

Chapter 1

Architectures and Cross-Layer Design for Cognitive Networks

Dzmitry Kliazovich* and Fabrizio Granelli†

DISI, University of Trento

Via Sommarive 14, I-38050 Trento, Italy

**granelli@disi.unitn.it*

†kliazovich@disi.unitn.it

Nelson L. S. da Fonseca

Institute of Computing, State University of Campinas

Av. Albert Einstein, 1251, Campinas SP, Brazil

nfonseca@ic.unicamp.br

Network evolution towards self-aware autonomous adaptive networking attempts to overcome the inefficiency of configuring and managing networks, which leads to performance degradation. In order to optimize network operations, the introduction of self-awareness, self-management, and self-healing into the network was proposed. This created a new paradigm in networking, known as cognitive networking.

This chapter describes state-of-the-art, as well as future directions in cognitive networking. Fundamental techniques for enabling cognitive properties, such as, adaptation, learning, and goal optimization processes are detailed in this text. A comparison of available research proposals leads to the design of a promising cognitive network architecture capable of incorporating cognitive network techniques. Finally, a discussion on the required properties of the cross-layer design for cognitive networks and deployment issues are specified.

1.1. Introduction

Requirements and expectations on network services have constantly evolved. The evolution of communication technologies, especially in the wireless domain, introduced a paradigm shift from static to mobile access, from centralized to distributed infrastructure, and from passive to active networking.

Technological advances have brought networking a step forward towards the goal of service provision on an “anytime, anywhere” basis, while ensuring instantaneous and secure communications. However, such innovation is bound by the constraints included in the original Internet (and TCP/IP) design, which does not include, for

example, mobility support, security, and active networking. For this reason, technological advancements were achieved at the cost of increased network complexity and limited performance.¹ The fundamental reason for performance inefficiency is the difficulty in configuring and managing networks — a task traditionally performed by network operators and technicians.²

Self-awareness, self-management, and self-healing characteristics have been proposed in order to optimize network operation, reconfiguration, and management, as well as to improve data transfer performance by bringing “intelligence” into the network, thereby creating a new paradigm known as *cognitive networking*, which is expected to become a key part of 4th generation wireless networks (4G).³

The term *cognitive* is related to the ability of a network to be aware of its operational status and adjust its operational parameters to fulfill specific tasks, such as detecting changes in the environment and user requirements. Cognition requires support from network elements (routers, switches, base stations, etc.), which should host active tasks to perform measurements to reconfigure the network. These characteristics are related to the paradigm of *active networks*,⁴ which differ from cognitive networks service in that they do not include a *cognitive process* that considers adaptation and learning techniques.

The main challenges in cognitive networks range from the limitations of wireless technologies to network complexity, heterogeneity, and Quality of Service.

Network complexity is a function of the number of nodes and alternate routes, as well as the number of communication mediums and protocols running in the network. The introduction of wireless links in the network increased its complexity, since it changed the notion of connectivity.

Wireless nodes communicate over radio channels, which are subject to frequent fading and signal interference. In wireless networks, nodes can join and leave the network in ad hoc manner. Furthermore, mesh type connections can be established.

In addition, mobility allows wireless terminals to dynamically change their location, as well as their point of attachment to the network core. Mobility affects path availability, making it difficult to reach stability in a reasonable timeframe. Consequently, network management and optimization must add functionalities for self-healing.

Heterogeneity: The Internet is composed of combinations of different transmission technologies and a large variety of applications and transmission protocols (see Fig. 1.1). However, there is no layer in the TCP/IP that accounts for heterogeneity.

One of the most widely used approaches for performance improvement is the division of a connection into segments, each optimized for a particular domain. Despite several attempts to address heterogeneous paths in the Internet,⁵ application performance will remain poor until the deployment of such capability on the majority of existing domains.

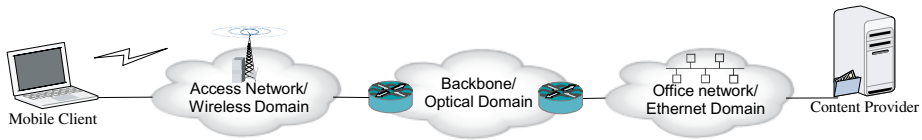


Fig. 1.1. Heterogeneous network path.

Enhancing performance requires awareness of underlining transmission technologies over the entire path between sender and receiver nodes. The optimization process should be distributed across multiple domains and optimization should achieve the goals defined at the connecting ends.

Quality of Service (QoS): Limited provisioning of delay bounds and bandwidth unavailability require the implementation of different reservation control mechanisms to support QoS requirements of different applications and users. Firoiu *et al.*⁶ provides a comprehensive survey of available QoS mechanisms from the technical and the business perspective. No complete QoS solution can be performed within a single protocol layer.^{6,7}

The two main driving forces for cognitive networks are technological and business oriented.

From the technological perspective, cognitive networking is envisioned as a logical evolution towards the definition of a unified QoS-aware environment, encompassing multiple technologies already available in the wireless network domain. The diversity of network configurations, involved technologies, and objectives dictated by the requirements of user applications is the main motivation behind cognitive networking.

From the business perspective, cognitive networks are envisioned as the way to increase profits for wireless service providers through cost reduction and development of new revenue streams obtained by the offer of heterogeneous wireless access solutions. The benefits enabled by cognitive networking include: the possibility to rely on common hardware and software platforms while supporting the evolution of radio technologies, development of new services, minimization of infrastructure upgrades, accelerated innovation, and maximization of return-on-investment through the reuse of already available network equipment.²

The business case offered by cognitive technologies provides network operators with the possibility to continuously analyze the configuration and the performance of a network segment, aiming at efficient service provision. Moreover, reconfiguration can be triggered by application requirements, policies, or billing plans. Cognitive networking offers extended sets of operation choices, creating new ways for interactions between network operators and end-users.⁸

The rest of this chapter provides fundamental concepts on cognitive networks. Novel cognitive network architecture is proposed after a discussion on current

research trends. Finally, cross-layer techniques for improving current proposals on cognitive networks are introduced.

