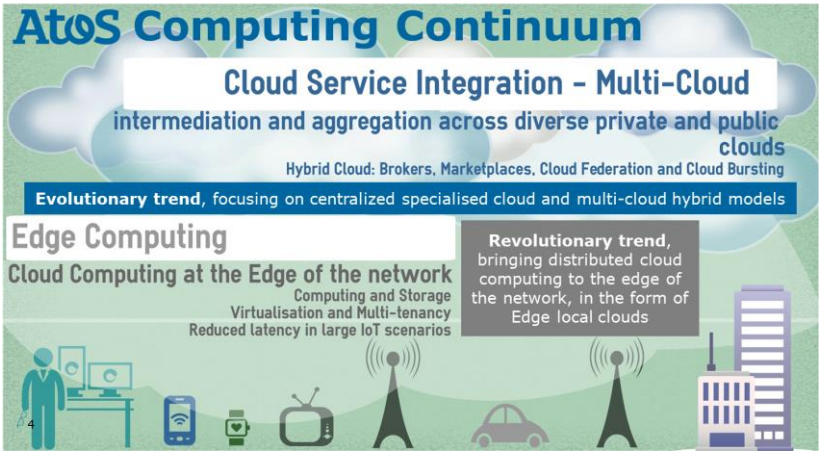
Atos Computing Continuum Vision

Cloud adoption is increasing across all types of organisations. IDC predicts "cloud IT infrastructure spending will reach $53.1B billion by 2019" while identifying that "Over 70% of heavy cloud users are thinking in terms of a "hybrid" cloud strategy". 451 Research's Market Monitor forecasts that "the market for cloud computing, including PaaS, IaaS and infrastructure software as a service will achieve $44.2bB by 2020". Gartner anticipates that "By 2020, over 50% of all new applications developed on PaaS will be IoT-centric".

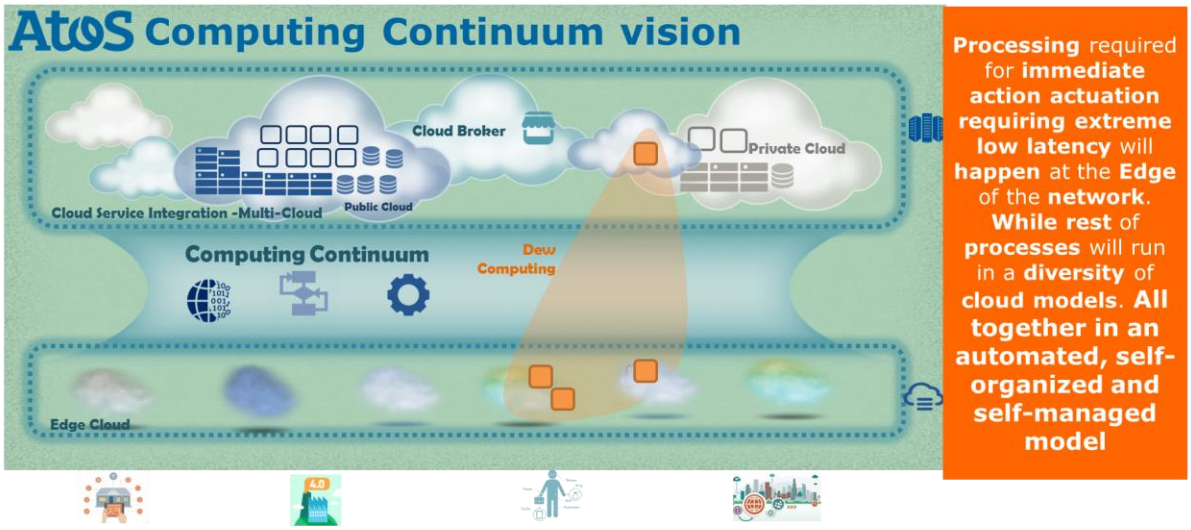
Two trends are emerging in this context, to which the Atos Cloud Computing Continuum vision focus its attention. First, a disruptive approach emerged due to the specific needs of IoT applications enabled by combining network with typical cloud principles in order to create decentralised and disperse cloud platforms, coined under the term "Edge Computing". An secondly, a more evolutionary trend focusing on multi-cloud hybrid models, that although been around now for a number of years, new technological developments together with advances in standardisation efforts and, in general, a more cloud computing cloud market, create the necessary ground for making these a reality in the coming years.



