A Cognitive Handoff

Holistic Vision, Reference Framework, Model-driven Methodology and Taxonomy of Scenarios

Francisco A. González-Horta, Rogerio A. Enríquez-Caldera, Juan M. Ramírez-Cortés, Jorge Martínez-Carballido
Department of Electronics
INAOE
Tonantzintla, Puebla, México
e-mail: fglez@inaoep.mx, rogerio@inaoep.mx, jmramar@inaoep.mx, jmc@inaoep.mx

Eldamira Buenfil-Alpuche
Faculty of Engineering
Polytechnic University of Guerrero State, UPEG
Taxco, Guerrero, Mexico
e-mail: eldamira@gmail.com

Abstract—Current handoffs are not designed to achieve multiple desirable features simultaneously. This weakness has resulted in handoff schemes that are seamless but not adaptive, or adaptive but not secure, or secure but not autonomous, or autonomous but not correct, etc. To face this limitation, in this paper we envision and develop a new kind of handoff system, which is context-aware in the sense that uses information from its external and internal environment: a cognitive handoff. Thus, the resulting cognitive handoff can attain multiple purposes simultaneously through trading-off multi-criteria and based on a variety of policies. We also discuss the difficulties of developing cognitive handoffs and propose a new model-driven methodology for their systematic development. The theoretical framework of this methodology is the holistic approach, the functional decomposition method, the model-based design paradigm, and the theory of design as scientific problem-solving. We applied the proposed methodology and obtained the following results: (i) a correspondence between handoff purposes and quantitative environment information, (ii) a novel taxonomy of handoff mobility scenarios, and (iii) an original state-based model representing the functional behavior of the handoff process.

Keywords—Cognitive handoff; handoff methodology; handoff scenarios

I. INTRODUCTION

A handoff is intended to preserve the user communications while different kinds of transitions occur in the network connection. Thus, a handoff is the process of transferring communications among radio channels, base stations, IP networks, service providers, mobile terminals, or any feasible combination of these elements [1]. Significant desirable handoff features mentioned in the review of the literature [2] are: seamless [3], adaptive [4], autonomous [5], secure [6], and correct [7]; however, many others can be found in the vast literature of handoffs: transparent, reliable, flexible, robust, balanced, immune, fast, soft, smooth, lossless, efficient, proactive, predictive, reactive, QoS-based, power-based, location-aided, time-adaptive, intelligent, generic, etc. Despite the rich variety of desirable handoff features the resulting handoffs they are not enough to face the challenges of the future Internet [8], [9], [10], [11] and two important problems remain unsolved: i) how can be combined different desirable features into a single handoff process so that it can achieve many purposes simultaneously? And ii) how to define every desirable feature so that ambiguity and subjectivity can be reduced?

This gap in knowledge about handoffs has produced a number of single-purpose schemes that successfully achieve one attractive feature but completely ignore others; e.g., seamless handoffs with poorly or null adaptation to handoff scenarios or technologies [12]; adaptive handoffs that do not consider any security goal [4]; secure handoffs that ignore user autonomy [6]; etc. Also, there is a growing confusion in literature about similar features; e.g., accurate-correct, fast-timely, smooth-seamless, robust-reliable, etc. In order to reduce misuse and ambiguity of these attributes is convenient to associate a qualitative property (purpose) and quantitative measures (objectives and goals) to each desirable feature. By doing so, we can qualify and quantify their performance individually or in comparison with others.

The development of handoffs achieving multiple desirable features has been “delayed” by the research community itself, despite it was advised since 1997 by Tripathi [8], because many authors preferred to focus on understanding and controlling very specific handoff scenarios (reductionist approach) instead of managing complex and generic handoff scenarios (holistic approach). However, recent handoff schemes, like the ones proposed by Altaf in 2008 [9] for secure-seamless-soft handovers, or Cardenas in 2008 [10] for fast-seamless handoffs, or Singhrova in 2009 [11] for seamless-adaptive handoffs, show a tendency towards cognitive handoffs.

Major contributions of this paper include:

1) A new holistic vision of handoffs.

Many handoff solutions follow a reductionist approach; i.e., they achieve one desirable feature, use a small amount of handoff criteria, and work only in very specific scenarios. Although these simplistic solutions provide understanding and control of particular situations, we have seen how they...
quickly become special cases of more general models. Thus, we claim that the handoff problem for the future Internet requires holistic solutions, achieving multiple desirable features, using a diversity of context information, and adapting to any handoff scenario.

2) A reference framework for cognitive handoffs.

We propose a new kind of handoff that is multipurpose, multi-criteria, context-aware, self-aware, policy-based, and trades-off multiple conflicting objectives to reach its intended goals. This paper provides the conceptual model and its first level of functional decomposition.

3) A model-driven methodology to develop cognitive handoffs.

This methodology allows to systematically develop cognitive handoffs using a comprehensive model-based framework. The proposed methodology is founded on a synthesis of holism, reductionism, functional decomposition, model-based design, and scientific problem-solving theory. As a result of deploying our methodology, we present a clear correspondence among cognitive handoff purposes and handoff environment information.

Besides, in order to test the resulting cognitive handoff when applying such methodology with the parameters associated to, and for a given scenario, we develop the following contributions:

4) A taxonomy of handoff mobility scenario.

This taxonomy gives a classification of handoff scenarios by considering all feasible combinations of several communication dimensions involved.

5) An original state-based model of the handoff process.

This state based model is represented by five-state diagram, which describes the control handoff process and the way all different stages are being coordinated before, during, and after the handoff.

The paper is organized in sections that correspond to the previously described contributions and therefore it starts in Section II with a distinction between Single-purpose and multi-purpose handoffs. Section III presents the holistic vision for the conceptualization of cognitive handoffs. Section IV presents our cognitive handoff reference framework and its specific characteristics. Section V describes the ad hoc model-driven methodology that we are using for developing cognitive handoffs. This section discusses the difficulties for developing cognitive handoffs and provides an overview of theoretical framework setting the basis of our methodology. Section VI shows the first results we obtained from applying the methodology. These results include: (a) the correlation between context data and desirable handoff features through the definition of handoff purposes, objectives, and goals; (b) the taxonomy of handoff scenarios derived from combining all the possible transition elements involved in handoffs; and, (c) a cognitive handoff state-based model that describes a general behavior of the control handoff process. Section VII presents a basic discussion on the applicability of preliminary results. Finally, Section VIII concludes the paper with a summary of contributions and future work.

II. SINGLE-PURPOSE VS MULTI-PURPOSE HANDOFFS

Dr. Nishith D. Tripathi in his outstanding thesis work published in 1997 [12] probably was the first author in considering a handoff that can simultaneously achieve many desirable features. His inspiring work served for many years as a basis for developing high performance handoffs; however, the complexities of handoff scenarios from 1997 to present days have changed significantly. For instance, the handoff concept changed from simple lower-layer transitions between base stations and channels to more elaborated cross-layer transitions among networks, providers, and terminals. The limited scope of Tripathi’s handoff concept has brought in consequence that his algorithms and models become today special cases of more general models. Holism is relevant in this way to provide a long-term solution for the handoff problem. Another author who describes several desirable handoff features is Nasser et al. [13] in 2006. Both, Tripathy and Nasser, described various desirable features, but they did not make any difference among features, purposes, objectives, and goals. A handoff model needs a clear distinction to such former concepts.

The holistic vision to the handoff problem has also been studied by Dr. Mika Ylianttila in his exceptional thesis work [14] published in 2005. He presented a holistic system architecture based on issues involved in mobility management areas (e.g., mobility scenarios, handoff strategies, handoff control, handoff algorithms, handoff procedures, mobility protocols, mobility parameters, performance measures, and handoff metrics). The work of Ylianttila improved the architecture of handoff issues that Pahlavan [15] published in 2000. However, these architectures have some drawbacks: i) they did not include the context management problem in their models; ii) they did not mention the tradeoffs that handoffs should consider in a multi-objective scenario; and iii) their architectures are based on types of issues and not in the functionality aspects of the handoff process.