1.2. Fundamental Concepts on Cognitive Networks

Interest in cognitive networks gained momentum in the research community just some years ago. One of the first works that underlined the relevance of the concept of cognitive networks dates back to August 2003.² Clark *et al.*² proposed a network knowledge plane with cognitive techniques, such as “representation, learning, and reasoning that allow the knowledge plane to be “aware” of the network and its actions.”

The following subsections define the cognitive network concept and related fundamental concepts.

1.2.1. Definition

The term *cognition* (from Latin, *cognoscere*, “to know”) is used in many disciplines to describe the phenomenon closely related to the concepts of knowledge, intelligence, and learning. In networking, cognition is primarily motivated by system complexity and difficulty to use simple decision-making elements, such as those based on the closed form system models.

The principle of cognitive networking was conceived in association with cognitive radio. Mitola *et al.*⁹ discussed the possibility of the adaptation of a network of cognitive radio devices where decisions are based on the knowledge obtained through the use of the unsupervised learning process.⁹ They defined the cognitive cycle as consisting of six processes: observation, orientation, planning, learning, decision making, and action. According to the authors, a cognitive system is considered to . . .

“employ model-based reasoning to achieve a specified level of competence in radio-related domains.”

As illustrated in Ref. 10, such processes can envisage a reconfigurable network node with observation and action functionalities, while the remaining functionalities (such as orientation, planning, decision making, and learning) could constitute a “cognitive engine.” As a result, a reconfigurable node constitutes a platform for the implementation of cognitive networks. Observations and actions should be flexible and bring the required degree of freedom for the design of cognitive elements. However, the precise concept of cognitive networking was first introduced by Clark *et al.*² who defined:

“the knowledge plane: representation, learning, and reasoning that allow the knowledge plane to be “aware” of the network and its actions in the network.” (by Clark et al.)

The knowledge plane should be designed to function properly in the presence of incomplete or misleading information, account for different high-level goals, and operate with future network technologies. These requirements cannot be satisfied using simple closed form expressions demanding complex or “cognitive” techniques. This definition is considered as one of the first for cognitive networking, due to its explicit reference of such notions as “knowledge” and “awareness” within a network-wide scope.

The scope of operation is what distinguishes a cognitive network from other systems, such as a network of cognitive radios. Cognitive radios focus on the optimization of wireless channel(s) access, thus limiting the scope to the node’s vicinity, while cognitive networking aims at network-wide optimization and end-to-end network-wide goals. The PhD dissertation of Thomas¹¹ introduces new goals to the definition given by Clark²:

“... a network with a cognitive process that can perceive current network conditions, and then plan, decide and act on those conditions. The network can learn from these adaptations and use them to make future decisions, all while taking into account end-to-end goals.”

The end-to-end user or application-defined goals are achieved through adaptation and cognition that involve all network elements (routers and switches) and communication techniques from the physical to the application layer across the data path.

1.2.2. Cognitive network fundamentals

The fundamental techniques enabling cognitive properties of networking algorithms can be summarized as the following functions: observation, analysis, decision making, and action.

In Fig. 1.2, these functionalities are presented as elements of a pyramid. The functional elements located closer to the pyramid foundation are typically more distributed across the network. For example, observation elements could simply keep track of node physical characteristics (such as signal level) and variables (such as the size of TCP windows). Alternatively, action elements could perform simple operations, such as tuning Network Interface Card (NIC) transmission parameters. These functional elements are typically passive, while most of the intelligence is located closer to the top of the pyramid. In this way, analysis and decision-making functions receive feedback obtained from the observation and issue action commands to the action elements.

Consequently, functional elements located closer to the bottom of the pyramid can be widely spread in the network, while those located closer to the top of the pyramid are commonly less spread.

Another important aspect of functional elements of cognitive networks is the property of scaling. For example, their behavior and implementation trend

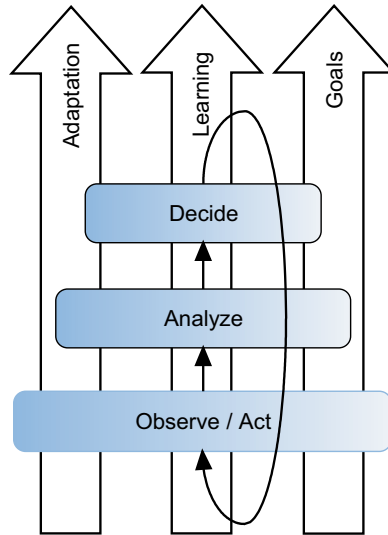


Fig. 1.2. Cognitive process essentials.

(centralized or distributed) is repeated on different scales, either at node or at network level. For example, at node level, observation and action elements correspond to intra-protocol layer agents able to control internal layer parameters, while the analysis or decision making process run as more generic processes, typically in parallel to the protocol stack.

At network layer, network nodes as a whole could be considered as observation and action elements, in which case measurements gathered at different protocol layers are aggregated at the node and sent as a single report to analysis. They can also be considered a decision-making engine, which, in the case of hierarchical implementation, provide aggregation and localization at network layer before passing information to more global entities.

Additionally, the pyramid of cognitive network functional elements is driven by three planes: adaptation, learning, and goals.

Adaptation is an essential part of all functional elements, enabling themselves to respond to changes in their environment. For example, the observation element could be requested to perform measurements of certain parameters within the protocol stack of the node and report it to the analysis module in a given interval of time. Then, the adaptive observation element could adjust this interval based on the operation scenario and network setup.

Commonly, small reporting intervals are used for parameters with frequently changed values, while large reports are considered for more static parameters. The observation element controls the reporting interval, which is the first instance of measurement.

However, high frequency of reporting increases signaling between observation and analysis elements. In this case, both observation and analysis are performed at the same network node and messages are transferred internally, causing overhead which could be neglected, while analysis is performed at another network node, signaling consuming relevant network resources.

The Learning plane is the core of the cognitive system. Learning enables functional elements of the pyramid to perform different actions based on past experience. An essential part of the learning process is the feedback adaptation loop, enabling the cognitive system to study reactions to the performed actions through observation, analysis, and decision making, yielding conclusions that are stored in the system and that can influence future decisions.

