Evaluation and testing methodology for evolving entertainment systems

Audrius Jurgelionis
University of Genoa
Biophysical and Electronic Engineering Department
Via Opera Pia 11a, 16145 Genoa
+39 010 3532780
jurge@elios.unige.it

Francesco Bellotti
University of Genoa
Biophysical and Electronic Engineering Department
Via Opera Pia 11a, 16145 Genoa
+39 010 3532780
franz@elios.unige.it

Wijnand IJsselsteijn
Eindhoven University of Technology
Game Experience Research Lab
Den Dolech 2
5600 MB Eindhoven
+31 40 2474455
w.a.ijsselsteijn@tue.nl

Yvonne de Kort
Eindhoven University of Technology
Game Experience Research Lab
Den Dolech 2
5600 MB Eindhoven
+31 40 2475754
y.a.w.d.kort@tue.nl

ABSTRACT
This paper presents a testing and evaluation methodology for evolving pervasive gaming and multimedia systems. We introduce the Games@Large system, a complex gaming and multimedia architecture comprised of a multitude of elements: heterogeneous end user devices, wireless and wired network technologies, and various multimedia and gaming services. It enables an easy access to the offered services through the use of a versatile system architecture. Testing and evaluating the design takes central stage in the project through an iterative process which has a great impact on overall design course and the final product in terms of assuring a proper quality of service and usability of the system. A comprehensive testing and evaluation methodology for newly evolving entertainment systems is developed after the adaptation of the metrics, used for video game and multimedia evaluation.

Keywords
Testing, evaluation, usability, user-centred design, usability evaluation, pervasive entertainment, multimedia, video games.

1. INTRODUCTION
The future home is the always-on connected digital home. By 2010, there will be over 420 million broadband households' worldwide [2]. With the standard set for super-high speed, always-on connection, the way people view entertainment has fundamentally changed and has created a new standard for media consumption. Consumers no longer expect their internet access to be only from a desktop PC- now they want it through the TV in the living room or in the palm of their hand, inside the house and on the go.

The new scenario will bundle video gaming capabilities into consumer electronics devices, such as set-top-boxes, DVRs (Digital Video Recorder), home entertainment systems, TVs, handhelds and other devices that are currently not considered as real gaming devices since they lack the necessary CPU power and graphical performance [2].

In this paper we present the testing and evaluation methodologies developed under the EU FP6 project, Games@Large (GaL), which has the aim to design a new platform for pervasive multimedia and gaming [1]. The testing and verification process is part of the iterative, spiral-life workflow model (user-centred design and incremental improvement based on feedback from user and expert evaluation of prototypes). The proposed testing methodology is based on literature studies of multimedia and video game testing and evaluation methods. These methods are adapted to GaL specific needs according to their relevance for testing GaL system elements.

In the next section of the document we briefly introduce the new framework for pervasive gaming and multimedia. The third section describes user-centred design aspects that are relevant to the design testing and evaluation. In the final section we describe the approach for pervasive system testing and evaluation, which is based on user experiences and perceptions about the system design. In parallel with user based testing and evaluation, necessary functionality tests are considered as well.

2. SYSTEMS FOR PERVASIVE GAMING AND MULTIMEDIA

The increasing number of broadband users, and a demand for quality and diversity in entertainment services, drives the development of new pervasive entertainment systems. Services should be easily accessible without limits on time and location. Within the GaL project we are developing and testing the architecture that is required to provide such service [1], [2].

GaL is not developing any of the offered content, either video games or multimedia; but it integrates and supplies them to the end user. Provision is achieved by the employment of networking technologies and integration of heterogeneous end user devices into GaL, see Fig. 1, infrastructure. The versatile GaL system architecture and effective data encoding/decoding (pre-rendering or video streaming approaches) techniques [2] facilitate the system's pervasiveness in terms of location, served users and hardware configurations.

The GaL user-centred design process is focused on the QoS/bandwidth/latencies requirements of each multimedia service. Moreover, the final test sessions will also include samples of such services in order to test the system in realistic conditions of use [2].