Besides the above related work, we use two criteria to classify handoff schemes that are approaching to cognitive handoffs: the number of desirable features they achieve and the amount of context information they use. Handoff schemes, like the ones proposed by So [16] and Zhang [17], achieve only one desirable feature using limited context information; they provide seamless handoffs between particular network technologies and specific mobility scenarios. The schemes proposed by Siddiqui [18] and Hasswa [19] use broad context information, but they are focused only in one feature (seamlessness).
Conversely, the solutions proposed by Sethom [6] and Tuladhar [20] provide seamless and secure handoffs on a variety of handoff scenarios because they use broad context. The schemes proposed by Singhrova [4] and Chen [21] achieve seamless and adaptive mobility, but they cannot adapt to any handoff scenario because they use limited context. Finally, the scheme proposed by Altaf [22] achieves seamless, secure, soft, and adaptive handoffs, but just between WiMAX and 3G networks because they use limited context.

The information the handoff process uses for making decisions, has been increasing as we want to deploy more intelligent handoff systems. Handoffs for the first generation wireless networks were called single-criterion, because they were mainly based on signal-strength or few criteria taken from the access network dimension. At the second generation, the need for improving the effectiveness of handoffs led to include more information to this process. Handoffs for 2G networks were known as multi-criteria. They included many criteria from distinct dimensions, but not from all dimensions; e.g., they might consider the battery load from the terminal dimension and the user speed from the mobility dimension, but ignore the fees from the service provider.

At the beginning of 3G networks, several handoff schemes were deployed using information or criteria from the entire external handoff environment, and they were called context-aware handoffs. In 2003, Prehofer et al. [23], defined a context-management architecture for addressing the problem of collecting, compiling, and distributing handoff context information. This remarkable work started a new stage in the development of handoffs. Part of this architecture was used by Pawar et al. [24] in 2008 for developing context-aware handoffs applied to mobile patient monitoring.

At the dawn of 4G networks, a new type of handoff solutions, called self-aware, started to use a variety of handoff performance parameters from its internal environment to self-adapt its behavior according to different performance goals. For the future networks, environment-aware handoffs, using information from both, the external and internal environment, will be deployed.

Despite these recent advances in context-management architectures and applications, the lack of a clear relationship between handoff context information and handoff desirable features is adding unnecessary complexity to the process of handoff. The handoff decision making process should be oriented to accomplish more than just one desirable feature. Therefore, we consider that in difference with current handoff schemes, a cognitive handoff is aware of its external and internal environment and optimally achieves multiple desirable features simultaneously.

Considering the tendency on handoff research, it will be common to observe in the near future a new generation of handoffs that can achieve many desirable features using broad handoff context information. In current literature, none architecture, model, or algorithm is reported to have this property.

Regarding the related work of standardization bodies, like the IEEE 802.21 and the IETF MIPSHOP, we observed that they are focusing in seamless heterogeneous handoffs; they are not taking into account the vast diversity of desirable features that handoffs could have. The IEEE 802.21 workgroup has approved three task groups to face very particular handoff scenarios: the IEEE 802.21a for security extensions to media independent handovers, the IEEE 802.21b for handovers with downlink only technologies, and the IEEE 802.21c for optimized single radio handovers. We believe they are following a reductionist approach, but they lack the holistic vision of cognitive handoffs. Emmelman [25] discusses ongoing activities and scopes of these standardization bodies.

III. THE COGNITIVE HANDOFF HOLISTIC VISION

First of all, we will describe in a holistic manner the vision of cognitive handoffs.

A. Origin of Single-Purpose Handoffs

The thoughtful study of handoffs started in the early 1990s with the first generation (1G) cellular networks (e.g., AMPS [26]). These networks provided seamless conversations while the mobile phone switched between channels and base stations. The decision to perform a handoff was made only on a signal strength basis, but the handoff execution should be imperceptible to users. For this reason, the AMPS system required that the handoff gap be no more than 100 ms to avoid the possibility of dropping a syllable of speech [26]. These traditional handoffs are single-purpose/single-criterion or seamless/signal strength.

B. Major Challenges in the Future Internet

1) Multidimensional Heterogeneity: A major trend in future communication systems is the coexistence of multiple dimensions of heterogeneity integrated into a seamless, universal, uniform, ubiquitous, and general-purpose network. This future Internet will be seamless if it hides heterogeneity to users, universal if it can be used by anyone with any terminal, uniform if it is an all-IP network, ubiquitous if it is available anywhere and anytime, and general-purpose if it can provide any service. We divide heterogeneity into five dimensions as illustrated in Fig. 1 and explained in the next paragraphs. The arrows going down from the service provider dimension to the user mobility dimension depict two different handoff scenarios created by instantiating objects in each dimension.

a) Diversity on service providers and operators: Offer different classes of services, billing models, security policies, and connection prices. They deploy different
wireless technologies around the world and make roaming agreements and alliances with other providers and operators.

![Diagram of network components]

**Figure 1. Multidimensional heterogeneity in the future Internet.**

**b) Diversity on service providers and operators:** Offer different classes of services, billing models, security policies, and connection prices. They deploy different wireless technologies around the world and make roaming agreements and alliances with other providers and operators.

c) **Variety of applications and services:** Intend to fulfill the distinct ways of human communication; e.g., voice, video, data, images, text, music, TV, telephony, etc.

d) **Several access network technologies:** Include wired and wireless access technologies [21]; e.g., Ethernet, Bluetooth, WiMAX, WiFi, UMTS, MBWA, IMT-2000, GPRS, GSM, EDGE, LTE/SAE, DVB-HS, etc. They differ in terms of electrical properties, signaling, coding, frequencies, coverage, bandwidth, QoS guarantees, mobility management, media access methods, packet formats, etc.

e) **Plethora of mobile user terminals:** Users can be humans, machines, or sensors. Terminals for machines are integrated parts of machines. Sensor terminals collect information from networked sensors [27]. Terminals for humans are mobile and multimode, equipped with telecommunication capabilities and different saving energy characteristics; they change its factor form from those looked like computers (laptops, netbooks) to those looked like cell phones (PDAs, smartphones).

f) **Numerous user mobility states:** Network terminals can be located anywhere – in space, on the ground, under the ground, above water, underwater, and they can be fixed in a geographic position or moving at any speed – pedestrian, vehicular, ultrasonic [27].

Nowadays, no handoff solution exists, which comprehensively addresses the entire scale of heterogeneity. Moreover, multidimensional heterogeneity has three main attributes: is inevitable, is the source of great amounts of context information, and produces an infinite number of handoff scenarios.

2) **Ubiquitous Connectivity:** It enables connectivity for anyone or anything, at any time, from anywhere. A myriad of wireless access technologies are spread across the entire world overlapping one another but avoiding interferences among them. Two requirements for ubiquitous connectivity are:

- a) to develop scalable architectures to integrate any number of wireless systems from different service providers [28] and
- b) to develop smart multimode mobile terminals able to access any wireless technology [29].

3) **Cognitive Mobility:** It allows roaming mechanisms where the user is always connected to the best available network, with the smaller number of handoffs, service disruptions, user interventions, security threats, and the greater number of handoff scenarios.

**C. External and Internal Handoff Environment**

We envision a cognitive handoff as a process that is both context-aware and self-aware. This implicates to make the handoff process aware of its external and internal environment. We borrowed the term ‘cognitive’ from Dr. Dixit vision of cognitive networking [30]. He defines cognitive networking as an intelligent communication system that is aware of its environment, both external and internal, and acts adaptively and autonomously to attain its intended goals. We believe cognitive handoffs not only should behave adaptively or autonomously to attain its intended goals, but also seamlessly, securely, and correctly.

On one hand, the external environment is directly related with all the external entities that provide a source of context information to the handoff process. These entities are users, terminals, applications, networks, and providers; a cognitive handoff should adapt to any kind of these entities. These entities maintain a strong cyclic relationship as follows: users interact with terminals, terminals run applications, applications exchange data through networks, networks are managed by providers, and providers subscribe users. The cyclic relationship of external entities suggests that all external context information emanates just from these five basic entities and no more; hence, if we ignore information of any of these entities, the handoff process will not adapt properly to all the scenarios. Therefore, a cognitive handoff should consider all the five entities.

On the other hand, the internal environment is another source of context and it is directly related with the behavior or performance of handoffs. This behavior directly depends on the desirable features of handoff. Next, we identified and describe five major desirable features, which are considered highly significant for the current and future scenarios.

**D. Multiple Desirable Features of Handoff**

1) **Seamlessness:** It means to preserve the user communications before, during, and after the handoff thus
reducing service degradation or interruption. Service degradation may be due to a continuous reduction in link quality, network quality, handoff quality, QoS guarantees, and energy savings. Service interruption may be due to excessive degradations or a “break before make” approach.