Two classes of learning can be performed by the cognitive system: informal and formal. Informal learning is implicit and can be performed by the same cognitive cycle on a regular basis during its normal operation. An example of such learning is the cognitive cycle implemented in a wireless node that increases the transmission rate when it perceives performance degradation. This is considered knowledge obtained by an informal learning procedure.

Formal learning is the process of knowledge transfer directly from the entity which obtained the knowledge. An example of formal learning in cognitive networks is that obtained by network nodes by gathering, aggregating, and broadcasting to other nodes of the network. Alternatively, it could be a simple exchange of knowledge between neighboring nodes.

The Goal plane provides a set of objectives that should be reached or approached by the cognitive optimization process. As suggested by Thomas *et al.*¹² the goals should be end-to-end and have network-wide scope. Optimization goals can be defined in a network-centric way by the network operator based on a chosen business model. They may correspond to personal preferences of an end user or can come from user application requirements (QoS). It is important that the goal plane resolves the conflicts between goals by taking into account their diversity and priority level and by providing a set of objectives using high-level languages.¹³ The specification of goals for cognitive network is a difficult task and it is currently poorly addressed in literature. The questions that should be answered are: How are the rules described? Who (regulatory authority, network operator, users, applications, etc.) specifies them? How goals are made conflict-free? How fast are these goals disseminated in the network?

In summary, the adaptation and the goals planes take observation, analysis, and the decision making process into account, in order to choose a proper set of actions to adjust the basis of cognitive process for achieving specified network goals. Furthermore, the learning process is responsible for the reconfiguration of cognitive process elements based on experience and future prediction.

Another important point is the existence of a quality control feedback channel, which for every optimization step should perform an analysis of resulting network

performance, compare it against the set of targeted objectives, and confirm that the optimization process is progressing in the right direction. Conversely, the cognitive process should backtrack; in the case of actions, it should verify that the system is one step behind its objectives. This process can be repeated at different scales in the network.

Cognitive network management should include well-defined operational metrics, allowing control of the optimization process and its possible backtracking procedures.

1.2.3. *Similarities to cognitive radio*

There has been some source of misunderstanding about the scope of cognitive networks and the scope of cognitive radios. Cognitive radio provides efficient and dynamic spectrum access by adaptively changing transmission and reception parameters to avoid interference with other communication systems. According to Ref. 14, the main functions of a cognitive radio are: *spectrum sensing*, aimed at detection of primary users and available spectrum “holes”; *spectrum management*, in order to select the best-suited frequency channel among spectrum holes; *spectrum sharing*, for fair coordination of spectrum usage with other cognitive nodes; and *spectrum mobility*, for leaving spectrum usage, in case primary licensed users are detected.

The main motivation behind cognitive radios is the need to enable dynamic spectrum sharing (as opposed to the traditional fixed allocation policy), while facing growing demands for high data rates and as a result, more flexibility and automation in the management of the available spectrum. Additional restrictions on the operational spectrum band are related to the nature of electromagnetic waves propagation and limitations on the size of antenna and mobile terminals.¹⁵ Practically speaking, cognitive radio technologies are designed to overcome this limitation by enabling operations in the portions of spectrum sensed to be free from other transmissions.

Cognitive radios focus on the optimization of spectrum usage for increasing spectrum utilization, both at the physical layer (spectrum sensing) and at the link layer (scheduling and coordination). Even when management involves a joint effort of several cognitive radio devices, optimization is pursued in the vicinity of a node.

Moreover, the scenarios targeted by cognitive networks are far more complex and network-intensive, including several communication technologies (wire line, optical, and wireless), as well as different network elements (nodes and routers), which may or may not be designed to cooperate. In this scenario, optimization procedures are distributed across different layers of the protocol stack and across different networks.

Therefore, the main difference between cognitive radios and cognitive networks lies in scope of the optimization performed; while cognitive radios operate locally

at the radio link level, cognitive networks aim at end-to-end optimization.¹² This consideration underlines the network-wide scope of a cognitive network, separating it from other local adaptation and optimization approaches.

However, the main common denominator between cognitive radios and cognitive networks is the definition of a cognitive process based on observations and measurements for optimizing reconfiguration of operating parameter.

1.3. Current Research Initiatives on Cognitive Network

Research on cognitive networks was embraced by several efforts, both in the United States¹⁶ and in Europe, being relevant within the European Sixth and Seventh Research Framework Programmes (FP6¹⁷ and FP7¹⁸).

Initially, research projects in cognitive networks considered Beyond-3G (B3G) network architectures, given the full control of the network core and the ease of including additional functionalities. The research projects in this category are E²R¹⁹ and m@ANGEL platform.²⁰

The E²R (End-to-end Reconfigurability) project¹⁹ capitalizes on the benefits of Next Generation Network (NGN) and exploits a wide range of network technologies, such as cellular, fixed or WLAN. The ultimate goal of E²R is an all-IP network fully integrated with reconfigurable equipment.²¹ However, the assumption of simultaneous reconfigurability support at all the layers for all the involved actors/devices represents a drawback and limits its incremental deployment.

The m@ANGEL platform²⁰ introduces a special approach for solving mobility problems in heterogeneous network environments with the support of cognition. The cognitive process is considered to be implemented in the access part of the network, between base stations and mobile users. The structure of the access network consists of two planes: the infrastructure plane, which includes reconfigurable elements (such as hardware transceivers, base stations, and the network core) and the management plane, composed of m@ANGEL entities. Each m@ANGEL entity is responsible for monitoring, resource brokerage, goals management, and reconfigurable element control functionalities.

A certain degree of cooperation is considered between m@ANGEL elements. However, this cooperation is usually performed within the scope of network elements located in neighboring cells and it is not propagated to the network core, somehow limiting the scope of a unified solution.

Differently than the B3G-focused approaches, researchers from Trinity College of Dublin presented a general framework for implementing the cognitive functionality.¹⁰ This work focuses on node architecture enabling reconfigurable properties, implying logical separation between network nodes and the cognitive engine running in the network. While the cognitive engine performs learning, orientation, planning, and decision-making functions, observation and action are left to the reconfigurable node.

