AMBER

Assessing, Measuring, and Benchmarking Resilience

FP7 - 216295

Project no.: 216295

Project full title: Assessing, Measuring, and Benchmarking Resilience

Project Acronym: AMBER

Deliverable no.: D3.2

Title of the deliverable: Final Research Roadmap

Preparation Date of Delivery: December 31st, 2009

Approval Date of Delivery: December 23rd, 2009

Organisation name of lead contractor for this deliverable: Universitá degli Studi di Firenze

Authors: A. Bondavalli, P. Lollini (editors), R. Barbosa, A. Ceccarelli, L. Falai, J. Karlsson, I. Kocsis, A. Kovi, H. Madeira, I. Majzik, L. Montecchi, A. van Moorsel, L. Strigini, M. Vadursi, M. Vieira

Participant(s): All partners

Work package contributing to the deliverable: WP3

Nature: R

Version: 1.0

Total number of pages: 76

Contract Start Date: 1 January 2008

Duration: 24m

Project Coordinator: University of Coimbra (Portugal)

Partners: University of Coimbra, Budapest University of Technology and Economics, Chalmers University of Technology, City University London, University of Newcastle upon Tyne, Universitá degli Studi di Firenze, and Resiltech

Project funded by the European Community under the Seventh Framework Programme

Final Research Roadmap

Authored by: A. Bondavalli, P. Lollini (editors), R. Barbosa, A. Ceccarelli, L. Falai, J. Karlsson, I. Kocsis, A. Kovi, H. Madeira, I. Majzik, L. Montecchi, A. van Moorsel, L. Strigini, M. Vadursi, M. Vieira

December 31st, 2009

Table of Versions

Version	Date	Authors	Version Description	Reviewers	Date of Approval
0.1	June 23, 2009	ALL partners	It is the final version of the Research Roadmap (D.3.2), based on the preliminary version (D3.1) with the following modifications: 1. figures 1-4 have been added; 2. chapter 6 has been added; 3. examples in chapter 6 are cross-referenced from tables 1-8; 4. a proposal for an additional need, and a couple of proposals for re-phrasing are highlighted in the text and commented.		
0.2	September 1, 2009	ALL partners	Inserted further examples of needs, challenges, objectives and actions.		
0.3	September 18, 2009	ALL partners	• Inserted new "Future Internet" scenarios in Chapter 2; • Reviewed the drivers of Chapter 3, and added new ones addressing the comments from the AB members; • First review of the examples of needs/ challenges/ objectives/		

			actions of Chapter 6.		
0.4	October 16, 2009	ALL partners	Revision of Chapter 6 based on the comments received. Revision of Chapter 4 to incorporate new responses received to the questionnaire		
0.5	October 29, 2009	ALL partners	• Fixed some minor comments; • Removed the Questionnaire chapter and summarized it in the introduction; • Added in the ToC two chapters. The first one will summarize the general AMBER top priorities. The second one will propose different ways/perspectives of reading the roadmap considering different roles and industrial domains, specifying the related top priorities.		
0.6	December 07, 2009	ALL partners	• Chapter 5 (“Examples of needs/challenges/objectives/actions”) moved to Appendix A. • Shortened the description of scenario “Network 2020: Information on the Move". • Inserted the contribution on the AMBER topmost priorities. • Inserted contributions on topmost priorities with respect to specific domains/roles/points of view.		
0.7	December 23, 2009	ALL partners	Deliverable finalized.		

Table of Contents

Summary and Conclusion

This document provides a roadmap for research in technologies for assessment, measurement and benchmarking (AMB) of the resilience of information, computer and communication systems. It is the result of the EU-funded AMBER Coordination Action, integrating the consortium experience in the field with the insights resulting from discussions and interviews with a variety of stakeholders about motivating scenarios, drivers and priorities.

In the following we first review the main insights established in this document.

Motivating Scenarios and Drivers

We describe a set of motivating scenarios that help understand the current needs and challenges in resilience assessment. These scenarios present viewpoints of industrial players, end users, system operators and regulators. Based on these scenarios, we identify opportunities and challenges that we believe will act as drivers for investment in improved resilience AMB technologies. In particular, we believe that the establishment of standardized and sound assessment technologies and benchmarks will be a catalyst for the acceptance of AMB solutions. If done well, it will lead to improved competition by providing easy to communicate measurable objectives for manufacturers, system integrators and users alike. In addition, the increasing demands by regulators as well as the continuing technological progress in software and hardware will create challenges to be addressed by research in the field.

Research Roadmap Outline

The research roadmap constitutes the main contribution of this document. It first provides a detailed list of research needs and challenges grouped in three categories: i) scientific and technological foundations, ii) measurement and assessment, and iii) benchmarking.

The **foundations** make the case for two types of research advances, which we could label as ‘back to basics’ and ‘holistic’. The desire to go ‘back to basics’ refers to the creation of a standardized set of sound but simple techniques and tools for assessment, based on, for instance, insights from metrology (the science of measurement). The ‘holistic’ view refers to the identification that the context (human, socio-economic, political) in which computer systems operate should be considered and assessment should take a holistic view, thus requiring the study of human factors, business impact and the integration of tools and/or arguments, as well as identifying practical limits for the applicability of each class of methods as a function of the environment. A two-prong strategy is therefore needed, on the one hand to keep advancing our assessment methods and techniques to deal with increasingly complex system deployments, and address hard theoretical problems, on the other to work towards standardized basic tools that through widespread use can dramatically change the way resilience is viewed and perceived.

The **measurement and assessment** category identifies a number of topics of acute interest and that are particularly challenging (in addition to topics already discussed above). In particular, this concerns on-line assessment for run-time system adaptation and optimisation (often referred to as self-adaptive systems), the quantitative assessment of security (that is, the ability of a system to withstand attacks and malicious interference), and the analysis of collected data in a structured and powerful a way.

Resilience benchmarking aims at providing generic, repeatable and widely accepted methods for characterising and quantifying the system (or component) behaviour in the presence of faults, and comparing the resilience of alternative solutions. A problem is that, typically, benchmarks may not be designed well enough and are prone to being ‘gamed’, in which case the benchmark may have unintended negative consequences. The research challenges identified in the benchmarking section are therefore about constructing benchmarks that are robust while easy to use. Questions that need to be addressed are how to subdivide application domains, how to create acceptance through standardization, how to include measurement and fault injection hooks into systems, etc.

In addition to the above research issues, we also identified the challenges we see in **education** as well as **standardization**. It has become apparent to us that to fulfil some of the potential of broadly applied resilience assessment, both these aspects need to be addressed. We already touched on the importance of standardized basic assessment techniques and standardized benchmarks. In addition, advances in education of assessment techniques are critical for computer system engineers to appreciate the power of quantitative assessment as well as the pitfalls of poorly conducted assessment.

1. Introduction

Assessment, measurement and benchmarking of resilience (of computer systems) are related concepts, but each of them reflects a different form of characterizing computer resilience and demands specific methods and tools.

These three terms may mean slightly different things to different communities, but the scope of the AMBER Coordination Action is easily defined as covering activities that involve *quantitative descriptions of the qualities of dependability and resilience*. Within this scope, we will use "assessment" to mean obtaining any kind of quantitative statement about these qualities, especially supporting decision-making; use "measurement" in its usual meaning in science and engineering, of mapping empirical observations to numbers in a rigorous manner, through rigorous procedures and calibrated instruments; and distinguish "benchmarking" activities as those in which the main purpose is ranking systems using a simple and standardized method, even at the cost of it being somewhat simplistic. We have generally excluded from consideration, or described only as the background for the research on which AMBER aims to foster debate and coordination, many activities that may be called "measurement" in a broad sense of the word, like formal proofs and checklist-oriented methods.

The very word "resilience" is not uniformly defined (or accepted) in what can be called the "resilience research community". "Resilience" tends to be a synonym for fault tolerance in a broad sense, but it is also used to convey ideas of dependability measures, sought for less narrowly defined and less static scenarios than assumed in currently established methods [1]. AMBER addresses the communities of researchers and users dealing with assessment of qualities related to these two broad concepts. Therefore, in this document we will use "resilience" as a generalisation of terms like dependability, security, fault tolerance, and so on, to encompass all attributes of the quality of "working well in a changing world that includes failures, errors and attacks".

The main objective of AMBER is the synthesis of a research roadmap on assessing, measuring, and benchmarking resilience, to be understood as a list of research directions that seem worth pursuing now, with associated priorities.

The rest of this document is structured as follows:

Chapter 2 proposes a set of ad-hoc scenarios (in the form of "stories") to exemplify the current needs and challenges in the fields of future internet, embedded systems, enterprise computing, supervision of IT infrastructure, safety certification and regulation.

In **Chapter 3** we consider the drivers that can influence the research needs on assessing, measuring and benchmarking resilience, as well as the effective transfer of resilience assessment best practices to European industry and the adoption of standards for resilience assessment and benchmarking.

The final research roadmap is then provided in **Chapter 4**, integrating the consortium experience in the field with the insights resulting from a long discussion of scenarios, drivers and inputs from stakeholders and experts. Four main areas have been identified: i) Scientific and technological foundations, ii) Measurement and assessment, iii) Benchmarking, and

iv) Education, training, standardization and take up. For each of them we identify recognised *needs*, *challenges* to be overcome to satisfy these needs, and *objectives* for specific *actions* to be performed in a short or medium term period. For many of these elements, extended descriptions are provided in Appendix A. Moreover, we provide a synoptic diagram of the connections between needs, challenges and (short and medium term) objectives.

Chapter 5 lists six top priority topics for research about Assessment, Measurement and Benchmarking, and two priority topics regarding education, training, standardization and take up of results in these areas, among those identified in Chapter 4. These priorities have been selected by consensus within the AMBER Consortium, and are presented with their justifications in terms of the possible general goals of research policy.

Finally, **Chapter 6** presents a series of alternate viewpoints about priorities within the research roadmap. Each viewpoint has been written by one of the partners in consultation with one or more industrial experts in a specific sector, to give examples of priorities that specific stakeholders would identify within the roadmap. The parts of this chapter are accordingly titled: “Embedded Systems”, “Transportation”, “Certification Authorities and Assessors”, “Future Internet technological platforms”, “Service Architectures, Platforms and Infrastructures” and “Enterprise Security”. By showing samples of viewpoints within each sector, this last chapter allows a reader to appreciate how the roadmap could be tailored to specific points of view, offering a particular perspective for reading the roadmap itself.

Appendix A presents expanded descriptions and discussions of a subset of the needs, challenges, objectives and actions from Chapter **4**, to clarify their meaning and justification and to provide a more concrete view of the underlying issues. These are cross-linked with the corresponding items in Chapter **4**.

Since the **coordination activities** that led to this final report are of particular importance, we discuss them in detail in this introduction (Section 1.1).

1.1. Coordination Activities

AMBER has been funded by the EU FP7 programme to bring together visions, ideas, knowledge and information from academia, industry, regulators and society at large, within the area of resilience assessment for information and communication technology (ICT). One of the main outputs of AMBER is a research agenda, or roadmap, for advances in European research and industrial practice in this area, which can be used as input to current and future European funding programmes. This document presents the AMBER roadmap. To solicit input to it, the AMBER consortium has conducted various coordination activities. In detail, these are:

- Debate among researchers and practitioners through a **panel** on “How Hard is Assessing and Measuring Resilience?” held at EDCC-7 (7th European Dependable Computing Conference) in Kaunas, Lithuania, May 2008.
- The **workshop** on “Resilience Assessment and Dependability Benchmarking” held at DSN08 (38th Annual IEEE/IFIP International Conference on Dependable Systems and Networks) in Anchorage, Alaska, USA, June 2008.The two-day **workshop** with the AMBER **Advisory Board**, where the 15 members of the Advisory Board illustrated their vision on challenges and solutions in resilience assessment. This workshop, held in Budapest, Hungary, on February 2008, was organized under the theme “Assessing, Measuring, and Benchmarking Resilience: Can we Really Do It in a Trustable Way?”.

- The **workshop on “Sha**ring Field Data **and Experiment** Measurements on Resilience of Distributed Computing Systems“, SRDS08 (27th International Symposium on Reliable Distributed Systems), October 5, 2008 (http://www.amber-project.eu/srds-ws/).
- The second **workshop** with the **Advisory Board**, held in London, UK, Jan**uary 14-**15, 2009.
- The **workshop** on “*Evaluation of Dependability and Resiliency*”, in conjunction with the winter 2009 meeting of the IFIP Workgroup 10.4 on Dependable Computing and Fault Tolerance, Cortina d’Ampezzo, Italy, January 27-31, 2009.
- The joint **workshop** “*Measurability of Trustworthiness of Complex ICT Systems and Services*” together with the Think-Trust CA, Brussels, Belgium, March 9, 2009.
- A **panel** session on the theme “*From Assessment to Standardized Benchmarking: Will it happen? What could we do about it?*” was organized in the frame of the 39th IEEE/IFIP Symposium on Dependable Systems and Networks (DSN 2009), Estoril, Lisbon, Portugal, June 29 – July 2, 2009.
- The third **workshop** with the **Advisory Board**, held in Florence, Italy, October 26-27, 2009.

The **state of the art** reported in the AMBER report by that name [1] was another important input to the roadmap, offering a detailed analysis of the AMBER-related challenges and needs identified in other **past research roadmaps**, specifically AMSD [2], GRID [3] and ReSIST [4].

In addition, to take into account as much as possible the stakeholders’ perspective in the definition of the research roadmap, we distributed the **AMBER questionnaire** (http://www.amber-project.eu/roadmap.php) to the AMBER Advisory Board members and to other influential colleagues both in academia and industry. The main results extracted from the analysis of the received responses have been considered in the roadmap definition. For example, among the major research challenges we have to face in the near future the respondents indicated the development of techniques and tools to: handle the complexity of target systems, deal with system adaptation, handle (measure) human errors (including design errors), support runtime monitoring and automated online data collection and analysis, automate the measurements and enable portability across the computing platforms, and sanitize and make dependability and resilience data reusable.

2. Motivating Scenarios

In this chapter we present a set of example scenarios to illustrate some of the challenges that need to be addressed in resilience assessment, measurement and benchmarking. The scenarios belong to the fields of future internet (Section 2.1), embedded systems (Section 2.2), enterprise computing (Section 2.3), supervision of IT infrastructure (Section 2.4), and safety certification and regulation (Section 2.5).

2.1. Future Internet

The two scenarios presented in this subsection show how the Future Internet could, and possibly will, shape the lives of all Europeans in or around 2020. They concern the green urban transport (Section 2.1.1) and the information on the move (Section 2.1.2).

2.1.1. Green Urban Transport

Scenario[1]. It is 2020. As problems caused by vehicular traffic increase, several cities are considering new mechanisms to control the emissions of pollutants. A proposed method is based on monitoring drivers' behaviour. Drivers are allowed to use personal or public transport, but this depends on the level of traffic congestion at the time; each driver has a personal carbon credit. Drivers can save their own carbon credit by using public transport or reducing the use of their personal vehicles; users pay more or less for travelling, depending on their carbon credit.

Several service providers propose their own ICT implementations of this mechanism. A large European capital city decides to introduce this service in the city centre, since the pollution levels and people's use of personal vehicles have become excessive. The Traffic and Mobility Office of the city thus needs to choose a service provider; the choice must be based on cost as well as trustworthiness qualities, like reliability, resilience to attacks and operational failures, ability to guarantee quality of service and to protect user data to ensure privacy. The Traffic and Mobility Office gives the technical tasks of evaluating the different service providers, identifying the best candidate, and defining a suitable combination of network infrastructure and wireless connections, to an external company. This company has access to usage and failure data concerning the various service providers, from other cities where the service is already in use.

Will this company be able to assess the level of trustworthiness of each service provider, so as to correctly select the best one for the needs of this specific city?

Open Challenges. This scenario highlights several challenges in the measurement and assessment of trustworthiness of future Internet-based applications. First of all we have a problem of requirements: what are the main attributes of the service, to be measured for assessing a Green Urban Transport application? Which aspects of security and trustworthiness are measurable and quantifiable, and which metrics are appropriate, for this application? If experimental evaluation (of subsystems or of small-size pilot implementations of the complete service) is sought, are there any reference fault loads and attack loads that are known to be appropriate for assessing resilience to attacks and operational failures? There is also the problem of compositionality of measurements: after estimating low-level metrics referring to subsystems (e.g. performance/reliability/security of the wireless network, of the wired infrastructure network and of vehicles' on-board subsystems) how can these measures be

[1] This scenario is partly inspired by *Application scenarios & functional requirements for wireless sensor & actuator networks in Future Internet (F. Forest)* presented at the 10th LETI Annual Review (24-25 June 2008)

aggregated to extract meaningful information about the trustworthiness of the overall system, that is, of the service? Even with access to usage and failure data about the competing service providers in other cities (in which the service is already used), how can the assessments obtained from these data, representing trustworthiness attributes of each provider's service in a different set of cities, be extrapolated to the new application environment (the specific city), or even meaningfully compared?

2.1.2. Network 2020: Information on the Move

Scenario[2]. This scenario is about seamless mobility: Matti moves effortlessly between his home, car and office interacting with family, friends and work colleagues as he goes, and always with his whole world of personal information at his fingertips. The scenario illustrates how in the Future Internet our personal information, content and services will be available to us anywhere, at any time. Our everyday environments will be context-aware: systems and devices will be able to sense how, where and why information is being accessed and respond accordingly. The Internet will be our personal global network. This new world of seamless applications, services and content requires a new network infrastructure, in which advanced features such as semantics and trustworthiness must be built in. Matti's world is more seamless than anything we have today. There is no lost connectivity, no waiting for logons, no poor quality content, no systems that don't talk to each other. What's more, all environments (Matti's home, office, car) are context-aware: systems and devices are able to sense how, where and why information and content are being accessed and respond accordingly.

Open Challenges. From the point of view of measurement, assessment and benchmarking of trustworthiness of future Internet-based applications, this scenario opens several challenges. First of all, trust and security are paramount in this scenario: they will be key enablers in realising the potential of the new online world, and measurement, assessment and benchmarking of trust and security will be necessary in order to be able to assure the quality of new future services as perceived by their users (also allowing a fair comparison of alternative services and avoiding mistrust). Secondly, developments such as cloud computing, social networks, and service mash-ups require new approaches to regulation of privacy at a Europe-wide scale. This should be complemented by research into privacy enhancing technologies (where data protective features and services are built in from the ground up) as well as into their assessment techniques that provide public characterization of the level of built-in privacy and security, together with accountability. To counter these strategies, and to create a level playing field, regulators and public policymakers should strive for open standards (including standard benchmarking of trustworthiness), open interfaces (including hooks for measuring and assessment), and federated architectures (reliable and secure interoperable platforms). Adherence to these principles should be a mandatory requirement for developers of public Internet services. Thirdly, resilience AMB technologies and large-scale testbeds are required to identify (and then reduce) the major technology-related roadblocks that may be in the way to the realization of this scenario: vulnerable architectures, lack of adaptability (also in quality of content and energy consumption), non-scalable connectivity and accessibility, performability bottlenecks, lack of resilience to attacks and operational failures.

[2] It is a shortened version of the scenario "Network 2020: Information on the Move" taken from Future Internet 2020, Visions of an industry expert group, which is an industry expert group report. This document was found on the website http://www.future-internet.eu, the European Future Internet Portal, a project initiative which hosts the Europe-wide debate on the Future of the Internet.

2.2. Embedded Sub-Systems in Automobiles

Scenario. Large automotive manufacturers rely on third parties for many aspects of their products – that is, they are 'Original Equipment Manufacturers' (OEM). As an example, assume that an automotive manufacturer has decided to offer an infotainment server as an optional feature for its next generation of cars. Such an infotainment server has interfaces to the car's backbone network to connect user panels, the wireless communication unit, and the GPS receiver. This poses high risks, and the OEM must make absolutely sure that an infotainment system supplied by a second company does not endanger the operation of the car.

In more detail, the infotainment server is required to implement a set of "sandboxing[3]" techniques that provide confined execution environments for downloaded applications. This must prevent hidden malware to access or manipulate restricted data and should ensure that software design faults (software bugs) in a downloaded application do not cause interruptions of the infotainment or system programs. It also continuously monitors the application for failures and must take appropriate action for benign as well as severe failures.

Assume now that the OEM has two offers from infotainment systems suppliers. How do the OEM's test engineers decide between offerings from different infotainment system suppliers?

Open Challenges. The OEM test engineers must measure both security and dependability properties since malicious attacks as well as accidental failures must be anticipated. However, the OEM test engineers have almost no standard ways of testing the two offerings. This is true even if we assume that both suppliers offer easily accessible and usable test infotainment systems. There are no widely accepted attack or fault loads, nor methods to derive such attack and fault loads. In addition, the products delivered by the suppliers do not have standard measurement hooks, fault/attack injection hooks or measurement data formats. Finally, there are no well-understood and reusable grading criteria for ranking of contending systems.

