

Future for European Grids: GRIDs and Service Oriented Knowledge Utilities

Vision and Research Directions 2010 and Beyond

Next Generation GRIDs Expert Group Report 3

January 2006

Commission Disclaimer

This document contains information provided by a group of independent experts convened by the European Commission with the objective to identify potential European Research priorities for Next Generation Grid(s) 2010 and beyond. The document does not necessarily reflect the view of the European Commission.

Authors' Institutional Disclaimer

The views represented here are those of the individuals forming the group and do not necessarily represent those of the organisations to which the individuals belong.

Contents

1. Executive Summary	4
2. Background and Context.....	6
2.1 Initial consultation activities	7
2.2 The NESSI Initiative.....	9
2.3 The i2010 Initiative.....	10
3. History.....	10
3.1 The Next Generation Grids Reports	10
3.2 Next Generation Grids	13
4. Vision.....	15
4.1 Introduction.....	15
4.2 The SOKU Concept	15
4.3 The SOKU Services.....	16
5. Illustrating the SOKU Vision.....	17
5.1 Introduction.....	17
5.2 Enterprise Scenario	17
5.3 End-user Scenario	20
5.4 Manufacturing/Industrial Scenario	20
5.5 Summary	21
6. Research Topics	21
6.1 Full SOKU Lifecycle Management	22
6.2 Trust and Security in Virtual Organisations	22
6.3 Adaptability, Dependability, Scalability.....	24
6.4 Raising the Level of Abstraction	24
6.5 Pervasiveness and Context Awareness of Services	25
6.6 Underpinning Semantic Technologies.....	26
6.7 Human Factors and Societal Issues.....	28
6.8 NGG2 Topics as yet unaddressed.....	28
7. Conclusions.....	29
References.....	31
Appendix 1 – Market/SWOT Analysis for Next Generation Grids.....	32
Market evolution.....	32
SWOT analysis	33
Appendix 2 – NGG3 Participant List.....	37

1. Executive Summary

In the past few years the Grid vision has started to materialise into concrete technologies and products, enabling new and more demanding applications and services which go beyond the borders of research labs. The convergence between Grids and Web Services, Grids and Semantic technologies, and emerging Service Oriented Architectures will enable the provision of computing, data, information and knowledge capabilities as utility-like services in the future, including services which intersect with the physical world through a wide range of computing devices.

However, there is a fundamental gap between the technology and its users – particularly those users who will be the main consumers of these services – if the vision is to be realised to drive European competitiveness. The technology is still too complex, requiring too much knowledge to utilise it and too much effort to manage it.

In line with the above technology convergence trends, the IT market is evolving dramatically, driven by the increasing need to reduce costs of operations management and make business processes more and more agile and effective. The ongoing paradigm shift in the IT market revenues from the sales of products towards the provision of on-demand services has led IT companies to develop or adapt their concepts and strategies towards the emerging paradigm of **providing IT services as a set of utility services**, in a fashion similar to traditional utilities. Recent market forecasts (e.g. the EITO report 2005) agree on the fact that the European market for IT services will increase considerably in the future; in particular, process management services will grow fastest, as demand for outsourcing IT management and applications rises. Grid technologies have the potential to drive the market evolution of the IT industrial sector “toward IT services”. According to Gartner Inc., many businesses will be completely transformed over the next decade by using Grid-enabled Web Services to share application and computing capabilities.

In the past two years, a group of high-level experts, named the Next Generation Grid (NGG) expert group, has developed a vision that has emerged as the **European vision for Grid research**. Driven by the need and opportunity of bringing Grid capabilities to business and citizens, the NGG vision underpins the evolution of Grid from a tool to solve compute- and data-intensive problems towards a general-purpose infrastructure enabling complex business processes and workflows across virtual organisations (VOs) spanning multiple administrative domains.

The NGG vision consists of three complementary dimensions: the end-user perspective where the simplicity of access to and use of Grid technologies is exemplified; the architectural perspective where the Grid is seen as a large evolutionary system made of billions of interconnected nodes of any type; and the software perspective of a fully programmable and customisable Grid.

In order to realise the Next Generation Grid vision, numerous research priorities have been identified in terms of properties, facilities, models, tools, etc., which have inspired national and international research programmes for Grid research.

Thanks to the substantial investments and the numerous initiatives launched at the Member States and European levels, Europe has succeeded in establishing a leading worldwide position in Grids worldwide. The consistent portfolio of Sixth Framework Programme (FP6) Grid research projects will further contribute to the realisation of the NGG vision, thus boosting European competitiveness in Grid technologies and applications.

In the last quarter of 2005 the Next Generation Grids expert group was reconvened to identify the gaps between the leading-edge of Grid technologies and the end-user. The previous NGG work had identified the need for two aspects: semantically rich facilities and a service-oriented approach. The convergence of the evolving NGG vision with the service-oriented vision of significant European industry stakeholders in NESSI (Networked European Software and Services Initiative, an industry-led Technology Platform that aims to provide a unified view for European research in Services Architectures and Software Infrastructures [NESSI]) naturally led the NGG3 group to define the scientific and technological requirements necessary to evolve Grids towards the wider and more ambitious vision of **Service Oriented Knowledge Utilities (SOKU)**.

The SOKU vision identifies a flexible, powerful and cost-efficient way of building, operating and evolving IT intensive solutions for use by businesses, science and society. It builds on existing industry practices, trends and emerging technologies and gives the rules and methods for combining them into an ecosystem that promotes collaboration and self-organisation. The benefits are increased agility, lower overhead costs and broader availability of useful services for everybody, shifting the balance of power from traditional ICT (Information Communications Technology) players towards intermediaries and end-consumers of ICT.

The need for developing the SOKU vision stems from the necessity of effectively bringing knowledge and processing capabilities to everybody, thus underpinning the emergence of a competitive knowledge-based economy. The SOKU vision builds on and extends the Next Generation Grids vision. It captures three key notions:

- *Service Oriented* – the architecture comprises services which may be instantiated and assembled dynamically, hence the structure, behaviour and location of software is changing at run-time;
- *Knowledge* – SOKU services are knowledge-assisted ('semantic') to facilitate automation and advanced functionality, the knowledge aspect reinforced by the emphasis on delivering high level services to the user;
- *Utility* - A utility is a directly and immediately useable service with established functionality, performance and dependability, illustrating the emphasis on user needs and issues such as trust.

The primary difference between the SOKU vision and earlier approaches is a **switch from a prescribed layered view to a multi-dimensional mesh of concepts**, applying the same mechanisms along each dimension across the traditional layers.

The SOKU vision encompasses several issues addressed in the new FP7-ICT pillar '**Software, Grids, Security and Dependability**' as defined in the European Commission's proposal for Seventh Framework Programme (FP7) of April 2005 [COM(2005) 119] and the European Commission proposal for a Council Decision concerning the Specific Programme "Cooperation" [COM(2005) 440]. In particular: "*architectures and middleware systems that underpin knowledge-intensive services, including their provision as utilities*", "*service-oriented, interoperable and scale-free infrastructures*", "*grid-like virtualisation of resources*", "*network-centric operating systems*", "*mastering emergent behaviours of complex systems*", "*improving dependability and resilience of large-scale, distributed and intermittently connected systems and services*".

The development of a consistent programme of research activities in the above areas is considered essential to consolidate and further strengthen the European competitiveness in Grid technologies and in other related areas, including software and service technologies and applications, trust, security and dependability technologies. In view of its broad scope and cross-disciplinary character, this research should be developed through different types of funding interventions in order to build a critical mass at the pan-European level.