2) Autonomy: This desirable feature is closely related to seamlessness. A handoff is autonomous, automatic, or autonomic when no user interventions are required during a handoff in progress. However, this does not mean that user interventions are not required in handoffs. It is good that users participate in the handoff configuration process by defining their preferences, priorities, or necessities; but, it is convenient that users can perform this activity offline to prevent any distraction during online communications.

3) Security: We say a handoff is secure if not new threats appear along the handoff process and security signaling traffic does not overload the network and degrades the communication services. This is a very challenging task, but if optimization techniques are used together with our model it could be shown that by minimizing handoff latency, authentication latency, and signaling overload, the risk of new threats appearance may be reduced.

4) Correctness: A handoff is correct if it keeps the user always connected to the best available network with the smaller number of handoffs; this is similar to the Gustaffson’s vision of ABC defined in [31]. We consider that the best network is the one that is sufficiently better and consistently better. Furthermore, correctness can bring other additional features to the handoff process:

- **Beneficial**: if quality of communications, user expectations, or terminal power conditions get improved after handoff.
- **Timely**: if handoff is executed just in time; i.e., right after target is properly selected and before degradations or interruptions occur.
- **Selective**: if it properly chooses the best network among all the available networks.
- **Necessary**: if it is initiated because of one imperative or opportunist reason.
- **Efficient**: if it selects the most appropriate method, protocol, or handoff strategy, according to the types of: handoff in progress, user mobility, and application.

These handoff attributes derived from correctness, take special relevance during the decision-making phase, where it must be decided why, where, how, who, and when to trigger a handoff.

5) Adaptability: An adaptable handoff should be successful across any handoff scenario. A handoff is successful if it achieves a balance of every desirable feature at a minimum level of user satisfaction.

### E. Structure of Handoff Context Information

The handoff context information is extensive, heterogeneous, distributed, and dynamic. It supports the whole operation of the handoff process and the achievement of multiple desirable features. Therefore, such context information should be arranged in a clear structure. Table I and Table II show the structure of handoff context information according to a pair of criteria: the source of context and the class of information respectively. The sources of context originated in the external handoff environment support context-awareness while the one originated in the internal environment (the handoff process itself) will provide self-awareness.

<table>
<thead>
<tr>
<th>TABLE I.</th>
<th>STRUCTURE FOR SOURCE OF CONTEXT INFORMATION</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>User context</strong>:</td>
<td>This context allows users to customize the handoff according to their own needs, habits, and preferences. It includes: user preferences, user priorities, user profiles, user history, etc.</td>
</tr>
<tr>
<td><strong>Terminal context</strong>:</td>
<td>Allows the deployment of QoS-aware handoffs, power-based handoffs, and location-aided handoffs:</td>
</tr>
<tr>
<td>(a) <strong>Link quality</strong>:</td>
<td>Received signal strength (RSS), signal to noise ratio (SNR), signal to interference ratio (SIR), signal to noise and interference ratio (SNIR), bit error rate (BER), block error rate (BLER), co-channel interference (CCI), carrier to interference ratio (CIR), etc.</td>
</tr>
<tr>
<td>(b) <strong>Power management</strong>:</td>
<td>Battery type (BT), battery load (BL), energy-consumption rate (ECR), transmit power in current (TPC), transmit power in target (TPT), power budget (PB), etc.</td>
</tr>
<tr>
<td>(c) <strong>Geographic mobility</strong>:</td>
<td>Velocity (Vel), distance to a base station (Dist), location (Loc), direction (MDir), coverage area (GCA), etc.</td>
</tr>
<tr>
<td><strong>Application context</strong>:</td>
<td>This context includes the QoS requirements of active applications: Lost packets (LP), delayed packets (DP), corrupted packets (CP), duplicated packets (DuP), data transfer rate (DTR-goddput), packet jitter (PJ), out-of-order delivery (OOD), application type (AppT), etc. The consideration of these QoS parameters makes provisions for application-aware handoffs.</td>
</tr>
<tr>
<td><strong>Network context</strong>:</td>
<td>This context is needed to avoid selecting congested networks (before handoff), to monitor service continuity (during handoff), and to assess the handoff success by measuring network conditions (after handoff): Network bandwidth (NBW), network load (NL), network delay (ND), network jitter (NJ), network throughput (NT), network maximum transmission unit (NMTU), etc.</td>
</tr>
<tr>
<td><strong>Provider context</strong>:</td>
<td>Connection fees, billing models, roaming agreements, coverage area maps, security management (AAA), types of services (data, voice, video), provider preferences, and provider priorities. A negotiation model may be required to equate the differences between service providers, network operators, and mobile users.</td>
</tr>
<tr>
<td><strong>Handoff performance context</strong>:</td>
<td>Call blocking (CB), call dropping (CD), handoff blocking (HOB), handoff rate (HOR), handoff latency (HOL), decision latency (DLat), execution latency (ExLat), evaluation latency (EvLat), handoff type (HOType), elapsed time since last handoff (ETSLH), interruptions rate (IR), interruption latency (IL), degradations rate (DR), degradation latency (DL), degradation intensity (DI), utility function (UF), signaling overload (SO), security signaling overload (SSO), improvement rate (ImpR), application improvement rate (AppImpR), user improvement rate (UsrImpR), terminal improvement rate (TermImpR), successful handoff rate (SHOR), imperative handoff rate (IHOR), opportunistic handoff rate (OHOR), dwell time in the best (DTIB), authentication latency (AL), detected attacks rate (DAR), online user interventions rate (OUIR), tardy handoff rate (THOR), premature handoff rate (PHOR), etc. His context allows users to customize the handoff according to their own needs, habits, and preferences. It includes: user preferences, user priorities, user profiles, user history, etc.</td>
</tr>
</tbody>
</table>
TABLE II. STRUCTURE FOR CLASS OF INFORMATION

| Handoff criteria: | Network discovery, decision-making, and performance evaluation. Some examples of handoff criteria include variables or parameters from the external/internal environment such as RSS, NL, BL, LP, HOL, Vel, connection price, etc. |
| Handoff metrics: | Mathematical models used to measure several significant tasks of the handoff process; for instance, the quality of links, the quality of communications, the quality of different networks, the quality and quantity of handoffs, the quality of different providers, the achievement of user preferences, the power budget of a mobile terminal, the geographic mobility of a user, etc. Handoff metrics may combine a variety of handoff criteria and help any specific handoff algorithm to make optimal decisions. |
| Performance measures: | Set of handoff metrics that are used to quantify performance of communications, performance of networks, performance of handoffs, and to evaluate the degree of achieving a handoff objective. |
| Handoff policies: | Users and providers define a series of policies to the handoff operation. Policies define and specify rules for making handoff decisions in any particular situation; for instance, what to do if the link quality drops below a level required for an acceptable service. User and provider may have different views of the handoff process; provider may be interested in QoS while user in connection charges. Both points of view must be consistently integrated into a single handoff policy management database. |
| Handoff constraints: | Conditions that must be satisfied in a particular handoff scenario and used to control the handoff operation by keeping performance parameters within specific limits. For instance, for a seamless handoff process, the delay has to be kept within certain boundaries; for real-time applications a delay of 50 ms could be acceptable, whereas non-real-time applications might accept delays as long as 3-10 sec [11]. |
| Handoff configuration: | Defines preferences, priorities, and other configuration parameters required to customize the handoff operation. Typically, the configuration information is organized in a handoff profile linked to a particular user, provider, and terminal and should be initially performed offline either by the user, the provider, both or an auto-configuration setup. But, depending on the type of handoff profile, different configuration parameters may be required to be initialized, e.g., thresholds, timers, hysteresis, weights, etc. |

F. Cognitive Handoff Reference Framework

Once we have established and justified the necessity for developing a new handoff system, we present our reference framework based on the statement that “a cognitive handoff should intend to achieve multiple desirable features and be aware of its entire environment by using information coming from multiple context domains”. Fig. 2 depicts this basic idea by interconnecting multiple desirable features with multiple context domains that we already explained separately in III.D and III.E.

The purpose of this model is to help people debate and discuss about the complexity of cognitive handoffs. Thus, topics of discussion would be related to level of complexity, correlation among desired features and context data, and the possibility of establishing handoffs as a multi-objective optimization problem as well as to give specifications for practical implementations. Used in this way this model is not intended for predicting, designing, or implementing cognitive handoffs, but for understanding and explaining such difficult and complex process.

Figure 2. Cognitive handoff conceptual model. The desired features to achieve determine the context data to use and vice versa.