We envisage this cloud computing continuum as the combination of complex multi-cloud architectures with Edge computing. This approach is a digital shockwave for existing computing environments forcing existing offerings that have emerged as part of a centralisation paradigm to evolve to a pure hybrid decentralised and autonomic model. This vision can realise on temporal infrastructures, created on demand in order to respond to a specific need.

Clearly not all the clouds are the same and each cloud provides different capabilities or attributes such as the computing capacity, the end to end latency provided, security and privacy which define the systems that can be hosted. We foresee this Cloud Computing Continuum as a digital infrastructure running edge and multiple cloud platforms in cooperation to run systems and connect entities.

- We believe that Edge computing within a this vision will evolve into a set of computing and storage platforms able to provide low latency and near real time response time with security capabilities focusing on physical entities;
- Within this vision, Cloud Computing will evolve into a set of cloud platforms with high computing power, able to process huge data volumes with complex algorithms to devise the intelligence and deep learning from the data collected by the edge platform. In this context, processing required for immediate action actuation requiring extreme low latency will happen at the Edge of the network. While rest of processes will run in a diversity of cloud models. All together in an automated, self-organized and self-managed approach.



In order to make this vision reality we envision four main research areas to be exploited: Research on Cloud Service Integration, the so-called Multi-Cloud; Research on Edge Computing, as well as, Research in the intersection among these – nowadays tagged as Dew Computing by some authors. These are finally complemented with additional research topics cross-cutting these lines as application field of the Computing Continuum vision. In the following sections we develop specific research topics in each of these areas.

Research Areas

Cloud Service Integration / Multi-Cloud

Improved mechanisms for QoS and SLA Management, These novel mechanisms have to consider Service composition and aggregation in multi-cloud and federation scenarios for Cloud marketplace and brokerage, as well as, legal assessment and security aspects. They will require of deeper monitoring mechanisms at level of application, service, virtual infrastructure and physical devices. These improved mechanisms for SLAs in complex scenarios have to be able to express the following requirements: (i) SLA composition of SLAs from end customers, brokers and cloud providers; (ii) Customer ability to define penalties on SLA violations; (iv) Customer ability to express conditions to agreement renegotiation or cancellation. (iv)Standard Service Level Objective specification to be shared among all stakeholders.

Development of Trust Models in Cloud: Traditionally, relationships between cloud stakeholders have been focused on cost-performance trade-offs, i.e., economical factors. However, in an open and highly dynamic environment of a Cloud ecosystem, relationships will be created on an on-off basis therefore with higher degree of anonymity between stakeholders.. Thus, on one hand, it is necessary to offer methods and tools to quantitatively assess and evaluate cloud stakeholders, e.g., audit and monitoring functions including analysis on service failure probability, risk of data loss, price of service, and probability of SLA violations. On the other hand, methods to measure stakeholder satisfaction are also important, e.g., individual and group perceptions, reputation of stakeholders, or individual previous experiences. Altogether, these mechanisms confirm dependability and reliability between members of the cloud ecosystem.

Scalability: Scalability has to consider heterogeneous and flexible vertical and horizontal scalability in a secure, dependable and auditable provider's ecosystem enabling operations across heterogeneous domains. Application's elasticity and scalability have to provide self-management, heterogeneity at level of resources considered including hardware accelerators, controlling and maximizing the application infrastructure's sustainability and reliability, going beyond oversimplified rules based

systems but also considering applications behaviour in deep, through application QoS and QoE prediction models, data usage analytics and new virtualization forms.

Multi-tenancy / Isolation: Multi-tenancy is a key factor in making the most of a Cloud environments. By supporting multi-tenancy, a platform can achieve economies of scale, a reduction in cost per user, and the ability to share continuous improvements across a wider community. However, it also demands introduction of isolation into its architectural design to ensure that the execution of an application does not have any impact on others with regards to security, performance, availability, and administration. While Cloud computing offers a paradigm-shifting technological solution for computational resources and software, concerns about data privacy and confidentiality, are certainly considered obstacles to the uptake of Cloud-based delivery models. Innovative mechanisms that overcome these concerns are necessary to further advance and uptake of Cloud.

Cloud Migration: Migration refers to the ability to integrate with legacy software and on premise assets (i.e. licensed software), as well as with third-party Cloud services (Cloud Orchestration). A part of considering the development of new services, Cloud providers have to enable the adaptation and combination of legacy and licensed software in addition to the composition of applications within larger contexts, including existing in-house applications and other third-party provided Cloud services. By using general purpose Clouds, applications will be able to execute software in hybrid-Cloud scenarios by means of distributed license validation mechanisms, as well as integration with existing services, enabling enterprise software to exploit the potential of a Cloud ecosystem, based on approaches that include new containerisation/virtualization methods.