The knowledge gaps apply to theoretical underpinnings, modelling and emulation of software faults and security attacks, injection and data collection techniques, as well as definition of useful measures.

2.3. Information Security Management in a Financial Enterprise

Scenario. A Chief Information Security Officer (CISO) of a large investment firm carries large responsibilities and must balance pressures and objectives from many angles. In this particular case, the CISO must decide about the Information Rights Management (IRM) solution that is appropriate for the company. IRM solutions are variations of digital rights management solutions, associating metadata to information such that management software can regulate access. The CISO's first objective is to determine whether such a system is worth the investment, and the question therefore is: how does the CISO assess the value of the IRM solution to the company?

To address this question, the CISO uses as starting point a variety of company security policies and government regulations to determine the requirements for an IRM solution. These pertain to often contradictory requirements, such as the need to protect data from

[3] *Sandboxing* means providing a strictly-controlled set of resources for a guest program to run in, such as a limited space on disk and memory. Many features, such as reading from input devices or accessing the host system, are restricted.

unauthorised access, allow for auditing of purposely retained data, and the need to avoid employees accessing or receiving information that would endanger the lawfulness of a financial interaction. In addition, the CISO must take into consideration the employees sustained productivity as well as the behavioural patterns of employees with respect to the use of technological security solutions. Finally, a cost-benefit and risk analysis would need to be carried out.

Open Challenges. The CISO is far less concerned with technology matters than the engineers in the first scenario, but the lack of well-established reusable tools, techniques and methodology is as much (if not more) a pain point as in the previous scenario. To carry out assessment of socio-technological solutions such as the deployment of IRM, the CISO would want a set of well-established tools and techniques that allow for integration of many concerns. The CISO needs input parameters about security offered by the IRM solution, thus requiring well-established security benchmarks. The CISO also needs analysis methods and techniques to associate IT decisions with the financial and business implications for the company. Finally, the CISO would like a knowledge base or other reusable way of identifying and taking into account human factors.

The knowledge gaps apply to dealing with the multi-facetted nature of socio-technical systems in an integrated fashion.

2.4. High-level Education for IT Administration

Scenario. We assume a fresh computer science graduate whose first job is in IT administration. The company he works for manages the IT for its customers and the young IT administrator is made responsible for the daily operation of the office applications of the customer enterprise. His tasks are relatively focussed, but become more challenging over time, and include elements such as configuration management, application monitoring, service level agreement management and eventually software purchasing. The core in all jobs of this young administrator is assessment, both of the existing applications and of possible newly acquired applications. The question this scenario asks is: how well did the computer engineering degree prepare the young graduate for the job?

The young graduate is fortunate enough to work for a company that is at the forefront of management technologies. A considerable tool suite is available to help the IT administrator with tasks such as network, server and service monitoring and for more advanced analysis such as root cause analysis of failures. The R&D division of the company develops advanced software tools for IT management, for instance using model-based approaches. These model-based approaches include UML design and domain-specific deployment models that include resilience properties in their abstractions. In addition, the various divisions for which the young administrator manages applications pose service level agreements (SLAs) that they want the IT systems to fulfil.

Open Challenges. As we mentioned, in this scenario we are especially interested in the question whether the young computer engineering graduate received the right training for his job. In particular, without proper fundamental training in assessment methods and techniques the IT administrator is in danger of misusing the existing set of elaborated management software tools (such as for monitoring, data processing and root cause analysis). Of increasing importance is the ability of an IT administrator to work with model-based abstractions, including UML in the design of application integration solutions, domain-specific abstractions for deployment and assessment, and model-based prediction to determine if SLAs can be met and operation can be optimised with respect to SLAs. The increasing virtualisation-based dynamism through service-oriented software and service provision solutions (including cloud computing) create further challenges in dealing with partners through contracts that include service level agreements.

The educational gaps are found in fundamental aspects of assessment and model-based abstraction.

2.5. Safety Certification and Regulation

Scenario. An assessor (working in an independent assessment body or in a regulatory agency) has to recommend or approve the use of a certain safety system for a dangerous plant in a regulated industrial sector. The proposed system is a combination of off-the-shelf platforms with industry-specific and plant-specific application software, developed with the help of off-the-shelf development tools. Accidents in this plant may have very serious consequences, so the assessment must give assurance of very low probability of accident, and very high confidence in this assessment is required before the plant can be allowed to operate with the proposed safety system. Following accepted practice, the assessor relies on aspects such as 'proven in use' evidence, quality of the development process and resilience of the architecture. But, how confidently can the assessor assess system safety using such evidence?

The assessor uses the common set of techniques and tools in his work, but this does not prevent him from being uncertain about the quality of the assessment. In fact, the value of "proven in use" evidence is difficult to assess: it is hard to know whether, for instance, the new safety system uses the same set of features of the off-the-shelf platform as the one from which statistics were derived, and how many failures may have gone unreported in the past. Another potential problem is the reliance on process quality evidence: even though the assessor accepts the practical constraints that dictate such reliance, he can feel particularly uneasy justifying the relation between process quality and system safety. A third significant difficulty is in deciding how much "credit" to give for resilience features in the architecture: while the architecture may include useful redundancy, statistically reliable information about its effectiveness is limited, both because it is intrinsically hard to obtain and because the vendors are often reluctant to provide it.

Open Challenges. The difficulties that the assessor needs to overcome indicate some hard open problems. More statistical evidence would be available if widely accepted and uniform ways for collecting data such as amount of use, profile of use and failure data were available, and if vendors were willing to make such data available. Even given abundant statistical data, the assessor needs methods and techniques for extrapolating resilience measures from such data. The issues include for instance: estimating the coverage of error detection and reporting facilities; characterising how process quality relates to resilience of a product; improving the modelling and empirical knowledge of how system architecture affects resilience, so as to characterise and, where possible, reduce the uncertainty of predictions based on probabilistic models. Last, we can notice two other open challenges: i) the need to increase awareness of the sources of uncertainty in the evaluation of computing systems and ii) the lack of uniform and standard ways to collect, filter and report results obtained in experimental evaluations. These methods should be based on sound statistical and metrological science.

The knowledge gaps are in the quality and appropriateness of the data available for assessing the resilience of systems, and the mathematical methods needed for sound inference given the inevitable limits of the data that can feasibly be collected.

3. Drivers

In this chapter we discuss the main drivers for expanding research activities on assessing, measuring and benchmarking resilience. Two important ever-present drivers are **complexity**, which often sprouts difficulties in project development, and **the pace of change**, which requires companies to rapidly understand the benefits and drawbacks of any new technology or development process, while maintaining legacy systems and obsolete technologies. This chapter focuses on the drivers that have recently emerged and that are expected to continue motivating the development of resilience-related techniques in the foreseeable future. The description is organized into the following four groups:

- Information technologies are global and pervasive (Section 3.1);
- There is a greater awareness for the environment (Section 3.2);
- Socio-economic factors affect the resilience market (Section 3.3);
- Technological innovation creates new issues (Section 3.4).

3.1. Information technologies are global and pervasive

New ways of distributing services. Recent initiatives promote new technologies and new forms of distributing services through the Internet. Economic factors boost the market for everything as a service, including software, aiming at a reduction of acquisition and maintenance costs. For an organization to move to this service model there is a strong need for assessing how good a given provider is, and how service failures affect the return on investment of such a move. Using software through the Internet, as pushed by cloud computing initiatives, will also open up opportunities for competition among different providers of cloud applications. This creates the need to benchmark, measure and forecast the service provided through these new means, given that decision-making requires factual data on resilience, availability, integrity, etc.

Emerging cyber-threats. Most enterprise systems cannot be considered secure unless they are dependable, and vice versa. For this reason, organizations are interested in evaluating the resilience of their infrastructures to attacks. There are numerous emerging cyber-threats, such as the increasing number of malware objects on the Internet and the growing concerns with botnets (groups of infected computers that are controlled by attackers), which may be used for data theft and other malicious intents. Here, AMB technologies can be used as instruments for deciding whether to migrate an organization to a newer version of a given software product, for evaluating whether an infrastructure requires improvements, etc.

3.2. There is a greater awareness for the environment

New technologies for a greener world. Technology is the means for guiding and monitoring green policies, with governments and businesses wishing to analyze energy usage and carbon footprints, to control traffic congestion, and so on. On the end-user's side, environmental and economical concerns motivate an increasing number of professionals to carry out their activities using computer systems at home – telecommuting. Those end-users, and the organizations they work for, are interested in benchmarking different Internet access providers (wired and wireless) as well as the resilience of online collaboration tools. The success of most green initiatives depends on the resilience of the underlying technologies. For this reason, governments, organizations and end-users would benefit from trustworthy techniques for resilience assessment and measurement.

Green computing. There are several ongoing initiatives to make efficient use of computer resources, to reduce the emission of greenhouse gases and to improve the environment. The energy consumption of computer systems may be reduced by centralizing processing power in a server and using thin clients as terminals. This provides also the means for reducing storage requirements, by storing only a single copy of each file and allowing multiple users to access it. The large-scale usage of this type of architecture and associated techniques (e.g., virtualization) is unprecedented. Resilience assessment and measurement are fundamental to guarantee the success of these initiatives, as it is necessary to optimize the balance between resilience and environmental impact of computer infrastructures.

3.3. Socio-economic factors affect the resilience market

Regulatory demands. Society is becoming increasingly dependent on large-scale ICT systems, as well as advanced embedded systems, which in the event of major service failures may cause not only significant economic loss, but also severe accidents or loss of vital government and public services. This dependence has increased the efforts to regulate many issues related to resilience and dependability, as regulators seek to protect the public. Data retention regulations are motivated by cyber-crime and terrorism; the Sarbanes-Oxley (SOX) act requires corporations to maintain and retain correct financial records, related to assessment through the need for monitoring, logging, auditing and analysis; the upcoming ISO 26262 standard for the automotive industry recommends fault injection as a means to assess the effectiveness of safety-related functions; the Basel 2 agreement has created a need for banks to apply quantitative forecast to “operational risk” (which includes risks from ICT failure). These are examples of the trend to regulate resilience and its assessment, which compels companies to adopt the necessary AMB technologies.

Human factors. It is well known that the dependability of complex IT systems relies to a large extent on human operators and their ability to handle failures and other critical events. Experience shows that outages of systems that have been designed to be highly resilient (e.g., telephone systems and large file servers) are often caused by operator mistakes. While human reliability analysis has a long history, there is a striking lack of adequate techniques for assessing and modelling users and operators in complex roles for a large range of IT-based systems.

3.4. Technological innovation creates new issues

Component-based and off-the-shelf products. Computer systems and systems-of-systems are often built using off-the-shelf products, and software is increasingly designed by decomposing a system into subcomponents that can be purchased from different suppliers. This way of developing systems is now widespread, including its use for building critical systems and infrastructures. Consequently, it is necessary to create and adapt AMB techniques to compare different suppliers regarding the resilience of their products. Since system integrators have less control over the development process, it is increasingly important to evaluate the offerings of multiple vendors of a given component.

Hardware and software reliability. Hardware failure modes are likely to change significantly with new field-programmable devices and new integrated circuit technologies, as these are increasingly susceptible to soft errors (data corruption), device aging, and variations in manufacturing processes. This will force chip manufacturers to add more fault tolerance to their circuits, thereby changing the way hardware failures are manifested at the system level. Regarding software, a few trends of interest are the increasing use of programming frameworks (that implement generic functionality and handle flow-control), new development processes (such as agile and model-based development), and automatic code generation, as well as the use of thread-level parallelism in multi-core programming. These advances are likely to change the rate and nature of software and hardware faults, calling for new fault models and new techniques for understanding how to mitigate their effects.

4. Final Research Roadmap

Considering the input coming from the other chapters, and based on the analysis of the state of the art reported in deliverable D2.2 [1], on the feedback received during the organized AMBER panels and workshops, as well as on the received responses to the AMBER questionnaire, in this chapter we provide the final roadmap for research in Assessing, Measuring and Benchmarking resilience (abbreviated in the following as AMB resilience).

According to the initial goals of the Coordination Action, we identify four main areas for investigation:

- Research area 1 (**Scientific and technological foundations**) addresses the foundations that underlie the other areas of research discussed in this roadmap. Among the foundational issues, we identify for example the need for sound metrology-based assessment principles, the handling of complex models and multi-faceted arguments, and the inclusion of human behaviour.
- Research area 2 (**Measurement and assessment**) deals with the challenges and research directions related to measurement and assessment activities as typically used to characterize a system alone, as opposed to ranking different systems. Measurement and assessment require sound and well defined methods, although these need not be standardized.
- Research area 3 (**Benchmarking**) targets benchmarking activities, which can be seen as the evolution of current resilience assessment techniques into more standardized approaches. Resilience benchmarks offer generic, repeatable and widely accepted methods for characterising the system behaviour in the presence of faults, and allow the comparison of the resilience of alternative solutions.
- Research area 4 (**Education, training, standardization and take up**) discusses the educational, training and standardization issues related to resilience AMB. Some of these issues identify actions that can be performed or supported by the research community, while some others are related to more general policy actions.

For each of these research areas, we specify:

- ***Needs*** (linked to the various drivers identified in Chapter 3), which are felt by stakeholders and research activities would aim to satisfy (although full satisfaction may be more an ideal state than a feasible objective).
- ***Challenges***, the most probable difficulties and obstacles to be overcome, in view of the context, the present state, the objectives and the nature of the problem to be solved.
- ***Objectives***, which identify either tangible results to be achieved or research directions to be followed. For each objective we specify:
  - The **short** (0-3 years) or **medium** (3-8 years) **term** in which the specified results should be achieved or progress in the research directions should be made;
  - The **actions**, i.e., specific activities that should be carried out to achieve the result or to pursue the research direction.

The timeline division between short and medium term objectives has not been applied to the "scientific and technological foundations" research area, since the research topics listed require incremental progress but over a long-term horizon, and thus a timeline division becomes useless or even impossible.

Topics in each research area are presented in two formats. First, two tables list needs and challenges, and objectives with the related actions, respectively. For many of these elements, extended descriptions have been provided in Appendix A. Then, at the end of each section, a synoptic diagram shows the connections between needs, challenges and (short and medium term) objectives.

Note that a single problem identified here at a general, abstract level may appear as multiple specific problems in the various domains of application (considering the specific technologies used and the types of attributes to be assessed), and solutions at the level of theoretical understanding will need specific implementations in each domain. For example, let us consider the following two concrete problems: i) estimating the probability of compromise of privacy of sensitive data on a consumer PC due to criminal attacks, and ii) estimating the error detection coverage of an internal software check against generic accidental faults in an embedded system-on-a-chip. The two share needs, for instance, for sound measurement approaches or for judging the validity of prediction, but the solutions will depend on the constraints and possibilities characterising the two concrete situations, for instance leading to different simplifying assumptions. An implication of this is that objectives listed here may well be reached at very different times for different domains. On the other hand, a method developed for one domain may well prove be portable to others and/or help in building a shared insight about the common fundamental problems and techniques for addressing them.

Please note that in every table of this section, there are one or more underlined elements. By clicking on them while holding the CTRL button will display an extended explanation or example for that need or challenge or objective or action. All these examples are listed in Appendix A. Below each example there is another hyperlink that permits the reader to go back to the original section.

4.1. Scientific and Technological Foundations

This section deals with needed advances in the scientific and technological foundations that underlie the other areas of research discussed in this roadmap. Among these foundational issues, we identify:

- Sound metrology-based assessment principles.
- Prediction with statistically “inappropriate” data, e.g., in the presence of change.
- Complex models and multi-faceted arguments.
- Inclusion of human behaviour.

In the following Table 1 we outline the needs and challenges.

Scientific and technological foundations
Needs • *“Interoperability”*: ability for practitioners to understand and use raw measurement data and predictions obtained by different people or organisations. • Validated methods for extrapolating measurements to predictions of system behaviour despite differences between the system in operation and its environment and the system/environment where the measurements were taken. • Integration of considerations related to *human behaviour* in the assessments of resilience of computer systems as affected by the behaviour of their users, system managers, and adversaries. • Improved “argumentation” processes, which correctly formulate, communicate and verify complex arguments combining “hard” evidence (measurement, mathematical

models) and “soft” evidence (judgement), with proper treatment of epistemic uncertainty, levels of confidence, and “unknown unknowns”.

- Investigation of principles for successful integration of resilience assessment, measurement and benchmarking technologies into different phases of the life cycle of IT systems, including methods for evaluating technical efficiency and economic impact.

Challenges

- Providing an agreed set of common measures (within a domain and whenever possible between domains).
- Providing a shared terminology and common accepted data presentation formats to support understanding of which data items can be compared (as representing the same resilience related attributes of systems).
- Understanding how environments change, how to detect and predict changes, and how differences in the environment (or variations in manufacturing or configuring different instances of a type of system) affect the important measures of resilience/dependability.
- Providing sound patterns of argument for extrapolation and bounding ranges of scenarios outside which any given extrapolation method should not be trusted.
- Improving the body of empirical evidence about the range of human behaviour in the various situations of interest and understanding the limits of predicting human behaviour.
- Selection of formalisms for correctly manipulating complex arguments with adequate trade-offs between intuitive readability and formal rigour, also clarifying the boundaries of the problem domains in which these methods can be trusted.
- Understanding the economy of building confidence in the resilience of an IT (Information Technology) system, and the key factors that motivate companies to adopt resilience AMB techniques.

Table 1: Scientific and technological foundations: needs and challenges

The following Table 2 details actions linked to the objectives. As previously mentioned, due to the incremental nature of the objectives we do not distinguish between short and medium term periods to achieve them.

Scientific and technological foundations	
Objectives	***Actions***
Identification of simple and widely acceptable resilience-related metrics.	• Elaboration of easy-to-use, practically measurable (domain-specific) resilience metrics (including metrics for security) and establishing common “operational” definitions for them, in domains where these are lacking. • Identification/understanding of the relationships among different resilience-related metrics, at the different levels of abstraction in a system’s description.
Adoption of sound measurement practices in domains where these are lacking.	• Extending the usage of the support of mathematical and scientific bases for the sound measurement practices – in particular, the conceptual framework provided by metrology science. • Selection of appropriate measurement methodologies for given measures and domains and development of measurement tools that are well-based from a metrology point-of-view.

Standard set of monitoring features and “hooks” to facilitate AMB and the development of related tools (at least for computer systems in specific domains).	• Exploitation of existing features and “hooks” (such as JTAG at processor level, NEXUS at board level, “standard” OS facilities like monitoring and logging, logging systems of web servers, auditing in database management systems, etc.). • Development of “hooks” for monitoring and measuring for open source software systems, especially to demonstrate the application of “hooks”. • Integrating the use of “hooks” and standard monitoring features in (new) AMB tools for given domains.
Interoperability in (re)using raw measurement data and predictions in specific domains of application.	• Elaboration of common data formats to support interoperability of raw measurement data. • Publication of (multi-dimensional) measures that characterize the differences between environments of use of systems.
Elaboration of *extrapolation* methods and tools to generalize observations (measures).	• Development of methods/tools to predict resilience characteristics under future usage profiles by extrapolating assessments derived from measurements of operation, attacks and failures obtained under a different profile, with rigorous estimation of the error thus introduced. • Development of methods/tools to predict the dependability of a design using the observed parameters of its components in previous designs. • Collecting bodies of domain-specific empirical evidence about the sensitivity of predictions to variations along the dimensions characterising the environment, in order to improve the methods and models used in prediction.
Integration of *human behaviour* in AMB resilience of computer systems.	• Improving the body of empirical evidence about the range of human behaviour in the various situations of interest. • Identifying the limits of predicting human behaviour. • Creating reusable models of human behaviour and perception with specified limits of application (in given application environments).
Elaboration of methods for manipulating *complex arguments*.	• Selection of formalisms for correctly manipulating complex arguments with adequate trade-offs between intuitive readability and formal rigour. • Clarifying the boundaries of problem domains in which these methods can be trusted. • Elaboration of argument patterns that link the differences between environments of use to uncertainty in prediction.

Understanding the economics and overall impact of resilience assessment on the lifecycle of IT systems.	- Establishing business or financial models to relate system resilience with its business impact. - Cost/benefit analysis of experimental and analytical resilience assessment techniques and their combinations (cost balancing). - Proposing effective feedback loops between design and resilience assessment activities/teams.