This report introduces the SOKU vision and recommends a structured and coherent approach to future research in Grid technologies, Service Oriented Architectures and Utility Services aiming at realising that vision. This approach is defined in the light of the most recent advances in adjacent areas such as distributed operating systems, software and services engineering, agents, peer-to-peer (P2P) architecture, knowledge and security technologies, and building on cross-disciplinary synergies and emerging European strengths.

Section 2 provides the technology background and the European research and innovation policy context. Section 3 is devoted to the history and main findings of the Next Generation Grids consultations. Sections 4 and 5 provide an overview of the SOKU vision and give some illustrative examples. A comprehensive set of research topics to be tackled to realise the SOKU vision is recommended in section 6. Section 7 concludes the report. Finally, the Appendices include a short analysis of the market trends and the NGG expert group membership.

2. Background and Context

The popularity of Grids has been growing very rapidly, driven by the promise that they will change dramatically the life of individuals, organisations and society as much as the Internet and the Web have done in the past decade. By enabling knowledge and computing resources to be delivered to and used by citizens and organisations as traditional utilities or in novel forms, Grids have the potential to give new impetus to the IT market and boost growth and competitiveness in many industrial and business sectors.

As the Grids concept matures from meta-computing (virtualised supercomputing over several physical machines) towards a pervasive Ambient Intelligence for wealth creation and improvement of the quality of life, the number of resources (or nodes) “abstracted” by the Grid increases dramatically and the range of node capability widens considerably. Addressing individual services as well as assemblies (composed atomic services) becomes commonplace, allowing ad-hoc dynamic creation of IT organisations to support business or societal applications. This has revolutionary implications for security, scheduling, data movement, code movement and concepts of ‘state’, messaging, transactions, rights management and payment. New infrastructural demands for performance, reliability, self-healing and self-management become apparent, coupled with the essential dimension of providing effective mechanisms by which users will interact with and gain benefit from this advanced capability.

The Grids environment must behave intelligently using knowledge-based techniques, including semantic-rich operation and interoperation. Long-cherished computer science principles are re-examined in the light of the new requirements. There are particular implications in managing the software complexity demanded by the requirements derived from the applications envisaged; this has aspects of software

theory, design, construction practice and also tools and environments to assist software development. This software complexity needs to be constrained through these tools to be applicable by semi-skilled and non-skilled developers and deployers of applications/services building on Grid technology. Using the Grid must be at least as easy as using a Web browser for it to achieve wide acceptability outside the academic and IT-literate commercial worlds.

The ongoing convergence between Grids, Web Services and the Semantic Web is a fundamental step towards the realisation of a common service-oriented infrastructure empowering people to create, provide, access and use a variety of intelligent services, anywhere, anytime, in a secure, cost-effective and trustworthy way.

In order to succeed, Grids need to leverage core standards including those for enabling services management, service-oriented infrastructures and services operations. Various consortia are involved in enabling the standards at the IT level for enabling better integration and development. This includes Standards Development Organisations such as IETF, W3C, GGF, OASIS, OMG, WS-I, DMTF, and FIPA. There are also ad hoc consortia, programmes and vendor groups, varying from groupings of major vendors to government supported projects.

Numerous Grid-related programmes and initiatives have been launched at the Member States and European levels in the past few years, giving Europe a worldwide leading position in Grid technologies and applications. The consistent portfolio of FP6 Grid research projects will further contribute to strengthen European competitiveness in emerging Grid-related areas.

This high momentum reached by Grid should be sustained and further evolved towards wider objectives, encompassing software and services, trust and security technologies, to ensure a higher impact of Grid research in terms of contributing to better products, processes and services, improving quality of life and increasing competitiveness and wealth creation.

It is essential that the new FP7-ICT Technology Pillar “Software, Grids, Security and Dependability” provides opportunities to develop a coherent programme of research activities, in order to consolidate and further reinforce European strengths in Grid technologies and in other related areas, including software and service technologies and applications, trust, security and dependability technologies.

2.1 Initial consultation activities

With the Communication of April 2005 “concerning the seventh framework programme of the European Community for research, technological development and demonstration activities” [COM(2005) 119] the European Commission has proposed the overall structure and research themes of FP7. One of the Technology Pillars proposed under the ICT theme of the Specific Programme “Collaboration” covers the research activities on “Software, Grids, Security and Dependability” (SGSD) with focus on “*dynamic, adaptive, dependable and trusted software and services, and new processing architectures, including their provision as a utility*”.

The Framework Programme text for the ICT Technology Pillar “Software, Grids, Security and Dependability” has been further refined in the European Commission proposal for a Council Decision concerning the Specific Programme “Cooperation” [COM(2005) 440], as follows:

“Technologies, tools and methods for dynamic and trusted software, architectures and middleware systems that underpin knowledge-intensive services, including their provision as utilities. Service-oriented, interoperable and scale-free infrastructures, grid-like virtualisation of resources, network-centric operating systems; open platforms and collaborative approaches for development of software, services and systems; composition tools; mastering emergent behaviours of complex systems; improving dependability and resilience of large-scale, distributed and intermittently connected systems and services; secure and trusted systems and services, including privacy-aware access control and authentication, dynamic security and trust policies, dependability and trust meta-models.”

This Technology Pillar builds on existing work on Grid Technologies, Software Technologies and ICT for Trust and Security, developed in FP6 under the responsibility of the three Directorate General Information Society and Media (DG INFSO) units F2, D3 and D4, respectively.

On 22 June 2005, the DG INFSO commission services involved in the preparation of the ICT Technology Pillar “Software, Grids, Security and Dependability” jointly organised a consultation workshop with a small group of experts representing the different constituencies. The main goal of this workshop was to discuss and expand the draft Specific Programme text with a view to obtaining the inputs needed to identify and shape up the priorities and strategy for a detailed research agenda.

The results of this workshop are available in the report “Software, Grid, Security and Dependability – Report of the workshop held in Brussels on 22nd June 2005” (see: http://www.cordis.lu/ist/trust-security/towards_fp7.htm). [SGSD]

In the following months several workshops and initiatives took place, aiming at deriving a coherent structure for the new Technology Pillar, addressing the key Science and Technology challenges in sufficient detail, bringing together the different constituencies embodied in the Technology Pillar and leveraging complementary contributions and viewpoints.

In the last quarter of 2005, Unit F2 “Grid Technologies” has reconvened the Next Generation Grids expert group in order to identify the gaps between the leading-edge of Grid technologies and the scientific and technological requirements necessary to evolve Grids towards the wider and more ambitious vision of **Service Oriented Knowledge Utilities (SOKU)**. The work of the expert group can be seen both as a follow-up of the former NGG consultations on the future of Grid research, as well as a main contribution to the wider consultation process on the FP7-ICE Pillar on “Software, Grids, Security and Dependability”. This report is the main outcome of the expert group meetings and discussions. A summary of the previous NGG consultations is presented in section 3, and Appendix 1 contains the SWOT analysis from NGG2, which illustrates the confluence with SOKU.

Further consultations are expected during the first half of 2006 which should lead to the first Work Programme in FP7, expected to be adopted in the second half of 2006 for first Calls to be published towards the end of 2006 or early 2007.

2.2 The NESSI Initiative

In parallel to the FP7 developments, the Networked European Software and Services Initiative (NESSI) was launched on 7 September 2005. NESSI is an industry-lead Technology Platform that aims to provide a unified view for European research in Services Architectures and Software Infrastructures. The mission of NESSI is to support the transformation of the European economy into a knowledge-based economy, by developing a visionary strategy for software and services driven by a common European research agenda.