All the above issues have not been addressed in the handoff literature; therefore, in effect, the purpose of this conceptual model is being achieved. Models like the one we present here are validated by credibility, and credibility comes from the way in which the cognitive maps are built and the clarity it represents most of the opinion’s experts [32]. In the next section we provide some advances towards the development of cognitive handoffs.

IV. COGNITIVE HANDOFF AT WORK

A. Cognitive Handoff and Complex Systems

Cognitive handoffs are complex adaptive systems because: (1) they exhibit a complicated hierarchical structure (e.g., a power saving system is part of a network discovery system, which is part of a handoff system, which is part of a mobility system, which is part of a wireless communication system, and so on, but also a power saving system is part of the decision system, which is part of the handoff system, and so on); (2) the whole cognitive handoff system achieves purposes that are not purposes of the parts (e.g., a cognitive handoff purpose is to maintain the continuity of services, but this purpose is not defined in any of the parts or subsystems of the cognitive handoff system); and, (3) the handoff environment is dynamic and therefore adaptability is a desired handoff feature.

B. Correlating Desired Features and Context Data

With respect on whether all previously described context data are necessary to describe limitations on the model; one has to realize that the usage of certain context parameters depends on the desirable features being implemented and the context data available in a moment will allow to accomplish or not a particular desired feature. Thus, we need to state a correct relationship or dependence between each desirable handoff feature and the subset of context data necessary to be accomplished. We made a correlation between desired features and context data by transforming desired features into purposes, purposes into objectives,
objectives into goals, and goals into context data. This correlation will be shown in Section VI.

C. Advances for a Practical Implementation

The cognitive handoff system, represented in Fig. 2 by the oval in the middle, can be expanded into several sub-systems by using a functional decomposition approach [33].

Fig. 3 shows the main functional sub-systems for cognitive handoffs represented in ovals: handoff control algorithm, network discovery, handoff decisions, handoff execution, handoff evaluation, and handoff context information management. We briefly describe them:

- **Handoff Control Algorithm**: This is the main director of the handoff procedure. The entity, which implements the control algorithm is called Handoff Control Entity (HCE). There should be one HCE in every user terminal and also there may be many others distributed across the network infrastructure. HCEs are agents that cooperate and compete to take a particular handoff to succeed.

- **Network Discovery**: This is the system for detecting and discovering available access networks. An available network is a reachable and authorized network considered for an eventual handoff.

- **Handoff Decisions**: The handoff decisions system is intended to answer the questions of why, when, where, how, and who should trigger the handoff. Typically, this system has focused only in where and when to handoff [34]. The holistic vision extends the scope of handoff decisions.

- **Handoff Execution**: This system is intended to change the physical and logical connection from one network to another, from one provider to another, or from one terminal to another. This change requires the most effective method, protocol, or strategy according to the current handoff scenario. The MIPSHOP group at IETF and the IEEE 802.21 standard are creating tools for implementing media-independent handoffs since 2003.

- **Handoff Evaluation**: This system measures the achievement of every desirable handoff feature and decides whether the executed handoff was successful or not. The evaluation results should be delivered after the handoff execution but within strict time constraints, thus this task is proactively distributed along the handoff process.

- **Handoff Context Information Management**: This system is intended to collect the distributed handoff context data, transform the data in information, and redistribute this information to the HCEs, which are responsible for making handoff decisions and control.

Discovery, decisions, execution, and evaluation systems can be viewed as sequential stages of the handoff process; however, the context manager is a background process, which permanently supplies the handoff control entities with fresh information about the handoff environment.

D. Cognitive Handoff Performance Measures

The performance evaluation of cognitive handoffs requires a performance metric for each handoff purpose and a graphical representation to visualize multivariate data [35]. These metrics combine mathematically several performance measures that are associated to every handoff purpose. It is possible that metrics can normalize heterogeneous data into a single value representing the performance of each handoff purpose. Moreover, metrics can also be designed as utility functions so that greater values are better and all values are on the same scale.

Fig. 4 exemplifies a radar graph comparing the performance of multiple handoff purposes simultaneously. We say that if all measures are within a boundary circle of acceptable quality, then the cognitive handoff is successful, otherwise the handoff is defective and outliers should be corrected.
E. Formulating the Cognitive Handoff as a MOP

Let $F$ be the set of desirable handoff features and $C$ be the set of context data. We say that a context variable $v_i \in C$ is *correlated* with a desired feature $f \in F$ if and only if a change on the value of $v_i$ impacts on the purpose of $f$. For instance, some changes on the value of SNR may degrade or improve the link quality and impact on the purpose of seamlessness that is to maintain the continuity of services; thus, we say that SNR is correlated with seamlessness.

Let $V_f$ be the set of correlated variables with $f$, where $v_i \in V_f \subseteq C$. We say that $v_i$ is *positively correlated* with $f$ if and only if increments on the value of $v_i$ produce improvements on the purpose of $f$ and, decrements on $v_i$ produce degradations on the purpose of $f$. For instance, increments on SNR improve the link quality, which improves the service continuity of seamlessness, and conversely, decrements on SNR degrade the link quality, which degrades the service continuity of seamlessness. Therefore, SNR is positively correlated with seamlessness.

\[ \uparrow \text{SNR} \Rightarrow \uparrow \text{LINKQUALITY} \Rightarrow \uparrow \text{SEAMLESSNESS} \]

\[ \downarrow \text{SNR} \Rightarrow \downarrow \text{LINKQUALITY} \Rightarrow \downarrow \text{SEAMLESSNESS} \]

We say that $v_i$ is *negatively correlated* with $f$ if and only if increments on the value of $v_i$ produce degradations on the purpose of $f$ and, decrements on $v_i$ produce improvements on the purpose of $f$. For example, increments on BER degrade the link quality, which degrades the service continuity of seamlessness, and conversely, decrements on BER improve the link quality, which improves the service continuity of seamlessness. Therefore, BER is negatively correlated with seamlessness.

\[ \uparrow \text{BER} \Rightarrow \downarrow \text{LINKQUALITY} \Rightarrow \downarrow \text{SEAMLESSNESS} \]

\[ \downarrow \text{BER} \Rightarrow \uparrow \text{LINKQUALITY} \Rightarrow \uparrow \text{SEAMLESSNESS} \]

The set $V_f$ is partitioned in two subsets $V_f^+$ and $V_f^-$ where $V_f^+$ is the set of variables positively correlated with $f$ and, $V_f^-$ is the set of variables negatively correlated with $f$.

Furthermore, every $v_i$ may have associated a weight $w_i$ depending of its priority where $w_i \in \mathbb{R}[0,1]$ and $\sum w_i = 1$. Let $V$ represent the vector of variables $V = (v_1, v_2, ..., v_m)$, then the *objective function* for the desired handoff feature $f$ is defined by

\[ D(V) = \sum (k + w_i) \log(v_i^+) - \sum (k + w_i) \log(v_i^-) \quad (1) \]

where $k$ is a scaling factor so that small changes on the context variables reflect big changes on $D(V)$, $v_i^+$ and $v_i^-$ are positively and negatively correlated variables of $f$.

In general, the objective function is such that $D(V) : \mathbb{R}^m \rightarrow \mathbb{R}$ and is a utility function that we want to maximize because, when desirable features get higher, they represent that they get at the best.

Thus, considering $K$ different objective functions $D_k(V)$ that we want to maximize simultaneously where some of them may be in conflict, the multi-objective optimization problem (MOP) can be stated as the problem of

\[ \text{Maximize} \{ D_1(V), D_2(V), ..., D_K(V) \} \]

\[ \text{constrain to} \quad V_L \leq V \leq V_U, \]

where $V_L$ and $V_U$ represent the vectors of lower and upper values of the tolerance range for each variable.