Node reconfiguration can be requested by the cognitive engine and performed by the Stack Manager component, which is the core of the reconfigurable node architecture. The stack manager builds a customized protocol from the layer components provided by the Component Inventory. Layer components are the software modules implementing functionalities of an entire protocol layer or a part of the layer (like a digital modulator, for example). They aim at interconnection with other layer components and communication with the stack manager.

This approach relies on the techniques to make the cognitive node capable of modifying or adjusting its protocol stack as a function of the dynamics of network environment. Moreover, the logical separation of the cognitive network primitives, such as learning or decision making outside reconfigurable nodes (into cognitive engine), can potentially limit the benefits from local optimization and can increase the amount of signaling traffic.

CogNet (Cognitive Complete Knowledge Network) is an ongoing research project funded by NSF.²² It proposes new cognitive network architecture designed to maintain layered abstraction of TCP/IP protocol stack. In CogNet, each protocol layer is extended with so-called Intra-layer Cognition Modules, which are software agents performing intra-layer monitoring, control, and coordination functions. Modules are interconnected through the Cognitive Bus, part of the Cognitive Plane, to coordinate the cognition modules and are implemented in parallel to the protocol stack.

A unique property of the proposed architecture is the fact that the cognitive functions implemented in intra-layer cognitive elements are distributed between different protocol layers. Such design simplifies the cognitive processes running in the network and reduces signaling overhead.

However, the performance of the proposed architecture seems to be highly dependent on Cognitive Plane operation, which is responsible for translation of end-to-end goals into objectives and configuration parameters at each layer. Consequently, the lack of proper coordination or intra-layer cognitive agents monitoring could lead to unpredictable performance results.

The CogNet project develops the IEEE 802.11²³ and CDMA²⁴ testbeds for gaining understanding and identifying relationships among different parameters at different layers in a real operation environment.

The cognitive network model proposed by Thomas *et al.*¹² is composed of three horizontal layers. The top level is responsible for specification and translation of user/application requirements into goals understandable by cognitive process.

Several cognitive processes can run in the immediate plane, with implementation potentially distributed between several network nodes. The cognitive process involves learning, knowledge, and decision-making and operates when information about the network is limited.

The bottom layer of the model corresponds to the Software Adaptable Network (SAN), consisting of modifiable network elements and sensors. The communication between modifiable elements and the cognitive plane is performed using the

software adaptable network API. Such an architectural solution brings modularity and flexibility into the design of modifiable elements.

Gelenbe *et al.*²⁵ proposed the idea of cognitive packet networks, which basically moves routing and flow control capabilities from network nodes into packets. Such packets, called cognitive packets, “route themselves” and learn to avoid congestion and avoid being destroyed. Each cognitive packet contains a cognitive map and a piece of code that is executed every time the packet arrives at the network node (router). Routing decisions are taken relying on the cognitive map, as well as mailbox messages left by other packets or by the network node.

The idea of cognitive packets bears similarities to the concept of active networking,²⁶ related to custom code execution. However, a unique feature of cognitive packets is their ability to change their behavior based on the state of the network.

Another approach to overcoming limitations of traditional IP networks was presented by Lake *et al.*²⁷ The Software Programmable Intelligent Network (SPIN) merges concepts from IP, PSTN, cellular, and ad hoc networks for overcoming the fundamental limitations of IP networks (such as in-band signaling and impact of long and nested feedback loops on network performance).

SPIN architecture consists of three planes interconnected by layer-2 transport infrastructure:

- *Forwarding plane*: This plane is responsible for switching and monitoring and it can provide connectionless packet forwarding, connection oriented packet forwarding, tag switching, and label switching. Additionally, it performs active and passive measurements, such as packet loss, jitter, bandwidth, and one-way latency.
- *Control/management plane*: It manages forwarding plane devices targeting data, forwarding optimization based on the received measurements. This plane also has the advantage of physical separation from the forwarding plane, including high availability, reliability, and fault tolerance.
- *Cognitive plane*: This plane resides on top of control/management and forwarding planes, providing intelligence for and administration of the entire system. It operates multiple functions dedicated to performing single tasks, including schemas for optimal routing and load balancing, as well as managing responses to legacy control protocols.

A brief comparison of the cognitive network proposals overviewed previously is provided in Table 1.1. The following parameters were used in the comparison: consistency with TCP/IP, stack reconfigurability, cognitive process, network support, and goals.

Consistency with TCP/IP gives the degree of modifications required to work on a standard layered TCP/IP protocol reference model; for example, how easy it is to deploy the proposed approach. Most of the approaches are not “TCP-friendly,” mainly due to the reliance on reconfigurable elements of the protocol

Table 1.1. Cognitive network research proposals comparison.

Approach	Consistency with TCP/IP	Reconfigurable elements	Cognitive process	Required level of support from the network	Goal
E ² R	No	Entire protocol stack	Centralized or partially distributed	High	All-IP B3G network with seamless mobility
m@ANGEL	No	Lower protocol stack layers	In access network	High in access network	Resource allocation and mobility management in heterogeneous networks
Sutton <i>et al.</i> ¹⁰	No	Entire protocol stack	No a part of reconfigurable node	Moderate	Reconfigurable node architecture with dynamic protocol stack
CogNet	Yes	Intra-layer modules	Distributed (at the node level)	Moderate	Design of generic cognitive network framework
Thomas <i>et al.</i> ¹²	No	Reconfigurable elements and network sensors	Distributed (at the network level)	Moderate	Combines properties of IP and PSTN network architectures using cognition
SPIN	No	Separate cognitive/management plane	Distributed in the network	High	

stack. The only TCP-friendly approach is CogNet, which was specifically designed to maintain TCP/IP layer abstraction. This is achieved by introducing a cognitive network interface at each protocol layer, which ensures smooth interaction between internal elements to the layer functionalities with the cognitive engine.

Reconfigurability specifies the required degree of reconfiguration needed by the proposed cognitive network framework. Solutions such as m@ANGEL, furnish reconfigurability at the lower protocol stack layers close to hardware. As a consequence, incremental deployment in existing networks is possible, while other proposals like E²R, Sutton *et al.*,¹⁰ and SPIN, require entire protocol stack reconfigurability.