NESSI introduces a new technological scenario, characterised by large distributed systems with many data collection points, services, and computers that evolve data into knowledge and help humans coordinate the execution of complex tasks. Large amounts of data will be generated by sensors, transmitted via wireless channels to ground stations, then moved through fast optical technology to powerful computational infrastructure, and the results are visualised on different devices according to the context of use. A crucial missing piece is a software infrastructure middleware facilitating a seamless and cost-effective composition of services in this new era of the Web. This software infrastructure should support pervasive and ubiquitous application scenarios where machines dissolve into the net and become a set of special purpose and domain specific appliances.

Achieving the initiative's objectives requires work supported by an integrated roadmap of research and development in the technology domains of infrastructure, service integration and semantics. These domains are inter-related, and aim to create the holistic framework where high-level services contextualise flexible low-level infrastructures. The infrastructure domain aims at the virtualisation of resources across servers, storage, distributed systems and the network. Service integration provides the tools and methods for configuration and composition of services in a modular and flexible way. Semantics is a key element for the transformation of information into knowledge and for effective machine to machine communication. Cross-domain aspects like end-to-end quality and reliability of services, trust and security, management services, interoperability and open standards will drive this holistic approach.

NESSI's approach is application domain independent, even if the outcome of the work on the roadmap could be configured and extended in domain specific platforms that could be re-combined in cross-domain platforms, creating a federation process. Autonomic capability will allow the solutions to adapt themselves to changes in the deployment environment conditions, in the business scenarios and in the user requirements.

Besides the technical contributions, NESSI will also interact with standardisation bodies, regulatory bodies and investment agencies to promote the development and / or address the needs of the area, will promote the take-up of technologies by business and industry, and will actively contribute to the dissemination and use of these technologies.

The initial NESSI promoters include Atos Origin, British Telecom, Engineering Ingegneria Informatica, HP, IBM, Nokia, ObjectWeb, SAP, Siemens, Software AG, Telecom Italia, Telefónica and Thales. However, the Technology Platform is open to the inclusion of additional committed players be they technology providers, integrators, solution or service providers, leading-edge end-users, innovative SMEs, universities or research centres.

Several members of the NGG expert group are actively involved in NESSI activities, in particular in the preparation of the NESSI Strategic Research Agenda. This shall facilitate the endorsement of the NGG's SOKU vision and findings by a larger industrial community, stimulating longer term cooperation between the several constituencies related to the FP7 ICT Pillar "Software Grids, Security and Dependability".

Further information on NESSI can be found on <http://www.nessi-europe.com>

2.3 The i2010 Initiative

On June 1st 2005, the Commission adopted the initiative "i2010: European Information Society 2010" [i2010] to foster growth and jobs in the information society and media industries. i2010 is a comprehensive strategy for modernising and deploying all EU policy instruments to encourage the development of the digital economy: regulatory instruments, research and partnerships with industry. The Commission's i2010 initiative is built around three main areas of focus:

- *Focus 1* – creation of a modern, market-oriented regulatory framework for the digital economy by combining all the regulatory instruments at the Commission's disposal;
- *Focus 2* – promote innovation and technological leadership by bringing the EU's research and development instruments into the game of digital convergence and setting priorities for cooperation with the private sector;
- *Focus 3* – promotion of an inclusive European Information Society, supported by efficient and user-friendly ICT enabled public services.

The convergence of Grids, Web Services, Semantic technologies, and emerging Service Oriented Architectures towards a service-oriented utility infrastructure has in particular significant potential to directly contribute to i2010's second focus area by stimulating European innovation in leading edge technologies and services. The European research strengths in these areas can be translated for industry (and especially SMEs) into services that deliver reduced costs, improved service levels and increased agility.

In addition European service providers are strongly positioned to benefit from a future Service Oriented Knowledge Utilities paradigm on three fronts: offering Grid services, developing and enhancing the underlying infrastructure, and managing hosted resources.

Further information on i2010 can be found on <http://europa.eu.int/i2010>

3. History

3.1 The Next Generation Grids Reports

In the first half of 2003 a group of high level experts was convened by the European Commission, Unit INFSO/F2, in order to produce a report entitled "Next Generation Grids, European Grid Research 2005-2010" [NGG1]. In this report, known as the NGG report, the experts pioneered the vision of the 'Invisible Grid', whereby the complexity of the Grid is fully hidden to users and developers through the complete

virtualisation of resources, and sketched the research priorities underpinning the realisation of the Next Generation Grids. According to the NGG report:

“A Grid provides an abstraction for resource sharing and collaboration across multiple administrative domains. The term resource covers a wide range of concepts including physical resources (computation, communication, storage), informational resources (databases, archives, instruments), individuals (people and the expertise they represent), capabilities (software packages, brokering and scheduling services) and frameworks for access and control of these resources (OGSA – Open Grid Services Architecture, The Semantic Web). At present multiple different Grid technologies co-exist, which stimulates creativity in the research community. Ultimately, however, we envision one Grid based on agreed interfaces and protocols just like the Web. Within that environment virtual organisations can co-exist, evolve and interact with each other in a secure way. This should avoid a proliferation of non-interoperable Grids, which would hamper the wide acceptance of Grid technology.”

The NGG vision, which has emerged as the European vision for Grid Technologies, consists of three complementary perspectives (see figure 1): the end-user perspective which exemplifies the simplicity of access to and use of Grid technologies; the architectural perspective where the Grid is seen as a large evolutionary system made of billions of interconnected nodes of any type; and the software perspective of a programmable and customisable Grid.

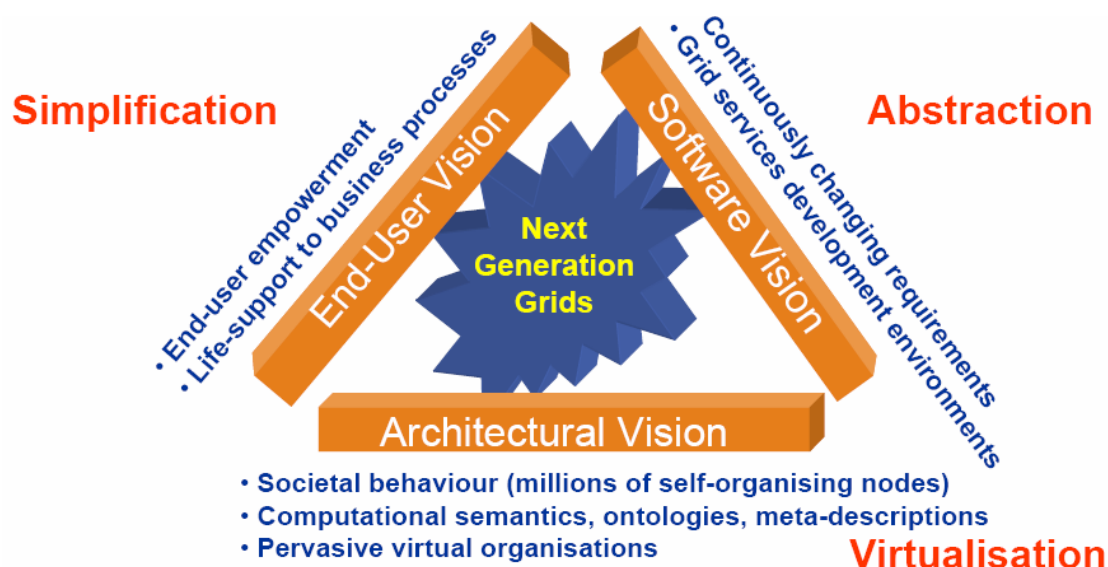


Figure 1: The NGG Vision

In order to realise the Next Generation Grids vision, numerous research priorities have been identified which have inspired national and international research programmes. They have been structured along the three categories: properties, facilities and models (see figure 2).

