Artificial Intelligence-based 5G Network Capacity Planning and Operation

J. Pérez-Romero, O. Sallent, R. Ferrús, R. Agustí
Dept. of Signal Theory and Communications
Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain
e-mail:[jorperez, sallent, ferrus, ramon]@tsc.upc.edu

Abstract— **The highly demanding requirements envisaged for future 5G networks together with the required support of new customers from vertical industries (e.g. e-health, automotive, energy) pose a big challenge for operators in 5G on how to balance investments, user experience and profitability. There will be the need to revisit the actual methodologies of network planning and operation, fully exploiting cognitive capabilities that embrace knowledge and intelligence to achieve a proper understanding of the network usage in multiple dimensions. In this respect, this paper presents a vision on how these planning and operation processes can rely on the inclusion of Artificial Intelligence (AI) concepts that will allow devising models to characterize the impact of many correlated inputs on specific operator objectives and to drive decisions for different processes.**

Keywords—5G; artificial intelligence; capacity planning; network operation

I. INTRODUCTION

As a next step in the evolution of mobile communication systems, research carried out by industry and academia is nowadays focused on the development of the new generation of mobile and wireless systems, known as 5th Generation (5G) that targets a time horizon beyond 2020. 5G intends to provide solutions to the continuously increasing demand for mobile broadband services associated with the massive penetration of wireless equipment such as smartphones, tablets, the tremendous expected increase in the demand for wireless Machine To Machine (M2M) communications [1], and the proliferation of bandwidth-intensive applications including high definition video, 3D, virtual reality, etc.

To cope with the abovementioned demands, requirements of future 5G system have been already identified and discussed at different fora. Requirements identified in [2] specify 1000 times higher mobile data volume per area, 10 times to 100 times higher number of connected devices, 10 times to 100 times higher typical user data rate, 10 times longer battery life for low power devices and 5 times reduced End-to-End latency [3]. In similar terms, in [4] different key challenges are established, such as providing 1000 times higher wireless area capacity and more varied service capabilities compared to 2010, saving up to 90% of energy per service provided, reducing the average service creation time cycle from 90 hours to 90 minutes, creating a secure, reliable and dependable Internet with a "zero perceived" downtime for services provision, facilitating very dense deployments of wireless communication links to connect over 7 trillion wireless devices serving over 7 billion people, and enabling advanced user controlled privacy. Finally, the Next Generation Mobile Networks (NGMN) has also set up the 5G initiative [5] to develop the end to end Mobile Network Operator (MNO) requirements to satisfy the needs of customers and markets beyond 2020.

In addition to extending the performance envelope of mobile networks, 5G is also envisioned to include by design embedded flexibility to optimize the network usage thanks to the inclusion of Software Defined Networking (SDN) and Network Function Virtualization (NFV) technologies. This will allow supporting current and new business models involving different types of customers and partnerships [5]. Specifically, 5G is expected to allow MNOs to better support customers from a number of vertical industries (e.g., e-health, automotive, energy). Moreover, 5G is expected to facilitate the establishment of partnerships on multiple layers, ranging from sharing the infrastructure (e.g., network sharing relationships among MNOs, delivery of Infrastructure as a Service, Platform as a Service, Network as a Service by assets providers), to exposing specific network capabilities as an end to end service and integrating partners' services into the 5G system through a rich and software oriented capability set. Facilitating such partnerships is believed to be pivotal for the deployment of 5G because MNO's large investments in CAPEX and OPEX nowadays are not being followed by sufficient revenue increase.

Furthermore, 5G networks will be fuelled by the advent of big data and big data analytics [6]. The volume, variety and velocity of big data are simply overwhelming. Nevertheless, nowadays there are already tools and platforms readily available to efficiently handle this big amount of data and turn it into value by gaining insight and understanding data structures and relationships, extracting exploitable knowledge and deriving successful decision-making.

The 5G challenge for MNOs is very much about how to balance investments, user experience and profitability. In this respect, and sustained on the pillars of network flexibility and data volume handling as key technology enablers, two main areas of improvement are identified: (1) The methodologies used for network capacity planning, which can result in significant CAPEX savings and (2) The methodologies used for network operation, which can result in significant OPEX savings. This paper claims that there is the need to revisit the actual methodologies of network planning and operation, fully exploiting cognitive capabilities that embrace knowledge and intelligence. Concerning (1), a proper understanding of the

This work has been supported by the Spanish Research Council and FEDER funds under RAMSES grant (ref. TEC2013-41698-R).