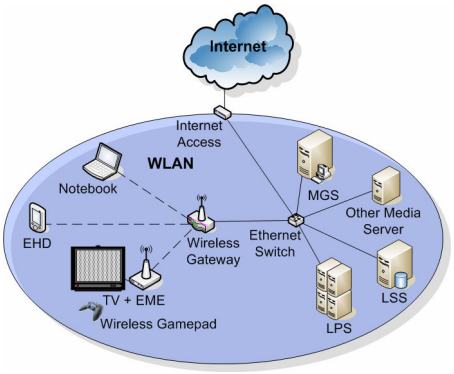

Figure 1. Games@Large framework.

In the next section we present user-centred design (UCD) principles and their relevance to the design and testing of pervasive entertainment systems.

3. THE IMPORTANCE OF UCD FOR PERVASIVE GAMING SYSTEMS

The design of interactive products and services [1] is ultimately about creating interactive experiences for people. Being user-centred means putting people first – designing products and services in such a way that they support people and are enjoyable to use and own. In particular, with increased functionality and complexity of technological products as well as a significant broadening of the user base, especially in the area of digital games, it has become crucial to involve potential end-users from an early stage onwards in the design process to ensure a product's usability and enjoyability. Thus, technical excellence, although necessary, is not sufficient in itself. User-centred design can help make a product or service fulfil real user needs, adapted to user capacities, limitations, and preferences, safe in its operation, and accessible to the largest possible range of users.

User-centred design of productivity-oriented applications has traditionally focused on usability as one of its most relevant quality metrics. Usability is defined as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" [4]. *Effectiveness* refers to whether or not a user is at all able to accomplish a certain task with a system. *Efficiency* is the extent to which a user is able to be productive with a system once he or she has learned how to use it. This is typically expressed in number of errors or time needed to complete the task. *Satisfaction* refers to the extent the user finds using the system to be subjectively pleasing. Nielsen [5] uses a somewhat broader conception of usability which also includes *learnability*, i.e., the extent to which a system is easy to learn so that the (novice) user can immediately be effective, and *memorability*, the extent to which a system is easy to remember, so that a casual user can return to it after some period of not having used it, without having to learn everything all over again.

In general, during usability tests a representative sample of potential end-users is asked to perform a representative set of tasks using the interactive product or service under study. They may be video-taped, observed, and/or timed on task performance.

Usability problems can be a serious showstopper for pervasive gaming systems such as GaL. For example, if a user is confronted with an interface that is too obscure to review the various game options, the user will opt out. If the user cannot adequately access an easy yet secure payment procedure, the user will opt out. If the user finds the mapping of the in-game functions onto the controls too difficult to operate, the user will opt out. If the user finds the game's responsiveness terribly slow, the user will opt out. In short, there are different levels at which usability problems may occur, all of which have the potential to be seriously detrimental to the overall user experience, resulting in user frustration, and consequently, in the user turning away from the product.

At present, there are no standard methods of assessing the full gamut of game-related experiences [3]. For the purposes of GaL, one of the measures we will employ is a newly developed questionnaire, the Game Experience Questionnaire, which covers, flow [6], [7] and immersion [7], [8] as well as a number of additional game-related experiences. This questionnaire will be employed as part of our iterative multilevel system testing approach, which is described in the next section.

4. TESTING METHODOLOGY

Given GaL testing specificity we will take the best suited UCD methods [3] as a point of departure and extend them according to the specific features of the GaL system [2]. Methods used for on-line game testing and evaluating of the effects of specific network parameters, [9], [10], as well as MPEG [12] and AVC/H.264 [11], [12], video verification are suitable for GaL testing purposes, and can serve as reference methodologies. User interface [13], [14], network and device tests are to be integrated in the multilevel GaL system testing. Thus a comprehensive methodology is developed for the purpose of verification of the different GaL components and the system as an integrated whole.

The GaL platform is intended for running interactive multimedia applications such as computer games. In contrast to video streaming or music, computer games are sophisticated and

sensitive applications in terms of varying network conditions and device performance parameters. The GaL framework is composed of both wireless and wired networks, different end devices (and control devices) and applications (middleware) built on the basis of specific protocols, which are the medium that connects different system parts and makes the whole system functional. An important aspect is the user interface, which provides the user access to a number of the system services. Thus, the whole system performance and usability is conditioned by a number of its elements. Therefore the GaL system is differentiated into several groups of primary elements [2]. Testing methods for each of the elements are considered and in the end the final suite of testing and evaluation methodology is predefined.

4.1 Multilevel analysis approach

In this chapter we present the multilevel GaL system analysis approach, which is considering the whole system testing and evaluation.