The NGG report 2003 complemented the description of the research scope and objectives of the FP6 2nd Call for the Strategic Objective “Grid-based systems for complex problem solving” reflected in the IST work programme 2003-2004.

Following the results of the evaluation of this Call, a number of projects which were launched in the course of 2004 addressed several of the research challenges identified by the NGG report. These actions are making significant progress beyond the state-of-the-art in Grids in terms of new architectures, middleware and generic services necessary to make the Grid an economically viable utility for industry, business and society.

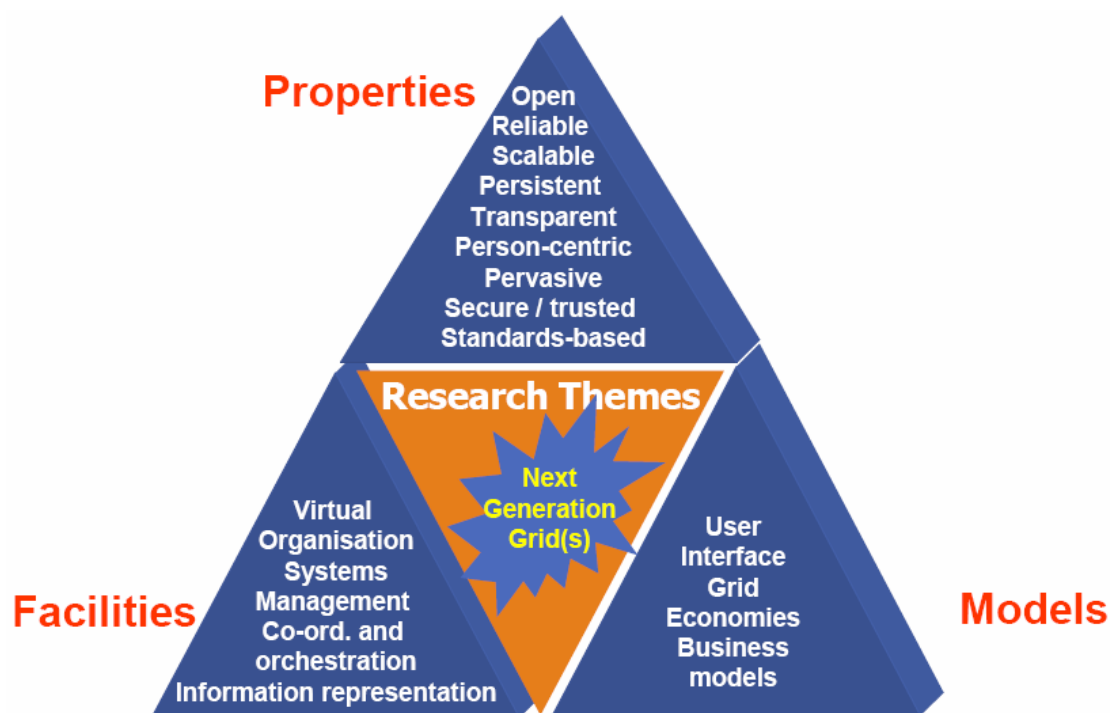


Figure 2: NGG Research Priorities

The NGG report 2004 [NGG2] identified additional requirements that arose in the light of one more year of experience by the experts working in the Grids domain. In particular, the shortcomings of existing Grids middleware were much better understood, and despite the development directions concerning OGSA providing some integration with service-oriented architectures, it was becoming clearer that applications in the Grids environment require a greater range of services than can be provided by the combination of currently evolving Grids middleware and existing operating systems. The report concluded that a new operating system environment was required to assure security, trust, performance and self-* properties¹, and indicated that the foundationware and middleware layers required more intelligence in order to provide the dynamics of the self-* properties required.

The report identified a need for semantically-rich knowledge-based services in both Grids Foundations Middleware and Grids Services Middleware both to improve the functionality but also to support applications on the Grids surface which require semantic and/or knowledge-based support. This led to the proposed architectural stack where the development of applications is supported by three abstraction layers

¹ Properties such as self-managing, self-optimising, self-configuring, self-healing, self-protecting and self-organising (also called autonomic systems).

consisting of Grids Service Middleware, Grids Foundations Middleware and Operating system.

This requirement led to the formation of a reconvened Next Generation Grids expert group known as NGG3 which has compiled this report.

3.2 Next Generation Grids

The complexity of software applications is growing rapidly under the time-to-market pressure. This implies that the future applications can no longer be built, from scratch, as monolithic entities as they have in the past. Today applications are no longer monolithic kernels running on a big computing platform, but rather dynamic collections of computing entities. In general, such applications will tend to be more and more multi-modular, written by several development teams using several programming languages, using multi-source heterogeneous data, mobile, and interactive. This will take us to a new style of application development based on software services. In this style programmers do not start from scratch but build new applications by reusing existing off-the-shelf services available in an ecosystem of services present in the Grid. The ecosystem is the set of connections among different services, competing for visibility and adoption, and cooperating within larger applications. When a complex service uses smaller services, this creates an association between the two of them. This concept can be easily compared to the well-known concept of the Web.

In this view, mainly for cost reasons, it is foreseeable that no single company or organisation would be interested in creating by itself very complex and diverse software applications. Future business and scientific applications will be built as a complex network of services offered by different providers, on heterogeneous resources, constrained by administrative problems when crossing the borders of different organisations.

Over several years the concepts of “Service Oriented Computing” (SOC) and “Service Oriented Architecture” (SOA) have generated a great deal of interest. The basic idea behind them is the encapsulation of coarse-grained application’s functionalities into services that are spread-out in the network. Those services should interact and establish contracts themselves using standards-based interfaces, enabling them to operate transparently between a diverse set of platforms and across organisational boundaries enabling the creation of dynamic virtual organisations. Although most definitions of SOA identify the use of Web Services in its implementation, one can implement SOA using any service-based technology.

In order to implement those concepts it is important that the resources present in the network are made visible, available and accessible in a standardised way to other potentially cooperating resource/services. This will require a complete virtualisation of the resources as Grid services. Standardisation efforts on service models, integration platforms, and business domain concepts based on XML and the Resource Description Framework (RDF) will accelerate the usage and spreading of service components for building Next Generation Grid services and applications.

The virtualisation of the resources is one of the most important aspects of the Next Generation Grid (NGG). It is necessary to raise the level of abstraction of the resources available in the Grid by virtualising them at a different level of abstraction, in order to enable Grid application designers and end-users to solve easily even more

complex problems. Next Generation Grid applications will be increasingly multi-disciplinary, collaborative, distributed, and most importantly, extremely dynamic (they will be assembled on-the-fly and could exist only transiently). This enables the creation of dynamically evolving virtual organisations, collaborative enterprises populated by dynamic participating components, services and information in order to achieve common objectives. The evolution of Next Generation Grids and SOKU is shown in figure 3.