Application deployment Automation: the complete automation of provisioning, configuration, and administration of resources is required in order to bridge the existing gap between the development of an application and its operation beyond DevOps. In order enable the co-existence of agile DevOps approaches in combination with Enterprise software definition mechanisms, new capabilities have to be created to deploy and promote application source code, software, and artefacts along the whole application lifecycle (considering different environments, such as development, testing, and production environments). It also has to offer automated deployment and configuration of application binaries, adaptation to use hardware accelerators, and application infrastructure components, such as OS, Application containers, DBMS, Load Balancers, etc. Finally, service provision based on instantiation from templates delivered in an agnostic way has to be provided for multi-Cloud deployments (considering multiple Cloud providers). This automation of application deployment should be achieved by the use of a declarative language at application level, which includes expected QoS and scaling rules. There are some approaches on that direction, (e.g. TOSCA, CAMEL), but they are still far from reaching a full automation covering all previous aspects. The goal of such a language is the homogeneity of a deployment declaration regardless of the underlying cloud (IaaS or PaaS). New intelligent deployment and in runtime re-deployment can be done taking into account the energy used and the way that applications are accessed. Some applications might be only used during day-time, others during the weekend. Algorithms can be developed trying to predict the behavior of the different applications deployed in the cloud, also rules can be set-up of when a deployed application has to be running and on-line. Virtual infrastructure can be re-deployed in the cloud infrastructure in order to minimize the physical machines to be used and automatically turned-off the ones that are not being used in order to improve both costs and energy expenditure.

Improved Cloud Monitoring: Application vs System metrics: nowadays there is a big gap between cloud providers and the user running its application in the infrastructure. Cloud providers' monitor their infrastructure e.g. cpu from physical, virtual machines, bandwidth in the switch, etc. but application developers has the data from the virtual infrastructure and its own metrics. A good performance in the applications in the virtual machine doesn't imply necessary this is happening at application level. Multi-tenancy might produce application performance misbehaviour. E.g a videoconference system might have jitter when the underlying network is saturated. To be able to monitor properly the execution of an application, mechanisms has to be created to be able to relate the application variables with the system variables.

Interoperability: Interoperability among Cloud providers is still an open issue. A wide range of Cloud standardization efforts exist today, these directly reflect the diversity and heterogeneity of acceptations of the Cloud service term. This makes that still comparability across services is a challenge that compromise further advance of multi-cloud service provisioning. In order to enable automatic establishment of chains of contractual relationships across multiple and heterogeneous cloud providers, it will be necessary to analyse and extend existing formalisms to describe Cloud service offers in order to enable service comparability. Comparability will enable further reusability of atomic services by participating in multiple complex cloud service provision, as well as enabling the optimization of cloud service compositions at operation by replacing services that do not provide the adequate QoS performance. Interoperability approaches have to cover all levels of the Cloud stack (IaaS, PaaS and SaaS) , its deployment units, the virtual infrastructure in its diverse forms, application portability and data portability, including metadata adaptation strategies.

Cloud Networking: The advent of SDN brings the idea of a fully scalable, end-to-end, software-based infrastructure. Multi-Cloud Federation requires extensions to the concept, techniques and primitives of cloud networking to embrace a) federated intra- and inter-cloud networking and b) application- and service-aware virtual networking, going beyond offering as a pure infrastructure service that is fully isolated from the cloud value chain it serves. But in addition to these, two additional needs are emerging: First, the demand for reduced time to service; and increasingly, the need to provide improved network management. While server virtualization has very much simplified to move compute loads around and across the clouds; still the management and configuration of the network policies, such as QoS settings, VLAN configurations, access controls and VPN configurations haven't achieved the same level of automatization.

Multi-Cloud Provisioning: There is the need for richer mechanisms for automatic selection of the most suitable Cloud provider. Usually the selection is based on the service requirements and the evaluation of provider's, including decision making mechanisms that will enable the optimal selection between candidate services and their according data for federation. To decide the best choice for the deployment of each service among the different Cloud offers, the algorithm can be optimized based on the type of service to be deployed and the different application execution requirements, such as level of trust in the provider, or based on previous application execution experiences, including eco-efficiency, risk, cost, security levels offered, legal constraints, and quality of service parameters (e.g. availability, performance, operation). A user may even want to launch a cross-site application that will be deployed in different providers at the same time. This can be beneficial for the distribution of the different components of a given application among different Cloud providers. For example, to improve fault tolerance in case of Cloud service disruption, to implement application load balancing features among different sites. This will enable the possibility of expanding the application capacity from one Cloud to another, to implement proximity policies regarding the location of service consumers, or for any other reason (to lower costs or improve security, for better ecological performance, or enhanced performance or security, etc.).