Table 2: Scientific and technological foundations: detail of actions

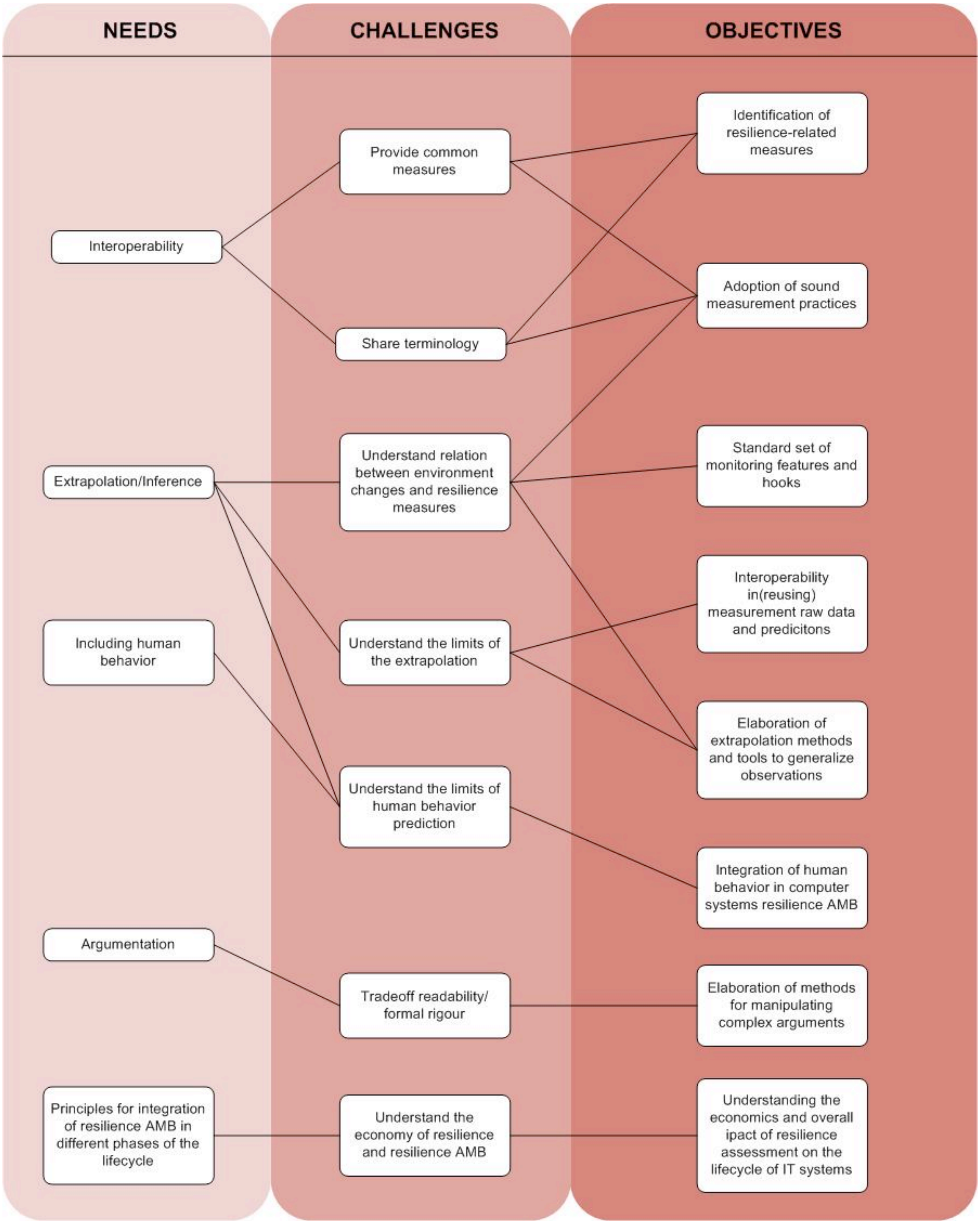
NEEDS
CHALLENGES
OBJECTIVES
Interoperability
Extrapolation/Inference
Including human behavior
Argumentation
Principles for integration of resilience AMB in different phases of the lifecycle
Provide common measures
Share terminology
Understand relation between environment changes and resilience measures
Understand the limits of the extrapolation
Understand the limits of human behavior prediction
Tradeoff readability/ formal rigour
Understand the economy of resilience and resilience AMB
Identification of resilience-related measures
Adoption of sound measurement practices
Standard set of monitoring features and hooks
Interoperability in(reusing) measurement raw data and predicitons
Elaboration of extrapolation methods and tools to generalize observations
Integration of human behavior in computer systems resilience AMB
Elaboration of methods for manipulating complex arguments
Understanding the economics and overall ipact of resilience assessment on the lifecycle of IT systems

Figure 1: Scientific and Technological Foundations: links among needs, challenges and objectives.

4.2. Measurement and Assessment

In this section we address measurement and assessment activities as typically used to assess individual systems. Among the main issues, we identify:

- Extensions of the capabilities of AMB methods and tools to cope with system complexity.

- Providing experimental evaluation techniques considering the perspective of metrology.
- Elaboration of processes and tools for dynamic evaluation and online assessment.
- Development of compositional, holistic frameworks for assessment.
- Development of (large-scale) test beds.
- Mapping resilience assessment, measurement and benchmarking into business and organisational impact.

In the following Table 3 we outline the needs and challenges.

Measurement and assessment
Needs • Practical, trustworthy and widely applicable tools for measurement and assessment in large-scale dynamic systems, adaptable and evolving infrastructures, and other domains where these are lacking. • Domain-specific frameworks that provide seamless integration of the results of different methods and tools for resilience assessment. • Test beds to help in assessing the resilience of complex applications and to improve assessment techniques. • Prediction of how user behaviour will impact resilience of newly designed or deployed systems, and predict the users' (and adversaries') perception of system resilience. • Analysis of the risk and business or organisational impact of resilience issues.
Challenges • Developing measurement and assessment tools for systems/measures for which these tools are currently lacking. • Providing domain-specific rules to compose/integrate different methods and tools for resilience assessment. • Managing/mitigating the complexity of models used in resilience assessment tools, both in model construction and in the model solution process. • Validating assessment models and controlling the impact of the approximations introduced in the assessment process. • Finding the proper trade-off between "reactiveness" (speed) of online assessment processes and "trustworthiness" (accuracy) of their results. • Providing measures of the uncertainty associated to measurements and assessments (to avoid overconfidence in assessment results). • Designing and developing test beds for emerging applications, e.g., cloud computing and collaborative services. • Reducing the cost of resilience assessment and measurement by developing methods and tools that are easy to integrate into existing development methods and tool chains.

Table 3: Measurement and assessment: needs and challenges

The following Table 4 details actions linked to the objectives, estimating the time at which they could be achieved.

Measurement and assessment	
Short term (0-3 years)	
Objectives	***Actions***
Identification of processes for the dynamic evaluation and online assessment of systems and infrastructures.	• Identification of the relevant environment conditions/properties to be monitored. • Elaboration of methods for (model-based) design and configuration of monitoring systems. • Adaptation and extension of existing data processing techniques to integrate data sources and identify hidden phenomena and relations on the basis of measured data. • Analysis of the proper balance between the accuracy of the results and the assessment time.
Development of efficient methods coping with model and size complexity.	• Development of methods and tools for resilience assessment, able to : ○ Avoid the generation of large/complex models; ○ Tolerate the generation and solution of large/complex models (i.e., optimizing the generation and processing of the models).
Development of experimental techniques for resilience assessment in systems where these are lacking.	• Development of experimental evaluation techniques that are characterized by low intrusiveness and reproducibility. • Application of assessment models to map low-level data collected during experimental evaluation onto the high-level relevant measures interesting for the user or evaluator. • Execution of experimental campaigns in real systems to characterize and demonstrate experimental measurement features and properties.
Medium term (3-8 years)	
Objectives	***Actions***
Development of (domain-specific) compositional framework for a holistic assessment process.	• Identification of the base types of measurement and modelling techniques and tools (applicable in a given application domain) and the possible interactions among them to provide realistic assessments. • Elaboration of techniques for assessing the combined impact of accidental faults and malicious threats. • Assessment of the impact of the approximations introduced in modelling on the resilience-related predictions sought.
Development of efficient on-line mechanisms to monitor the environment conditions of the system and to dynamically evaluate and assess its resilience.	• Development of methods to dynamically assess resilience-related properties. • Application of automatic "machine learning" methods to improve/refine the assessment of resilience. • Elaboration of methods to online update/augment assessment models on the basis of measured data. • Tuning of the proper balance between results' precision and assessment time.

Development of large-scale test beds for applications.	• Identification of important application areas that are in need of practical test beds. • Design and implementation of the large-scale test beds and elaboration of methods for its operation and maintenance.
Development of business and economic models for risk analysis.	• Development of appropriate models for mapping IT resilience into business or organisational impact. • Development of appropriate risk analysis models with the contribution of economists, business analysts and risk analysts.
Development of methods for resilience assessment through automated analysis.	• Elaboration of notations (extensions of design languages) to support automated assessment. • Development of methods for automatic generation of resilience assessment models from design models. • Identification and (if needed) development of domain- and process-specific methods and corresponding tools that can be integrated into automated tool chains.

Table 4: Measurement and assessment: detail of actions

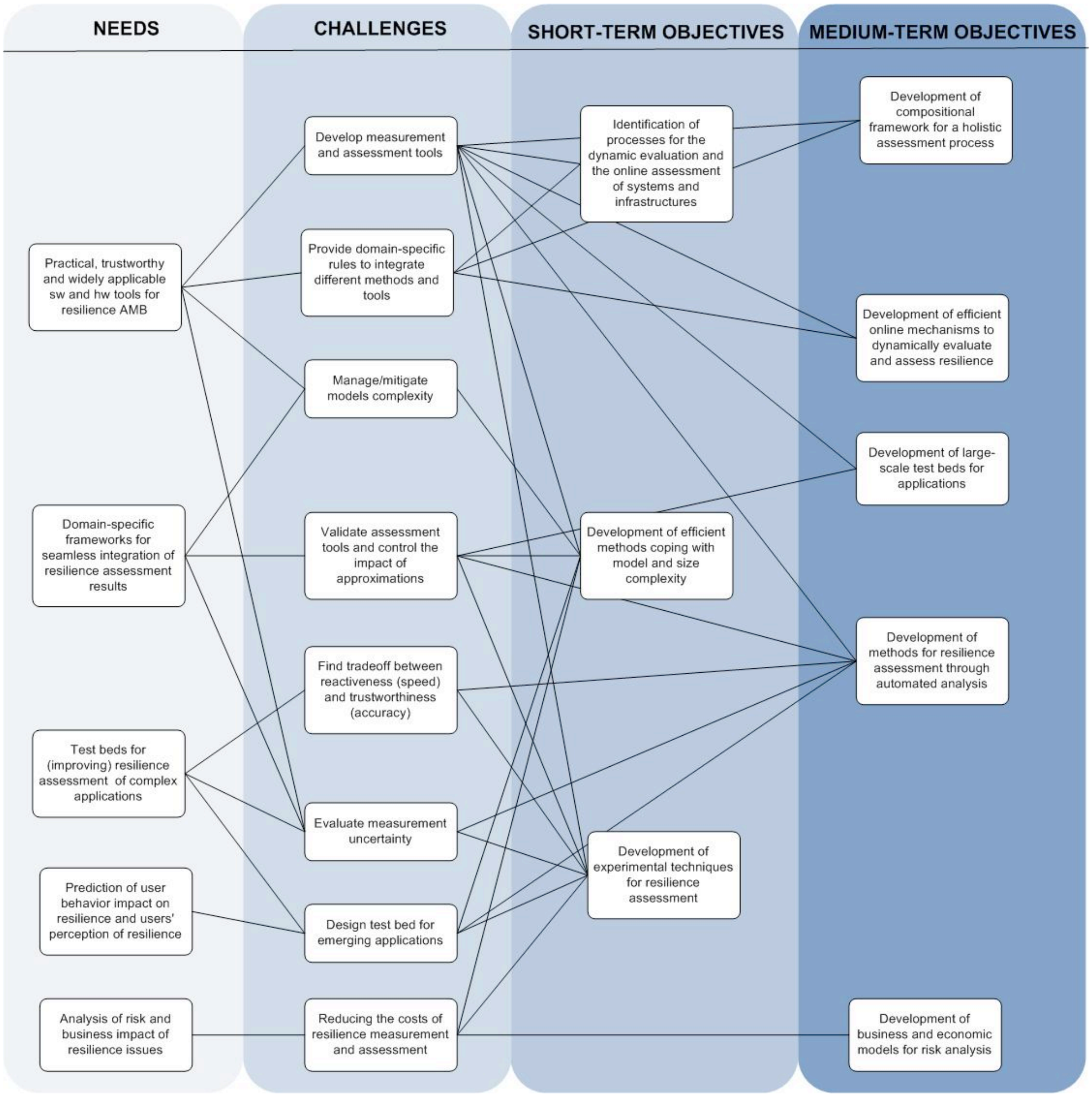
NEEDS
CHALLENGES
SHORT-TERM OBJECTIVES
MEDIUM-TERM OBJECTIVES
Practical, trustworthy and widely applicable sw and hw tools for resilience AMB
Domain-specific frameworks for seamless integration of resilience assessment results
Test beds for (improving) resilience assessment of complex applications
Prediction of user behavior impact on resilience and users' perception of resilience
Analysis of risk and business impact of resilience issues
Develop measurement and assessment tools
Provide domain-specific rules to integrate different methods and tools
Manage/mitigate models complexity
Validate assessment tools and control the impact of approximations
Find tradeoff between reactiveness (speed) and trustworthiness (accuracy)
Evaluate measurement uncertainty
Design test bed for emerging applications
Reducing the costs of resilience measurement and assessment
Identification of processes for the dynamic evaluation and the online assessment of systems and infrastructures
Development of efficient methods coping with model and size complexity
Development of experimental techniques for resilience assessment
Development of compositional framework for a holistic assessment process
Development of efficient online mechanisms to dynamically evaluate and assess resilience
Development of large-scale test beds for applications
Development of methods for resilience assessment through automated analysis
Development of business and economic models for risk analysis

Figure 2: Measurement and Assessment: links among needs, challenges and objectives.

4.3. Benchmarking

The needs, challenges and research directions related to the benchmarking activities are addressed in this section. Among the main issues, we identify:

- Availability of agreed resilience benchmark measures.
- Elaboration of adaptable benchmarking processes.
- Development of benchmark frameworks (reusable benchmark components and tools).
- Integration of benchmarking with the design methodologies.
- Proper maintenance of benchmarks to avoid negative effects[4].

[4] Introducing benchmarks in an area where there is a lack of objective data for decision making has obvious positive potential. However, benchmarks need to be the *right benchmarks*, and for some applications there may be no right benchmark. The risk of negative effects arises from the fact that benchmarking imposes simplifications (that may not be synchronized with the changing demands): the benchmark environments do not necessarily match the environments of the majority of users, and benchmarks may give more prominence to some measures than appropriate for users. Inappropriate benchmark measures may distort design processes: designers may overrate the design aspects that affect the benchmark compared to their true importance for users.

In the following Table 5 we outline the needs and challenges.

Benchmarking
Needs • Agreed, cost effective, easy to use, fast and representative enough resilience benchmarks for well defined domains. • Benchmark frameworks (components and tools) able to be reused to create benchmarks in different benchmarking domains. • Inclusion of adequate design methodologies to facilitate benchmark implementation and configuration in future components, systems, and infrastructures. • Uniform (standardized) benchmarking process that can be applied by independent organizations to offer certification of the resilience of COTS products (like in the case of standards compliance testing).
Challenges • Defining benchmark domains (components, systems, application domains) in order to divide the problem space in adequate/tractable segments. • Defining key benchmark elements such as measures, workload, faultload, attackload, models, to ensure the necessary benchmark properties (e.g., representativeness, portability, scalability, accuracy, repeatability, evolvability) that allow agreement on benchmark proposals. • Coping with highly complex, adaptable and evolving benchmark targets (components, systems and services). • Dealing with human factors in the definition and execution of resilience benchmarks. • Assuring proper validation of resilience benchmarks in order to achieve the necessary agreement to establish benchmarks. This implies the validation of the different benchmark properties (e.g., representativeness, accuracy, repeatability, portability, scalability, etc). • Assuring reusability of benchmark frameworks (components & tools) to create benchmarks in different benchmarking domains. • Defining and agreeing on a domain-specific resilience benchmarking process that can be accepted by the parties concerned (supplier, customer and certifier) and can be adapted to different products in the domain (e.g., in a product line).

Table 5: Benchmarking: needs and challenges

The following Table 6 details actions linked to the objectives, estimating the time at which they could be achieved.

Accordingly, the benchmark adoption process needs to be informed by research, and benchmark maintenance needs to include the adaptation of benchmarks to changing demands and phasing out of benchmarks that have become inadequate.

Benchmarking	
Short term (0-3 years)	
Objectives	***Actions***
Agreed resilience benchmark measures.	• Validation of candidate measures for different benchmarking domains, with the goal of making these measures universally accepted. • Defining lists of relevant features/properties to be checked/verified in the target system, for the benchmarking domains where the verification of properties through analytical means is prevalent.
Validated reference faultloads (i.e., sets of faults that are representative of specific domains) and corresponding injection tools (that allow easy implementation and portability of the faultloads) to be used in the development of resilience benchmarks.	• Finding whether representative types of faults exist through field studies and analysis. • Definition and validation of reference faultloads (considering the required benchmark properties) for different benchmarking domains and classes of targets. • Development of tools to inject reference faultloads in different classes of benchmark target systems.
Reference attackloads and injection tools to be used in the development of security benchmarks.	• Finding whether representative types of attack patterns and security vulnerabilities exist through field studies and analysis of information available. • Definition and validation of reference attackloads for different security benchmarking domains and classes of targets. • Development of tools to inject reference attackloads in different classes of benchmark target systems.
Cost effective, easy to use, and fast enough resilience *benchmark prototypes* for well defined domains.	• Definition of resilience benchmark prototypes for specific domains. • Validation of the benchmark prototypes. • Execution of benchmarking campaigns in real systems to characterize and demonstrate benchmark features and properties.
Medium term (3-8 years)	
Objectives	***Actions***
Concrete resilience benchmarks for specific domains.	• Development of concrete examples of resilience benchmarks (possibly as an evolution/standardization of benchmark prototypes).

Reusable benchmark components & tools to facilitate the development of benchmarks in different benchmarking domains.	• Identification of the layers for the different classes of benchmark targets where benchmark components and tools can be integrated/reused[5]. • Definition of standard interfaces for the integration of available components and tools. • Development of techniques for using existing "hooks" for monitoring and injecting faultloads and attackloads. • Development of tools and configurable components (e.g., parameterized workloads, faultloads, attackloads) that can be reused through their standard interface. • Development of configurable benchmark management engines to integrate available components and tools and manage the steps of a benchmark execution. • Development of methods and tools for the integration and configuration of benchmark components on the basis of the specification of the benchmark targets.
Resilience benchmark prototypes for highly complex dynamic systems.	• Identification of benchmark measures for dynamic systems. • Development of benchmark prototypes based on online monitoring and continuous benchmark measure integration.
A uniform (agreed) measurement-based, quantitative process to certify the resilience of products in a given domain.	• Identification of domains in which uniform benchmarking processes can be defined. • Agreement on the benchmarking processes in relevant industry and user community. • Specification of re-usable and adaptable benchmarking processes. • Developing concrete examples and demonstrations.

Table 6: Benchmarking: detail of actions

[5] Examples of layers: operating system, virtual machines, middleware, and key components such as web servers and database management systems, etc.