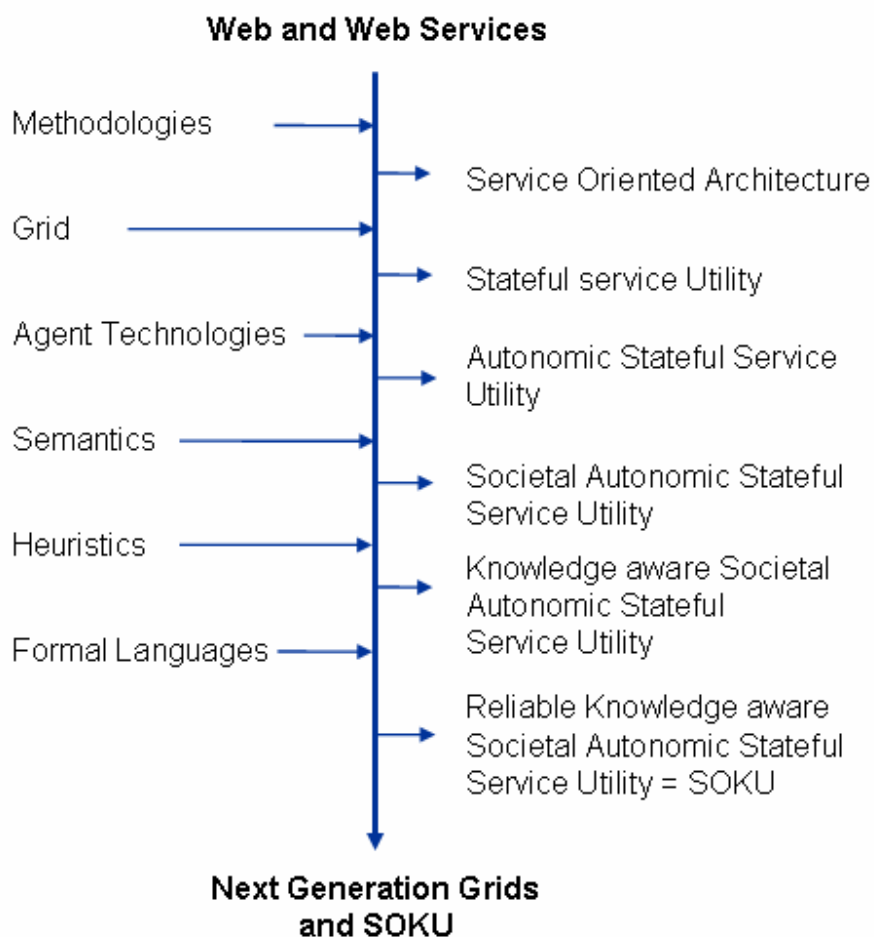


Figure 3: Evolution of Next Generation Grids and SOKU

This implies that current static Grid infrastructures will not be adequate for such applications, and the envisioned scenarios could be possible only because Next Generation Grids and SOKU will provide all the capabilities needed for the dynamic management of the services distributed across the available resources. In this light we can expect that, in a very near future, there will be thousands of open-market services available on the Grid. Service creation will consist of selecting, coordinating and deploying services chosen from this large software market: the problem will be to find the best services that fit the requirements and with the best performance/price trade-off. Dynamic service deployment will allow services to be added or upgraded without “taking down” a site for re-configuration and will allow the VO to respond effectively to changing resource availability and demand including “flash crowds” (i.e. sharp increases in the numbers of users attempting simultaneous access to resources, perhaps in response to an event).

4. Vision

4.1 Introduction

The NGG3 expert group considered the requirements of Next Generation Grids outlined above and, while respecting the vision of NGG1 and NGG2, realised that a novel approach – beyond the layered service-oriented architectures – was needed to meet those requirements. The SOKU concept was discussed and it was realised that, while supporting generally a new paradigm for construction of ICT systems, SOKU was applicable particularly to Next Generation Grids. Simultaneously it was realised that the desirable properties of Grids developed in NGG1 and NGG2 would enhance the initial SOKU concept making it more widely acceptable.

The SOKU vision builds on and extends the Ambient Intelligence vision developed as a driving concept for the 6th Framework programme [AMI] and partially relates to the Global Computing initiative [GC].

4.2 The SOKU Concept

This section introduces a European vision and technological requirements towards the realisation of Service Oriented Knowledge Utilities, a new paradigm for service delivery and software infrastructure, for the next decade.

The Service Oriented Knowledge Utility concept:

- is a flexible, powerful and cost-efficient way of building, operating and evolving ICT-intensive solutions for use by businesses, science and society;
- builds on existing industry practices, trends and emerging technologies;
- provides the rules and methods for combining services into an ecosystem that promotes collaboration and self-organisation;
- brings benefits of increased agility, lower overhead costs and broader availability of useful services for everybody, shifting the balance of power from traditional ICT players towards intermediaries and end-consumers of ICT.

The SOKU vision technically builds on a natural evolution and combination of concepts from Web Services, Grid technologies, the Semantic Web, distributed analytics and self-organising systems that have reasonably broad international industry acceptance.

The name captures three key notions:

- *Service Oriented* – the architecture comprises services which may be instantiated and assembled dynamically, hence the structure, behaviour and location of software is changing at run-time;
- *Knowledge* – SOKU services are knowledge-assisted (‘semantic’) to facilitate automation and advanced functionality, the knowledge aspect reinforced by the emphasis on delivering high level services to the user;
- *Utility* – A utility is a directly and immediately useable service with established functionality, performance and dependability, illustrating the emphasis on user needs and issues such as trust.

In the same way as the Web is the combination of multiple webs and the Grid of multiple grids, we envisage the Service Oriented Knowledge Utility as composed of multiple Service Oriented Knowledge Utilities.

The primary difference to earlier approaches is a switch from a prescribed layered view to a multi-dimensional mesh of concepts, applying the same mechanisms along each dimension across the traditional layers. There is a strong focus on exposing only user level concepts and the expectation of being able to use the provided services with the same dependability, safety, ease and ubiquity as existing utilities such as power or water.

4.3 The SOKU Services

SOKU services are distinguished from services in a typical SOA (such as Web Services) because they are described by explicit, machine-processable knowledge and they also work with explicit, machine-processable knowledge:

- SOKU services are *semantically described*, i.e. annotated with machine-processable metadata which facilitates their automated use. This enables them to be dynamically composed and configured, and for them to adapt automatically, providing self-management and autonomic behaviour. A SOKU service may itself consist of collections of services which are statically or dynamically orchestrated.
- SOKU services also work with semantically described content and semantic descriptions, i.e. they *process knowledge* – they may contain and use it, consume it, or produce it. This leads to a more generic set of services which are configured to the task at hand using explicit representations of the appropriate vocabularies, schema or ontologies.

The notion of knowledge in SOKU is interpreted broadly but typically refers to explicit symbolic knowledge – which can also be described as ‘actionable’ knowledge. Information becomes knowledge when context and interpretation are added, such as through explicit representation and sharing of schema or ontologies.

A layered architecture might typically place middleware layers between the underlying infrastructure and the applications. However, resources and services at any level may be semantically described and may contain, consume, use or produce knowledge. Hence SOKU is not a layer in itself, but we may classify any services within an SOA as various kinds of SOKU. This is illustrated in figure 4, which deliberately does not imply any additional structure.

Note that a Next Generation Grid involves both SOKU and non-SOKU services. The SOKU vision does not mandate that all services have SOKU properties, but rather that SOKU services are used within the service-oriented architecture. Services which use the semantic descriptions, for example to support the creation and operation of SOKU (the SOKU lifecycle) are by definition processing knowledge and may themselves be SOKU.

SOKU services will enable the creation and orchestration of dynamic Virtual Organisations (VO) across multiple domains, ensuring secure and trusted knowledge sharing. They will involve the management of VO membership and communities, allocation of knowledge resources as well as supporting the lifecycle of the knowledge.