In order to comprehensively evaluate the overall system performance and usability [3], and to assess the user perceived quality of service [2], the four-level scheme (Fig. 2) of the testing is specified. The evaluation process is focusing on the evaluating of user perceptions and experience (user level analysis). The entire system functionality and usability is considered in the user level analysis. Additionally, some specific tests for particular GaL components are necessary in order to obtain useful information about the performance of the separate parts of Game@Large system. This information can be compared with other level testing results and will facilitate the assessment of the system; it will also provide quantitative information to the designers.

The derivation of the multilevel analysis approach is done according to the GaL specificities and studies conducted on distinct GaL elements testing methods [2].

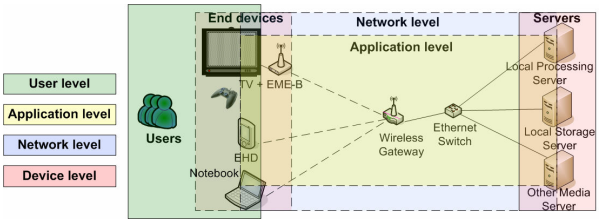

Figure 2. Multilevel testing approach.

The parameters, and specifications given below are general to all testing levels, and they are defined appropriately for every system prototype and target environment:

- parameters: number of concurrent users application type/types (i.e. 4 concurrent processes, where 2 videogames and 2 multimedia applications run), number of tests for obtaining a statistical distribution of the test results, network parameters, device parameters,
- specifications: device specifications (CPU, RAM, graphic card), display specifications (resolution, colour depth), source video game, source video sequence,
- user demographics and environmental properties (test participants' age, gender, education, etc.),
- network emulation tools : NISTNet or similar [9]

4.1.1 User level analysis

GaL system evaluation and testing is concentrating on the user perceptions and experience analysis. A special questionnaire will be utilized for this purpose. Further in this section we describe key issues that are covered in the questionnaire. General requirements for testing process are specified as well.

4.1.1.1 User Interface

GaL graphical user interface (GUI) testing is focusing on usability evaluation at user level analysis.

GUI testing is performed at different environments, for laptops, Enhanced Handheld Devices (EHDs), and TVs. Specific GaL properties [2], [3] considered at user-level analysis include: installation (ease-of-use, performance, time, and additional steps or knowledge required to accomplish the installation), connection to GaL server from different devices (time, easiness), ease (simplicity) of control and browsing (with gamepad, EHD), game browsing (menus and game selection displaying), fonts, colours, images (quality rating for different devices), authentication (performance testing for different authentication ways [user login/password, biometric, etc.]), personalization, user registration, and billing availabilities check, the correctness of presentation of the data on the end device screen.

4.1.1.2 Video games and Multimedia

A number of experiments will be conducted on each environment test-bed using pre-defined test-maps and video sequences. Player perceptions and comments after/during the experiments are observed.

Experimental user evaluation is performed through the research data which is collected with various methods. Qualitative data reflecting the user experience is collected with a questionnaire filled in by each test user after the experiment. The questionnaire contains an appropriate number of statements on which the users are asked on a 5-point scale (mean opinion score) 1 (disagree completely) –5 (agree completely), and open-ended questions [2].

Some individual users and a group of a few users will also be interviewed after the experiment. Test users are informed about the observation before the experiment.

The following requirements are taken into account while performing tests: all users should be widely familiar with the game and respective test-map, users are allowed to familiarize themselves with the game play map with no loss and latency (normal frame rate and resolution) before collecting any data, an amount of loss or latency (and/or change of frame rate, resolution) is induced from the experimental range and the experiment is run.

Users are thus "blind" to the amount of loss or latency to prevent the knowledge of the network conditions from biasing their play. After the experiment is completed, the operator should archive the data for later analysis, modify the amount of loss or latency, and the users should repeat the experiment.

User level analysis represents qualitative system evaluation data, while quantitative data is collected in the application, network, and device level analysis, which are described in the next sections.

4.1.2 Application level analysis

Application level analysis focuses on the quantitative information analysis obtained during particular testing stages (data can be obtained during user-level tests or by separately running particular tests). These tests enable the quantitative evaluating of the effects of network and device parameters in application performance. Below some examples of possible performance measurements are given for GaL user interface and provided gaming and multimedia services.

Application performance at installation time and launch time can be measured for GUI (i.e. time to install, time to launch the service).