Economy: Cloud computing has been recognized as a bridge between distributed systems and economics. Cloud computing providers offer a number of services to users based on different business models, Cloud pricing schemes are based on incurred resource consumption, known as pay-as-you-go models, although some works recognize lack of fairness for these models due to interferences delivered from the multi-tenancy nature of Cloud Environments. Furthermore, depending on the type of service offered there are options of charging per CPU cycles, I/O writes, Gbs of storage reserved, bandwidth use or combing some of the aforementioned like reserving complete Virtual Machine instances for a period of time (e.g. monthly-based charging). Further research is required to elaborate on pricing schemes and pricing, including dynamic and bind to SLA that include application profiling and real-time resource usage monitoring.

Cloud-marketplaces: Businesses are using cloud, hosting and managed services to facilitate growth, not just cut costs. For business critical applications to move to the cloud, however, significant challenges to widespread adoption continue to remain, mostly around the security, assurance and compliance: data protection, control, availability and resilience. Harmonising levels of service quality and assurance, compliance requirements, supply chain relationships and streamlining procurement and contract management underpins cloud marketplaces and sector specific clouds. In order to build a Cloud market for Europe over a Cloud ecosystem where freedom of choice prevails, stakeholders

compete on differentiation on Service delivery., Businesses can use Cloud services to fulfil processes without compromising assurance and compliance and they maintain visibility and control in governing their business applications and processes in the Cloud. In this vision, European businesses can create high-assurance application and business services efficiently and at lower costs and manage complex supply networks from heterogeneous providers while always staying in control of their assets in the Cloud and meeting sector specific compliance and quality objectives. To develop the key technological and business innovations underpinning this vision by enabling businesses to rapidly assemble, deploy and operate their business application services in the Cloud. This is predicated on achieving and maintaining desired levels of trust, transparency and governance for Cloud assets. To achieve this, innovations have to enable businesses to enhance their applications and services on private or public Clouds of their choice with pre-accredited capabilities addressing service qualities or business critical functions of a generic nature and/or of significance for specific vertical markets.

Eco-efficiency: Cloud computing has received considerable attention, as a promising approach for delivering ICT services by improving the utilization of data centre resources. In principle, cloud computing can be an inherently energy-efficient technology for ICT provided that its potential for significant energy savings that have so far focused on hardware aspects can be fully explored with respect to system operation. Further research it is required at levels of:

- Energy monitoring, including new metrics and KPIs, scheduling optimisation, consideration of storage energy consumption as well as application management on energy efficient hardware accelerators.
- Maximize the integration of renewable energy sources: The inclusion of renewable energy coming from different types of sources is still and open challenge nowadays, the intelligence and control mechanisms that manage the resources deployed on top of a cloud may need to take into consideration this information to adapt (if possible) the computational load over time to accomplish a better utilization of the renewable sources when they are available.
- QoS for Energy Efficiency Data Centers: Taking into consideration the energy efficiency goals that a Cloud provider aims to reach, new SLA terms and negotiation of enforcements will need to be considered to satisfy the requirements imposed in the agreements between the different entities. Not only new agreement terms need to be considered here but also guarantee that the monitoring services available in the Cloud provide enough information to enable continuously monitoring of these agreement terms.
- Holistic management of Cloud workload considering thermal and power aspects. We use to see workload, thermal and power management systems working independently in a Cloud infrastructure. Top layer mechanisms capable to optimally coordinate the information collected from each of the management systems need to be considered in order to be able to provide recommendations or actuation policies for a better energy efficient management of the Cloud Infrastructure. In addition, to coordinate these three management pillar within the Cloud, it is also important to understand the operational scale (virtual resource, server, group of servers, rack, room) of each of these management systems and normalize the exchange of information between them, if required.

Edge (Fog) Computing

Edge(Fog) Management and heterogeneity: Management of potentially thousands/millions of small diverse devices and sensors in an Edge computing set-ups will require of new management styles, potentially decentralized and able to scale to degrees that nowadays are unprecedented in existing cloud architectures.