F. Tradeoffs between Conflicting Objectives

A cognitive handoff is designed to achieve multiple purposes, objectives, and goals simultaneously. In the space of handoff objectives, we can distinguish between those with complementary nature and those with competitive nature. Complementary objectives can be simultaneously optimized without any conflict between them, but competing objectives cannot be simultaneously optimized, unless we find compromised solutions, largely known as the tradeoff surface, Pareto-optimal solutions, or non-dominated solutions [36]. We describe several tradeoffs to consider in a multi-objective handoff scheme:

a) (Max. DTIB and Min. HOR): There is a tradeoff between maximizing the time to stay always best connected (DTIB) and minimizing the number of handoffs (HOR). The conflict arises because in a dynamic environment the best network is changing frequently and stochastically; thus, to maximize DTIB is necessary to make frequent handoffs as soon as a new best is available. This increase in the number of handoffs creates a conflict with minimizing HOR.

b) (Min. DLat and Max. SHOR): This tradeoff is between minimizing the handoff decisions latency (DLat) and maximizing the number of successful handoffs (SHOR). The conflict emerges because the less time elapsed to make decisions will necessary lead to reduce the number of successful handoffs. For example, in case of imperative handoffs, DLat is reduced but this may lead to select an incorrect target because the selection time is also reduced.
c) (Max. Sizeof-ContextInfo and Min SO): This is a tradeoff between minimizing the handoff signaling overload (SO) and maximizing the amount of handoff context information to be managed by the handoff control entities. The conflict arises because broad handoff information is required to attain multiple desirable features, but this will increase the amount of signaling traffic in the network.

d) (User and Provider Preferences): Several conflicts may appear due to differences between provider and user preferences. For instance, providers may prefer networks within its own administrative domain while users may prefer networks with lower charges even if they are owned by other service providers; users may prefer a Mobile Controlled Handoff (MCHO) while providers may prefer Network Controlled Handoffs (NCHO). Conflicts like these require a balance between different interests. Handoff protocols like Mobile Assisted Handoff (MAHO) and Network Assisted Handoff (NAHO) try to balance the handoff control [9].

V. MODEL-DRIVEN METHODOLOGY FOR DEVELOPING COGNITIVE HANDOFFS

Next, we are going to describe the methodology to develop cognitive handoffs.

A. Difficulties for Developing Cognitive Handoff

The simple idea of achieving multiple purposes simultaneously is challenging even for humans. Moreover, if the intended purposes represent opposing situations, which all of them are desired, then even humans need a way to balance the different purposes in conflict; e.g., the conflict between doing the job accurately and doing it quickly. In optimization theory, multi-objective optimization states that improvements to a single purpose can be made as long as the change that made that purpose better off does not make any other purpose worse off. This is called a Pareto improvement. When no further Pareto improvements can be made, then the solution is called Pareto optimal [36].

Typically, a decision-maker chooses one optimal solution according to his preference. Therefore, the first difficulty in developing cognitive handoffs arises because there are many purposes, objectives, and goals all of them in conflict that need to be tradeoff.

A second significant difficulty emerges when numerous sources of environment information need to be considered to achieve the desired multiple purposes. Six sources of context we consider include: user, terminal, network, provider, application, and handoff process. Such sources produce context data that need to be collected, transformed, and distributed at the different handoff control entities (HCEs). The challenge is how to manage large amounts of unsorted high-dimensional data that have very complicated structures and at the same time reducing the signaling traffic overload produced by this task.

The last significant difficulty is originated by the different transition elements involved in the handoff process. These elements include radio channels, base stations, IP networks, service providers, user terminals, and all the feasible combinations. This variety of elements produces a large amount of scenarios that need to be considered for an adaptive handoff scheme.

B. Theoretical Background

First, we state the basis for establishing our methodology.

1) Holism and Reductionism: Holism and reductionism are two complementary and opposing approaches for analyzing complex systems [37]. They represent different views of the relationship between the whole and the parts. Holism states that parts cannot explain the whole, the whole states the behavior of parts; i.e., it is necessary to understand how the entire handoff system determines the behavior of its components. Conversely, reductionism states that parts can explain the whole, then the behavior of parts determine the behavior of the whole. We have seen how reductionist handoff schemes achieve its goals in specific scenarios but they quickly become special cases of more general models. Holistic models are more complex models that pretend to consider all the individual parts and to understand the purposes of the whole.

2) Model-based Design: The model-driven paradigm has emerged as one of the best ways to confront complex systems. As it was clearly expressed by Dr. Hoffman [38], models can capture both the structure of the system (architecture) and behavior (dynamism). Model-based systems engineering [39] helps to address complexity by raising the level of abstraction, enabling developers to view system models from many perspectives and different levels of detail while ensuring that the system is consistent. The Systems Modeling Language (SysML) [38,39] is becoming an accepted standard for modeling in the systems engineering domain. Using SysML for modeling helps to reduce ambiguity in models. In fact, models can now show the dynamic behavior of systems, including how they transition between states and how the system behaves overall.

3) Functional Decomposition: refers to the process of resolving a functional relationship into its constituent parts in such a way that the original function can be reconstructed from those parts by function composition. The process of decomposition [40] is undertaken for the purpose of gaining insight into the constituent components.

4) Design as Scientific Problem-Solving: In his inspiring paper, Braha [41] showed the similitude between the systems design process and the solving-problem process.
Therefore, we developed his foundation and proposed a methodology establishing a general procedure that starts with a problem statement and ends up with the solution deployment. This theory views the problem statement as the initial state and then, by searching through a state-space, reaches a goal state representing the solution.

C. Design and Development Procedure

The steps involved in a form of top-down procedure are:

1) Stating the problem: Develop a handoff procedure that can optimally achieve multiple desirable features simultaneously. The handoff procedure should be implemented for operating in real scenarios with multiple dimensions of heterogeneity. Then, as part of the problem:

   a) Identify and analyze the required system functions: Study the desirable handoff features that need to be implemented and determine the purpose, objectives, and goals associated to every feature. Associate a clear and single purpose to every desirable feature. Decompose each purpose into one or more objectives by identifying the performance parameters that help to quantify the achievement of every purpose. In the same way, divide every objective into one or more specific handoff goals, using optimization values and handoff context data and

   b) Determine the needed handoff context information: Establish what handoff criteria, handoff metrics, performance measures, handoff policies, handoff constraints, and handoff scenarios are needed to achieve every desired purpose. Study the availability, locality, dynamicity, structure, and complexity of the variables, policies, and constraints to use.

2) Design a subsystem structure or model-based framework: State a cognitive handoff conceptual model, i.e., identify all external context information as well as all internal context information with the highest abstraction level. Whilst internal data constitutes self-awareness, external data constitutes context-awareness of the handoff process. Then, using functional decomposition divide up the conceptual model into a number of sub-models. Every sub-model corresponds to a particular sub-problem that functionally is part of the whole handoff problem. The structure of the system may be represented with a hierarchy of models or framework enclosing the parts of the whole system organized through functional relations. Models in this framework describe the system behavior in an accurate and unambiguous way if one uses a finite set of states and a set of transition functions, thus to ease this part: Identify the associated system states and phases. These dynamic models can be formally represented using finite automata, Petri nets, timed automata, etc. [42]. The states or phases of the handoff process should describe a general behaviour rather than specific details of particular sub-models.

3) Execute the models: Execution of models allows verification and validation of such models. This is the difference between just drawing pictures and making pictures “live” as it was pointed out by Hoffmann in [38]. However, verification and validation should not be confused. Model verification means to test if the model satisfies its intended purposes or specifications. Model validation tests if the model provides consistent outcomes that are accurate representations of the real world. We use three strategies for these tasks: simulation, prototyping, and analysis. Whatever the strategy we choose, model testing or model checking [43] requires the use of a formal notation; e.g., modelling languages for simulation, mathematic and logic for analysis, and programming languages or middleware for model prototype implementation. If a model cannot be properly validated or verified, then it must be redesigned within the framework.

4) Implementation stages: Once all the models in the framework have been individually tested, the design problem now reflects a well-structured solution. A detailed design can now be generated considering the entire framework of models. This whole system design should be implemented in a whole system prototype. The final prototype is ready to be tested in-situ; should any failure occur during testing, then a review of the conceptual model or any sub-model in the framework should be performed.

5) Solution deployment: The cognitive handoff solution is ready to operate on a real handoff environment. The solution system (cognitive handoff) provides a simultaneous accomplishment of the multiple purposes defined by the handoff problem. Each purpose should be associated to quantitative objective functions to measure the degree in, which every handoff purpose was achieved.

VI. APPLYING THE MODEL-DRIVEN METHODOLOGY

Now, we are going to apply the previously proposed methodology to develop cognitive handoffs.

A. Purposes, Objectives, Goals, and Context Data

The handoff context information is extensive, heterogeneous, distributed, and dynamic. It supports the whole operation of the handoff process and the achievement of multiple desirable features. From the external and internal vision of the handoff environment, we have identified five external sources of context information (creating context-awareness) and one internal source, which is the handoff process itself (creating self-awareness):

1) User context: This context includes the user preferences, user priorities, user profiles, and user history and it is used to respond to user needs, habits, and preferences.