978-1-4673-6540-6/15/$31.00 ©2015 IEEE

network usage in multiple dimensions (time, space, type of user, type of application, etc.) will result in better investment decisions regarding how the network is expanded. Concerning (2), it can rely on the capability to smartly and dynamically analyse the huge amount of data that the MNO will have available and on increasing the degree of automation, by making the network more self-autonomous and intelligent with respect to how the actual network is parametrised.

In this context, this paper supports the idea that Artificial Intelligence (AI) mechanisms, which intend to develop intelligent systems able to perceive and analyse the environment and take the appropriate actions, will fully fertilise in the 5G ecosystem. While many seeds can be found in the literature both from an academic/theoretical perspective (e.g., connected to the so-called Cognitive Networks) and from a practical perspective in current 3G/4G networks (e.g., connected to the so-called Self-Organising Networks), a truly intelligent network means much more than this and the 5G era is the proper time for AI-based networks to happen.

Based on the above considerations, this paper intends to present a vision of how the different planning and operation procedures of future 5G networks can be built upon the pillar of AI concepts. For that purpose, the paper will discuss in Section II the general functionalities of the proposed AI-based framework, while Sections III and IV will provide details on the applicability of this framework to the network capacity planning and operation processes, respectively. Finally conclusions will be summarized in Section V.

II. AI-BASED FRAMEWORK FOR 5G NETWORK PLANNING AND OPERATION

Fig. 1 illustrates the components of the AI-based framework for 5G network capacity planning and operation envisaged here. The framework operates based on input data from its environment and smartly processes it in order to come up with the appropriate actions to be executed on the network and/or to assist MNO's decisions through appropriate insights and decision support systems.

A. Data Acquisition and Pre-processing

MNOs have traditionally operated with complex, disparate sets of data, with useful information residing in multiple systems such as customer relationship management systems, network management systems, billing, inventories, network elements, service management systems, deep packet inspection devices, application-specific databases, etc., [7]. Therefore, the challenge for an efficient 5G network planning and operation is to smartly analyse and correlate all these different data sources. In general, gathered input data can belong to different categories:

- Network data: It characterizes the behaviour of the network. Some examples include radio related measurements made by terminals and base stations, usage statistics at the different network nodes and routes, network performance indicators, Quality of Service (QoS) measurements, network configurations, etc.
- User data: It characterizes the profiles of the subscribers accessing the network. Examples include demographics, billing and subscription plans, devices and capabilities, used applications, etc.
- Content data: It characterizes the type of information associated to the applications that use the network.
- External data: This includes data from outside the MNO systems, such as information services (e.g., weather forecast, airport flights, road traffic, public transportation services, football game schedules, etc.), social networks, etc. These sources can be relevant to extract conducts or patterns from the users in front of specific events, which could eventually impact on the mobile data traffic requirements for the network at a given point and space.

The proposed framework relies on the application of data mining techniques over the collected data in order to distil all the available information and identify meaningful models and patterns that will drive the subsequent decisions. In this respect, the first step to be executed is the data pre-processing, where the data coming from multiple heterogeneous sources is prepared for mining. This is done through different tasks that include: data cleaning to remove noise and inconsistent data; data integration to combine multiple data sources; data selection to choose the relevant data for each specific analysis; and data transformation where data are transformed or consolidated into forms appropriate for mining by performing e.g. summary or aggregation operations [8].