Across Edge execution orchestration: Edge set-ups are envisaged to be spread covering wide geographic areas. For serving applications and services that make use of these distributed set-ups, mechanisms for deployment, provisioning, placement and scaling service instances across execution zones in the distributed Edge set-ups are necessary.

Security and privacy: Edge computing, similarly to traditional cloud, is envisaged as multi-tenant, and therefore actual set-ups will require of specific isolation mechanisms so to avoid security and privacy concerns.

Interoperability and standardization: Current status of Edge computing developments very much relies on specific vendor solutions. In order for these to interoperate among them and with traditional clouds, new standards would have to appear to manage the expected scale of edge set-ups and the interoperability of devices and sensors.

Off-loading optimization: Edge computing aims at using cloud computing techniques on storage and processing of data mobile devices. Edge it is used to overcome limitations with regards constrained hardware resources, battery lifetime and connectivity issues, etc. that aim by means of off-loading computation techniques to extend limited resource capabilities. These considers diverse architectures that range from off-loading to specific service, to traditional datacentre clouds; to local Cloud or Cloudlets; and mobile devices, trying to take advantage of the ever growing mobile devices capacity. Diverse works consider off-loading techniques, four models are considered: off-loading to a Server, off-loading to virtual infrastructure in a pre-defined Cloud provider, off-loading to a dedicated infrastructure located in the vicinity and off-loading to another subrogate mobile device. The dynamicity of this selection and possibility of portability of off-loading workloads across diverse of these infrastructures is not yet considered by existing works as well as, QoS service mechanisms enabled by trustworthy mechanisms.

Edge computing communication topology: usually every mobile component at the edge will have direct communication (e.g. 3G, 4G, HSPA) with the upper computing level to offload their processed data. Depending on where we place the mobile component it might not have option to have a direct communication with an upper level, but it might have communication with other components nearby. Most devices nowadays are able to set-up its own Wi-Fi, Bluetooth. The mobile devices without direct communication can take profit and establish communication using another edge component that will act as a bridge to offload the information. New mechanisms to dynamically auto-configure communication topology and make it fault tolerant should be created.

Support for heterogeneous and geographically distributed systems: Modul-based approaches and reuse of existing technologies (libraries, frameworks and tools) are encouraged. The solutions will need to support different type of architectures topologies including those based on heterogeneous processors. Geographical implications need to be taken also into account from legal and security perspectives.

DeW Computing: Where the Edge meets the Cloud

Edge / Fog Service Management: Fault tolerance, availability and performance are service management aspects still to be addressed in Edge computing. These are of specific interest due to the nature of devices and its volatility it in addition to this need of considering additional aspects for QoS management and heterogeneity in resources, integration of special devices including hardware accelerators, FPGAs and GPUs, that considers frequent loss of connectivity and low bandwidth. Additional research of availability issues related to ad-hoc and distributed clouds together with the scalability problem, associating it to ability of these systems to sustain the provision of services and resources for large number of clients.

Admission Control Similarly to some aspects of Edge Service Management, still Admission control has to be further developed in Edge computing. The mechanisms to decide whether to accept or not a service to be executed on a infrastructure in Edge computing has an additional complexity not traditionally considered in Cluster and Cloud computing systems. Given that one cannot assume that mobile resources will be available. Admission control, in this context, would have to take into account analysis of historical information on resource in order to determine if a service can be accepted, and which would be the most adequate placement among available resources taking into account service execution requirements and resource heterogeneity. Management of volatility of mobile resources and the availability issues derived from this fact is the main challenge. Besides, additional research has to consider limited and highly demand resources and mobility of applications and users.

Data abstractions and movement in Cloud and Edge computing: The need of developing data intensive applications able to manage more and more data coming from distributed and heterogeneous sources effectively, quickly, correctly, and securely is increasing. Cloud computing paradigm – as it is now adopted – is not fully appropriate to store and analyse such data. Indeed, although scalability and reliability are properties that make cloud-based storage favourable, latency, security, and compliance are hampering the adoption of this type of data-centre centric approach. At the same time, Fog Computing has emerged as a paradigm promising to fully exploit the potential of the edge of the network involving traditional devices as well as new generation of smart devices, which can process data closer to where they are produced and/or consumed. In this scenario, developers are more and more interested in developing data-intensive applications, but managing data scattered on a heterogeneous and distributed environment needs to deal with the intricacies of the underlying complex infrastructure composed by smart devices, sensors, as well as traditional computing nodes. Conversely, developers must be focused on defining, which are the relevant data, their format and quality, and how to process them, without dealing with details such as how to gather data, where to store or process them.