Cognitive process implementation ranges from centralized to distributed implementations. Centralized implementations are able to provide better control and optimization properties, while distributed ones lead to reduced operational complexity and more failures. Most of the proposals combine centralized and distributed implementation for the cognitive process, attempting to achieve an optimal trade off.

Required level of support from the network means that the cooperation from different network elements (switches and routers) is required by the approach in order to work properly. Approaches like E²R, m@ANGEL, and SPIN rely on a high level of cooperation from network elements, while other solutions like Sutton *et al.*,¹⁰ CogNet, and Thomas *et al.*¹² reduce the level of requested cooperation to a moderate level. In any case, it is clear that cognitive network frameworks tend to break end-to-end Internet principle by adding intelligence to the network core rather than keeping it at the end nodes.

Goals involve configuration, optimization of data flow, and its performance metric.

1.4. A Reference Cognitive Network Architecture

Presented in this section is a proposal for cognitive network architecture, shown in Fig. 1.3, derived as a combination of the key concepts from other research initiatives previously presented. The objectives are to maintain consistency with the TCP/IP protocol stack, to be simple in managing reconfigurable elements, to have a distributed cognitive process, to need a minimum level of network support, and to optimize end-to-end performance.

All these imply that cognitive network elements should be implemented in a transparent and incremental way to the existing protocol stack.

To operate with a standard protocol stack, each protocol layer is enhanced with a small software module able either to obtain internals to the layer information (observation) or to tune internal parameters (action). The information sensed at the protocol layers is delivered to the cognitive plane implemented at the cognitive node. This cognitive plane runs data analysis and decision making processes.

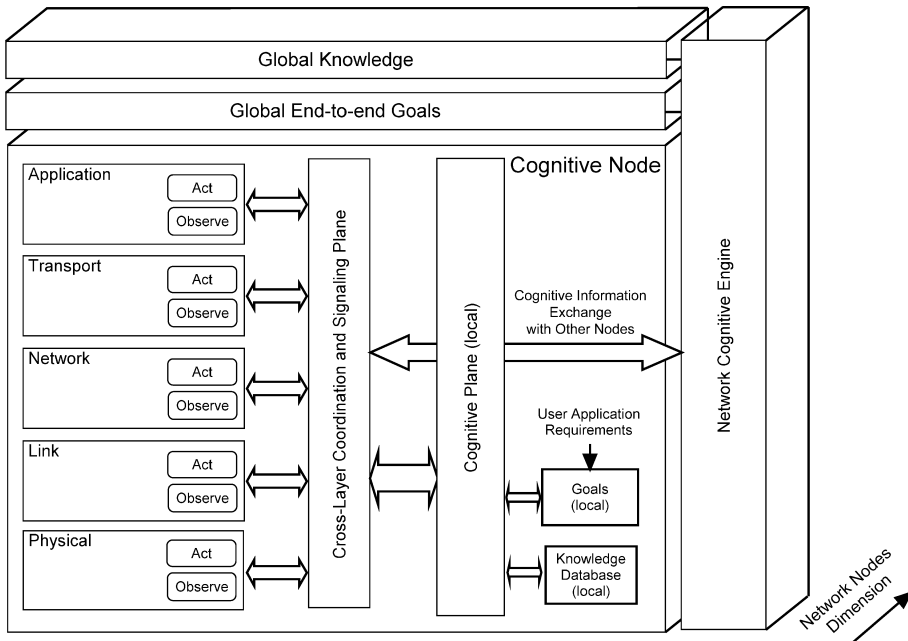


Fig. 1.3. Cognitive network framework.

Results of data analysis could lead to information classified as knowledge, storable in the local knowledge database.

The decisions made by the cognitive plane at the node aim to optimize the protocol stack and are driven by the Goals specified in the local database. The scope of these goals is local (at node level). Most of them are generated by the demands and QoS requirements of user applications running at a given cognitive node.

While goals and knowledge databases are directly connected to the cognitive plane of the node and allow instant information exchange, the cognitive plane communication with the protocol stack is performed by the Cross-layer Coordination and Signaling Plane (CCSP). CCSP is responsible for providing a proper way for signaling information delivery. An example of such functionality is the immediate delivery of parameter values. Another example is the use of a predefined threshold. Completely different signaling methods are required for parameters at different layers associated with a particular packet traveling. In summary, CCSP should provide optimal signaling information delivery and interconnecting elements of the cognitive node architecture.

The proposed architectural concepts are related to cognitive process running in a single network node. However, as defined in previous sections, the main property of cognitive networking is its network-wide scope. Therefore, the network of cognitive nodes is driven by a Network Cognitive Engine (NCE), which is capable of

communicating with cognitive planes of different network nodes coordinating and managing them. NCE is responsible for harvesting cognitive information available at cognitive nodes. This information includes local goals of the node and applications demands, the knowledge obtained by the node, or direct values of specific protocol stack parameters. NCE information harvesting, which corresponds to the observation function, could be performed on a scheduled basis or by using instant requests. Moreover, information could be node related or parameters associated with a particular flow transmission.

The analysis of information gathered from cognitive nodes helps the NCE to construct global knowledge and goals after definition and upon every adjustment, which are reported back to cognitive nodes so that they can adjust their appropriate local databases and, as a result, their behavior.

A main characteristic of the cognitive network architecture is scalability, assured by the use of a combination of centric (at node level) and distributed (at network level) techniques. In particular, at node level the core cognitive techniques, such as data analysis, decision making, and learning, are concentrated in the cognitive planes of the nodes and implemented in a centralized manner. Furthermore, observation and action software add-ons to the protocol layers serve only as instruments and cognitive planes are typically “non-intelligent” ones. This constitutes one of the main differences of cognitive network architecture presented in Ref. 28, which is that it adopts cognitive processes inside a single protocol layer. Distributing cognitive process among the protocol layers (especially the learning and decision making functions) would require complex algorithms for synchronization and coordination between intra-layer cognitive processes. Alternatively, it seems that a single centralized cognitive process at node level brings a simpler solution, while implementation of cognitive process at network layer (CNE) must be distributed or clustered implemented.