For video games, the relevant user performance metrics will depend on the type of game and test-map (time to complete the task or course; scored points, precision, etc.). Taking user performance as the main dependent variable, the application level variables of interest include: resolution, frame rate, latency, packet loss, and jitter.

4.1.3 Network level analysis

Among other things, a better understanding of network game/multimedia traffic can help us design networks and architectures that more effectively accommodate network game/multimedia traffic footprints. Furthermore, careful empirical measurements of network games/multimedia can provide the data required for accurate simulations, a typical tool for evaluating network performance.

The following measurements can be performed: bitrate versus time for different network conditions, cumulative distribution functions of inter-packet times for different network conditions. (client to server and server to client), video game traffic analysis, multimedia traffic analysis, transmission rates for different encoding rates, SNR measurements (for wireless network).

4.1.4 Device level analysis

Device performance (server and client) measurements are performed for different types of games (test-maps) and multimedia. Device tests allow quantitative evaluation of the effects of device performance in the GaL system. Observed parameters are: CPU usage, memory usage, gamepad vs. keyboard performance at application and user levels, EHD, notebook and EME/TV (Enhanced Multimedia Extender) performance at application and user levels (considered device controls, displays, computational capabilities, etc.)

5. CONCLUSIONS

The present paper describes a testing and evaluation methodology for a multi-level assessment of pervasive entertainment systems, and particularly the system architecture proposed under the Games@Large project.

The GaL system and prospective entertainment systems consists of multiple elements and is intended to serve a number of different users in terms of cultural belonging and environmental dependence. To develop a flexible, versatile, and useful system to the end users/stakeholders, it is important to start thinking of system testing already at an early stage in the project.

The differentiation of the GaL framework and services, and multilevel system testing approach, which basically focuses on the end user perceptions and evaluation of the system, will facilitate the testing of the system at different target environments and will enable a proper evaluation of its usability. Network, device, and application level tests provide quantitative test results, which are evaluated and supplied to system designers, thus advising them and facilitating the correction of the errors and improvement of the design.

6. ACKNOWLEDGMENTS

This article was composed within the process of testing and evaluation methodologies development for EU FP6 project, Games@Large.

7. REFERENCES

[1] www.gamesatlarge.eu

[2] A Jurgelionis, P. Eisert, J. P. Laulajainen, *Testing and evaluation of new platforms for delivery of distributed gaming and multimedia*, Web3D 2007, Perugia

[3] W. IJsselsteijn,Y. de Kort, K. Poels, A. Jurgelionis, F. Bellotti, *Characterising and Measuring User Experiences in Digital Games*, International Conference on Advances in Computer Entertainment Technology, June 2007, Salzburg

[4] ISO 9241-11,. *Guidance on Usability*. Geneva, Switzerland: International Standards Organisation, 1998

[5] Nielsen, J., *Usability Engineering*. Cambridge, MA: AP Professional, 1993

[6] Csikszentmihalyi, M.,. *Beyond Boredom and Anxiety*. San Fransisco: Jossey-Bass, 1975

[7] P. Sweetser, P. Wyeth, *GameFlow: a model for evaluating player enjoyment in games,* 2005, ACM Press, New York

[8] Ermi, L. & Mäyrä, F., *Fundamental components of the gameplay experience*: *Analysing immersion*. In: S. de Castell & J. Jenson (eds.), Changing Views: Worlds in Play. Selected papers of the 2005 Digital Games Research Association's Second International Conference

[9] T. Beigbeder, R. Coughlan, C. Lusher, J. Plunkett, E. Agu, M. Claypool *The effect of loss and latency on User Performance in Unreal Tournament 2003*

[10] M. Claypool, K. Claypool and F. Damaab, *The Effects of Frame Rate and Resolution on Users Playing First Person Shooter Games*

[11] T. Oelbaum, V. Baroncini, Thiow Keng Tan, Charles Fenimore, *Subjective quality assesment of the emerging AVC/H.264 video coding standard*, 2002-2003

[12] Fenimore Ch., Baroncini V., Oelbaum T., Tan Thiow Keng, *Subjective testing methodology in MPEG video verification*, Applications of Digital Image Processing XXVII. 2004

[13] http://www.methodsandtools.com, *GUI testing checklist*

[14] http://members.tripod.com, *Software Quality Assurance and Usability Testing*