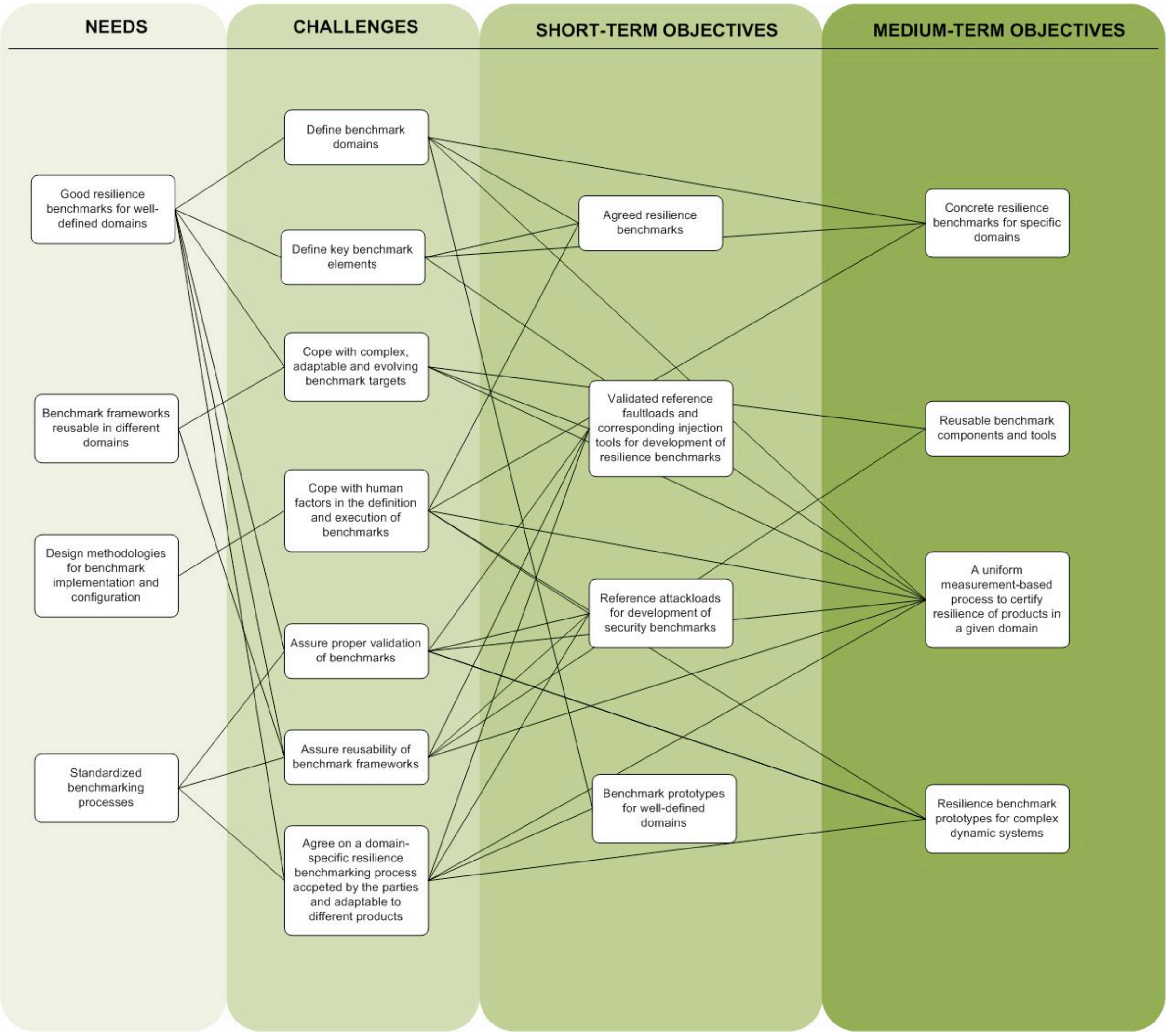
NEEDS
CHALLENGES
SHORT-TERM OBJECTIVES
MEDIUM-TERM OBJECTIVES
Good resilience benchmarks for well-defined domains
Benchmark frameworks reusable in different domains
Design methodologies for benchmark implementation and configuration
Standardized benchmarking processes
Define benchmark domains
Define key benchmark elements
Cope with complex, adaptable and evolving benchmark targets
Cope with human factors in the definition and execution of benchmarks
Assure proper validation of benchmarks
Assure reusability of benchmark frameworks
Agree on a domain-specific resilience benchmarking process accpeted by the parties and adaptable to different products
Agreed resilience benchmarks
Validated reference faultloads and corresponding injection tools for development of resilience benchmarks
Reference attackloads for development of security benchmarks
Benchmark prototypes for well-defined domains
Concrete resilience benchmarks for specific domains
Reusable benchmark components and tools
A uniform measurement-based process to certify resilience of products in a given domain
Resilience benchmark prototypes for complex dynamic systems

Figure 3: Benchmarking: links among needs, challenges and objectives.

4.4. Education, Training, Standardization and Take Up

This section focuses on the educational, training and standardization issues related to the assessment, measurement and benchmarking of resilience. Among the main issues that should involve the actions of the research community, we identify:

- Promoting proper and fair resilience AMB practices, in the form of, for example, state-of-the-art reports, cookbooks, and success stories.
- Preparation of courses and trainings to assure the availability of sufficient workforce with experience in AMB methods.
- Contribution to the establishment of the proposal and promotion of benchmark standards (in different application domains).

These issues may not identify well-defined research actions (i.e., objectives of research projects), thus this section can be considered as an (important) complementary part of the research roadmap.

In the following Table 7 we outline the needs and challenges.

Education, training, standardization and take up
Needs • Availability of a sufficient workforce with an education in AMB methods for resilience to allow vendors, customers and other parties to take advantage of available assessment methods and benchmarks: o General understanding (for engineers/managers/users) to choose resilience AMB methods and use their results in decision making. o Specialist training for developing and applying the AMB methods. • Commonly accepted AMB methods, for those purposes for which sound methods are technically feasible. • Standards for benchmarking (and for the assessment methods behind them) of resilience aspects in different domains. • Availability of assessment and benchmarking results to the parties (vendors, customers, authorities, researchers, consumers) that can benefit from them, in sectors where benchmarks are mature. • Dissemination of research results, including benchmark prototypes, showing that resilience benchmarks are technically achievable and cost effective. • Availability of comprehensive state-of-the-art reports (including research gaps, limitations and success stories) on resilience assessment techniques. • Availability of a cookbook (or cookbooks for different domains) on resilience assessment and benchmarking. • More widespread use of sound and objective assessment methods even in sectors where benchmarks and cookbooks are not or cannot be developed.
Challenges • Reaching consensus on assessment methods, for various problems and domains: o Understanding the state-of-practice in resilience assessment which is made difficult by the confidential nature of resilience data and resilience assessment methods. o Creating sufficient mutual understanding among practitioners of the technical, psychological and sociological specialties involved in interdisciplinary aspects like human behaviour. • Creating generally accepted courses on resilience assessment, experimental evaluation of computing systems, and methodologies for benchmarking to promote widespread and shared understanding of principles, potential and limits of resilience AMB methods. • Understanding the value of assessing and benchmarking resilience in the business domain and in the user/customer community. o Results of proper and fair assessment to become first class information for systems and services available to consumers and general public. • Achieving acceptance of resilience benchmarks considering the risk of premature acceptance of benchmarks.

Table 7: Education, training, standardization and take up: needs and challenges

The following Table 8 details actions linked to the objectives, estimating the time at which they could be achieved.

Education, training, standardization and take up	
Short term (0-3 years)	
Objectives	***Actions***
Definition of specific syllabus and course material for resilience assessment.	• Identification of knowledge and skills required to measure and assess resilient systems (partly on the basis of surveying potential employers' practices). • Identification of core taught subjects for specific courses (university, short and industrial courses). • Specification and preparation of syllabus and course material including: o Fundamental theories (like measurement theory, model based evaluation). o Applications of fundamental theories to assessment, safety and dependability cases. o Experimental evaluation of computing systems. o Application and capabilities of tools. o Methodologies for benchmarking. • Make the course and support materials (including tool related materials) available in the form of cookbooks.
Promotion of proper and fair resilience assessment practices for specific classes of systems/services.	• Promotion of measurement practices for the different domains in academic, industry, and standardization bodies, with the goal of making these universally accepted. • Distribution of assessment results. • Promotion of the inclusion of procedures for assessment and benchmarking in standards. • Publication of success stories and proof cases (of business/market value) about the application of assessment and benchmarking activities. • Planning education campaigns to help managers decide whether to invest in education for AMB resilience.
Identification of the conditions under which standard dependability benchmarks would be beneficial.	• Research in economics and social sciences about the effects of various types of benchmarks and standard assessment methods, drawing on the history of other industrial areas with comparable technical, social or market situations. o Perform market analysis about the value of public dependability measures in various application domains. o Identify application areas with the need for regulation to apply dependability measures.

Establishment of a first forum (from industry and/or user community) for the proposal and promotion of resilience benchmarks for specific benchmarking domains[6].	• Dissemination and promotion of the benchmark prototypes to influence the development of standard benchmarks inspired by the prototypes. • Promotion of the value of resilience benchmarking in the industry, user communities, special groups and standardization bodies to encourage the establishment of resilience benchmarking forums. • Dissemination of research results, including benchmark prototypes, showing that resilience benchmarks are technically achievable and cost effective.
Medium term (3-8 years)	
Objectives	***Actions***
Gain widespread industrial and user acceptance of benchmark standards (in different domains).	• Creating a consortium (of vendors, buyers, regulatory authorities with an interest in a certain application/technology domain), or get in contact with existing consortia, for the definition and acceptance of benchmarks (with industry support specific for the domain). • Based on the benchmarks identified above, promoting standard hardware and software benchmarks tools (instruments) that allow for widespread proliferation of the benchmark.
Promotion of proper and fair assessment as routine procedure to characterize systems and services by customers and general public.	• Contributing to standards and policies that force/promote publicity of assessment and benchmarks results. • Promoting benchmarks results to motivate the routine use of resilience assessment and to increase the trust in it.

Table 8: Education, training, standardization and take up: detail of actions

[6] Standard, de facto resilience and security benchmarks can appear as a result of spontaneous adoption of benchmark prototypes by the user community.

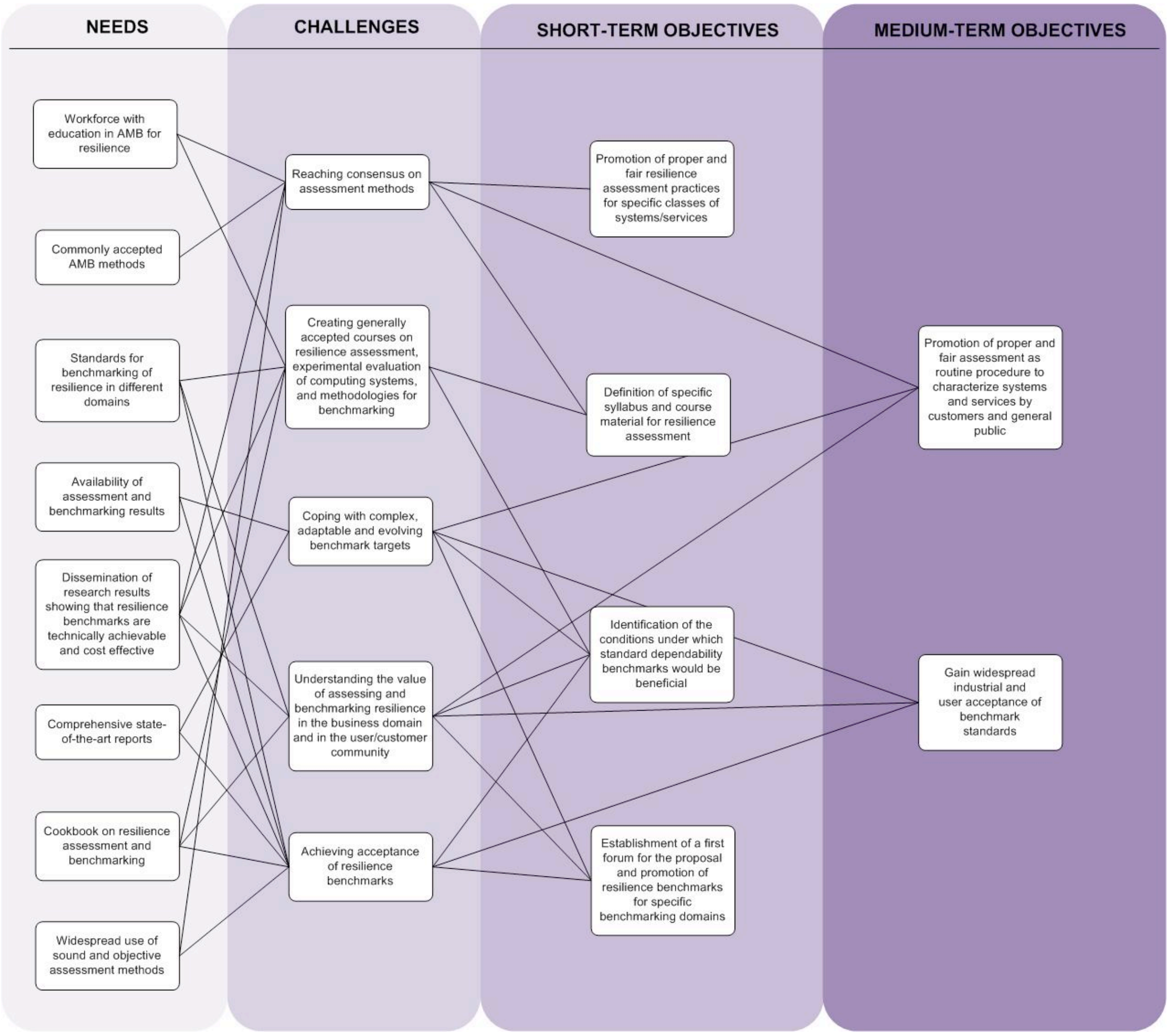
NEEDS
CHALLENGES
SHORT-TERM OBJECTIVES
MEDIUM-TERM OBJECTIVES
Workforce with education in AMB for resilience
Commonly accepted AMB methods
Standards for benchmarking of resilience in different domains
Availability of assessment and benchmarking results
Dissemination of research results showing that resilience benchmarks are technically achievable and cost effective
Comprehensive state-of-the-art reports
Cookbook on resilience assessment and benchmarking
Widespread use of sound and objective assessment methods
Reaching consensus on assessment methods
Creating generally accepted courses on resilience assessment, experimental evaluation of computing systems, and methodologies for benchmarking
Coping with complex, adaptable and evolving benchmark targets
Understanding the value of assessing and benchmarking resilience in the business domain and in the user/customer community
Achieving acceptance of resilience benchmarks
Promotion of proper and fair resilience assessment practices for specific classes of systems/services
Definition of specific syllabus and course material for resilience assessment
Identification of the conditions under which standard dependability benchmarks would be beneficial
Establishment of a first forum for the proposal and promotion of resilience benchmarks for specific benchmarking domains
Promotion of proper and fair assessment as routine procedure to characterize systems and services by customers and general public
Gain widespread industrial and user acceptance of benchmark standards

Figure 4: Education, training, standardization and take up: links among needs, challenges and objectives.

5. Topmost priorities identified by the AMBER project

This chapter presents a short list of topics, among those listed in Chapter 4, that are seen by the AMBER consortium as having the highest priority for a possible research programme. This list was selected by consensus, through rounds of debate within the consortium. Each partner contributed its knowledge of specific stakeholders and of the links between required scientific advances. An agreed constraint was to keep the final list short: there are six top priorities from the research areas of “Scientific and technological foundations”, “Measurement and assessment”, “Benchmarking”; plus two concerning “Education, training, standardization and take up”.

We have not ranked the priorities within this short list. We have instead identified, for each item in the list, the more general goals, in terms of desired changes in the landscape of application of Assessment, Measurement and Benchmarking (AMB) of resilience, which it aims to satisfy.

Ranking within our list, for the purpose of a funded research programme, would depend on a necessary political choice between these more general goals, which are listed below.

General Goals

1. Extension and combination of AMB methods to ensure that resilience assessment integrates security issues together with accidental faults, design faults with physical faults, human behaviour with machine behaviour, even in very complex systems.
    - *This integration is necessary now, and becoming more essential, with the increasingly complex and integrated systems that characterise “future Internet” scenarios and generally the Information Society.*
2. Addressing the fundamental difficulties in quantitative assessment for high consequence, low probability events: predictive value of past experience, combination of diverse evidence, uncertainties about the models and assumptions used.
    - *This goal is crucial for society: these difficulties affect the assessment of systems and infrastructures with great societal value but also great potential risk. Advances that expand the range of applicability of AMB in these critical areas would also have beneficial fall-out for the less critical applications.*
3. Empirical validation of the practices already developed in measurement, modelling and benchmarking, so that industry has a basis for steering its own investment regarding these techniques.
    - *Adoption of new techniques throughout industry requires companies to invest in adapting and implementing the techniques. But this requires sufficient empirical evidence of how effective each technique is, and within which constraints.*
4. Making the current practices of measurement, assessing and benchmarking more rigorous, e.g. via better use of the established principles of metrology, better matching of the choice of metrics to the measurement needs, and widely.
    - *A more rigorous approach in applying AMB techniques would significantly increase the benefits they offer and reduce the risks from inappropriate application.*
5. Building AMB techniques or tools for specific systems and application areas where these are currently inadequate.
    - *In some applications, for instance in Future Internet scenarios with their characteristics of dynamicity, large scale, heterogeneity, developing AMB techniques and tools poses new research challenges, beyond the application of known principles and solution*

6. Gaining widespread acceptance of AMB practices and results.
    - *Promoting AMB practices in the form of, for example, state-of-the-art reports, cookbooks, and success stories, and disseminate the AMB results to the parties that can benefit from them would contribute to the achievement of this goal.*

List of high priority items

Investigation Areas	Topmost Priority Topics	Related General Goals
Scientific and technological foundations	Validated methods for extrapolating measurements to predictions of system behaviour despite differences between the system in operation and its environment and the system/environment where the measurements were taken.	2, 4, 5
	Investigation of principles for successful integration of resilience assessment, measurement and benchmarking technologies into different phases of the life cycle of IT systems, including methods for evaluating technical efficiency and economic impact.	1, 4, 5
	Improved “argumentation” processes, which correctly formulate, communicate and verify complex arguments combining “hard” evidence (measurement, mathematical models) and “soft” evidence (judgement), with proper treatment of epistemic uncertainty, levels of confidence, and “unknown unknowns”.	1, 2, 4
Measurement and assessment	Development of efficient on-line mechanisms to monitor the environment conditions of the system and to dynamically evaluate and assess its resilience.	1, 5
	Practical, trustworthy and widely applicable tools for measurement and assessment in large-scale dynamic systems, adaptable and evolving infrastructures, and other domains where these are lacking.	1, 5
Benchmarking	Validated reference faultloads (i.e., sets of faults that are representative of specific domains) and corresponding injection tools (that allow easy implementation and portability of the faultloads) to be used in the development of resilience benchmarks.	3, 5
Education, training, standardization and take up	Dissemination of research results, including benchmark prototypes, showing that resilience benchmarks are technically achievable and cost effective.	6
	Promotion of proper and fair resilience assessment practices for specific classes of systems/services.	6

6. Tailoring the roadmap to specific roles and industrial domains: specific examples

In Chapter 5 we listed the research (as well as education/training) topics agreed, within the AMBER Consortium, to have highest priorities for the progress of Assessment, Measurement and Benchmarking of resilience. This chapter, by contrast, proposes a set of viewpoints about priorities as seen from the perspective of specific industrial domains or professional roles, thus providing alternative ways of reading the roadmap. We considered the following perspectives:

- Embedded Systems (Section 6.1),
- Transportation (Section 6.2),
- Certification Authorities and Assessors (Section 6.3),
- Future Internet technological platforms (Section 6.4),
- Service Architectures, Platforms and Infrastructures (Section 6.5), and
- Enterprise Security (Section 6.6).

Within each industrial domain and each role there will be a variety of actors with different interests and opinions. From this range of different positions, this chapter documents a sample of opinions of senior experts. The discussion provided for each "perspective" in the above list is the outcome of rounds of interaction between AMBER members and AMBER Advisory Board members working in these different domains, aiming to identify the items of the roadmap that may get priority in the respective domain.

6.1. Embedded Systems

This section describes the research priorities for the embedded systems domain, specifically addressing the perspective of a systems integrator. The role of a systems integrator is to bring together components manufactured by several companies into a complete product, such as an airplane, a power plant or a satellite. One important concern in this domain is to be able to predict the robustness of a design as early as possible in the development. This calls for techniques to extrapolate measurements from previous designs to new ones, taking advantage of the field data usually collected by mature companies. To facilitate adoption, those techniques must be easily integrated with existing methodologies, such as model-based development. Therefore, two important research problems are:

- Validated methods for extrapolating measurements to predictions of system behaviour despite differences between the system in operation and its environment and the system/environment where the measurements were taken.

- Reducing the cost of resilience assessment and measurement by developing methods and tools that are easy to integrate into existing development methods and tool chains.

It is fundamental for an integrator to understand the resilience of components and subsystems that are purchased from suppliers. Simple metrics such as the mean time between failures are insufficient, and suppliers are gradually expected to provide more information on the failure modes and failure rates of their components. This can be achieved, at least in part, if suppliers perform resilience benchmarking and make the results available. A relevant step in this direction is:

- Development of concrete examples of resilience benchmarks (possibly as an evolution/standardization of benchmark prototypes).

Manufacturers of safety-critical systems must provide arguments sustaining the safety of their products. In some industries these arguments are documented and compiled into “safety cases”, which enable all stakeholders to examine the available evidence. It is important for manufacturers to have high confidence in the argumentation. Otherwise, they are forced to incorporate more redundancy in their designs, in order to remain on the safe side. Thus, an important issue is:

- Improved “argumentation” processes, which correctly formulate, communicate and verify complex arguments combining “hard” evidence (measurement, mathematical models) and “soft” evidence (judgement), with proper treatment of epistemic uncertainty, levels of confidence, and “unknown unknowns”.

When a systems integrator acquires a hardware module from a supplier and installs its own software on it, there is a need to assess the final configuration. The supplier can ease this assessment by equipping the hardware modules with:

- Standard set of monitoring features and “hooks” to facilitate AMB and the development of related tools.