Edge Workload management: Encapsulation of monolithic and interactive workloads on top of heterogeneous systems: Considering different types of workloads (monolithic or interactive) as well as the different processors types where these workloads can be computed, the final encapsulation solution may vary. System able to deal with different encapsulation approaches (Virtualization vs. Container) will be required to prepare the workloads depending on the final execution environment. Mechanisms capable to balance between high-performance processor and low power processor according to the final objectives of the workload should be taken into consideration.

Development areas for Computing Continuum, Mavericks Cloud Research

Swarm management among IoE, Edge and Cloud Computing: Swarm intelligence refers to the combined behavior of decentralized, massively distributed and self-organized systems. With the number of devices comprising the IoE foreseen to achieve unprecedented figures, and making the consideration that many of these individual devices will have very limited compute capabilities, it emerges the idea of Intelligent Swarm management. In which each limited devices (resources) would be complimented by connection to other objects in a community, therefore creating dynamic eco-systems encompassing cyber-physical devices, edge and clouds, each of these adding to the collective capability and insight. This will enable service provisioning to shift from the existing pre-defined and fixed orientation to a pure on-demand, opportunistic and ad-hoc approach. In this novel approach, applications will rely on dynamic, limited, distributed resources, which are governed under different domains and may have possibly conflicting interests. This approach aims to take advantage of all available highly heterogeneous resources (i.e. edge, clouds, communicating objects, sensors and smart devices) so to provide a service-based environment that allows for collecting, dynamically creating and managing these massive, diverse, unconnected and highly distributed resources. These distributed service networks are envisaged as opportunistic, which are created on-the-fly in order to respond to an specific need, coordinated in an open and distributed self-* runtime model. In this context, business models that regulate the participation of different entities in the service networks will be necessary together with the definition of the necessary mechanisms that incentive participation. Other aspects to be taken into account are the following: reservation, adaptive selection, conflict resolution and techniques to consider the volatility and uncertainty introduced due by real-world dynamics should be considered to enable efficient and reliable service provision.

Cyber-physical Cloud/Edge computing: Cyber-physical systems (CPS) have been traditionally designed in order to make use of resources available in-house. Increasingly CPS, although keeping the strict properties of safety and mission critical, are looking at Cloud and Edge solutions, seeking for reduction of cost and enhanced scalability. The so-called Cyber-physical cloud computing aims to address these needs by providing "a system environment that can rapidly build, modify and provision cyber-physical systems composed of a set of cloud computing based sensor, processing, control and data services". The emergence of this trend raises new challenges, both for CPS and Cloud computing technologies. On one side, factors such as: dynamic aggregation / disaggregation of behaviors, consolidation, economic incentives, auto-scaling of resources would have to be considered in CPS design and operation. On the other side, Cloud and Edge computing would require to advance on specific aspects such as QoS management, reliability and fault management, data-clustering, security, safety, ethics and real-time support in order enable the specific need of CPS. In addition to this, specific emphasis would have to be done the need of integration of computing accelerators and data analytics tools in Edge and Cloud in order to cope with specific data processing capacities required by CPS, Edge and Cloud interaction.

Cloud Robotics: Cloud Robotics, and generally speaking Smart Machines, is an area that increasingly is getting attention from researchers and industry worldwide. The initiatives such as RoboEarth and Open source robots reflect this fact. Advances in machine learning, artificial intelligence, image recognition and speech demonstrate that today knowledge work is on its way to be automated. This builds the necessary ground for more and more smart machines to participate in our everyday life and in industrial processes while, at the same time, they become better, faster and cheaper, according to Moore's law. Robots, in order to perform a task such as moving a part, recognizing or grasping an object require a significant amount of processing power as well as contextual and stored data and information. This is even expected to increase with the advent of more sophisticated systems like humanoid-like robots, which will require of carrying powerful computing units and large batteries to power them. The foreseen involvement of Robotics into the business dynamics of at least one-third of the industries in the developed world in the future will require of bringing together Cyber-physical Cloud Robotics to expand their capacities with computing power at the Edge and Cloud. In such environment, physical world of smart machines and robots will need of cyber offloading capacities in the Cloud thus offering computing power, data and knowledge. This will require of the Edge/Cloud to offer dynamic and real time response to fast processing (i.e. for simulation and optimization tasks in image recognition processes), as well as, access to vast amounts of data for simulation and algorithm training. The consideration of Robot as a service or as a Cloud Resource in this environment will bring additional challenges to the intersection between Cloud and Robotics.