A similar approach is considered for aggregation of all kinds of signaling data. Data such as that observed at the local protocol stack, user application requirements, and knowledge obtained by cognitive plane are aggregated at node level before being delivered to the CNE.

1.5. Cross-Layer Design for Cognitive Networks

An essential element of cognitive network architecture presented in the previous section is the cross-layer coordination and signaling plane, which is recognized for providing information exchange between the cognitive primitives, observation, and action performed, either at the node or across the network. For example, in Ref. 28, the availability of the Cognitive Bus, which is responsible for signaling information exchange between different layers of the protocol stack, is considered. However, communication techniques implemented by the Cognitive Bus are left out of the scope of this text.

In the following sections, an overview of the available approaches for cross-layer signaling and techniques most appropriate for cognitive networking is given.

1.5.1. *Cross-layer design proposals*

Information exchange between two or more layers of the protocol stack raises important issues concerning the implementation of different cross-layer solutions inside the TCP/IP protocol reference model, their coexistence, and interoperability.²⁹

The principles behind implementation of common cross-layer signaling models are directed towards rapid prototyping, portability, and efficient implementation of cross-layer entities, while maintaining TCP/IP modularity.³⁰

In this framework, several cross-layer signaling architectures have been proposed:

Interlayer signaling pipe is one of the first approaches used for implementation of cross-layer signaling,³¹ allowing the propagation of signaling messages layer-to-layer along with packet data flow. Signaling information, included in an optional portion of packet headers, follows the packet processing path to another in the protocol stack, either in a top-bottom or a bottom-top manner.

An important property of this signaling method is that signaling information inserted into a packet header can be associated with this particular packet either at the ingress or at the egress path of the protocol stack.

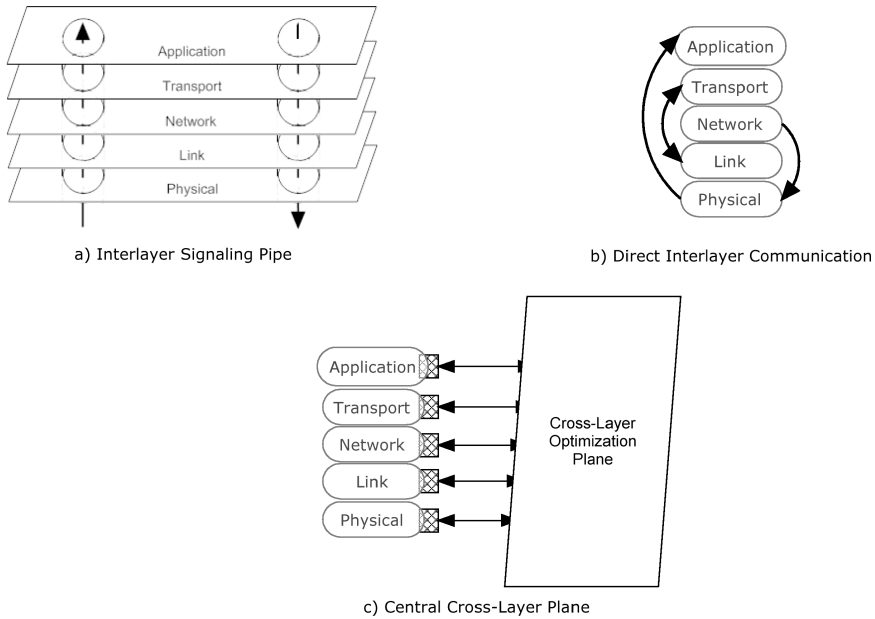
The main disadvantage of the Interlayer Signaling Pipe method is the limitation of the propagation of the signaling information to the direction of the packet flow, making it unsuitable to cross-layer schemes which require instant signaling.

Direct Interlayer Communication is another signaling method which was proposed in Ref. 31 to improve the interlayer signaling pipe method by the introduction of signaling shortcuts performed out of band. In this way, the proposed Cross-Layer Signaling Shortcuts (CLASS) approach allows non-adjacent layers of the protocol stack to exchange messages, skipping processing at every adjacent layer.

Along with reduced processing overhead, CLASS avoids insertion of signaling information into packet headers, which makes it suitable for bidirectional communication. Signaling messages use the Internet Control Message Protocol (ICMP).^{32,33}

Despite the advantages of direct communication between protocol layers and a standardized way of signaling, the ICMP-based approach involves operation with heavy protocol headers (IP and ICMP), as well as significant protocol processing overhead. Moreover, it appears to be limited to request-response actions, while more complicated signaling should be adapted to handle asynchronous events. To this aim, a mechanism which uses callback functions can be employed. This mechanism allows a given protocol layer to register a specific procedure (callback function) with another protocol layer, whose execution is triggered by a specific event at that layer.

Central Cross-Layer Plane, implemented in parallel to the protocol stack, is probably the most widely proposed cross-layer signaling architecture. Implementation



of this signaling method could be as simple as a shared database accessed by all the layers.³⁴ More advanced implementations introduce signaling interfaces as each protocol level internally provides access to the internal protocol layer parameters and functions. Communication with the central cross-layer plane is maintained by using a predefined set of API functions.^{30,35}

1.5.2. Cross-layer design for cognitive networks

Most of the existing cross-layer signaling proposals employ cross-layer signaling between different layers within the protocol stack of a single node. However, as emphasized in Ref. 12, true cognitive networking should maintain a network-wide scope with cognitive process operating based on end-to-end goals. Consequentially, most of the cross-layer signaling approaches currently available in literature are not appropriate to support cognitive networks which require network-wide propagation of cross-layer signaling information. In addition, there is the question of how cross-layer signaling can be performed.

A set of proper techniques required for network-wide cross-layer signaling is discussed next.

Among the overviewed methods, an encapsulation of signaling information into packet headers or ICMP messages can be considered appropriate. Their advantages, underlined in the single-node protocol stack scenario, become more significant for network-wide communication. For example, the way of encapsulating cross-layer signaling data into optional fields of the protocol headers does not produce any

additional overhead and keeps the association of signaling information with a specific packet. However, this method limits propagation of signaling information to packet paths in the network. For that reason, it is desirable to combine packet headers signaling with ICMP messages, which are well suited for explicit communication between network nodes.