2) Terminal context: This context domain includes the following evaluating parameters: (i) Link quality: Received Signal Strength (RSS), Signal-to-Noise Ratio (SNR), Signal-to-Noise-and-Interference Ratio (SNIR), Bit Error Rate (BER), Block Error Rate (BLER), Signal-to-Interference Ratio (SIR), Co-Channel Interference (CCI),
Carrier-to-Interference Ratio (CIR), etc.; (ii) Power management: Battery Type (BT), Battery Load (BL), Energy-Consumption Rate (ECR), Transmit Power in Current (TPC), Transmit Power in Target (TPT), and Power Budget (PB); (iii) Geographic mobility: Terminal Velocity (Vel), Distance from a Base Station (Dist), Geographic Location (Loc), Moving Direction (MDir), and Geographic Coverage Area (GCA). All these evaluating parameters allow the deployment of QoS-aware handoffs, power-based handoffs, and location-aided handoffs.

3) Application context: It includes the QoS requirements of running applications; Lost Packets (LP), Delayed Packets (DP), Corrupted Packets (CP), Duplicated Packets (DuP), Data Transfer Rate (DTR- goodput), Packet Jitter (PJ), Out-of-Order Delivery (OOD), Application Type (AppT).

4) Network context: This information is necessary to select among networks (before handoff), to monitor service continuity (during handoff), and to measure network conditions (after handoff) thus they are: Network Bandwidth (NBW), Network Load (NL), Network Delay (ND), Network Jitter (NJ), Network Throughput (NT), Network Delay (ND), Network Load (NL), Network Delay (ND), Network Maximum Transmission Unit (NMTU).

5) Provider context: Information about connection fees, billing models, roaming agreements, coverage area maps, security management (AAA), types of services (data, voice, video), provider preferences, and provider priorities.

6) Handoff performance context: This information forms the self-aware part of our cognitive model and allowing evaluation of its performance. Call Blocking (CB), Call Dropping (CD), Handoff Blocking (HOB), Handoff Rate (HOR), Handoff Latency (HOL), Decisions Latency (DLat), Execution Latency (ExLat), Evaluation Latency (EvLat), Handoff Type (HOTYPE), Elapsed Time Since Last Handoff (ETSHL), Interruptions Rate (IR), Interruption Latency (IL), Degradations Rate (DR), Degradations Latency (DL), Degradations Intensity (DI), Utility Function (UF), Signaling Overload (SO), Security Signaling Overload (SSO), Improvement Rate (ImpR), Application Improvement Rate (AppImpR), User Improvement Rate (UsrImpR), Terminal Improvement Rate (TermImpR), Successful Handoff Rate (SHOR), Imperative Handoff Rate (IHOR), Opportunistic Handoff Rate (OHOR), Dwell Time In the Best (DTIB), Authentication Latency (AL), Detected Attacks Rate (DAR), Online User Interventions Rate (OUIR), Tardy Handoff Rate (THOR), and Premature Handoff Rate (PHOR).

Once we have identified the context data from all the context sources and the desired handoff features that we wish to implement, then, we assign a qualitative purpose to every desired feature and, a set of quantitative objectives and goals to every handoff purpose. Tables III and IV summarize such previous description.

<table>
<thead>
<tr>
<th>Desired Handoff Features</th>
<th>Qualitative</th>
<th>Quantitative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seemlessness</td>
<td>Maintain continuity of services or preserve user communications</td>
<td>Reduce DR, DL, DI, IR, IL</td>
</tr>
<tr>
<td>Autonomy</td>
<td>Preserve handoff operation independent of users</td>
<td>Reduce OUIR</td>
</tr>
<tr>
<td>Security</td>
<td>Maintain a constant level of security along the handoff</td>
<td>Reduce SSO, DAR</td>
</tr>
<tr>
<td>Correctness</td>
<td>Keep user always connected to the best network with minimal handoffs</td>
<td>Reduce HOR, Increase DTIB</td>
</tr>
<tr>
<td>Adaptability</td>
<td>Keep success of all handoff objectives across any scenario</td>
<td>Multi-objective optimal balance Increase SHOR</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Other Desired Properties of Cognitive Handoffs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desired Handoff Features</td>
</tr>
<tr>
<td>--------------------------</td>
</tr>
<tr>
<td>Necessary</td>
</tr>
<tr>
<td>Selective</td>
</tr>
<tr>
<td>Efficient</td>
</tr>
<tr>
<td>Beneficial</td>
</tr>
<tr>
<td>Timely</td>
</tr>
</tbody>
</table>
These tables represent a relevant preliminary result of the applicability of cognitive handoff methodology. On one hand, they help to reduce the ambiguity and confusion on the usability of similar handoff features because every desirable handoff feature is defined in qualitative terms (purpose) and quantitative terms (objectives and goals). On the other hand, they help to correlate context data with desirable features. For instance, from Table III, we observe that RSS is correlated with seamlessness, IL with autonomy, AL with security, etc. This correlation is intended to select the context data that is needed to support every handoff purpose.

B. Taxonomy of Handoff Mobility Scenarios

A second significant result obtained from the proposed model-driven methodology is a new taxonomy of handoff mobility scenarios derived from combining all the possible transition elements involved in handoffs; i.e., channels, cells, networks, providers, and terminals. This taxonomy depicts all different kinds of handoffs that are possible in real networks.

Nowadays, no handoff solution exists, which comprehensively addresses the entire scale of heterogeneity. Multidimensional heterogeneity is the reason for the large number of handoff scenarios. If we define a handoff scenario as an array \((d_1, d_2, \ldots, d_n)\) where \(d_i\) is an instance of \(D_i\), the \(i\)th dimension of heterogeneity and there are \(|D_i|\) different ways to instantiate the \(i\)th dimension, then by the multiplication principle there will be \(|D_1| \times |D_2| \times \ldots \times |D_n|\) possible handoff scenarios. However, for the user mobility dimension, the array (location, velocity, direction) may have distinct values at any instant along the path with infinite paths crossing the network; therefore, the number of possible mobility scenarios is infinite. Despite of such infinite scenarios, it is important to make a classification of handoffs according to the elements involved during the transition.

The complexity and treatment for a handoff depend on the type of transition that is occurring. A handoff will require of services from distinct OSI model layers depending on the elements involved in the transition. For example, a handoff between channels of the same cell is a layer 1 handoff; a handoff between cells (base stations) is a layer 2 handoff; it is homogeneous if cells use the same wireless technology, otherwise is heterogeneous; a handoff between IP networks is a layer 3 handoff; a handoff from one provider to another or between user terminals will demand the services of layers 4-7.

Fig. 5 depicts the hierarchical structure of a mobile Internet in a four-layer design (core, distribution, access, and mobile). We will use this figure to explain a handoff hierarchy that involves channels, cells, networks, providers, and terminals.

The mobile Internet is divided into independent administrative units called Autonomous Systems (AS). An AS is a network administrated by a single organization or person. The Internet is a network of autonomous systems.

Fig. 5 presents two autonomous systems called ISP1 and ISP2 for two distinct service providers. Every ISP uses a very high-speed core network where main servers are located. Providers divide their distribution networks, physically and logically, into a number of IP networks, subnets, or VLANs (Virtual LANs), where the types of services and users are separated. Each IP Net includes a group of base stations (BS) or access points with the same or different wireless access technology. Base stations get distributed across a geographic area to offer mobile communication services. Each base station controls a cell that may have a group of channels to distribute among the associated terminals or a single channel that is shared among several associated terminals.

In Fig. 5, BS2 illustrates a layer 1 handoff when the mobile terminal (MT) changes its connection between channels ch1 and ch2 without changing of BS, IP Net, ISP, or MT. A layer 2 handoff is illustrated between BS1-BS2, BS3-BS4, BS5-BS6, and BS7-BS8. Note that layer 2 handoff changes from one channel to another and from one base station to another, but keeps the same IP Net, ISP, and MT; however, if the cells involved are heterogeneous, then the handoff is vertical, otherwise is horizontal. A layer 3 handoff is depicted in BS2-BS3 and BS6-BS7. Note that layer 3 handoff changes from one channel to another, from one cell to another, and from one IP network to another, but preserves the same provider and the same terminal; the layer 3 handoff may be heterogeneous, like in BS2-BS3, or homogeneous, like in BS6-BS7. We represent a layer 4-7 handoff, in BS4-BS5, when MT changes its communications from one channel to another, from one cell to another, from one IP Net to another, and from one ISP to another, but the user keeps the same terminal.