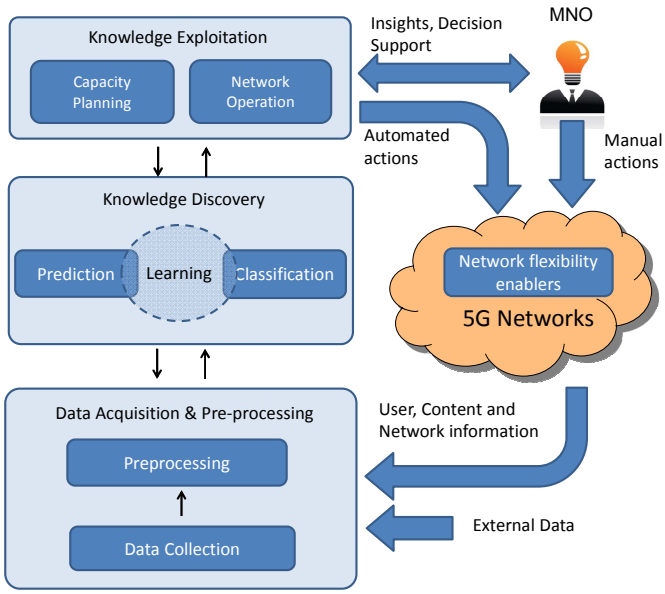

Fig. 1. AI-based framework for 5G network planning and operation

B. Knowledge Discovery

The Knowledge Discovery stage performs inference on the pre-processed data in order to build models that reflect the relevant knowledge that will drive the operation and planning decisions. The core of this stage is the use of machine learning tools to build such models.

The goal of machine learning is to build computer systems that can adapt and learn from their experience [9]. Machine learning techniques are usually subdivided into three big categories. The first one is supervised learning that consists in

learning from training examples provided by an external supervisor in the form of pairs of inputs and desired outputs, so it is not valid for interactive problems where the desired behaviour is not known. The second category is the unsupervised learning, consisting in learning to represent particular input patterns (e.g. independent samples of an underlying unknown probability distribution) in a way that reflects their statistical structure. The third category is the Reinforcement Learning (RL), which consists in learning the behavioural model of the environment (i.e. the network) through the dynamic interaction with it. The model outputs are translated into actions executed on the environment that provides as a result a reinforcement signal. This reinforcement signal uses to be a reward that encodes the success of the action's outcome, so the learner seeks to learn the model that relates inputs and outputs to maximize this reward. There exist in turn different categories of RL mechanisms, namely dynamic programming, Monte Carlo methods and Temporal-Difference (TD) learning. Among them, TD methods are particularly appropriate in the framework considered here as they do not need a model of the environment dynamics and can update the decision making policy without waiting for the final outcome of a number of actions. Some examples of these TD methods include Q-learning, Sarsa and Actor-critic RL [10].

The application of machine learning tools leads to the identification of models or algorithms that can be applied to perform specific functionalities based on the input data, usually organized in tuples. Each tuple is described by a set of attribute values (e.g. a set of features of the network at a given time). Two of these functionalities that are expected to play a key role to support the planning and operation of 5G systems are classification and prediction.

Classification is the process of finding a model or function that describes and distinguishes data classes or concepts [8]. This model (i.e. the classifier) is then used to determine the class or category to which a given tuple belongs. If the possible classes are known in advance, the classifier can be obtained from a supervised learning algorithm that analyses a set of training tuples with known classes are known. In contrast, if classes are not predefined or if there are not training tuples with known classes, a clustering process using unsupervised learning techniques is used to derive the possible classes.

Classifiers can be represented in various forms depending on the used technique: (i) Decision Tree Induction techniques consist in finding a flow-chart structure where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes. They adopt a greedy approach in which decision trees are constructed in a top-down recursive divide-and-conquer manner, starting from a training set which is recursively partitioned into smaller subsets as the tree is being built. (ii) Bayesian classification is a statistical technique that evaluates the probability that a given tuple belongs to a class based on the attributes associated to this class [11]. (iii) In Rule-based classification the classifier is represented by a set of if/then rules obtained from decision trees or directly from the training data. (iv) Fuzzy Logic extends rule-based classification by allowing "fuzzy" thresholds to be defined for each class. Then, the classification is given in terms of a value between 0 and 1 that represents the degree of membership that a certain attribute value has in a given category. (v) Classification by backpropagation is based on neural networks, i.e. a set of connected input/output neuron-like processing units with weighted connections between them. The weights are learnt by iteratively processing a set of training tuples comparing the neural network's output with the actual known target value or by applying RL if no training data is available. (vi) Support Vector Machines (SVM) [12] use a nonlinear mapping to transform the original training data into a higher dimension. Within this new dimension it searches for the linear optimal decision boundary to separate the tuples of the different classes. (vii) The k-nearest neighbour (k-NN) classification techniques [13] are based on comparing a certain test tuple with a set of training tuples, with the purpose of finding the set of k training tuples (or neighbours) that are closest to the test tuple.