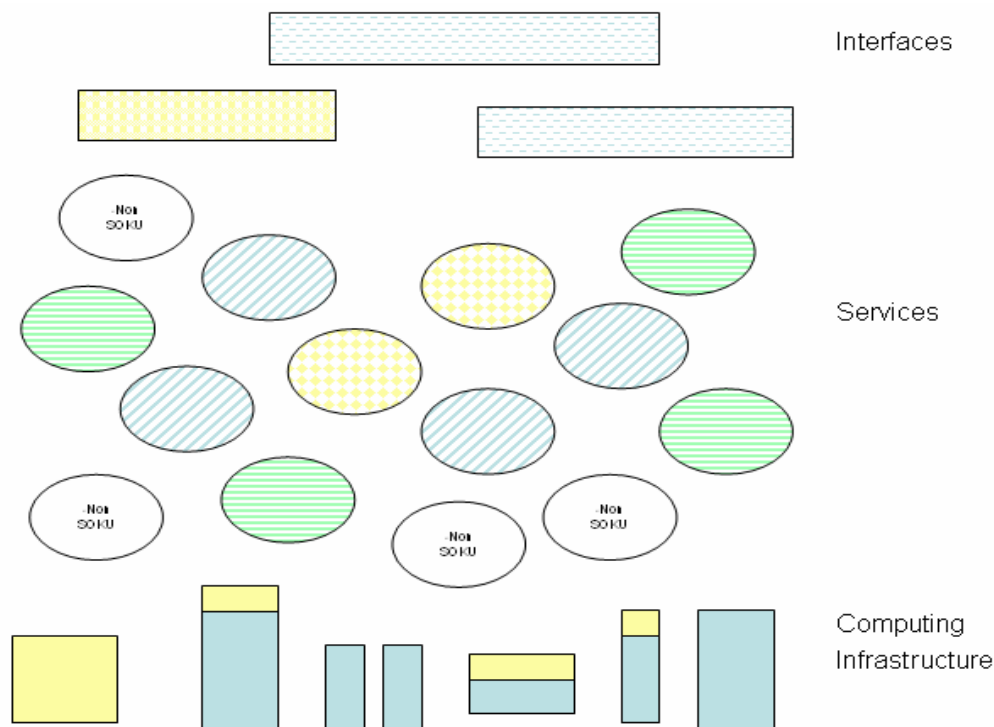


Figure 4: SOKU services may be used at any level

5. Illustrating the SOKU Vision

5.1 Introduction

The NGG infrastructure currently being developed provides virtualisation and abstraction of basic services including network, computation and data resources. Currently applications may be coupled closely to those resources. We envisage increasing virtualisation such that higher-level services bridge from this NGG infrastructure to the applications required by user or business. In other words there are interfaces that hide the implementation at a higher level than at present, behind which there occurs automatically facilitated assembly of services and resources. These services are Service Oriented Knowledge Utilities and are cooperative, self-*, dynamically adaptable and recomposable – in short ‘services with sociability’. There are many benefits to this approach: it enables resources to be shared, improving efficiency of assets, and to be combined in a flexible manner, facilitating management. This is illustrated in the following three scenarios.

5.2 Enterprise Scenario

The role of software, and particularly the degree to which partial or complete automation of business processes has led to increases in productivity and efficiency for the global economy, has increased dramatically in the last decade. Most of the gains that could be had from localised optimisations have been taken in the

industrialised nations and increasingly enterprises² are exploring more and more complex collaborative and interconnected scenarios to find new opportunities for improvements. The technological answer of the IT industry to the shifting emphasis from local optimisation and hence computation to collaborative and distributed semi-global automated business processes is an ongoing shift to service-oriented architectures.

Enterprise applications are heavily tailored to reflect business strategies, business models and business processes of companies. As businesses depend more and more on responding quickly to market changes, rapid adoption of changing business strategies and new business models is increasingly becoming a critical source of their competitive advantage. The interconnected nature of the new environment forces constant evolution and corresponding change management processes to the software solutions of companies and therefore effectively the business knowledge built into those. While service-oriented architectures may be the currently agreed upon direction to tackle this challenge conceptually, Grid technology is likely to be the best contender for an agreed upon strategy for managing the underlying computing infrastructure.

The increasing flexibility of Enterprise Applications and their growing complexity impose new demands on the underlying system infrastructure, which includes hardware, operating system and application platforms. The consequences of changes on the application level with respect to resource consumption and performance measures are no longer easily predictable at the infrastructure level. Static pre-planning of IT resources and fixed allocation of IT resources will soon have to become a thing of the past. With IT as a key driver for economic processes, IT itself needs to be run as a business with the same reasons driving for flexibility and collaboration among players. The most economic and efficient way of operating IT infrastructures is automated and collaborative, through sharing and dynamic assignment of resources.

Instead of assigning fixed resources to individual technical services supporting individual business functions, and cumbersome reassigning of these resources as requirements and load patterns shift, data centres of the future will have a clear abstraction between resources and resource consumption, making all assignments dynamic. There still will be data centres, for mission critical functions need guaranteed minimum available resources, but what constitutes location, distribution, boundary or resource ownership of such a “centre” in the future will look quite different. This is illustrated in Figure 5.

The observations above not only hold for large scale enterprises (LSE) that normally operate their own data centres but also for small and medium enterprises (SMEs) that in general cannot afford their own IT department and therefore use application provisioning and hosting services. In fact with Enterprise Applications becoming a kind of commodity software within the next decade, application service providers face the challenge to serve thousands of systems for their customers which makes the automation of hosting environments even more important.

² We interchangeably use “enterprises” for both private enterprises and public bodies such as governments in the following as long as they are driven by the same economic arguments.

In the past the data centre management has been focusing on the perspective of the individual services; because of the architectural changes a shift towards a higher level perspective is needed in which data centre resources are combined as required through policy and automation – this is the Grids notion within the data centre context. Looking at it from this angle the difference between an enterprise data centre and an application hosting environment, which could be seen as a shared data centre is mainly given by the fact that the first one is a single administrative domain whereas the latter one is an administrative super-domain providing a shared IT landscape in which several sub-domains run their own applications.

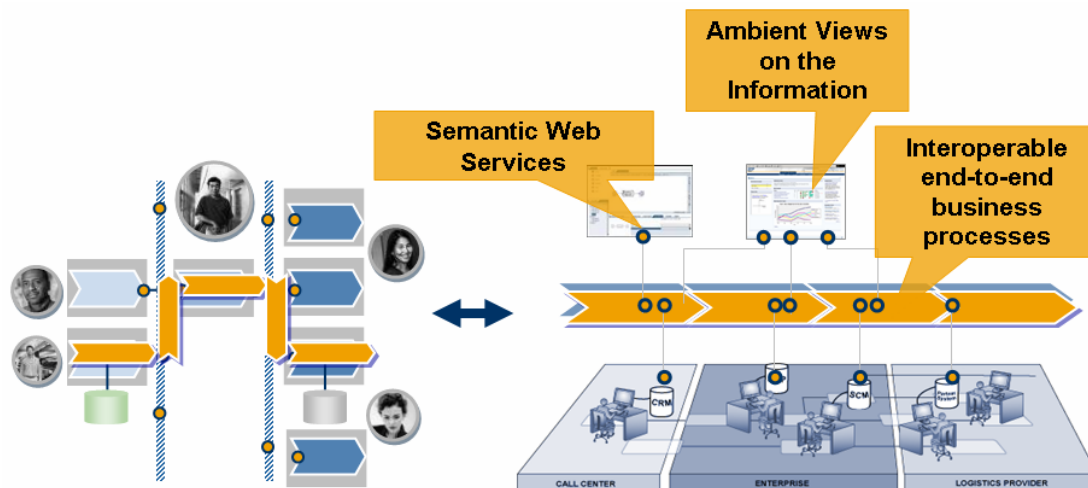


Figure 5: Services Architecture and scalable infrastructures to deliver future applications

IT landscapes of the near future will consist of entirely virtualised resources that can be dynamically assigned and managed. Virtualisation of compute resources is reasonably well understood already, though a lot of management challenges still remain. The virtualisation and management of storage and communication resources, in particular their performance, security and trust aspects are largely unexplored. In particular the information life-cycle management with all its aspects of persistence, distribution, reliability, provenance and privacy concerns in such a dynamic environments will pose major research challenges in the future.