Lastly, companies learn from accumulated experience how to improve the resilience of their systems. Field failure data is fundamental to enable this process. However, it is often difficult to obtain and log the necessary data, particularly when using commercial off-the-shelf components. The embedded systems industry would therefore benefit from ways to understand the circumstances under which components fail. To this end, progress should be made in:

- Development of efficient on-line mechanisms to monitor the environment conditions of the system and to dynamically evaluate and assess its resilience.

6.2. Transportation

The commercial transport industry is developing ever more complex Intelligent Transport Systems (ITS) to enhance the range of features provided to drivers, fleet managers and other stakeholders. The present focus is on enabling vehicle connectivity to numerous other systems and infrastructures, including traffic management centres, local authorities, weather stations, toll offices, other vehicles and the Internet.

The growing complexity of systems in this domain needs to be met with appropriate (and possibly new) methods and tools for assessing, measuring and benchmarking resilience. One important challenge lies in reducing the cost and time required to perform certification, verification and validation activities. To this end, the most relevant foundational issues are:

- Investigation of principles for successful integration of resilience assessment, measurement and benchmarking technologies into different phases of the life cycle of IT systems, including methods for evaluating technical efficiency and economic impact.

- Adoption of sound measurement practices in the transportation domain.

- Standard set of monitoring features and "hooks" to facilitate AMB and the development of related tools.

Transportation systems are becoming more dynamic and capable of adapting to changes in the operation environment. There is a lack of cost-efficient resilience assessment techniques for such systems, especially techniques that are able to cope with the unpredictability of the environment and the customized configuration of each vehicle. For these reasons, there are two main issues in measurement and assessment that should be addressed:

- Development of efficient on-line mechanisms to monitor the environment conditions of the system and to dynamically evaluate and assess its resilience.

- Development of experimental techniques for resilience assessment in ITS.

In the long term, transportation systems are expected to provide greater flexibility, allowing for instance third-party applications to be downloaded onto onboard units. Such runtime changes to the system's configuration would benefit from ways to benchmark the impact of software changes on the overall resilience. Two relevant steps in this direction are:

- Cost effective, easy to use, and fast enough resilience benchmark prototypes for the transportation domain.

- Validated reference faultloads (i.e., sets of faults that are representative of specific domains) and corresponding injection tools (that allow easy implementation and portability of the faultloads) to be used in the development of resilience benchmarks.

Lastly, the transportation industry faces very rapid changes in processes and technologies, leading managers and engineers to invest in self-learning. Such initiatives require teaching material to be made available. Thus, an important step would be:

- Definition of specific syllabus and course material for resilience assessment.

6.3. Certification Authorities and Assessors

The main goal of an assessor (or of a certification authority) is to check if the dependability requirements (e.g., in terms of safety, security, resilience) of a specific system are satisfied or not. This is a very crucial work that usually concerns safety-critical systems where a failure can lead to catastrophic consequences, so the assessor should rely on a set of techniques and tools which allows him to trustfully rely on the quality of the assessment process and on the outcomes produced. In other words, there is a general need of

- practical, trustworthy and widely applicable tools for measurement and assessment in large-scale dynamic systems, adaptable and evolving infrastructures, and other domains where these are lacking.

Another important issue is that, in the assessment process, the system's behaviour is usually predicted using data collected in the past, sometimes related to different environments or even to similar systems. To take correct decisions, corrective factors should be known to extrapolate future system's behaviour from slightly different scenarios, or at least indications should be available about the uncertainty of predictions. In other words, there is the need to

- validate methods for extrapolating measurements to predictions of system behaviour despite differences between the system in operation and its environment and the system/environment where the measurements were taken.

As detailed in the scenario concerning safety certification and regulation (see Section 2.5), there are several open challenges that still need to be managed and that currently prevent an assessor from being certain about the quality of the assessment, so both *short term* and *medium term* research activities are considered important. With respect to the AMBER roadmap, the focus will be on the "Measurement and Assessment" and "Scientific and Technological Foundations" areas rather than on "Benchmarking", since the objective is not comparing different systems or different products to select the most dependable one, but just to assess the dependability properties of a critical system to allow the claim they are above the required threshold.

The main *medium term* objective is the

- Development of (domain-specific) compositional framework for a holistic assessment process.

  The complexity of current critical system, in terms of heterogeneity, evolvability, largeness, dynamicity, inhibits the application of well-proven traditional methods "as they are", but requires the development of an assessment framework where the synergies between different evaluation techniques and tools are exploited to provide realistic assessments. A related challenge is the

  - provision of domain-specific rules to compose/integrate different methods and tools for resilience assessment,

  which concerns the expressive power of the formalisms (for efficient modelling) as well as the complexity that the supporting solution tools can handle. The elaboration of proper rules to divide the problem and then compose/integrate the results of the different methods and tools used to solve the sub-problems is a possible encouraging approach to attack this challenge.

To achieve this medium term objective, particular effort should be put on:

- The identification of the base types of measurement and modelling techniques and tools (applicable in a given application domain) and the possible interactions among them to provide realistic assessments;

- The assessment of the impact of the approximations introduced in modelling on the resilience-related predictions sought.

Moreover, a number of *short term* (thus preliminary) actions and objectives should be pursued:

- Identification of simple and universally accepted resilience-related metrics.

- Development of efficient methods coping with model and size complexity.

Besides these research activities, there are other issues seen of primary importance in education as well as standardization. Among them:

- the need of comprehensive state-of-the-art reports (including research gaps, limitations and success stories) on resilience assessment techniques.
- the availability of a cookbook (or cookbooks for different domains) on resilience assessment and benchmarking.
- the short term objective of defining specific syllabus and course material for resilience assessment.

6.4. Future Internet technological platforms

The most important theme in the Future Internet domain is currently related to the proliferation of *digital identities* and their *trustworthin*ess.

The services that will constitute the basis of the Future Internet should offer simple access methods, and the problem of guarantying security in the usage of these services should be always considered.

Future Internet will have characteristics of high dynamicity, with the necessity of maintaining a high level of resilience. Future Internet systems will be characterized by needs of adaptability to different environment conditions. Future Internet systems will also be evolving systems. In a context like this, to obtain off-line trustable (measurement) results of resilience will be difficult to enhance the capabilities of on-line monitoring systems will thus be of uttermost importance.

We identified the following six points of the AMBER final research roadmap as the topmost research priorities in this domain.

- Development of efficient on-line mechanisms to monitor the environment conditions of the system and to dynamically evaluate and assess its resilience.
- Standard set of monitoring features and “hooks” to facilitate AMB and the development of related tools (at least for computer systems in specific domains).
- Coping with highly complex, adaptable and evolving benchmark targets (components, systems and services).
- Designing and developing test beds for emerging applications, e.g., cloud computing and collaborative services.
- Reusable benchmark components & tools to facilitate the development of benchmarks in different benchmarking domains.

- Understanding the value of assessing and benchmarking resilience in the business domain and in the user/customer community.

6.5. Service Architectures, Platforms and Infrastructures

Service architectures, platforms and infrastructures become increasingly vulnerable as they become more complex, thus they need to be made resilient to attacks and operational failures. Minimizing the risk associated with loss of IT services or user data needs precise *metrics* and *models* for evaluating risk, augmenting *tools* with resilience evaluation capabilities in the design phase, and rigorous resilience *measurement* and *real-time assessment* techniques in the operational phase. These needs are reflected in the following objectives and actions selected by the domain experts.

Various *resilience metrics* are used in an increasing extent in IT systems. However, only a few metrics can be used all over the lifetime of an IT system in a consistent way, due to the lack of proper definitions and a related metrological foundation. There is a significant ambiguity in the definitions used and in the specification of the validity of measures and benchmark results, thus reducing the portability and reusability of results used in quantitative evaluation. In case of security, adaptation of metrics to evolving threats is largely unsolved. Accordingly, the first priority is the following:

- Elaboration of easy-to-use, practically measurable (domain-specific) resilience metrics (including metrics for security) and establishing common "operational" definitions for them, in domains where these are lacking.

Resilience *predictive capability* is useful for system developers, designers, architects, and IT practitioners who assemble the hardware and software into even more complex systems, and the clients who purchase and operate the systems. Methods to evaluate resilience metrics prior to build, at the various levels of build, are especially useful, permitting design change before design commitment. In addition to failures, it would be valuable to assess the resilience to planned change activities such as release and version updates, hardware or software configuration changes, etc. The ability to quantify resilience along the dimensions of the application environment so as to more accurately predict for a specific use case of the application is highly desirable:

- Elaboration of extrapolation methods and tools to generalize observations (measures).

The correspondence between IT based *risks* and business (or more generically environmental) processes is of high concern. The qualification and quantification of risks must heavily dependent upon their implications on business processes using the IT systems under analysis. Resilience measures need to focus on the risks with the highest impacts on the supported processes. In the area of IT, because of inadequate or non-existent models and tools, assessing business risks is not always done or done with insufficient rigour. If there were trustworthy and accurate models for evaluating risk associated with loss if IT service, it is expected that they would be extensively used by researchers (to demonstrate risk reduction capabilities), designers, consultants and clients (expectation is that clients would invest in resilience if risks were clearly understood). The development of models for risk analysis has to be extended with the elaboration of corresponding methods and tools. The related research priorities are the following:

- Understanding the economics and overall impact of resilience assessment on the lifecycle of IT systems.

- Development of business and economic models for risk analysis.

Real-time assessment of operational system resilience is a critical need. Even when a system has been properly configured for resilience, changes can take place which reduce or even eliminate protection measures. If there is no notification of the degraded state, outages may occur that the system was assumed to be guarded against. Advanced analytics to evaluate system state and to identify states where resilience is exposed have great value. Initially, real-time assessment should provide for issue notification and recommendations on corrective actions. Over time, automation of the required responses can be expected. In large-scale dynamic systems new, efficient model-based online assessment methods are needed. Accordingly, the following research is of high priority:

- Development of efficient online mechanisms to monitor the environment conditions of the system and to dynamically evaluate and assess its resilience.

Aimed more at the system design stage than the operational phase, the development of methods and tools for resilience assessment would be beneficiary. These would be used by system and application architects, that is, those who design and build complex IT systems at various levels. Augmenting design tools with resilience evaluation capabilities is an excellent objective as is resilience assessment of design models:

- Development of methods for resilience assessment through automated analysis.

The advances in resilience assessment and measurement shall face (and serve) the paradigm shift observable in this domain nowadays: details of computing resources are abstracted from the users who no longer need control over the technology infrastructure. Virtualization and other technologies enable convenient, on-demand access to a shared pool of configurable computing resources as outsourced services. In this context, assessment and measurement of resilience is closely related to Service Level Agreement / Quality of Service management and charging of customers. Accordingly, the used techniques have to be precise, rigorous, and at the same time widely accepted and agreed among service providers and customers.

6.6. Enterprise Security

CISOs need to make tools to make security investment decisions. Currently, they base their decision making on trends in the industry, general intuition of good practice, etc; they also need to be salesmen to sell their proposed decisions to management. The tools CISOs would want fall in two camps: (i) tools that allow objective decision-making and (ii) ways to share knowledge (an interesting aspect not covered by AMBER much). The hope of a CISO is not that much in expert systems that act as oracle and divine answers, but in tools that provide objective suggestions, communicate the important aspects to consider, etc. A CISO is very much aware of the importance in understanding human tendencies, often economically driven, possibly influenced by training, and often driven by personal tendencies such as risk averseness. Although a CISO understands that assessment is a key ingredient in making decisions in more objective fashion, assessment per se is not a goal they pursue: it's the decision making that matters, not the assessment.

Priorities:

- Integration of considerations related to human behaviour in the assessments of resilience of computer systems as affected by the behaviour of their users, system managers, and adversaries.

- Practical, trustworthy and widely applicable tools for measurement and assessment in large-scale dynamic systems, adaptable and evolving infrastructures, and other domains where these are lacking.

- Finding whether representative types of faults exist through field studies and analysis. (In this case, faults should be read as ‘attacks’. The fact that it is field studies and analysis is irrelevant, any means are okay.)

References

[1] A. van Moorsel et al., “State of the art (final version)”, EU FP7-216295 AMBER project, deliverable D2.2, July 2009.

[2] AMSD. AMSD: A Dependability Roadmap for the Information Society in Europe, Part 3 - Towards a Dependability Roadmap. AMSD Deliverable D1.1, 2003.

[3] Grid Consortium. ICT Vulnerabilities of Power Systems: A Roadmap for Future Research. Deliverable D11. December 2007.

[4] ReSIST NoE, Deliverable D13: From Resilience-Building to Resilience-Scaling Technologies: Directions, September 2007.

[5] A. Bondavalli, P. Lollini (editors) et. al., “Preliminary Research Roadmap”, EU FP7-216295 AMBER project, deliverable D3.1, September 2008.

Appendix A – Examples of needs/challenges/objectives/actions

This appendix presents examples of the needs, challenges, objectives and actions discussed in the previous chapters, aiming to clarify their meaning and to provide a more concrete view of the underlying issues. Each item is linked to the respective item of Chapter 4. The examples have been grouped in the same four research areas as detailed in Chapter 4.

A.1. Scientific and Technological Foundations

A.1.1. Needs

A.1.1.1 Validated methods for extrapolating measurements to predictions of system behaviour despite differences between the system in operation and its environment and the system/environment where the measurements were taken.

This need has overwhelming social importance because in applications with high social impact (mass transportation, nuclear/chemical plants, weapon control, financial transaction infrastructure and other pervasive applications of ICT), we need to predict before operation the level of risks over a long lifecycle, in a variety of operation conditions. But the same need for extrapolation underlies most uses of AMB in information and communication technology: the need is to predict the level of resilience or dependability, from the moment of assessment onwards; but measurements concern the past. Differences between the measured past and the future to predict arise because, for instance, hardware components in new copies of the system may differ from those measured in manufacturing details and thus reliability and robustness against various stresses; software reliability was measured under certain usage profile, but this will evolve as the system is in use; security threats evolve; etc. The differences may be subtle and difficult to measure and yet critical.

The current trends in ICT make the need more urgent: software design may evolve rapidly in deployed systems; the future Internet will see increasingly more large-scale, geographically distributed, multi-organisation, open application systems and infrastructures, in which continuous evolution is the norm. Change is already a threat to meaningful probabilistic risk Assessment, and will be more so in the future.

Thus the need is pervasive, and progress in for the high social impact systems would also help more mundane applications, and vice versa.

While mathematical models deliver insight into the causes and extent of extrapolation errors, the extrapolation problem cannot be solved by mathematics. We need empirical validation of any proposed extrapolation methods (including the standard method of assuming that the future will mirror the past) and proposed corrective factors, with indications of how accurate predictions have been shown to be, depending on the types of system, scenarios of use, and constraints on their possible changes.

[*back to section 4.1 on Scientific and Technological Foundations*]
[*back to section 5 on AMBER Topmost Priorities*]
[*back to section 6.1 on Embedded Systems*]
[*back to section 6.3 on Certification Authorities and Assessors*]

A.1.1.2 Improved "argumentation" processes, which correctly formulate, communicate and verify complex arguments combining "hard" evidence (measurement, mathematical models) and "soft" evidence

(judgement), with proper treatment of epistemic uncertainty, levels of confidence, and "unknown unknowns".

Most important decisions involving ICT (e.g., operational risk assessment for a bank, probabilistic risk assessment for licensing a chemical plant, pricing of outsourcing for IT services, licensing of a medical technology, etc) include *elements* of AMB to estimate quantitative measures of resilience. But system-level estimates that are crucial for the decision, e.g., probability of losses exceeding a threshold, expected deaths over the lifetime of a plant, probability of costly SLA violations, also depend on judgement about issues like, e.g.:

- How likely is it that assumptions made in modelling will be violated, and how severe could the consequences be? Which unnoticed assumptions have not been analysed?
- The risk due to software failures or security breaches has really been guessed by checking conformance to standards - what level or risk is thus being accepted?
- There was testing on a reduced scale installation - how relevant is it for the full scale system?
- Expert opinion played an important role; what risk is being taken in trusting these experts?

These factors may amount to a substantial part of the risk to be considered. But the part of the decision process that is mathematically documented and verifiable is usually limited to deriving the implications of given models, assumptions, and estimates of important parameters. That is, the quantitative approach is applied to *aleatory* uncertainty (e.g., about when a component with known reliability will fail), but *epistemic* uncertainty (e.g. about the true reliability of that component) is not represented. For instance, quantitative estimates are seldom accompanied by statements of confidence.

Epistemic uncertainty extends to the models used, i.e., our very understanding of a system: e.g., whether events that have been assumed to be practically impossible are actually likely, unsuspected links exist between apparently independent components and events, and so on.

A fully mathematical treatment of epistemic uncertainty is feasible in theory, but not (or not yet) in practice. Yet the current alternative is inadequate: decision makers, even if concerned with these un-quantified uncertainties, will reach decisions in which their weightings for these factors are not explicit, and thus their reasoning is impossible to audit and verify, even for themselves. In the worst cases, the AMB components divert attention from the main risks and dilute (psychologically and organisationally) responsibility for critical decisions.

Research must aim at gradual but concrete progress in including more of these uncertainties in *explicit* form in practical processes of assessment and decision. Correct quantitative treatment is desirable where feasible, but progressive improvements may start with more explicit treatment of the assumptions underlying quantitative assessments, of the confidence in claims and of the risks that residual doubts imply for each option open to decision makers.

[*back to section 4.1 on Scientific and Technological Foundations*]
[*back to section 5 on AMBER Topmost Priorities*]
[*back to section 6.1 on Embedded Systems*]

A.1.1.3 Investigation of principles for successful integration of resilience assessment, measurement and benchmarking technologies into different phases of the life cycle of IT systems, including methods for evaluating technical efficiency and economic impact.

The overall aim of this research is to investigate and develop principles for how to successfully integrate resilience assessment, measurement and benchmarking technologies into the development life cycle of IT systems. Traditionally, research in AMB technologies has focused on the technical challenges involved in measuring, assessing and benchmarking dependability and security. Few, if any, research efforts have comprehensively addressed the process perspective of these technologies. In general, process-oriented research is holistic in its nature as it involves studies of interactions between technical, organizational as well as economical aspects of system development and system operation. We believe that there is an urgent need for research that addresses the process dimension of AMB technologies in order to facilitate their adoption by industry. Important goals of this research would be to:

- provide concrete evidence of the technical benefits of AMB technologies and their ability to improve system resilience in practice,
- develop concrete guidelines for managers and project leaders on how to integrate and utilize resilience AMB in different industrial settings,
- demonstrate and investigate the economic benefits and drawbacks of using resilience AMB during system development and system operation.

Without such evidence and guidance, it will be difficult, not to say impossible, to gain a widespread use and acceptance of AMB technologies. Reaching these goals will require comprehensive empirical studies of how system resilience is addressed and handled during development and operation by different organizations. In particular, to be able to scientifically state whether one approach of ensuring resilience is better, or more economical, than another, it will be necessary to conduct comparative studies where independent investigators (most likely academic researchers) empirically study actual development processes, or operational procedures, in several organizations within a given application domain. Such studies could involve the applications of well established empirical methods such as controlled experiments, surveys and post-mortem analyses, but may also require development of new empirical techniques specifically tailored to investigating the economical aspects of resilience assessment, measurement and benchmarking. While this kind of research is needed for all strands of resilience assessment techniques, we believe it is fundamentally more challenging, and more urgently needed, for techniques that involve measurement or assessment of security and human reliability.

[*back to section 4.1 on* Scientific and Technological Foundations]
[*back to section 5 on* AMBER Topmost Priorities]
[*back to section 6.2 on* Transportation]

A.1.2. Challenges

A.1.2.1 Understanding how differences in the environment of a system, or variations in manufacturing or configuring different instances of a type of system affect the important measures of resilience/dependability.

This vast set of challenges arise from the need “Extrapolation” and are about improving the quality of prediction based on empirical measures. For many of the ways in which the system of interest or its environment may differ from those for which the measurements were obtained, there are no proven methods for assessing (predicting) their effect. So, suppose for instance that a software component with known past reliability is included in a new system. We know from both theory and experiments that its reliability may change drastically compared to what was measured in the past. The open questions include:

- Is there a probabilistic model that maps measures of the differences between the new and old system to predictions of the reliability of this component in the new system? (the answer in this case is “yes”).
- Is it feasible to apply this model in practice to predicting this new value of component reliability? (the typical answer is no”: the parameters are too difficult to estimate).

- Is it feasible at least to determine whether the reliability change caused by the difference between the old and new system should invalidate the previous prediction, or how it should alter confidence in it? (the typical answer is again “no”).

Similar questions arise regarding the effects of all other changes, and need to be answered in terms of knowledge (mostly yet to be obtained) specific to the type of change and the measures of interest, e.g. attack resistance of software in different attack mixtures, error rate given different environments, fault and error propagation distributions in different hardware implementations with different software loads.