Prediction intends to find models to anticipate future values of a certain parameter (e.g. predict the evolution of the traffic demand in a certain geographical area based on identifying some regular trends of this demand or of other correlated indicators, etc.). Prediction can cover also the spatial domain by extrapolating predictions to other geographical areas with similar characteristics or profiles.

Prediction models in the time domain usually exploit the trend analysis of input data in terms of four major components, namely long-term movements that indicate the general direction in which a time-series graph is moving over a long interval of time, cyclic movements that refer to oscillations about a trend line, seasonal movements that are systematic or calendar related, and irregular or random movements that characterize the sporadic motion of time series due to random or chance events. Models such as Auto-Regressive Integrated Moving Average (ARIMA) or Seasonal ARIMA to include seasonal factors are usually applied [14] for capturing these components. Support Vector Regression (SVR) is a relatively novel and promising prediction methodology in which a function is estimated using observed data which in turn trains a SVM [15]. It is particularly useful in predicting time series generated from non-linear systems, as it does not make any assumption on the underlying model of the time series, and it does not need prior knowledge about existing seasonalities.

C. Knowledge Exploitation

This stage will make use of the extracted knowledge (e.g. the predictions of the traffic in time/space, the identified behaviours of the users, etc.) to decide a set of actions and/or recommendations to be applied to the network. In the framework of this paper, these actions are related with both the network planning and the network operation processes. The network planning refers to the provision of the necessary network resources (e.g. base stations) to cope with the existing traffic demand in accordance with the strategic objectives of the MNO. In turn, the network operation processes include the actions executed on the deployed resources in order to control and optimize the network behaviour. Examples of these operation processes include the configuration of the resources allocated to a given node, the activation or deactivation of certain nodes in the network, the configuration of the spectrum to be used by a network node, etc. Depending on the degree of automation for each considered process, the results of this stage

can be directly (autonomously) implemented in the network (e.g. in case of a parameter reconfiguration) or they can just be presented as recommendations for decision support to the MNO who will finally make the corresponding manual action (e.g. in case that a new small cell needs to be provisioned to cope with a traffic increase in a certain area).

D. Network flexibility enablers

The introduction of the AI-based framework should be facilitated by the use of the SDN and NFV technologies, which are becoming foundational components within 5G.

NFV refers to the software implementation of network functions running on general purpose computing/storage resources [16]. Therefore, NFV brings flexibility to modify network configuration and/or topology in near real time. Moreover, NFV enables infrastructure sharing between multiple tenants (i.e. MNOs) that could lead to multiple virtual networks running on top of the same infrastructure and satisfying the profile and requirements of each MNO separately. In turn, SDN consists in decoupling network control plane from the data plane, which enables the use of software to control network functions and policies [17]. SDN involves the centralization of network intelligence in software-based controllers, facilitating decision-making based on a global view of the network. As such, it enables the implementation of the network control and optimisation functions as applications running on top of an SDN controller that provides a programmatic interface (northbound interface) to the network. In turn, the SDN controller will communicate through the southbound interface carrying out the control of the different network devices. Different architectures such as OpenFlow, CAPWAP or CloudMAC are being considered for supporting SDN functionalities [18][19]. The capability of SDN to relocate functionalities from hardware into software by providing configurable interfaces increases the programmability of the hardware and enables the implementation of NFV.

The adoption of SDN and NFV technologies allows faster and easier network deployment, configuration, and update of network functions. As a result, it is an ideal means to gather massive amounts of input network data and to quickly enforce the decisions made by the AI-based framework.

III. AI-ENHANCED PLANNING PROCESSES

An accurate planning of the necessary network resources (e.g., base stations) is a difficult process. Planning methodologies involve numerous parameters subject to significant uncertainties, variability in time and space. Traditionally, this has led to overprovisioning and, consequently, excess CAPEX. The vision for 5G is that there is tremendous room for enhancements through the formulation of more advanced and intelligent processes.