In order to run IT resources as a collaborative business, there need to be intelligent ways of both abstracting away the complexities underlying such systems as well as keeping them understandable and amenable to analysis for optimisation and adaptation to shifting priorities. In a world in which software is no longer installed on a fixed IT system, licensing and royalties for software and services certainly needs new thought and may require adapted legislation and practice.

To bring together these services into meaningful enterprise applications requires a degree of interoperability in how and what they convey. Furthermore, bringing them together automatically requires agreement about how they are described in a machine-processable way. These are the key qualities of Service Oriented Knowledge Utilities.

5.3 End-user Scenario

To illustrate the consumer utility aspects and the advantage of distributed storage, access and resource ownership, we consider a public food supply chain security application. This is an example of an information utility for the consumer which involves extensive on-demand data integration and computation and which respects privacy.

Producers, manufacturers and distributors and sellers of food items are required to produce a chain of provenance of food being sold to the public. The tracking of relevant goods through the production and supply chain can be automated by appropriate connection of auditing systems into the production systems and machine readable tagging of items (typically using RFID, Radio Frequency Identifiers).

At the store level, the tagging of items needs to be broken when an item is sold to an end-consumer, to protect the privacy of consumers.

Now consider a service that should allow consumers to check for hazardous food items they may have purchased, while still protecting both the privacy of consumers as well as isolating potentially sensitive business data between potentially competing commercial entities.

Naively establishing a single database in which purchased items and consumer records are stored would violate the privacy requirements of both suppliers and consumers. The actors could instead store encrypted sales records in public grid storage. When a potentially hazardous item is discovered, consumers can check against the service whether the stores they frequent actually have sold any of the items, by entering the public keys of the store and items and matching via their private keys. Since these algorithms are quite compute-intensive, but the checks are only needed when there is an alert for a given class of items, a grid based solution would be the most cost-efficient way of providing such a service.

The underlying techniques include zero-knowledge proof technology and privacy preserving algorithms such as homomorphic encryption, and public key encryption augmented with pseudo-random additional records to foil statistical attacks.

5.4 Manufacturing/Industrial Scenario

Product manufacturers in industries such as automotive, aerospace and others have multiple design and engineering challenges when integrating a complex collection of mechanical, electrical, electronics and software components together for design and production of products (e.g. vehicles, engineering systems, and others). Consider a scenario with design engineers, manufacturing specialists, third-party design specialists and suppliers involved in collaborative design and engineering of complex products. Design engineers entrusted with new vehicle design work with sophisticated software tools and applications to enable new component design and perform product (e.g. vehicle) system validation and verification. They work with design-specialists and third-party suppliers who offer efficient and new ways of validating designs and specifications. Specification documents are shared across private or public Internet between the manufacturers and the design specialists. Once the design is validated, the list of components and subsystems are approved and submitted for prototype building and purchasing for mass manufacturing. The components are then sourced and validated through suppliers across the supply-chain. The components under specification are procured by the manufacturing and purchasing engineers, and then

shipped to production assemblies to enable rapid production and release of the vehicles.

The main challenge is enabling the integration of complex tools and applications (e.g. CAD and PDM applications) that are involved in design of complex products such as automotive vehicles. The design process involves complex validation and verification processes which require multiple tools, applications, product data systems, and third-party vendors to interact, collaborate and enable the design and engineering of new products or enhance existing products. The challenge is in enabling advanced process, information and application integration into a seamless system that can improve the process efficiencies by 10-20% and save the manufacturer millions of Euros spent per day in labour and process costs. Currently the processes in most industries are disconnected because of the complexity of the product and the tools used. Collaboration across the Internet is achieved in an ad-hoc fashion with high process costs. The integration of critical processes and applications within and across the Enterprises for new product design is of utmost importance in faster cycle times for releasing new products.

A potential solution can be constructed using service-oriented principles, appropriate interfaces for tools and applications, and metadata for the critical product information. In addition, knowledge-based middleware and grid functions are required to enable seamless, context driven collaboration between the various parties involved in the design and engineering cycle of vehicle design and production. The design of such a system requires both top-down and bottom-up design of fundamental service components and service interfaces for enabling collaborative design, engineering and manufacturing of complex products; e.g. vehicles.

5.5 Summary

The three scenarios presented in this section illustrate the broad application of Next Generation Grids and SOKU to meet real user requirements. The first illustrates the power of the virtualisation and interoperability provided by Grids and SOKU within the enterprise context, the second shows the role of Grid in delivering public information services ('knowledge utilities') which respect ownership and privacy issues, and the third shows how these approaches benefit collaborative processes within industry. They illustrate three traditional dimensions of Grid computing – coordinated computation, the datagrid and support for distributed collaboration – and in each case this is enhanced by the adoption of the SOKU concept which provides semantically described services and the delivery of knowledge to users. The scenarios also reinforce the requirements which were explored through the scenarios in NGG2, such as the need for self-*

6. Research Topics

The confluence of the Service Oriented Knowledge Utility paradigm and Next Generation Grids is a compelling approach to our future IT architecture. It encompasses several important domains including foundations of service-oriented computing, service-oriented architecture, grid and utility computing, business process management, business integration, etc. In order to realise the vision there are a number of challenging research topics that need to be addressed. The remaining part of this document gives a short description of eight interrelated topics that have been selected due to their importance.

6.1 Full SOKU Lifecycle Management

State of the art in Grid technology claims to enable businesses to respond rapidly to change by helping them consolidate their computing resources through their virtualisation. This, in theory allows businesses to respond dynamically and rapidly to changes in functional requirements as a matter of course during day to day operations.

Take up of Grid technologies is nevertheless hindered by a constant need for programming efforts stemming from a lack of basic service lifecycle functionality required to easily create and maintain Grid services.

SOKU infrastructure will require support for automating the full SOKU lifecycle making use of a semantically rich information representation thereby enabling support for context-awareness and Ambient Intelligence.

In particular the SOKU lifecycle should enable the following:

- On the fly service creation and seamless and scalable deployment enabling large scale services roll-out and roll-back, dynamic migration and autonomous reconfiguration/adaptivity. Diverse resources, ranging from PDAs to supercomputers, from small files to large databases, from small and efficient services to complex and multidisciplinary applications should be supported.
- Robust, efficient and semantically aware discovery of services based on proven technologies, as well as new approaches providing access to services on-demand. Particularly interesting issues include versioning support, peer-to-peer approaches and tight integration with composition functionality.
- Composition and composition control of services forming a self-organising ecology based on semantics. Running such virtual organisations will require advanced orchestration and choreography functionality based on modelling of a number of factors (e.g. enterprise physical/logical organisations, job sequence / data results sequence, dependencies between jobs, etc.)
- Management of functional and non-functional properties and requirements, which include performance, quality of service (QoS), dependability and security aspects. In particular, these mechanisms, techniques and tools must support suitable ways of describing, negotiating, using, proving and ensuring such properties and/or requirements.
- Support for multiple “economy models” for the grid. In particular, such economy models must be able to support reliable and scalable accounting, billing, secure access to resources and the like.