[*back to section 4.1 on* *Scientific and Technological Foundations*]

A.1.2.2 Providing a shared terminology and common accepted data presentation formats to support understanding of which data items can be compared (as representing the same resilience related attributes of systems).

Systematic collection of resilience data is challenged by the lack of shared understanding of the semantics of the resulting data sets. On this subject we propose to devise and propose a de facto agreement using default data collection in the main operating systems, and raise vendors and industry awareness to the need of an inter-vendor wide data collection standard. As an example, consider core dumps in Linux or MS Vista. When redirecting these core dumps to a repository, the format of the data is as specified by the operating system, which is dependent on the OS brand. A de facto standard on the data collection and the dictionary structure for that data would remove the problem of semantic differences in data, since new data sets can be expressed in relation to that *de facto* standardised data models. A big challenge in this matter lies in the cooperation level required from vendors and industry and in providing additional tools for relating other data sets to the OS data models. To reach this level of cooperation and vender-wide standardization, there will be needed a strong commitment into researching which models are best suited to both detailed information and brand-independence, as well as which, what and how to collect the relevant data.

[*back to section 4.1 on* *Scientific and Technological Foundations*]

A.1.3. Objectives

A.1.3.1 Understanding the economics and overall impact of resilience assessment on the lifecycle of IT systems.

It has been realised that economics is an important determining factor in the making of IT security investments. The reasoning is that security can always be improved further, but at a certain cost. Security decisions are therefore by definition trade off decisions between adding security or saving investment, operating costs and opportunity costs from the reduced functionality. Interestingly enough, this same observation has been far less prevalent in resilience as a whole, that is, security as well as dependability (availability, reliability, etc).

Recent year’s revolution in service provision over the Internet has influenced the way how organizations treat the dependability of their solutions. For the telecom industry, and many sectors of the computing industry, making their services highly available and reliable has always been just as important as the security aspects. As markets become worldwide open, other, formerly less concerned businesses start to consider dependability as a key factor for productivity and efficiency and to improve customer satisfaction.

Despite this trend, most of the organizations tend to focus more on security and dependability is only important up to a certain point. The question to research is why this is the case, if this is justified, and whether resilience (like security) should be considered from an economic perspective as much as from a technological one.

[*back to section 4.1 on Scientific and Technological Foundations*]
[*back to section 6.5 on Service Architectures, Platforms and Infrastructures*]

A.1.3.2 Adoption of sound measurement practices in domains where these are lacking.

In several fields of science, sound measurement practices are already available and used. Metrology offers concepts and mathematical tools to characterize results obtained by experimental measurements; these concepts are already largely used in several fields of science. Recently, increasing interest is being paid to quantitative evaluation of computing systems, based on measurements, with special attention to the evaluation of Quality of Service (QoS) metrics of systems and infrastructures. It should be noted that the attention to this evaluation is usually devoted only to the output of the measurements, whereas little or no attention is paid to properly characterize the measurement systems and express measurement results according to measurement theory. In many domains, measurements are not conducted in less rigorous or ad-hoc ways; concepts of metrology are not yet widely diffused in computer systems experimental evaluation.

One of the areas of computer systems where measurements are especially important is where services are backed up by Service Level Agreements (SLA). When the certain measures are defined in these documents, the methods and processes of the measurements also need to be specified. However, still in these cases certain properties, like measurement uncertainty, confidence level or resolution, are not regarded in the definitions, leading to possible misunderstandings.

Sound measurements practices, as the ones proposed by metrology, currently are not widely adopted in experimental assessment of computing systems. Adopting these practices will allow:

- to obtain more trustable results (e.g. using concepts from metrology like *uncertainty* of a measurement);
- to enhance our capabilities of comparing in a trustable way results obtained in different experimental evaluation campaigns and using different tools (e.g. using concepts from metrology like *compatibility* of results).

[*back to section 4.1 on Scientific and Technological Foundations*]
[*back to section 6.2 on Transportation*]

A.1.4. Actions

A.1.4.1 Selection of appropriate measurement methodologies for given measures and domains and development of measurement tools that are well-based from a metrology point-of-view.

Measurement methods should be characterized according to sound metrological principles. Thus, measurement methods and the results stemming from them can be compared on the basis of standard and objective parameters (note that this does not directly imply that a relation of order among different tools can straightforwardly and universally be derived). The knowledge of the main “metrological” weaknesses of methods can lead to their improvement. For example, evaluating uncertainty and associating a confidence level to the measurement result gives more strength to decisions taken on the basis of these results.

Two different ways can be followed to undertake this action:

- Select a domain and a measure of interest for which a well-known measurement method and a related tool exist. Analyze the way it has been designed and experimentally tested, identifying which metrological principles have (or have not) been followed for its design and testing. Then, if improvements are needed, try to enhance it, if possible, or enhance its experimental validation on the basis of the fundamental metrological concepts that are lacking.

- Select a resilience measure of interest in a domain where measurement methods and tools are lacking. Design a measurement method and implement a tool which is based on solid metrological principles with respect to the measurement data collection, their analysis and the presentation of the results. Conduct a large measurement campaign to characterize the method under various operating conditions.

[*back to section 4.1 on Scientific and Technological Foundations*]

A.1.4.2 Establishing business or financial models to relate system resilience with its business impact.

In this research, we adopt economic models to describe the trade offs made in resilience investment making. We base our work on macro-economic models that provide relations between indicators such as unemployment and inflation, and aim at deriving similar models for key resilience properties, such as security versus productivity loss. Such models will give us deep insights into high-level trade-offs between various system and business properties and metrics. (Loosely based and expanding on existing ongoing work in the UK Technology Strategy Board (TSB) Trust Economics project that aims to develop modelling methodology to help IT consultants and IT staff make security investment decisions that are financially optimal for their company.)

[*back to section 4.1 on Scientific and Technological Foundations*]

A.1.4.3 Exploitation of existing features and ”hooks”.

Existing monitoring features and hooks (e.g., standard OS facilities, logging systems of web servers, auditing in database management systems, standard test ports such as JTAG and Nexus) can be used to build tools for helping developers to assess, measure and benchmark resilience. A focused work package in a research project, or a small research project, could tackle this problem by developing such tools while ensuring that the resulting techniques are generic and widely applicable.

The first step would be to define a process for assessment, measurement and benchmarking during the different phases of development. Such a process would specify the resilience-related activities that should take place during development, possibly taking the V-model as the basis. One should then identify the minimum set of features and hooks required by a tool supporting this AMB process and define a standard API for monitoring and supervising the execution of target systems (possibly some parts of the API would have to be domain-specific). The API should provide the means to collect logical and temporal data on target systems, according to the attributes/properties that are relevant for any given domain. Furthermore, the API should provide mechanisms for triggering actions on target systems, such as the execution of faultloads. The outcome of this project would be an AMB process and a standard API to be used by tool vendors.

[*back to section 4.1 on Scientific and Technological Foundations*]

A.1.4.4 Development of methods/tools to predict the dependability of a design using the observed parameters of its components in previous designs.

It is often costly to run a set of fault injection experiments to assess (with sufficient confidence) a new design without any additional information or analysis. An interesting research avenue is to assess program vulnerabilities and error coverage by combining experimental techniques with analysis methods (e.g., static analysis or symbolic execution). Recent efforts such as symbolic fault injection show some potential benefits that should be further researched. As an example, one can envisage a scenario where field data regarding a set of components is taken as input to a static analysis tool that examines a new design (using similar components with differing parameters) and estimates, for instance, error coverage.

Another important aspect of this area is the way complexity is handled. Usually systems are composed of numerous components with a great number of connections and relations. Simple analysis techniques on the source code or on application models cannot be applied as the space (memory) and time required to compute a huge model is usually orders of magnitude larger than what is allowed. A promising approach for this problem is the application of abstractions that narrow the set of relevant search paths, thus reducing the resource requirements of the analysis. Unfortunately, no ultimate abstraction exists that can help in all cases. Therefore, research should be focused on finding appropriate methods to reveal the relations between system attributes that allow for simplification.

[*back to section 4.1 on Scientific and Technological Foundations*]

A.1.4.5 Development of methods/tools to predict resilience characteristics under future usage profiles by extrapolating assessments derived from measurements of operation, attacks and failures obtained under a different profile, with rigorous estimation of the error thus introduced.

Taking, for instance, a software component to be re-used in a new environment or system, it is easy to show that its reliability may change radically compared to what was observed in the past. There are ways to predict the new reliability value given certain assumptions, e.g., that the reliability values *conditional on* specific classes of demands on the software are invariants, and that the probabilities of demands coming from each class is known for both the old and the new environment. At least the first assumption is generally false, because some demands in a class cause failures, some do not, and the probabilities of these two subclasses of demands will generally vary between environments. Yet, predictions based on these assumptions might prove to be sufficiently accurate in certain classes of situations (using "theoretically inadequate" models that yet perform "well enough" for practical use is a common situation in engineering). Thus, a research project in this area needs to focus on a specific category of predictions (for a specific type of system and/or application area), identify some such predictive model (with its assumptions) and compare its predictions, *a posteriori*, with the measurements on the actual behaviour that the prediction was about. Thus it will empirically demonstrate whether in the chosen category of predictions either the assumptions generally hold true or – a more likely result – the model using them generally provides useful predictions despite being provably false. A special case is that of identifying additional correction factors to describe differences between the old and the new environment for the new prediction, which can be incorporated in a refined model that in turn can prove to be accurate enough.

Possible useful results could be: identifying such useful models, with rough estimates of probability of correctness over the category of predictions examined, and/or of the probability distribution of errors; or demonstrating that some model is *not* generally reliable.

[*back to section 4.1 on Scientific and Technological Foundations*]

A.2. Measurement and Assessment

A.2.1. Needs

A.2.1.1 Practical, trustworthy and widely applicable tools for measurement and assessment in large-scale dynamic systems, adaptable and evolving infrastructures, and other domains where these are lacking.

Tools for measurement and assessment are needed to support developers, assessors and other stakeholders in making decisions objectively. Developers of large-scale dynamic systems and infrastructures are facing a lack of effective tools to guide their effort. Similarly, assessors and certification authorities require objective information when appraising whether a critical system or infrastructure meets its dependability requirements or not.

Since traditional methods for measurement and assessment existing in mature domains are insufficient to cover a wide range of new systems and infrastructures, it is fundamental to develop tools supporting novel measurement and assessment techniques. Associated to this, there may be a need to develop new techniques, if the large body of existing research proves insufficient. However, the priority is to develop practical tools capable of meeting the growing complexity and dynamicity of computer systems.

[*back to section 4.2 on Measurement and Assessment*]
[*back to section 5 on AMBER Topmost Priorities*]
[*back to section 6.3 on Certification Authorities and Assessors*]
[*back to section 6.6 on Enterprise Security*]

A.2.2. Challenges

A.2.2.1 Providing domain-specific rules to compose/integrate different methods and tools for resilience assessment.

Resilience assessment is a very challenging task in complex distributed systems that are large, dynamic, adaptive and/or mobile. A typical example of such systems can be found in the automotive domain, where the computing power of mobile devices and the quality of communication technologies have developed enough to allow the deployment of complex services including critical ones (like hazard warning between vehicles). In these domains the well-proven traditional methods cannot be applied “as is”. For example, the perceived availability or success probability of these dynamic services cannot be evaluated by simply applying traditional modelling techniques, like Markov chains or Stochastic Petri nets, due to the inherent variability of the system structure, the heterogeneity of the system, and the large number of possible scenarios. These factors together raise challenges concerning the expressive power of the formalisms (for efficient modelling) as well as the complexity that the supporting solution tools can handle.

Integration of various evaluation methods is an encouraging approach to handle this challenge: the assessment problem is decomposed into several sub-problems, and individual formalisms and evaluation techniques are applied to analyze the resulting sub-problems. For example, in a distributed mobile system potential sub-problems could be the evaluation of the dependability of the hardware/software architecture of nodes, the characteristics of the network traffic, and then the success of the activities of a mobile user. In case of these different sub-problems different solution methods (like analytical, simulation-based or experimental methods) can be applied.

In this context the main research challenge is the elaboration of proper rules to divide the problem and then *compose/integrate the results of the different methods and tools* used to solve the sub-problems. One example of a potential composition rule is the so-called "solution feedback": solution of a specific sub-problem is used as input to solve another sub-problem possibly using a different technique (e.g., the network delay as parameter in a user level model is obtained utilizing the simulation based solution of the network level traffic model).

The elaboration of proper and re-usable composition rules is difficult due to the great variety of potential models, solution techniques and properties that should be computed. Unfortunately, the composition of non-functional properties is not straightforward: several system level properties are emergent, i.e., the properties at the system level do not have the same meaning for subsystems, and vice versa. This is typically the case in distributed systems that are divided into subsystems by functionality. Accordingly, more intricate domain-specific composition rules have to be elaborated.

The success of this approach would result in an efficient re-use of existing methods and tools and at the same time evaluation of "bigger", more complex systems would be possible. A notable example of such a composition/integration approach is the contribution of the HIDENETS project (http://www.hidenets.aau.dk/) for the evaluation of large critical mobile applications.

[*back to section 4.2 on* *Measurement and Assessment*]
[*back to section 6.3 on* *Certification Authorities and Assessors*]

A.2.2.2 Designing and developing test beds for emerging applications, e.g., cloud computing and collaborative services.

Emerging applications such as cloud computing and collaborative services are characterized by a large distribution of computing resources. It is well known that distributed systems are more difficult to design and verify than centralized ones, due to parallelism among processes, random and unbounded communication delays, partial failures (failures of components), lack of strict clock synchronization among nodes, etc. These factors are especially challenging when considering the large scale and heterogeneity of platforms at the heart of emerging applications.

An example of a problem that may be addressed in a research project is the design and implementation of a debugging environment for large-scale distributed applications. One of the main research challenges in building such a test bed is that it is often impossible to replay the exact sequence of events leading to an erroneous state. This calls for the development of practical and efficient test beds for large-scale distributed systems, supporting debugging, execution of test cases, monitoring, etc.

[*back to section 4.2 on* *Measurement and Assessment*]
[*back to section 6.4 on* *Future Internet technological platforms*]

A.2.2.3 Developing measurement and assessment tools for systems/measures for which these tools are currently lacking.

The evaluation of computer systems can be performed using different approaches, generally classified into three categories: analytic, simulative and experimental. Each category shows different peculiarities, which determine the suitability of the method for the analysis of a specific system aspect.

Interest in quantitative evaluation based on measurements is increasing. Experts in the field of resilience recognized experimental measurements as an attractive option for assessing an existing system or prototype, because they are highly accurate information of the system executing in its real usage environment directly from its observation. Great interest is currently devoted also to the use of data taken from real system execution (e.g. field data and experimental measurement) as input data usable with other quantitative evaluation methods (e.g. modelling and simulation), following an *holistic approach* for the evaluation

To develop measurement and assessment tools for computer systems and to characterize them using metrological concept is a challenge of uttermost importance. In fact, no standard measurements and assessment tools are available in many domains (e.g. adaptable/evolving systems and infrastructures). Moreover experimental campaigns are currently designed ad-hoc each time they are needed, rather than applying proven and tested patterns as in mature engineering measurement methods used in other contexts (e.g. electronic, mechanical, …). In the experimental evaluation campaigns, attention is usually paid only to the final results of the evaluation, without attention to the way in which measurements are performed. Researches including computer science and metrology topics are currently needed in order to define *standards* and *verified tools* for experimental assessment of computer systems taking into account metrology concepts and best-practices. This kind of research will allow to obtain more trustable information about the quality of service of computing system as it is perceived by its users and it will allow to compare in a easier and more trustable way computing systems using experimental measurements through the definition of appropriate benchmarking techniques.

[*back to section 4.2 on Measurement and Assessment*]

A.2.3. Objectives

A.2.3.1 Development of efficient on-line mechanisms to monitor the environment conditions of the system and to dynamically evaluate and assess its resilience.

On-line monitoring is the enabling technology to allow AMB in practice, especially in large-scale, dynamic systems that are usually too complex for a complete resilience assessment before deployment. Dynamic online assessment allows evaluating the system status and identifying states in which notification or planned corrective actions are needed. Environmental or internal changes may cause degraded protection against failures or attacks even in systems that were originally properly configured. Accordingly, the ability to identify and trigger corrective actions is of utmost importance. It is a tangible goal in several domains, like Embedded Systems (see Section 6.1), Transportation (see Section 6.2) and Future Internet technological platforms (see Section 6.4).

Specific context: Monitor and Assess a Trustworthy Internet of Services
The current situation and the expected evolution of service infrastructures that are going to form the Future Internet show a clear trend towards extreme complexity, high dynamicity and a very large scale of components/elements and services.
This evolution has consequences that must be understood and seconded by any approach to improve the resilience of the infrastructure and the trust that can be put on its proper behaviour. Among these is the observation that the need to quickly react/adjust to changes, high dynamicity and the large scale of Future Internet require full exploitation of new approaches such as adaptive/dynamic online monitoring.
In this context, one relevant goal is to develop flexible and dynamic mechanisms and risk-based methodologies capable to attentively monitor such infrastructures and to dynamically adapt them at varying operating and environmental conditions and to varying requirements, aiming to provide and maintain an acceptable level of service so to respond to threats and vulnerabilities. To meet the objective it should be also developed a flexible monitoring system that can be itself adapted according to changes in the infrastructure that are observed, or predicted to occur. One of the advantages for such monitoring system is that it could be tuned in anticipation of potential anomalies in order to generate detailed data about the potential anomaly that may occur.

[*back to section 4.2 on* *Measurement and Assessment*]
[*back to section 5 on* *AMBER Topmost Priorities*]
[*back to section 6.1 on* *Embedded Systems*]
[*back to section 6.2 on* *Transportation*]
[*back to section 6.4 on* *Future Internet technological platforms*]
[*back to section 6.5 on* *Service Architectures, Platforms and Infrastructures*]

A.2.3.2 Development of business and economic models for risk analysis.

Existing business and economic models do typically not include a whole-system view. For example, they do not include system dynamics and/or human behavioural models. A research project could address this issue by adopting modelling techniques that are well-known in resilience engineering (in particular, stochastic and Markov processes) and use them to model key human and technology dynamics that influence resilience technology acceptance and use. Reward measures can be used to express business and economic risks. Several of the main challenges in this research are to make model components reusable, especially in relation to the human factors, and to obtain useful and realistic parameter to plug into the models. (This research item is loosely based and expanding on modelling research in Newcastle and UIUC on trust economics - see www.trust-economics.org.)

[*back to section 4.2 on* *Measurement and Assessment*]
[*back to section 6.5 on* *Service Architectures, Platforms and Infrastructures*]

A.2.4. Actions

A.2.4.1 Development of methods and tools for resilience assessment, able to: i) Avoid the generation of large/complex models; ii) Tolerate the generation and solution of large/complex models (i.e., optimizing the generation and processing of the models).

Specific context: Large scale Complex Critical Infrastructures

1. Motivation for the action: Large scale Complex Critical Infrastructures (LCCIs) are complex and highly interdependent systems, networks and assets that provide essential services in our daily life. They span a number of key sectors, including energy, finance, authorities, hazardous materials, telecommunications, information technology, supply services and many others.

With our increasing dependence upon such critical infrastructures, an unavoidable expansion in complexity is observed since these sectors are continuously called to provide new services and products to a growing population. Moreover, while in the past these infrastructures were used to provide services mostly in isolation, with very limited interconnections with each other, so they could only be impaired locally, nowadays several infrastructures cooperate in the provision of services. Despite such complexity, it is paramount that they be reliable and resilient to continue providing their essential services. Hence, there is the need to evaluate them to assess their degree of resilience/trustworthiness. In this perspective, model-based evaluation is commonly used to support the analysis of dependable computer systems, which is required to identify vulnerabilities, interdependencies and interoperabilities between systems, to understand what specific assets of the addressed critical infrastructure are utmost critical and need to be protected the most.

2. Research steps proposed: The main objective of this action is then to define stochastic modelling approaches for the quantitative evaluation of dependability-related attributes for LCCIs. A major difficulty in the project lies in the complexity of the modelled infrastructures in terms of high number of components, multiplicity of interactions and types of interdependencies involved, which results in the well-known state-space explosion problem since the dimension of the state space grows exponentially with the number of parts.

To address this problem, a research direction is to adopt both hierarchical composition and decomposition/aggregation approaches, exploiting their synergies and complementarities.

- Structured model composition approaches build the system model from the composition of submodels describing system components and their interactions. Important issues are how to abstract all the relevant information of one level to the upper one and how to compose the derived abstract models.
- Decomposition/aggregation approaches attack complexity decoupling the overall model in simpler and more tractable submodels, and the measures obtained from solution of the submodels are then aggregated to compute those concerning the overall model. Instead of applying the decomposition approaches at the model-level only, by decoupling parts of a pre-existing model, the decomposition approach should be first applied at the conceptual-level, thus identifying a set of interacting conceptual submodels in accordance with specific pre-defined criteria. The conceptual-level decomposition allows defining conceptual submodels that can be implemented exploiting a multi-formalism/multi-solution approach. In fact, each conceptual submodel can be implemented using the formalism that better represents its behavior, and can be solved using an appropriate, efficient solution technique. Therefore, the evaluation framework could include analytic stochastic models (such as Markov models or high-level specification formalisms, based for example on Stochastic Petri Nets and extensions), simulations as well as experimentations.