Models can be derived to conduct a decomposition of the expected traffic demand in time, space and QoS levels. In time and space dimension there will be fluctuations on the required capacity, sometimes associated to more predictable conditions (e.g., seasonal variations in certain areas -winter/summer-, events -with various durations ranging from hours to several days-, changes between working days and weekends, day and night, and even intra-day variations due for example to traffic jams) and sometimes associated to less predictable or unpredictable conditions (e.g., large disasters, emergency for a traffic accident or black spots, social groups behaviors such as demonstrations, weather conditions causing closed airports or closed roads, incidences in public transportation network such as traffic jam caused by an accident, traffic diversion due to work on roads, etc.). Regarding QoS levels, the traffic decomposition will exploit correlation with the space and time dimensions as there can be inferred an association with user groups (e.g., in an area with lots of pubs for the youth, one can expect during the happy hour a high traffic associated to lower QoS profiles). Traffic characterisation can also lead to mobility patterns to extract information about daily travels,

The planning of the network to cope with a certain traffic demand can be sustained not only on deploying owned infrastructure, but more advanced and dynamic elements can be considered, such as small cells as a service (SCaaS) (i.e., small cells owned, deployed and operated by the new player), dynamic RAN sharing (i.e., arrangements between multiple players are done to share the capacity of a network) or nomadic cells (i.e., small cells on wheels/rails, owned either by the MNO or by a third party).

At the same time, in light of the more advanced spectrum management models, which in general comprise licensed, light licensed and unlicensed components, the planning process in the 5G context involves as well the provisioning of the spectrum resources to be exploited at a given time and location. Spectrum trading (leasing and/or transfer of any licensed/light licensed spectrum usage rights) with other MNOs, through secondary markets or spectrum brokers/aggregators may also be a relevant mechanism to exploit when pursuing efficiency in 5G networks.

Classification techniques applied up to cell level according to the type of traffic that it is being generated there and prediction techniques, anticipating the evolution of the traffic demand will be pillars in these novel methodologies. The knowledge extracted (e.g., accurate prediction of future –either at day, week, month, year scale- spectrum needs) can be applied to spectrum trading and obtain the necessary spectrum resources at lower prices. In some situations, clustering strategies can be applied to identify groups of cells that exhibit similar characteristics, leading to the classification of the target geographical areas. Then, the accurate prediction of the traffic profile at the area level can provide a clear guidance of the most suitable deployment strategy (i.e., relying on SCaaS, deployment of a nomadic cell, etc.) from a more global perspective. Forming judgments or subjective beliefs about the likelihoods of certain outcomes or the frequencies of certain events (e.g., the likelihood that the traffic demand at a certain area will be high) will also support the decision making on the proper deployment strategy (or mix of strategies).

Prediction could also be applied to the spatial domain by extrapolating predictions to other geographical areas with similar characteristics or profiles (e.g. predicting the impact of a new tram line in a given city by extrapolating what happens in the actual tramline areas of another city, etc.).

IV. AI-ENHANCED OPERATION PROCESSES

Once the network resources have been provisioned, MNOs

need to monitor and control them to ensure its optimized behavior in relation to desired MNO objectives. This will require the analysis of the huge amount of data generated by the network, to devise models that assess how these objectives are being fulfilled and drive the required corrective actions.

By correlating traffic and subscriber information, the MNO can achieve a better understanding of the load behavior in the different cells. Classification models applied to these data can be used to categorize the cells depending on their radio conditions, available backhaul technologies, load and service requirements in different periods of time, while predictions will allow assessing the expected evolution of these load conditions on a cell basis. Based on these learnt models smart dynamic load balancing approaches can be implemented, extending the range of unloaded cells to take traffic load away from a neighboring overloaded cell based on profitability of subscribers, traffic pattern, location, etc. Similarly, classification applied to both cells and users can be used in real time to perform the association of users to cells or technologies depending on their subscription levels, the applications they are using, the traffic loading on the different cells and the cost of getting traffic to or from its end point.

Classification can be used to categorize the performance of the deployed cells for different combinations of radio measurements and network counters observed along the cell lifetime. For example, one could identify which combinations of signal strength, SINR, etc. and load levels have led in the past to bad/medium/good service performance in a given cell. This will allow anticipating future bad performance situations when these combinations are detected, and proactively taking the necessary actions before problems occur.

The application of clustering mechanisms can be useful to group cells that exhibit similar characteristics, leading to the identification of geographical areas in which configuration actions applied at the area level, i.e. jointly to a set of cells, can be more effective to solve bad performance situations than individual solutions applied at the cell level.