One of the early examples of cross-network cross-layering is the Explicit Congestion Notification (ECN) presented in Ref. 36. It realizes in-band signaling approach by marking in-transit TCP data packet with congestion notification bit. However, due to the limitation of signaling propagation to the packet paths, this notification needs to propagate to the receiver first, which echoes it back in the TCP ACK packet outgoing to the sender node. This unnecessary signaling loop can be avoided with explicit ICMP packets signaling. However, it requires traffic generation capability from network routers and it consumes bandwidth.

An example of the adaptation of Central Cross-Layer Plane-like architecture to the cross-network cross-layer signaling is presented in Ref. 37. This reference suggests the use of a network service which collects parameter values related to the wireless channel located at the link, as well as at the physical layer and the provisioning of this information to adaptive mobile applications.

A unique combination of local and network-wide cross-layer signaling approaches called Cross-Talk is presented in Ref. 38. CrossTalk architecture consists of two cross-layer optimization planes, where one is responsible for the organization of cross-layer information exchange between protocol layers of the local protocol stack and their coordination. The other plane is responsible for network-wide coordination, considered the aggregation of cross-layer information provided by the local plane. It serves as an interface for cross-layer signaling over the network. Most of the signaling is performed in-band, using the packet headers method, making it accessible not only at the end host but at the network routers as well. Cross-layer information received from the network is aggregated and then can be considered for the optimization of local protocol stack operation based on global network conditions.

Main problems associated with deployment of cross-layer signaling over the network, also pointed in Ref. 39, include security issues, problems with non-conformant routers, and processing efficiency. Security considerations require the design of proper protective mechanisms, avoiding protocol attacks attempted by non-friendly network nodes, which furnish incorrect cross-layer information in order to trigger specific behavior. The second problem addresses misbehavior of network routers. It is pointed out that in 70% of the cases IP packets with unknown options are dropped in the network or by the receiver protocol stack. Finally, the problem with processing efficiency is related to the additional costs of the routers hardware for cross-layer information processing. While it is not an issue for the low-speed links, it becomes relevant for high speed ones where most of the routers decrement only the TTL field to maintain a high packet processing speed.

1.5.3. *Co-existence and integration*

An important challenge of the design of cognitive network solutions corresponds to the implementation of cognitive network primitives in current networking environments, given the limitations of the widely spread TCP/IP stack and of proper solutions of the implementation in “non-friendly” network environments.

Indeed, cognitive networking relies on active network nodes which cooperate for observation, analysis, decision making, and action elements. However, an assumption that all the network nodes are friendly to cognitive networking can be verified. Moreover, some of the nodes capable of cognitive networking may not be willing to cooperate.

The possibility of deployment of cognitive networks in widely IP networks depends on the fundamental characteristics of the cognitive network solutions overviewed in Sec. 2.2.3.

One of the main issues is the consistency with TCP/IP node between the network node implementing cognitive network functionalities and an ordinary network node running unmodified TCP/IP stack. A potential solution to obtain compatibility is the introduction of a cognitive network interface at each protocol stack layer to provide access to the internal layer parameters and functions to the cognitive engine.

The limitation of reconfigurable elements to lower protocol layers only will help avoiding modification of the core of the protocol stack implemented in the operating system, which obviously will facilitate the deployment. However, this approach may lead to limitations in the design of cognitive network algorithms leading to the reduced performance.

Another important factor affecting co-existence with TCP/IP and incremental deployment of cognitive networks is related to the level of support required from the network elements. In the case of high reliance required implementation, it is possible only in networks (or parts of networks) with cognitive-friendly switches and routers, requiring the deployment of additional equipment. This is not always possible, especially in the widely deployed IP networks. However, it is feasible in isolated networks such as the cellular network environment.

1.6. Conclusions and Future Directions

Network evolution towards self-aware autonomous adaptive networking resolves inefficiency of network configuration and management. In order to optimize network operation, reconfiguration, management, and improving performance, it has been proposed to introduce self-awareness, self-management, and self-healing properties by bringing “intelligence” into the network, creating a new paradigm in networking, referred to as cognitive networking.

This chapter provides a detailed survey of state-of-the-art and future directions in cognitive networking by defining fundamental techniques, enabling cognitive

properties and unveiling details of adaptation, learning, and goal optimization processes.

Comparison of available research proposals motivated the design of a promising cognitive network architecture capable of fully implementing cognitive network techniques.

Finally, discussion on the required properties of the cross-layer design for cognitive networks and corresponding deployment issues are discussed.