The encryption schemes and data representation formats change from one provider to another, thus higher layer services are required. Inside the cell for BS5 we depict a handoff between terminals where the user transfers the whole session (current state of running applications) from terminal MT-A to terminal MT-B. Handoffs between terminals can be done for terminals within the same cell or different cells, within the same IP network or different IP networks, within the same provider or different providers. The terminal handoff depicted in BS5 keeps the same cell, same IP Net, and same ISP.
C. Cognitive Handoff State-Based Model

By applying the second step of the model-driven methodology, design a subsystem structure, we created a cognitive handoff conceptual model and its first decomposition model both illustrated and discussed in Section III. Following the reductionist approach, we now focus on a major component of the handoff system, the cognitive handoff control system. At this stage, we designed a state-based model whose purpose is to understand the general behavior that should have the handoff control system. Thus, this model represents our third main result obtained from following the methodology.

Fig. 7 shows a five-state diagram modeling a general control handoff process. The states are: (1) Disconnection, (2) Initiation, (3) Preparation, (4) Execution, and (5) Evaluation. This model describes a generic control handoff system coordinating the stages before, during, and after the handoff.

We describe each state briefly:

1) **Disconnection**: is the initial state and one of the two final states. Here, the terminal is disconnected but discovering available networks. The process will stay here while there are no available networks.

2) **Initiation**: in this state the terminal is connected to the best available network and communications flow normally. This is another final state. The process stays here while there are no reasons (imperative or opportunistic [45]) to prepare for a handoff. If current connection breaks and no other network is available, then the process goes back to the disconnection state.

3) **Preparation**: as soon as a better network appears, the process changes to the preparation state. Here is where properly the handoff begins. This state decides why, where, how, who, and when to trigger the handoff. The handoff in progress can be rolled back to initiation if current link becomes again the best one.

4) **Execution**: once a control entity decides to trigger a handoff, there is no way to rollback; the handoff will be performed. This state knows the current and destination networks, the active application to be affected, and the strategy or method to use.

Fig. 6 presents a process diagram that generates the complete taxonomy of handoffs by following the different paths from the upper node to the lower nodes.

Every handoff type in this taxonomy should be complemented or further classified according to many other criteria by using the handoff classification tree of Nasser et al in [44].
5) **Evaluation:** once the link switch is made, the control entity enters the evaluation state. This state recombines the measures for every objective function taken before and during the handoff, with new samples taken after the handoff to determine its successfullness. The evaluation latency is adjusted to a stabilization period [46].

Let us develop the state-based model as follows.}

### VII. Discussion

In this research, we have shown a new methodology to systematically develop cognitive handoffs, which are expected to be in operation in the mobility scenarios of the future Internet. Such methodology is based on a sound theoretical framework including: methods for analyzing complex systems, the model-based systems engineering, the functional decomposition approach, and the scientific problem-solving theory. There are five stages in the proposed methodology: 1) state the problem, 2) design a model-based framework, 3) execute the models, 4) implement a prototype, and 5) deploy the solution. Thus, we have presented three main results obtained from applying the first two stages of the methodology: i) a cascade relationship of desired features, purposes, objectives, goals, and context data; ii) a taxonomy of handoff mobility scenarios; and iii) a generic state-based model for a cognitive handoff control system.

Furthermore, there are some other issues that require detailed discussion: (a) the complexity of a cognitive handoff system, (b) the evaluation of cognitive handoff models, and (c) the implementation of cognitive handoffs.

#### A. Cognitive Handoff Complexity

In Section III we showed two main properties of complex systems that are also present in cognitive handoffs: the hierarchic structure of systems and the property of emergence. Now, in this section we provide other reasons of why cognitive handoffs are complex software systems: (1) Cognitive handoffs exhibit a rich set of behaviors: reactive, proactive, deterministic, non-deterministic, context-aware, self-aware, etc.; behavior is determined by the particular desirable features associated to handoffs. (2) Cognitive handoffs can be stated as multi-objective optimization problems. (3) Cognitive handoffs are driven by events in the physical world; e.g., the user mobility, the user preferences, the provider services, the coverage areas, etc. (4) Cognitive handoffs maintain the integrity of hundreds or thousands of records of information while allowing concurrent updates and queries. (5) Context information is extensive, heterogeneous, dynamic, and distributed. (6) Cognitive handoffs control real-world entities, such as the switching of data flows through a large set of available networks, providers, and terminals. (7) Handoff management has a long-life span; handoffs will exist in all future wireless networks. (8) Handoff management is a key issue for wireless industry and standardization bodies. Grady Booch in [46] provides further discussion on the attributes of complex software systems.

#### B. Evaluation of Cognitive Handoff Methodology and Models

Now, as a result of applying our proposed methodology, one gets a set of models that are different in purpose (intentions), usability (applicability), notation (language), and abstraction (hierarchy).

Methodology and each model must be evaluated, either by quantitative evaluation, which comprises the definition of criteria and metrics intended to measure one specific property or, conversely by a qualitative evaluation, which is related to credibility that comes from the way in, which the cognitive maps are built and the clarity it represents the opinion’s of most experts [48].

In relation to a qualitative evaluation of the methodology, one requires to think on the stages proposed by the development process, the kind of activities to accomplish in each stage, the strength of its theoretical basis, the kind of lifecycle in the development process, etc. Meanwhile, corresponding quantitative evaluation, metrics should be applied to all associated parameters in the stages of the process.

With respect to evaluate models, we made a clear distinction in Section II.C between verification and validation. The verification tests if the model satisfies its purpose, whilst validation tests if the model outcomes are representations of reality. During the development process of a new system, special purpose models are built to support the understanding that goes on during the development and no hard data emerge from such models, thus, they can only be verified, but not validated.

It is worth to notice that in this paper, we deal with a specific kind of model belonging to those known as soft models [48]. Soft models are intended to understand rather than predict and therefore verification is the way to qualitatively evaluate such models. Specifically, the theoretical framework in Section II.B has solid and proven bases.

#### C. Cognitive Handoff Implementation

We envision the implementation of cognitive handoffs as a network of distributed agents cooperating and competing to take any type of handoff to success. We distinguish between agents for controlling the handoff process (HCEs) and agents for managing the handoff context data (CMAs). The CMAs are responsible for recollecting the context data and updating the handoff information base at the HCEs. CMAs are located in user terminals and distributed in different layers of the network infrastructure. HCEs are located also in every user terminal and at the network access layer; HCEs perform a handoff control process like the one depicted in Fig. 3. Thus, let us develop the state-based model as follows.
A dynamic ordered list of available networks (ANL) is organized from best to worst, according to the value of desirability calculated for every network. The desirability metric is a utility function combining a broad set of network selection criteria. The best network is the one with highest desirability. The value of desirability for the \( n \)th network, named \( D_n(V) \), may have a geometric or stochastic distribution depending on the dynamic nature of context variables used as selection criteria, and arranged in a criteria vector \( V = (v_1,v_2,\ldots,v_n) \). We use Equation (1) to represent a general mathematical model for the desirability function:

\[
D_n(V) = \sum (k + w_i) \log(v^+_i) - \sum (k + w_j) \log(v^-_j)
\]  

(3)

The set of decision variables \( (v_1,v_2,\ldots,v_n) \) fetched for the \( n \)th available network is partitioned in two subsets: \( V^+_i \) and \( V^-_j \); where \( V^+_i \) is the set of criteria that contribute to the desirability (e.g., NBW and NT) and \( V^-_j \) is the set of variables that contribute to the undesirability (e.g., NL and ND). \( w_i \) and \( w_j \) are weights corresponding to each variable such that \( w_i \in \mathbb{R}[0,1] \), \( \sum w_i = 1 \) and \( k \) is a scaling factor so that small changes in the context variables reflect big changes in \( D_n(V) \).

For geometric distributions, a proactive handoff strategy may anticipate handoff decisions and for stochastic distributions a reactive handoff strategy with thresholds, hysteresis margins, and dwell-timers may prevent unnecessary handoffs. The control handoff process illustrated in Fig. 3 shows a reactive and deterministic procedure; reactive, because the process starts the preparation for a handoff until another network with higher desirability is present and, deterministic, because it is always possible to determine the current state of the process within one of five states.