3. Potential effects: The overall objective throughout the whole project will be the advancement of knowledge in the field of model-based dependability assessment of LCCIs, which will allow understanding and analyzing vulnerabilities, interdependencies and interoperabilities between different critical infrastructures and between the several heterogeneous subsystems each infrastructure is composed of. Following such results, steps can be taken to mitigate vulnerabilities revealed in critical assets.

[*back to section 4.2 on Measurement and Assessment*]

A.2.4.2 Elaboration of methods for (model-based) design and configuration of monitoring systems.

Motivation: Service and system management operations rely on on-line monitoring and data gathering as well as the analysis of the collected (on-line and historical) data. These operations aim at discovering (the effects of) various threats, and supporting the corresponding root cause and impact analysis. There is a mature market of generic software frameworks and specialized solutions that support data collection and analysis. However, a significant portion of the deployment effort associated with the application of them is the *configuration* for the given environment. The operator has to define what and how shall be monitored in the infrastructure and at the service layers, how events shall be aggregated, how errors propagate, and what is the proper action to avoid service disruption.

Research goals: In system development the integration of *model driven design and assessment approaches* (i.e., the systematic use of models as primary artifacts throughout the engineering lifecycle) has already proven that it can significantly enhance the confidence in non-functional properties (like availability, reliability, performance) of systems and services. These advances applied in the field of IT infrastructure management promise the synthesis of trustworthy and fine-granular supervisory configurations. The research steps involved in this action could include the development of (i) a model based *design methodology for automated configuration of monitoring systems* (i.e., what to measure), (ii) model-based *data processing techniques* to process data collected by the monitoring system and to extract information that is relevant at the level of system management actions, (iii) *methods to update initial analysis models* on the basis of the processed measurement data, and (iv) methods to *validate the effects of system management actions*. These actions together make up a model-based framework that can uniformise and automate the administrative actions throughout the whole system.

Results: The model-based approach in the design of monitoring systems, data processing and prediction of the effects of system management actions offer optimized operation of the IT infrastructure with respect to the achieved QoS. The previously manual tasks like the configuration of the monitoring subsystem and the evaluation of the actions can be performed in an automated and more trustworthy way.

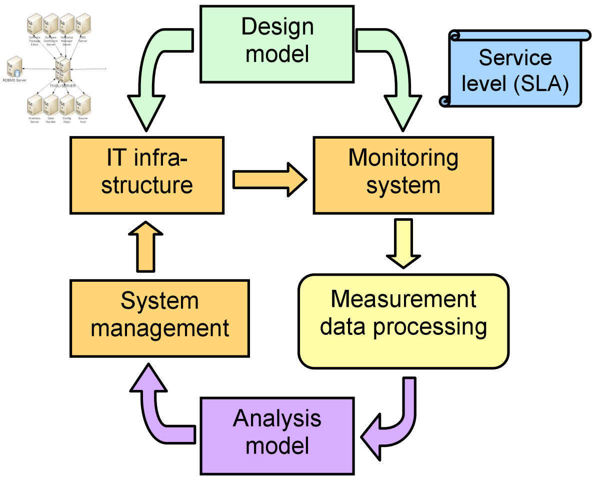

Figure 5: Model-based synthesis and operation of IT infrastructure management

[*back to section 4.2 on Measurement and Assessment*]

A.2.4.3 Identification and (if needed) development of domain- and process-specific methods and corresponding tools that can be integrated into automated tool chains.

This action is exemplified by discussing a specific objective, the development of automated tools for robustness testing (as a resilience assessment approach).

Motivation: Robustness, as an attribute of dependability, characterizes the correct operation of systems in the presence of exceptional inputs and stressful environmental conditions. Robustness faults activated by these inputs and conditions result in incorrect operation (e.g., crash, deadlock, malfunction), thus estimating the robustness of services is a prerequisite to improve resilience and reduce the risk of failure. The classical approach to estimate robustness is *robustness testing*, which is typically an ad-hoc, manual and time-consuming task.

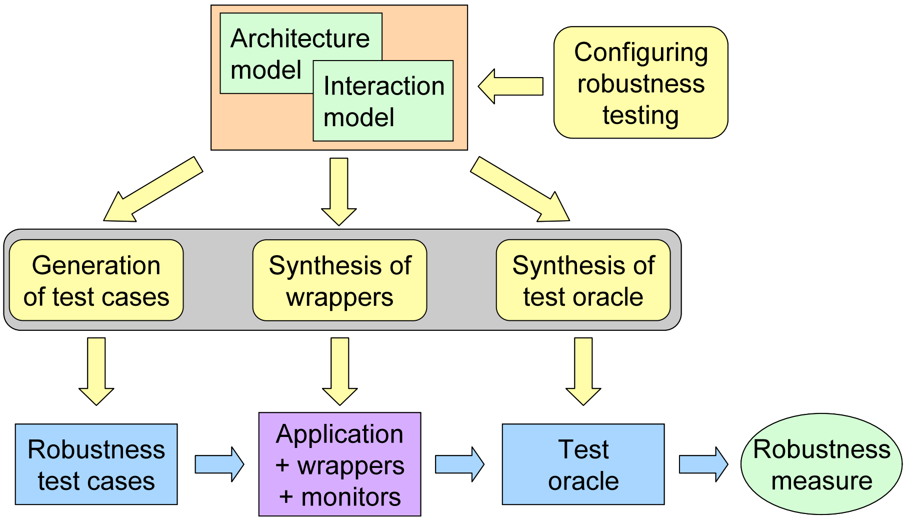

Figure 6: An example tool chain

Research goals: The opportunities of model-driven engineering can be utilized in the phase of generating robustness test cases and configuring the robustness testing process. Tools can be developed to generate type-specific robustness test cases for individual components/services on the basis of the model of the interfaces, and to cover interaction faults between services on the basis of interaction models that describe the typical scenarios. To go further, test wrappers and monitors can be automatically synthesized on the basis of the architecture model and the abstract format of the test cases. The tools developed for test generation, wrapper and oracle synthesis, configuration and evaluation can form an efficient tool-chain.

Results: The tool-supported techniques allow for the reduction of the efforts required for the assessment process, and at the same time offer more thorough testing. The latter is due to the fact that the test cases are derived and executed systematically, therefore, reducing the probability of undetected robustness faults. Better techniques allow development companies to provide more resilient components and services, and allow system integrators to qualify external services.

[*back to section 4.2 on Measurement and Assessment*]

A.2.4.4 Execution of experimental campaigns in real systems to characterize and demonstrate experimental measurement features and properties.

1. Motivation for the action: current and future computing systems are complex systems. Having access to data (e.g. field data) obtained by the real execution of computing systems is of uttermost importance since it makes able i) to use this data as input for other kinds of analysis (simulative and analytical ones); ii) to crosscheck results obtained using other kinds of analysis. The execution of experimental campaigns will also allow to characterize, using concepts from metrology, the tools used for the instrumentation of the system; this characterization will allow to obtain trustable results and to compare results obtained in different experimental campaigns (e.g. using metrological concepts like compatibility and uncertainty).

2. Research steps proposed: To perform this action it is first of all necessary to choose one or more domains in which results of execution of experimental campaigns can be of interest. To make an example, we can choose large-scale Internet based systems. The choice of the domain impacts the choice of tools and techniques usable to instrument the system and to collect experimental measurements (e.g. operating systems-based logging, specific mechanisms of data collections and so on). The tools used to collect experimental measurements can be characterized using metrological concepts, e.g. making available the collection of detailed information about measurements (e.g. uncertainty of measurements). The results of experimental campaigns can be used to characterize the tools/methods employed; this characterization should be based on metrological concepts.

3. Potential effects: Potential effects of this action are the following ones: i) Experimental results obtained from experimental campaigns are useful to characterize the system under test and as input data for other kinds of analysis (cross-fertilization of the results); ii) Experimental results can be used to characterize the tools and methods used from a metrological viewpoint, (using concepts like *uncertainty* and *compatibility*).

[*back to section 4.2 on Measurement and Assessment*]

A.2.4.5 Elaboration of techniques for assessing the combined impact of accidental faults and malicious threats.

It is currently possible to create probabilistic models (e.g., Markov or Petri-net based models) to predict, for instance, the availability of a system subject to both attacks and component failures. As usual, limits to the usefulness of the models come from the difficulty in estimating parameters and in deciding whether the model assumptions are accurate enough for practical purposes. In the case of combining security and non-security risks, additional difficulties arise from the fact that forms of uncertainty in the two areas may be different (e.g., uncertainties about the parameters of physical components *vs.* about the adversaries' intentions and strategies), the means that experts in the two areas will use to describe these uncertainties also differ, and security experts whose input is necessary may not be familiar with the probabilistic approach as used in dependability engineering. So, risks include over-simplistic assumptions in the model (e.g., stochastic processes of attack that misrepresent real strategies) as well as in the analysis of uncertainty about its predictions. Research projects in this area would be typically case studies in which modelling is applied to specific systems and progress is made with respect to any one of these difficulties.

[*back to section 4.2 on Measurement and Assessment*]

A.2.4.6 Assessment of the impact of the approximations introduced in modelling on the resilience-related predictions sought

Typical research projects in this area will select a predictive model, or family of models, that an industrial sector could feasibly adopt (that is, one that requires skills that are available and is affordable to apply), assess the size and statistical distribution of errors it introduces in predictions, and provide rules for estimating the error (generally, the probability distribution of the error) in a specific prediction. The methods for this assessment could be any combination of modelling (e.g., if the model M being assessed can be proved to be pessimistic, the errors it introduces could be bounded by its difference with a model M' that is known to be optimistic) and empirical measurement (statistics of the true values of the measures predicted, or of the magnitude of error terms identified by modelling).

[*back to section 4.2 on Measurement and Assessment*]
[*back to section 6.3 on Certification Authorities and Assessors*]

A.3. Benchmarking

A.3.1. Needs

A.3.1.1 Benchmark frameworks (components and tools) able to be reused to create benchmarks in different domains.

Defining a resilience benchmark is a difficult process. The division of the application spectrum into well-defined domains is necessary to cope with the huge diversity of systems and applications and to make it possible to make choices on the definition of benchmark components as most of the components are very dependent on the benchmark domain. However, it seems obvious that there are some components and tools that may be reused across several (if not all) domains (e.g., experimental control tools, data analysis tools, software fault injection tools, measurement probes). This reuse will facilitate the definition of new benchmarks. Additionally, it will allow using components and tools for which key properties (e.g., representativeness, repeatability, portability) have already been validated. In summary, in the same way performance benchmarks follow a typical benchmarking framework that includes three key components (measures, workload, and rules), the goal is to devise frameworks for resilience benchmarking that can be applied across different domains.

[*back to section 4.3 on Benchmarking*]

A.3.2. Challenges

A.3.2.1 Dealing with human factors in the definition and execution of resilience benchmarks.

In many situations, human behaviour is a determinant factor in the dependability of ICT: interactive software exhibits different failure modes and frequency depending on how it is used (even two users producing identical documents on a word processor may experience different failure patterns depending on the features and commands they use), complex hardware-software systems require maintenance, configuration and human-directed error recovery; security problems are created by attackers and often involve categories of active human victims and active "defenders". Humans affect resilience in multiple ways: they create problems by their mistakes or other actions that violate design assumptions; they detect problems created by hardware/software failures, mitigate them or recover from them, more or less effectively (even by interfering, for better or worse, with the automated defensive mechanisms built into the machines they deal with). In this variety of roles and behaviours, we can expect that different people will act differently in the same situation, and the same person may have a variety of reactions in the same (or apparently the same) situation. All this creates difficulties with the *repeatability* and *representativeness* of a benchmark that includes humans as part of the workload and/or of the faultload, and/or of the system to be benchmarked. One may improve repeatability by a strict selection and training protocol for users involved in benchmarking, but this adds cost and creates problems with achieving representative benchmarks. Recorded traces or simulations of human behaviour may be used, but then the benchmark may become less *portable* between different ICT systems (of the same type: e.g., different mobile telephones, automobiles, word processors, aircraft cockpits): they may have different interfaces that require or cause real people (but not the traces or simulations) to behave differently. We could define multiple benchmarks for stereotypical categories of users (defined by nationality, level of expertise and so on), but these may become too many to be manageable, before they are sufficiently representative of the variety of real behaviours; or focus a benchmark on an "average" or on a "typical" user, with the risk of being neither representative nor informative (we expect that slightly unusual user behaviours may create the most challenges for designers), or vice versa on including a few “extreme” types. Another concern is that the population of users evolves. They learn to use some new technology, or new “tricks” on existing technology, or they “unlearn” their skills with other technologies. They trust some technologies more, or less, over time, which affects their ability to use them and tolerate their failures. Keeping the human behaviour component of benchmarks relatively static brings the risk of the benchmarks becoming, over time, unrepresentative of real use; but rapid evolution of benchmarks would be expensive and create confusion about what the current standard is. Likewise, evolving user interfaces may make it difficult to assess newer and older technologies, even if meant for the same function, via a common benchmark. Last, a special kind of *intrusiveness* of measurement is an issue: people behave differently from their norm when they know they are under test, and reducing this effect may constrain how well a benchmark achieves other goals.

[*back to section 4.3 on Benchmarking*]

A.3.2.2 Coping with highly complex, adaptable and evolving benchmark targets (components, systems and services).

Future generations of systems will include features to provide automatic adaptation to the environment and will have the capability to evolve depending on the environments and user needs. This will result in highly complex and extremely dynamic systems. Defining resilience benchmarks able to characterize complex systems and able to adapt to different states of the system and to its evolvability capabilities is a major challenge. In practice, benchmarking complex systems may require complex benchmarks and benchmarking dynamic systems entail using dynamic benchmarks. The goal is to include in resilience benchmarks measures and other components able to deal with highly complex, adaptable and evolving targets. This will require dynamic benchmarking components that may need constant update and refactoring.

[*back to section 4.3 on Benchmarking*]
[*back to section 6.4 on Future Internet technological platforms*]

A.3.3. Objectives

A.3.3.1 Validated reference faultloads (i.e., sets of faults that are representative of specific domains) and corresponding injection tools (that allow easy implementation and portability of the faultloads) to be used in the development of resilience benchmarks.

Reference faultloads are fundamental to develop resilience benchmarks that can be agreed upon by vendors and buyers of computer systems. One goal of resilience benchmarking can be to rank competing systems according to some measures that are considered relevant (e.g., service availability). When such measures are estimated through benchmarking, the result should be as close as possible to measurements collected in real operation. The challenge, which could be addressed by a research project, is therefore in identifying faultloads that are representative of real faults, in the sense of supporting benchmarking that produces correct rankings.

Such a project could, for example, investigate the ability of diverse faultloads to predict the order in which real systems (or advanced prototypes) are ranked. To this end, one could define a comprehensive list of fault models, including for instance hardware faults, software faults and operator faults described in the literature. Using fault injection, a set of target systems could be ranked with respect to relevant measures of resilience – one ranking for each fault model. The same measures should then be measured having the same systems in operation, affected by real faults. The correct ranking (i.e., the one observed in operation)

should be compared to the rankings obtained using the diverse fault models, in order to determine which ones have the best predictive ability; and a final faultload should then be composed from these.
The result of this project would be a faultload that could be used for ranking systems with respect to a given set of resilience measures. Such a faultload would most likely be specific to a certain domain, and in principle its suitability would depend on the work load of the system. The research challenge would be in identifying faultloads that are nonetheless appropriate (produce correct rankings) for a practically useful range of workloads, and of systems differing from those targeted by the project. Complete success is unlikely – , but practical fault loads that have validated as viable for well defined ranges of applications would be a valuable outcome, useful both in benchmarking as defined above and for use by system developers, for guiding their development effort.

[*back to section 4.3 on Benchmarking*]
[*back to section 5 on AMBER Topmost Priorities*]
[*back to section 6.2 on Transportation*]

A.3.3.2 Agreed resilience benchmark measures.

Motivation: Measures are the key factor in designing of a resilience benchmark. In fact, the definition of all the other components of a benchmark depends on the measures to be obtained. Devising new measures is easy: what is difficult is to devise relevant measures and measures that are in fact useful for making pertinent comparisons. To be useful, measures should be easy to apply and understand by the benchmark users. As the users tend to ignore benchmarks that are difficult to implement or execute, the measures must be simple to obtain. In addition, the measures must allow an easy comparison of systems or components within the benchmark domain.

Research goals: The goal is to conduct a set of actions towards the identification of relevant resilience benchmarking measures for different domains. This implies identifying measures for different domains, validating those measures, and conducting dissemination actions with the goal of making those measures universally accepted foa category of systems and applications. Additionally, it is necessary to define guidelines on how to identify and validate new resilience-related measures and on how to foster agreement on those measures.

Results: The result is a set of agreed resilience benchmarking measures for different domains, including both quantitative and qualitative measures. A key result regarding qualitative benchmarking is the definition of relevant features/properties to be checked/verified in systems in the benchmarking domains where the verification of properties through analytical means is prevalent. Another result is an approach towards the identification and validation of resilience measures, including how to create agreement on those measures.

[*back to section 4.3 on Benchmarking*]

A.3.4. Actions

A.3.4.1 Development of tools to inject reference faultloads in different classes of benchmark target systems.

This example focuses on the development of a model based tool environment for configuring and executing software fault injection experiments.
Motivation: The resilience of IT applications is typically measured by controlled fault injection experiments that are able to reveal the behaviour of the application in case of faults. This way error handling and fault tolerance solutions can be assessed. The configuration of

fault-injection experiments determines what kind of faults shall be injected; where to inject faults; when the injected faults should be activated; what is the workload of the system during the experiments. The classic fault injection tools are mostly configured by using scripts or parameter files. The preparation of these tool-specific scripts and the parameterization of the injector are low-level manual processes. The efforts required to implement these tests can be significantly reduced by using the model of the application and a set of customizable *faultload patterns* for various faults like hardware faults, software faults or human faults. Model-based configuration of fault injection experiments promises that the tester does not have to learn a specific language for this purpose. She/he can configure the faults and the corresponding monitoring facilities in a friendly graphical framework (that allows the configuration through easy-to-use visual objects and graphical facilities), thus she/he can focus on the important questions mentioned above (location of faults to be injected, activation rate etc.). Moreover, model-based synthesis tools can be developed for the automatic implementation of the modifications required for instantiation, injection and activation of faults and observers.

Research goals: The research steps for this action include: (i) the development of *modelling language extensions* to configure the fault injector tools and the monitoring systems; (ii) the design of a *fault-pattern library* to provide a repository of customizable faultloads; (iii) the elaboration of *synthesis methods* that support the automatic injection of the configured fault patterns into the target systems. In case of legacy or off-the-shelf components when design model is not available, reverse engineering solutions can be used to construct the architecture model of the application that allows the configuration of the experiments.

Results: Using the tool framework, fault injection based experimental evaluation can be configured on the basis of the design model, this way fitting to the other model based verification activities like analytical evaluation and model based testing. The experiments are repeatable and the results of the same monitoring configurations are comparable.

[*back to section 4.3 on Benchmarking*]

A.3.4.2 Definition of standard interfaces for the integration of available components and tools.

In this research, we identify the critical tools and components needed for reusable benchmarking and provide a precise and open specification of its interfaces. The first problem we try to address is an agreed upon identification of the key components, such that these components allow standardisation of its interfaces and ensuing composition of the tools and components.

[*back to section 4.3 on Benchmarking*]

A.3.4.3 Execution of benchmarking campaigns in real systems to characterize and demonstrate benchmark features and properties.

1. Motivation for the action: using methods and tools for experimental evaluation of real systems makes possible to define appropriate benchmarking techniques and methods. Benchmark based on well-defined (from a metrological viewpoint) experimental evaluation methods and tools offers more trustable results. The metrological characterization of experimental evaluation tools and benchmarking techniques is important since it will make possible to compare in an easy and trustable way results obtained using different methods.

2. Research steps proposed: first step is to choose one or more domains in which results of execution of benchmarking campaigns can be of interest. The choice of the domain impacts the choice of techniques usable to collect experimental results and to inject specific workload and faultload in the system under test. Secondly, results must be obtained from execution on real systems. The results of benchmarking campaigns can later be used to characterize used tools and methods; this characterization should be based on metrological concepts.

3. Potential effects: Potential effects of this action are the following ones: i) Benchmarking results obtained from campaigns on real systems are directly usable since they are useful to characterize the systems under test; ii) Benchmarking results can be used to *validate* the tools and methods used, e.g. through characterization of tools/methods from metrology like uncertainty and compatibility.