References

1. C. Barakat, E. Altman and W. Dabbous, On TCP performance in a heterogeneous network: A survey, *IEEE Communications Magazine*, **38**(1), 40–46 (January 2000).
2. D. D. Clark, C. Partridge, J. C. Ramming and J. T. Wroclawski, A knowledge plane for the Internet, *ACM SIGCOMM*, Karlsruhe, Germany (August 2003).
3. *Clash of the Titans — WiMAX and 4G: The Battle for Convergence is Joined*, Maravedis Market research and analysis (July 2006).
4. D. L. Tennenhouse, J. M. Smith, W. D. Sincoskie, D. J. Wetherall and G. J. Minden, A survey of active network research, *IEEE Communications Magazine*, **35**(1), 80–86 (January 1997).
5. D. Kliazovich, F. Granelli, G. Pau and M. Gerla, APOHN: Subnetwork layering to improve TCP performance over heterogeneous paths, *IEEE Next Generation Internet Design and Engineering (NGI)*, Valencia, Spain (April 2006).
6. V. Firoiu, J.-Y. Le Boudec, D. Towsley and Z.-L. Zhang, Theories and models for Internet quality of service, *IEEE Proceedings*, **90**(9), 1565–1591 (2002).
7. H. Zhu, M. Li, I. Chlamtac and B. Prabhakaran, A survey of quality of service in IEEE 802.11 networks, *IEEE Wireless Communications*, **11**(4), 6–14 (2004).
8. P. Demestichas, G. Dimitrakopoulos and J. Strassner, Introducing reconfigurability and cognitive networks concepts in the wireless world, *IEEE Vehicular Technology Magazine*, **1**(2), 32–39 (2006).
9. J. Mitola, *Cognitive Radio: An Integrated Agent Architecture for Software Defined Radio*, PhD thesis, Royal Institute of Technology (2000).
10. P. Sutton, L. E. Doyle and K. E. Nolan, A reconfigurable platform for cognitive networks, *1st International Conference on Cognitive Radio Oriented Wireless Networks and Communications*, 1–5 (June 2006).
11. R. Thomas, *Cognitive Networks*, PhD dissertation, Virginia Tech (June 2007).
12. R. W. Thomas, D. H. Friend, L. A. DaSilva and A. B. MacKenzie, Cognitive networks: Adaptation and learning to achieve end-to-end performance objectives, *IEEE Communications Magazine*, **44**(12), 51–57 (December 2006).
13. P. Surana, *Meta-Compilation of Language Abstractions*, PhD Thesis Dissertation, Northwestern University, Illinois (2006).
14. I. F. Akyildiz, W.-Y. Lee, M. C. Vuran and S. Mohanty, Next generation/dynamic spectrum access/cognitive radio wireless networks: A survey, *Computer Networks*, **50**, 2127–2159 (May 2006).
15. W. Krenik and A. Batra, Cognitive radio techniques for wide area networks, *Design Automation Conference*, 409–412 (June 2005).
16. *National Science Foundation — Funding opportunities*, available at <http://www.nsf.gov/funding/aboutfunding.jsp>.

17. *Sixth Framework Programme Research* — FP6 home page, <http://ec.europa.eu/research/fp6/> (2002).
18. *Seventh Framework Programme Research* — FP7 home page, <http://cordis.europa.eu/fp7> (2007).
19. D. Bourse, S. Buljore, A. Delautre, T. Wiebke, M. Dillinger, J. Brakensiek, K. Moessner, K. El-Khazen and N. Alonistioti, *The End-to-End Reconfigurability (E²R) Research*, SDR Forum Technical Conference, Orlando, USA (2003).
20. P. Demestichas, V. Stavroulaki, D. Bosovic, A. Lee and J. Strassner, m@ANGEL: autonomic management platform for seamless cognitive connectivity to the mobile internet, *IEEE Communications Magazine*, **4**(6), 118–127 (June 2006).
21. D. Bourse and K. El-Khazen, End-to-end reconfigurability (E²R) research perspectives, *IEICE Transactions on Communications, Special Section on Software Defined Radio Technology and Its Applications*, 4148–4157 (2005).
22. Cognet Project, <http://adaptive5.ucsd.edu/cognet/>.
23. P. Baumgart, *Building an IEEE 802.11 Testbed for the CogNet: Cognitive Complete Knowledge Network*, available at <http://adaptive5.ucsd.edu/cognet/Documents>.
24. R. Rao and B. S. Manoj, Cognitive cellular network testbed development, Internal report, 2007, available at <http://ece-classweb.ucsd.edu:16080/fall07/ece291/Descriptions/cog.%20wifi-Fall07.doc>.
25. E. Gelenbe, Z. Xu and E. Seref, Cognitive packet networks, *11th IEEE International Conference on Tools with Artificial Intelligence*, 47–54 (November 1999).
26. S. F. Bush and A. Kulkarni, *Active Networks and Active Network Management: A Proactive Management Framework*, Kluwer Academic/Plenum Publishers, New York, Boston, Dordrecht, London, Moscow (2001).
27. S. M. Lake Sr., Cognitive networking with software programmable intelligent networks for wireless and wireline critical communications, *IEEE Military Communications Conference (MILCOM)*, 1693–1699 (October 2005).
28. B. S. Manoj, R. R. Rao and M. Zorzi, *Architectures and Protocols for Next Generation Cognitive Networking*, Frank H. P. Fitzek and Marcos D. Katz (eds.), *Cognitive Wireless Networks Concepts, Methodologies and Visions Inspiring the Age of Enlightenment of Wireless Communications*, Springer (November 2007).
29. V. Srivastava and M. Motani, Cross-layer design: A survey and the road ahead, *IEEE Communications Magazine*, **43**(12), 112–119 (2005).
30. V. T. Raisinghani and S. Iyer, Cross layer feedback architecture for mobile device protocol stacks, *IEEE Communications Magazine*, **44**(1), 85–92 (2006).
31. Q. Wang and M. A. Abu-Rgheff, Cross-layer signaling for next-generation wireless systems, *IEEE Wire-less Communications and Networking (WCNC)*, 1084–1089 (2003).
32. A. Conta and S. Deering, *Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification*, RFC 2463 (1998).
33. J. Postel, *Internet Control Message Protocol*, RFC 792 (1981).
34. K. Chen, S. H. Shah and K. Nahrstedt, Cross-layer design for data accessibility in mobile ad hoc networks. Wireless personal communications, Special Issue on *Multimedia Network Protocols and Enabling Radio Technologies*, **21**, 49–75 (2002).
35. K. M. El Defrawy, M. S. El Zarki and M. M. Khairy, Proposal for a cross-layer coordination framework for next generation wireless systems, *ACM International Conference on Communications and Mobile Computing*, 141–146 (2006).
36. K. Ramakrishnan, S. Floyd and D. Black, *The Addition of Explicit Congestion Notification (ECN) to IP*, RFC 3168 (2001).

37. B.-J. Kim, A network service providing wireless channel information for adaptive mobile applications: I: Proposal, *IEEE International Conference on Communications (ICC)*, 1345–1351 (2001).
38. R. Winter, J. H. Schiller, N. Nikaein and C. Bonnet, CrossTalk: cross-layer decision support based on global knowledge, *IEEE Communications Magazine*, **44**(1), 93–99 (2006).
39. P. Sarolahti and S. Floyd, *Cross-layer Indications for Transport Protocols*, Internet draft draft-sarolahti-tsvwg-crosslayer-00.txt (2007).