When groups of users move, the distribution of network load can change dramatically from one cell to the other, leading to congestion situations in certain cells. This situation can be critical in highly populated scenarios. Anticipation of these overload situations can be obtained by deriving models that correlate the mobility patterns, external events and radio measurements. Through these models, the AI-based framework can identify the actions to be executed (e.g. offloading certain users to other cells/technologies, accessing unlicensed spectrum to extend the capacity in some specific cells and at some specific periods of time, etc.).

The result of the abovementioned processes can lead to recommendations for the MNO on specific actions to be executed on the network. In addition, thanks to the flexibility offered by SDN/NFV in the 5G network, which allow fast and easy configuration and updating of network functions, a higher degree of automatism can be expected where some of these decisions are autonomously enforced on the network. In the latter case, the control of the proper operation of these autonomous systems becomes itself a challenging task for the MNO who needs to assess the behavior of each solution and to identify possible malfunctions that could severely impact on the network behavior. This can be achieved in the framework of Fig. 1 through the capability of learning models that correlate reconfiguration orders with network measurements and service performance indicators.

V. Concluding Remarks

This paper has presented a vision of how the capacity planning and operation procedures of future 5G networks can rely on the application of AI concepts in order to cope with the associated complexity and challenging requirements of these networks while at the same time reducing the CAPEX/OPEX incurred by MNOs. A framework has been presented that processes input data from very different sources to extract, through learning-based classification and prediction models, the relevant knowledge that should drive the planning and operation decisions. Details on the applicability of this framework to different processes have been discussed.

References

[1] Ericsson White Paper "More than 50 billion connected devices", February, 2011, http://www.ericsson.com/res/docs/whitepapers/wp-50-billions.pdf

[2] METIS 2020 project, http://www.metis2020.com

[3] M. Fallgren, B. Timus, (editors), "Scenarios, requirements and KPIs for 5G mobile and wireless system", Deliverable D1.1. of the METIS project, May, 2013.

[4] The 5G Infrastructure Public Private Partnership, http://5g-ppp.eu

[5] R. El Hattachi, J. Erfanian (editors) "NGMN 5G White Paper", NGMN Alliance, February, 2015

[6] Ericsson, "Big Data Analytics", White paper, August, 2013.

[7] A. Banerjee, "Advanced Predictive Network Analytics: Optimize your Network Investments and Transform Customer Experience", White Paper, Heavy Reading, February, 2014.

[8] J. Han, M. Kamber, "Data Mining Concepts and Techniques", 2nd edition, Elsevier, 2006.

[9] R.A.Wilson, F.C.Keil, The MIT Encyclopedia of the Cognitive Sciences, MIT Press, 1999.

[10] R.S. Sutton, A. G. Barto, Reinforcement Learning: An Introduction, MIT Press, 1998

[11] N. Friedman, D. Geiger, M. Goldszmidt, "Bayesian Network Classifiers", Journal Machine Learning, Vol. 29, No. 2-3, November/December, 1997, pp. 131-163.

[12] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition", Data Mining and Knowledge Discovery, Vol. 2, No. 2, June, 1998, pp.121-167.

[13] N. Bhatia, "Survey of Nearest Neighbor Techniques", International Journal of Computer Science and Information Security, Vol. 8, No. 2, 2010.

[14] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, "Time Series Analysis: Forecasting and Control", 3rd edition, Prentice-Hall, 1994

[15] N. I. Sapankevych, R. Sankar, "Time Series Prediction: Using Support Vector Machines: A Survey", IEEE Computational Intelligence Magazine, May, 2009.

[16] ETSI GS NFV 002 v1.1.1 "Network Functions Virtualisation (NFV); Architectural Framework", October, 2013.

[17] B. A. Nunes, M. Mendonça, X-N. Nguyen, K. Obraczka, T. Turletti, "A Survey of Software- Defined Networking: Past, Present and Future of Programmable Networks", IEEE Communications Surveys and Tutorials, February, 2014.

[18] H. Wen, P.K.Tiwary, T. Le-Ngoc, Wireless Virtualization, Springer, 2013

[19] ONF, "Software-Defined Networking: The New Norm for Networks", White Paper, April, 2012.