[*back to section 4.3 on Benchmarking*]

A.3.4.4 Finding whether representative types of faults exist through field studies and analysis.

In dependability benchmarking, fault injection, or similar pursuits, defining the "load" (set or probability distribution) of the faults (in the broad sense: permanent damage, internal bugs, attacks or misuse) to be artificially caused (or simulated by causing internal errors like those that the faults would cause) is possibly the most difficult step. It would be ideal to find a load that, when applied experimentally in order to estimate a chosen measure of resilience, *M*, gives a measured value that is a reasonably good estimate for the value that *M* will have in any future situation. But in reality the system will generally be more robust against certain faults than others, so that each fault load implies a different value of *M*.

A desirable situation is then one in which classes of faults can be identified such that for any specific class *i* the system being assessed has a characteristic value, M_i, of the measure. In the general case in which each fault in class *i* has different effects, the definition of each class is complemented with a probability distribution of faults within the class, and this distribution proves to be realistic. So, repeating the measurement yields estimates of M_i that are close together; extending the period of measurement to include more faults from this class only causes the estimates to converge to M_i; the value of M_i in a future real operational situation is predictable. It is then possible to trust that the value of *M* in a new environment will be an appropriate weighted sum (the weights being the probabilities of the classes of faults; the classification chosen needs to be such that ranges of these probabilities can be estimated with some confidence) of the M_i estimates obtained in the lab. In an even better scenario, the choice of fault classes seems to apply to a whole category of systems.

At the opposite end of the spectrum of possible scenarios, the extreme undesirable situation is one of total unpredictability: each new fault observed changes the estimated value of *M*; the series does not converge, no matter how we try to classify the faults in apparently similar classes. Some conjecture that this may be the case for operator errors and for malicious attacks in certain situations. The same might be true for design faults, at least "high level" ones (e.g., due to a requirement error). The only way to answer these basic doubts is by experiment. Research in this area needs to proceed by case studies, each focused initially on a system and then - if successful for that single system –on a category of systems. Obviously, lessons learned about mobile telephone handsets would not necessarily translate to aircraft flight instrumentation or to web server farms. In the process, the researchers would develop practical methods of fault injection and measurement that work for the systems and measures addressed. The purpose of each project would be to discover a way of classifying faults into classes that produces "reasonably constant" values of M_i and *M* in a combination of laboratory and real-use situations for one system, and then for a category of them; and to document the magnitude of the variations in these values (that is, the accuracy to be expected when using the method of weighted sums for assessing *M*). It is likely that very few projects of this type would be funded during any generation of technology, and they would focus on types of systems where dependability has great value, but the knowledge gained would be useful to much broader industrial sectors: if a research project reported that, for instance, it demonstrated an appropriate classification for attacks on a category of consumer mobile devices, it would lend some confidence in benchmarks (that used the same classification) for other mobile consumer devices; if it reported that it could not find such a classification (i.e., all attempted classifications produced wide variations in the measures of each M_i parameter), others would rightly put less trust in methods that implicitly assume M to be predictable.

[*back to section 4.3 on* Benchmarking]
[*back to section 6.6 on* Enterprise Security]

A.3.4.5 Definition and validation of reference faultloads (considering the required benchmark properties) for different benchmarking domains and classes of targets.

Motivation: A faultload represents a set of faults and stressful conditions that emulates real faults experienced by systems in the field. The specification of the faultload is clearly the most complex problem in the definition of a resilience benchmark. In fact, among the main components needed to define a benchmark, the faultload is clearly the most obscure one due to the complex nature of faults. In addition, the faultload specification is complex as it is very difficult to guarantee that the faultload fulfil key properties like representativeness, portability, scalability, etc.

Research goals: The goal of this action is to define and validate reference faultloads, taking into consideration different domains and different classes of target systems. The idea is to conduct field studies on the identification and characterization of the faults that typically affect systems in the field. Different studies should, obviously, be conducted for different domains and classes of targets. As it is virtually impossible to conduct studies for all existing targets and domains, it is also a goal of this action to propose a standard approach for the definition and validation of new faultloads.

Results: The result of the action is a set of reference faultloads that can be used for the definition of new dependability benchmarks (each faultload will include the most relevant faults considering the systems in a given domain). A standard approach for the definition of faultloads for resilience benchmarking will also be a result of the action.

[*back to section 4.3 on* Benchmarking]

A.3.4.6 Agreement on the benchmarking processes in relevant industry and user community.

Motivation: The key aspect that distinguishes benchmarking from existing evaluation and validation techniques is that a benchmark fundamentally represents an agreement that is accepted by vendors, buyers, users and regulatory authorities with an interest in a certain application/technology domain. This technical agreement is in fact the key that turns a benchmark into a standard. In other words, a benchmark is something that the user community and the computer industry accept as representative enough of a given application domain to be deemed useful and to be generally used as a (standard) way of measuring specific features of a computer system and, consequently, a way to compare different systems.

Research goals: The goal of this action is to seek ways of establishing agreement in relevant vendor and user communities regarding the resilience benchmarking process. In the same way performance benchmarking boosted due to the support of major industry players, resilience benchmarking success will depend on the support of industry and users. This will involve the identification of the players that may have huge impact on fostering resilience benchmarking, recognition of the goals and needs and industry and users, and adaptation of the benchmarking process to those needs and goals.

Results: The result of the action is the definition of how the benchmarking process should be in order to get the agreement for industry and user community. Another key result is a set of guidelines describing how resilience benchmarks should be defined, validated and marketed to achieve agreement.

[*back to section 4.3 on Benchmarking*]

A.4. Education, Training, Standardization and Take Up

A.4.1. Needs

A.4.1.1 Availability of assessment and benchmarking results to the parties (vendors, customers, authorities, researchers, consumers) that can benefit from them, in sectors where benchmarks are mature.

There is a need for approaches to make benchmarks accessible, to vendors, academics and consumers. In particular, one can provide a service, accessible over the web, which publishes data on a regular basis about the resilience of the main software and hardware used on the home and in office automation. This software includes operating systems, handhelds such as iPhones and laptop hardware characteristics. The desire is to convey results in ways meaningful to the targeted customer of the information, and also be reasonably representative for the actual resilience of the considered systems.

[*back to section 4.4 on Education, Training, Standardization and Take Up*]

A.4.1.2 Standards for benchmarking (and for the assessment methods behind them) of resilience aspects in different domains.

In many domains benchmarking and assessment standards are already available, e.g. for aspects of resilience, safety, etc., of railway, automotive, air traffic, control systems, etc. .There are, however, several domains in which no standard of this kind is available. In particular, benchmarking and assessment standards are currently unavailable in many non-safety critical contexts. As an example let us consider Future Internet. In Future Internet applications, security and privacy aspects will need to be considered; for these aspects currently no standards exist for benchmarking and for the assessment methods behind them. Research in this direction will need to be performed. Having access to standard benchmark for security and trust (and standard assessment methods) will allow to evaluate in a precise way and to compare results obtained using different systems from the point of view of "quality of service" (in a more general meaning) as perceived by the users. The definition of the standard assessment methods should be based also on the characterization of used tools and methods from a metrological viewpoint.

[*back to section 4.4 on Education, Training, Standardization and Take Up*]

A.4.1.3 Dissemination of research results, including benchmark prototypes, showing that resilience benchmarks are technically achievable and cost effective.

For various reasons, industry is not accepting resilience AMB technologies as universally as one would hope. Even though a widespread understanding of the importance of computer system security and dependability seems to exist, no widespread use of AMB technologies has taken off. The first necessary step in resolving this issue is the widespread dissemination of results from the research community, demonstrating the feasibility of using AMB technologies in cost-effective manner.

A possible example of such dissemination is the proliferation of text books on resilience issues, the emergence of standardisation efforts in dependability benchmarking, and intensified training of engineering students in AMB technologies for resilience.

[*back to section 4.4 on Education, Training, Standardization and Take Up*]
[*back to section 5 on AMBER Topmost Priorities*]

A.4.2. Challenges

A.4.2.1 Understanding the value of assessing and benchmarking resilience in the business domain and in the user/customer community.

The research community has elaborated methods to assess/benchmark the resilience of systems and services, but the demonstration of the value of these results in the business domain and in the user community is not convincing. It is clear that education and the presentation of relevant examples may help understand the value of resilience assessment and benchmarking, however, there are several challenges that are exemplified below both from end-users' and business players' viewpoint.

First, it is difficult to translate the results of resilience assessment/benchmarking to user experience. Users are not easy to convince that resilience of services (as demonstrated by relevant benchmarks) is a distinctive feature. They are typically amazed by new functionality of devices, and rarely realize that having the most advanced functionality without resilience will result in disappointment if these functions are vulnerable or often crash. A related challenge is that the resilience measures are mathematically precise, but they are not formulated in a convincing and "user-friendly" way (e.g., the difference between 99% and 99.99% availability is easier to understand if we formulate these as 3.5 days vs. 1 hour average service outage in a year).

Similarly, it is not easy to translate the results of resilience assessment/benchmarking into business terms. In several domains having a certain minimum level of resilience is a requirement for successful business (for example, in case of IT service providers service availability is specified in Service Level Agreements, and the value of assessment is direct: downtime can result in considerable financial and reputation loss). In other areas, however, the consequences of improper resilience are less evident. Moreover, business goals are diverse: in companies, profit is the most significant factor, while in case of a public service effective operation (e.g., faster service or more lives saved) is the main goal. Accordingly, several convincing examples shall be elaborated that demonstrate the value of putting efforts into assessing the resilience of the services. The challenge of education is to prepare decision makers for understanding both the metrics of resilience and the related assessment methods, in order to be able to translate the results into better business decisions. The difficulty of these activities is highlighted by the fact that in the last few years the scope of the classic dependability notion was extended, and neither the metrics of the advanced notions nor the related assessment and benchmarking methods are well understood. The research to find easy-to-use metrics and assessment methods shall be followed by educational activities that promote the new results and help to understand the effects.

[*back to section 4.4 on Education, Training, Standardization and Take Up*]
[*back to section 6.4 on Future Internet technological platforms*]

A.4.3. Objectives

A.4.3.1 Promotion of proper and fair resilience assessment practices for specific classes of systems/services.

The objective of the promotion of fair resilience assessment practices is exemplified from a users' viewpoint.

Providing assessment results (like results of performance tests, usability tests) is a popular approach to characterizing products or services offered for the public (users or customers). However, comparing the assessment results is usually a difficult task. For example, the results of resilience assessment depend not only on the test inputs but also on the test environment including various aspects like the system configuration, background workload, fault load, user profile etc. The different service or application providers typically use different test environments and the testing conditions are not necessarily made public together with the test results. Naturally, each provider tries to tune its test environment (representing the "typical usage") to get better results in case of its service or application than the "similar" test results of the competitors. This way, users are not sure (sometimes even wrong) how to take into account these results in their decisions. Accordingly, the objective is to provide education about what constitutes "proper and fair" assessment practice, illustrating these notions via examples of practices that are inappropriate or unfair, and promoting good practices. Also, the publication of assessment results should be facilitated by means of public data repositories, benchmark applications (that may be executed by the users) and regular surveys for classes of target systems or services. (For example, there are widely used performance testing applications for personal computers, and there are reference scenes that are used for testing a digital camera. Similar practices are rarely promoted for assessing the resilience attributes of IT applications, although access to results of fair resilience assessment would be very useful for service consumers who could compare the provided services easily and could evaluate the appropriateness of one solution or the other.)

[*back to section 4.4 on Education, Training, Standardization and Take Up*]
[*back to section 5 on AMBER Topmost Priorities*]

A.4.3.2 Identification of the conditions under which standard dependability benchmarks would be beneficial.

Those who can decide whether to encourage or discourage the adoption of a standard need first to judge whether the standard would be beneficial in its intended area of application. For instance, governments can use their influence to affect standards: they may use persuasion, technical regulations, competition law, and the clout they have on some markets, as large buyers. But before using their influence to support or oppose a certain standard, they have to decide whether it would be in the best interest of society (or even of themselves as buyers, or of the industrial sector affected). This is just as true about standards for dependability benchmarking as about any others. Standardising dependability benchmarks has the potential advantages of simplifying decisions in purchasing, design of systems-of-systems, and sometimes regulation and licensing. These advantages create a natural constituency for standardisation. On the other hand, there are risks. For instance, having a standard that produces simple (and inevitably simplistic) rankings between systems may reduce the degree to which decision makers are held responsible, or feel responsible, for difficult choices between systems to purchase, or to license for safety-critical uses; designers and their managers may be encouraged to shift their focus from improving actual benefits for users to improving scores on benchmarks, and decision makers to shift their attention from (expensive and always uncertain) empirical observation of the former to (simple and cheap) dependence on the latter; the standard may thus legitimise lack of actual technical progress. As with any standard, there is also a risk that entrenched technical solutions and companies will dictate the benchmarks, thus reinforcing their dominance and discouraging innovation; also, powerful players may distort the standard-setting process to achieve gains for themselves.

The objective of research in this area is to combine the pertinent insights from social sciences – classical microeconomics (the behaviour of self-interested decision makers), behavioural economics (taking into account the psychology of decision making under uncertainty), studies of organisational behaviour and so on – with the specific technical knowledge about the feasibility and predictive power of benchmarks in a given application area, and knowledge about the power relationships among players in the specific market, to predict or analyse the effects of benchmarks on design habits and markets. The gain to be achieved is better informed decisions about when a sector of application is suitable for standardisation of benchmarks, and especially avoiding some of the major risks outlined above. Research would have to consider case studies, and through them demonstrate how to asses the future effects of adopting a dependability/resilience benchmark.

[*back to section 4.4 on Education, Training, Standardization and Take Up*]

A.4.4. Actions

A.4.4.1 Identification of knowledge and skills required to measure and assess resilient systems (partly on the basis of surveying potential employers' practice).

Specific context: Tuning employers' expectations and employees' skills

1. Motivation for the action:

In some cases, private companies and industries perceive a significant gap between the knowledge and skills they require to employees, and what employees can actual offer; this is true in many different sectors, including the resilience measurement and assessment area. At the base of this gap there are several social, political and institutional factors that do not allow a close collaboration/interaction between industries and research institutes (including universities).

2. Research steps proposed:

Research institutes and industries should tightly interact to identify the actual knowledge and skills required to measure and assess resilient systems, as they both play a very important and complementary role.

- At one side, industries should provide a list of required knowledge and skills, also surveying the current employees. For this purpose, individual interviews and ad-hoc questionnaires could be used to extrapolate the actual knowledge and skills of the employees considering their position and role.
- At the other side, research institutes should support the requirements' definition, both refining them and proposing new ones based on the last available research results. This second aspect has a great importance for fostering the effective transfer of resilience assessment/measurement best practices to European industry and to influence/promote the standardization bodies and processes.

The final goal of this process is then to obtain a set of non-ambiguous and well specified requirements on knowledge and skills, which represent both industries and research institutes' views.

3. Potential effects:

The overall effect will be to reduce the current gap between employers' expectations and employees' skills concerning the assessment and measurement of resilient systems. In more detail, the employees could use the produced requirements list as criteria to select university courses or PhD/post-doc activities, so to be more linked to the actual needs of companies in this research area. At the same time, the employers will have a more skilled set of candidates applying for a given position, which will lead both to a reduced start-up time of the employees and to an increased level of base of knowledge. Finally, the tight cooperation between research institutes and industries will also contribute to the effective transfer of resilience assessment/measurement best practices to European industry.

[*back to section 4.4 on Education, Training, Standardization and Take Up*]

A.4.4.2 Research in economics and social sciences about the effects of various types of benchmarks and standard assessment methods, drawing on the history of other industrial areas with comparable technical, social or market situations. - Perform market analysis about the value of public dependability measures in various application domains. - Identify application areas with the need for regulation to apply dependability measures.

In this research we conduct a business study about the costs, benefits and risks of benchmarks, gaining depth of understanding of the social, regulatory and business processes that govern the acceptance of such technologies. The research will be based on literature of comparable issues of measurement and benchmarking, in technology as well as social sciences. It includes market analysis as well as an analysis of the regulatory aspects of benchmarking.

[*back to section 4.4 on Education, Training, Standardization and Take Up*]

A.4.4.3 Identification of core taught subjects for specific courses (university, short and industrial courses).

Experimental methodologies for resilience assessment, measurement and benchmarking are typically based on fault injection. Therefore, fault injection is an example of a subject that can be identified as adequate for a specific course. Such a course should cover the most important techniques and tools, and should give students contact with experimental data. To this end, it would be attractive to use the studies available in the AMBER Raw Data Repository for analysis during the course. Students are often taught assessment methods in resilience-related courses, but rarely use them in practice, due to the complexity, high cost, and time required for setting up and running experiments. The main goal of the AMBER Raw Data Repository is to promote interaction among practitioners and research groups, but it may also be used for teaching fault injection in courses at different levels. This concrete action is an example among many "core taught subjects" that can be relevant for training both in academia and industry.

[*back to section 4.4 on Education, Training, Standardization and Take Up*]

A.4.4.4 Promotion of measurement practices for the different domains in academic, industry, and standardization bodies, with the goal of making these universally accepted.

1. Motivation for the action: Recently, increased attention is being paid to the experimental measurement-based evaluation of computing systems. The promotion of these practices is important in order to obtain the goal of making these universally accepted. In order to reach this goal, a characterization of measurement-based techniques and monitoring tools from a metrological point of view is needed. In fact this characterization will allow identifying standard ways of defining the metrics of interest and representing measurements results.

2. Research steps proposed: Proposed steps to perform this action are as follows:

- First of all the design and implementation of appropriate computer-based tools and the definition of appropriate mathematical methods to make, collect and analyze measurements results of computer based systems;
- Characterize tools and methods from a metrological viewpoint;
- Specify and execute appropriate experimental campaigns (both assessment and benchmarking campaigns) in order to verify the correctness of the defined and developed tools/methods;
- Large dissemination of the results in academic communities of interest; these results should clearly demonstrate the potentialities of the methods/tools;
- Exploitation in industrial communities of the new methods and tools obtained in these researches; the exploitation can be done through integration of the methods and tools in standard activities of verification and validation. In order to be usable with success, it is necessary to demonstrate that the new methods and tools, with respect to traditional methods, a. enhance the capabilities of analysis and b. reduce the time needed to perform V&V.

3. Potential effects:
The main effect of this action is the availability of standard and universally accepted tools and methods for the experimental evaluation of computing systems on different domains.

[*back to section 4.4 on Education, Training, Standardization and Take Up*]

A.4.4.5 Creating a consortium (of vendors, buyers, regulatory authorities with an interest in a certain application/technology domain), or get in contact with existing consortia, for the definition and acceptance of benchmarks (with industry support specific for the domain).

Motivation for the action. Computer systems industry holds a reputed infrastructure for performance evaluation and the set of benchmarks managed by TPC (Transaction Processing Performance Council [1]) and by SPEC (Standard Performance Evaluation Corporation [2]) are recognized as the most successful benchmarking initiatives. In fact, the success of performance benchmarking was strongly influenced by the role played by these two major organizations. To foster the definition and use of resilience benchmarks it is necessary to create a consortium of major players or to get in contact with existing consortia open to the inclusion of resilience metrics in their benchmarking initiatives.

Research steps proposed. Proposed steps to perform this action are as follows:

- Identify existing consortia eager to support and define resilience benchmarks for different domains (obvious options are TPC and SPEC).
- Contact the identified benchmarking consortiums to disseminate the resilience benchmarking and foster the definition of new resilience benchmarks or the inclusion of resilience metrics in existing benchmarks[7].

[7] The AMBER consortium has already started this work. In fact, an AMBER representative has been invited to attend a TPC (Transaction Processing Performance Council) meeting in Las Vegas, NV, USA, December 7-10, 2009. During the meeting four invited talks were given: 1) “Dependability Benchmarking: Should TPC support it?” (for the TPC-H subcommittee); “From Performance to Dependability Benchmarking: A Mandatory Path” (organized by the TPC Public Relations subcommittee); and “Dependability Benchmarking: A challenge for TPC…” (for the TPC General Council). During the meeting several companies approached the AMBER representative to know more about the topic. One decision of TPC was to study the creation of a working group to start working on the definition of a benchmark proposal focused on resilience metrics.

- As an alternative, get in touch with major vendor, buyers, regulatory entities, etc, with interest on resilience aspects in specific domains. In the same way major vendors formed TPC and SPEC, we believed it is possible to create a consortium to support resilience benchmarking initiatives.

Potential effects. A potential result of the action is the support from an existing consortium to include resilience metrics in its benchmarking initiatives. An alternative is a new consortium (formed by major players) dedicated to the definition and support of resilience benchmarks for specific domains. This achievement will foster research and use of resilience benchmarks.

[*back to section 4.4 on Education, Training, Standardization and Take Up*]