Fig. 8 and Fig. 9 depict geometric distributions of desirability with different handoff strategies. Fig. 8 shows a proactive strategy where the handoff preparation starts before the target network improves the current connection. Fig. 9 shows a reactive strategy where handoff preparation starts after the target network has improved the current connection.

The darken line over the desirability functions illustrate the current connection. The performance parameters APREP, AEXEC, AEVAL, and AVHO depict the latencies for the different stages: preparation, execution, and evaluation. Configuration parameters include \( \Delta \) (hysteresis margin), desirability threshold (Thsup, Thinf), and dwell-timer (SP). Relative Desirability measures are (\( \Delta R_s \)), which are equal to \( |D_{curr} - D_{best}| \).

The available network list (ANL) is a data structure located at the HCEs, but continuously updated by the CMAs. When the ANL is empty, the terminal goes to the disconnection state (State 1) and stays there while such list is empty. CMAs are continuously discovering new networks and ordering the list from the highest desirable networks to the lowest desired networks.

The change from disconnection state to initiation state (State 2) occurs as soon as new networks are available. The HCE selects the best available network from the list and connects the terminal to it. The State 2 is the Always Best Connected state because the terminal will stay connected to the best network as long as no other available network improves the current connection.

The change from initiation to preparation (State 3) occurs when a new network is improving or has improved the current network. Handoff decisions, in State 3, start by identifying a reason to begin the preparation for a handoff (why). Next, selecting the target network (where). Then, deciding what strategy, method, or protocol to choose (how). Then, deciding what HCE will be responsible to trigger the handoff (who), and finally, deciding the best moment to trigger the handoff (when). The chosen handoff strategy,
method, or protocol depends on the current handoff scenario (as those depicted in Fig. 5) and the type of handoff in progress (as those illustrated in Fig. 6).

The decision to trigger a handoff in one terminal changes the control process from preparation to execution (State 4). The trigger handoff decision activates a procedure to change the data flows of an application from one access network to another, within specific handoff and time constraints. The switching mechanism takes a time \( \Delta \text{EXEC} \) to complete.

Once the switching process is completed, the HCE enters to the evaluation state (State 5). This is an important stage of feedback to the handoff control process. At this stage, the HCE has a constrained period of time to decide to accept or reject the recently executed handoff. One condition for handoff success occurs if the new current connection is the best available connection, but others include measuring the objective functions, associated to every handoff purpose, and if all these measures are within a boundary region of acceptable quality, then the cognitive handoff is successful, otherwise it is defective and outliers should be corrected.

VIII. MODEL RESULTS

So far we have described a challenging handoff optimization problem and we have created a series of models to study the problem. Moreover, we proposed both: a computational model that offers a heuristic solution to the problem and a methodology to implement cognitive handoffs. Therefore, now we are interested in a simulation instrument that can help us to validate the behavior of a given specific handoff algorithm over a variety of handoff scenarios based on time, space, or both and measure particular performance quantities. To this end, we created a Relative Desirability Handoff Algorithm with hysteresis, dwell-timers, and two thresholds in order to make a terminal stay most of the time on the best network, while it performs the fewer number of handoffs on most handoff scenarios.

Fig. 10 shows an example that considers our particular cognitive handoff algorithm and a user defined valid handoff scenario. The handoff scenario consists of two networks, one that changes abruptly and rapidly and another that changes smoothly and slowly. Lower and upper thresholds are defined within the visual area, \( L = -1 \) and \( U = 4 \), separating the graphics into three handoff regions.

The bottom thick line depicts the current network passing through different handoff states: initiation (black), preparation (blue), evaluation (pink), disconnection and execution (red).

![Figure 10. Visual outputs of handoff simulator with additional visual aids](image-url)

Each test in the virtual instrument displays graphically the behavior of the handoff algorithm and yields handoff performance data which, are collected in a structured file. The handoff collected data include handoff performance measures and the handoff scenario.

We design a nondeterministic experiment for collecting representative samples of input handoff scenarios which, will be used to test our proposed algorithm. The algorithm performs a cognitive handoff from the current network to the best candidate network in order to stay in the best available connection most of the time; i.e., increase DTiB. Simultaneously, this algorithm tries to perform the fewer number of handoffs because each handoff entails some overload to communications; i.e., decrease nEHO. However, these tasks are in conflict, they cannot be improved simultaneously. As a result, this algorithm makes a balance between increasing DTiB and decreasing nEHO. The way of doing this balance is by delaying the execution of a handoff until it becomes really necessary, i.e., until the candidate network becomes sufficiently and consistently better.

This algorithm obtains three performance measures (rTiB, rEHO, rBHO) which, are associated to the particular handoff scenario under analysis. Values for rTiB \( \geq 50\% \) or rEHO \( \leq 50\% \) are considered good or acceptable results.

In this experiment, we ask three users to define at their own will several statistically valid scenarios User “A” made 32 trials, user “B” 84, and user “C” 133, which, gives a total sample size of 249 tested scenarios.
Since for each input scenario, the instrument records and measures three handoff performance parameters: rTiB, rEHO, and rBHO, then the space of handoff results will be composed of data obtained from each test. By observing the distribution of sample data within the space of results, we may compute the degree of achievement of each performance goal.

Fig. 11 shows a scatter diagram of 133 random sample points obtained by user “C”. The graphic presents 17 samples in the space for very good results and very good balance which, represent the percentage of 12.78%. It includes 95 (17+78) samples in the space for good results and good balance, which represent the 71.43% of the sample size; and, 121 (17+78+26) samples in the space for good results which, represent a hit rate of 90.98%. The diagram also illustrates 12 sample points located in the space for bad results, representing a 9.02%. The random experiment of 133 samples meets all the percentage goals: for good results (90.98% > 90%), for good results and good balance (71.43% > 50%), and for very good results and very good balance (12.78% > 10%).

Table V presents a summarization of results taken from the testing experiment of the handoff instrument. This table compares the percentages of sample points falling in each region of handoff results with the different random samples obtained from the experiments. It can be seen that the hit rates in all testing cases meet the handoff performance goals.

The handoff simulation instrument produced, in average, a rate of “good” results above 90% or a rate of “bad” results below 10%, a rate of “good” results and “good” balance above 50%, and a rate of “very good” results and “very good” balance above 10%.

Therefore, all these results provide evidence that support the correctness of our proposed algorithm based on our cognitive handoff model and methodology as well the usefulness of the taxonomy to properly define scenarios.

IX. CONCLUSION AND FUTURE WORK

Handoffs are an integral component of any mobile-wireless network from past, present, and future. Handoffs are transitions that change the data flows from one entity to another, where these entities may be radio channels, base stations, IP networks, service providers, and user terminals. The handoff process should exhibit several desirable features beyond seamlessness and should consider more context information beyond the signal strength. This is a common requirement to face the handoff scenarios of the future Internet.

The existing handoff schemes are not able to achieve a variety of attractive features and managing arbitrary amounts of context information. Therefore, we proposed a conceptual model to create handoffs of this kind. We characterized a cognitive handoff to be multipurpose, multi-criteria, context-aware, self-aware, and policy-based.

We claimed that our cognitive handoff model is holistic because it considers all the transition entities that may be involved in handoffs, all the external and internal sources of context, and considers many significant desirable features.

Using a functional decomposition approach, we divided the functional behavior of a cognitive handoff into six general modules: control algorithm, network discovery, handoff decisions, handoff execution, handoff evaluation, and context management. Each module has assigned a purpose to every feature and decomposed each purpose into objectives and goals. We applied the cognitive handoff model to define its performance parameters and significant tradeoffs between conflicting objectives.

Table V  SUMMARY OF TEST RESULTS USING THE HANDOFF INSTRUMENT

![Figure 11. Scatter diagram for rTiB vs. rEHO 133 observations made by user “C”](image-url)
We proposed a new model-driven methodology for developing cognitive handoffs. We applied the proposed methodology and obtained a clear relationship between handoff purposes and handoff context information, a new taxonomy of handoff scenarios, and an original state-based model of a generic control handoff process.

We continue developing and integrating the models generated by the cognitive handoff methodology. A future work is to organize such models in a comprehensive framework of models representing the functional issues for the whole cognitive handoff process. Further work is needed to study the availability, locality, dynamicity, structure, and complexity of variables, metrics, polices, and constraints involved in cognitive handoffs.

ACKNOWLEDGMENT

F.A. González-Horta is a PhD candidate at INAOE Puebla, Mexico, and thanks the financial support received from CONACYT Mexico through the doctoral scholarship 58024.

REFERENCES