6.2 Trust and Security in Virtual Organisations

A VO, which can be seen as a social network that includes services and resources, is characterised by a notion of membership and roles, an agreement to share and collaborate, and mechanisms for achieving this. VOs range from persistent distributed organisations based on established trust relationships, such as resource sharing in large multi-institutional endeavours, to dynamic, ad hoc on-demand service assembly and provision amongst parties with little or no established relationships and partial or uncertain trust. In general these issues cut across both the human members and the digital entities and agents which constitute VOs. New questions arise beyond the trust of information, such as what it means to trust a service, agent or a workflow, and there may be a tension between virtualisation and trust, such as when a user

insists on a certain service provider or specific resources. New methods are required to analyse trust in these new circumstances.

Trust operates at various levels. At the highest level it involves cooperative activity in government, business or academia. At this level there are complex webs of contracts, memoranda of understanding, cooperation agreements and so on. The representation in a machine-understandable form of these cooperation-underpinning documents is a real research challenge. The interoperation of such documents is even more of a challenge – not least in gaining consensus on the syntax and semantics of such documents. Automated mediation systems are required and present a formidable challenge, not least because of different international legal frameworks within which they have to operate, and of the flexibility of business practice in different regions or different areas of commerce. Business depends on an understanding between the partners, this in turn is commonly based on ‘bona fides’ commonly not written but based on verbal recommendation or assurance. Representation of this aspect of business is a significant challenge.

At a lower level, formal contracts, service-level agreements and the documentation associated with trading such as orders, invoices, delivery notes etc require similar formalisms and consensus in order to permit the free-flow of commercial activity. These lower-level documents may be a lesser research challenge (not least because the EDI framework exists and particular frameworks exist for particular groupings of stakeholders) but improvements are required to adapt to the new environment. For example, automated brokers and trading agents need to function effectively, involving automated negotiation of SLAs, handling bundles of services, and dealing with QoS and non-functional properties.

In all these cases the security aspect is important. Many agreements depend on protection of the information from others for commercial or political advantage. On the other hand it is necessary for appropriate authorities (police, auditors) to be able to access certain information under appropriate conditions. We envisage a multi-domain environment in which entities have multiple identities and multiple roles. Designing a security framework in this context, including identity management, authorisation, authentication, digital rights management and temporal aspects, requires much research and development. Associated problems include safe digital signatures, particularly against advanced decryption techniques based on quantum computing methods, document encryption techniques, protection of metadata, detection of intrusion and its management – particularly in a heterogeneous environment. Designing security systems which are resistant to ever more inventive criminal activity is yet another challenge.

Hence there are issues of trust and security across the lifecycle of VOs, from business risk assessment to the mechanisms for contractual agreements and handling of compensation. Underlying this are technical challenges including identification and identity management in the dynamic and complex environment of VOs, handling of roles throughout complex workflows, mechanisms for traceability and provenance, and verification in order to trust what is not under direct control. At the end users have to be comfortable with the trust and security environment provided – or at least comfortable enough to feel that any risk is outweighed by benefits. This goes beyond technical issues and includes also sociological and psychological aspects, particularly in the context of a fast-changing environment.

6.3 Adaptability, Dependability, Scalability

New advances in networking and computing technology have produced an explosive growth in networked applications and information services. Applications are getting more complex, heterogeneous and dynamic. The combination of new parameters results in application development, configuration and management which will break present computing paradigms and are often inspired instead by biological systems which can manage themselves and react without external intervention. This is the basic paradigm behind *self-* systems*. Clearly, SOKUs belong to this kind of complex applications and the self-* paradigm will therefore be central for the design and deployment of this new kind of services.

The *adaptability* and *dependability* issues are related to the research on (statistical) availability caused by redundancy and replication, and to the research on self-healing, self-configuration and self-management as part of the self-* paradigm. The former research field includes a variety of topics tied to the Peer-to-Peer research, especially availability modelling, replica placement, distributed search mechanisms, and (statistical) quality-of-service guarantees. The later research field includes modelling of the resource capabilities, of system state (on different aggregation level), automated planning and decision making, and questions of control-loops treated e.g. in the control theory.

In addition to the above mentioned fields of Peer-to-Peer research, *scalability* issues will require the re-design of large parts of the existing Grid middleware, in order to adapt it to non-centralised methods of discovering, monitoring and managing resources. The approaches to handle this challenge are emerging in the domain of distributed systems, hand-in-hand with the ongoing paradigm shift from client-server architectures to Peer-to-Peer systems.

Performance in Grids is inherently related to issues of scalability and adaptability. In order to manage thousands of nodes efficiently, future Grids require built-in mechanisms ensuring automatic adaptation to new conditions such as load increase or resource deprivation. Also required is increased dependability (as compared to that available in current systems) on the level of individual nodes, clusters, and domains.

It is therefore desirable that a service-oriented approach behind SOKU does not rely on the pure client-server paradigm of today's SOA (identity of a service and its localisation on a single server and/or a centralised service discovery). SOKU should rely on a Next Generation Grid infrastructure that takes the benefit of Peer-to-Peer systems to further distribution of functionality such as scheduling, resource discovery, billing, and others. This will ensure that dependability, scalability and adaptability are intrinsic properties of next generation architectures.

6.4 Raising the Level of Abstraction

Substantial research efforts, greater than in the past, have to be invested to raise the level of abstraction of future generation grid systems at all the levels. In particular, this is necessary to raise the level of abstraction in such a way that the users/programmers are provided with higher level programming models and tools, as well as with better management abstractions. Such programming models, tools and abstractions must actually be able to relieve the programmers from most of (possibly all) the burden involved in the direct management of the specific, demanding and error prone grid related issues. These research efforts should be specifically finalised

to the design of innovative, advanced programming tools empowering rapid development, prototyping, debugging and (functional and performance) fine tuning of generic SOKU services. Research efforts have also to be dedicated to the development of models and tools for plain and SOKU service composition either to build full applications or to build (and provide users with) new services and SOKUs. Overall this would raise the level of abstraction as perceived from the applicative programmer viewpoint.

At the system programmer level, the level of abstraction must be raised by providing the user with better operating systems aimed at directly and efficiently supporting grid related activities at the kernel level, or by designing suitable and efficient abstract/virtual service containers supporting the virtual execution of SOKU services across multiple distributed nodes with different target architectures.

In the former case, higher level and more convenient OS level mechanisms and abstractions should be provided to the system programmers. They would be provided through actual network centric operating systems, dealing with proper management of network/service features. These new network centric operating systems can be derived from existing ones through proper additions and integrations as well as designed from scratch.

In the latter case, proper, clean and efficient abstract/virtual service containers have to be designed that can be used to run identical service implementation code and process identical data on nodes with different target architectures (e.g., differing for CPU and/or operating system) in a seamless manner. The abstract/virtual service containers should be designed in such a way that several levels of compatibility are provided, which can be used to target several kind of different devices, ranging from embedded/wearable ones to desktop or high performance architectures. This concept of service container can be provided through a lightweight middleware layer on top of existing virtual machines supporting the execution of portable bytecode (like existing Java Virtual Machines) or through new grid abstract/virtual machines supporting more suitable, easily and efficiently portable, instruction sets. In both cases, service containers instantiated at each node of the distributed grid would cooperate to deal with management of SOKU properties and perform the actions required to allocate or dynamically migrate service implementation code and data whenever it is required.

In both cases, common and compact (that is possibly binary, non XML based) data formats have to be designed that allow data to be seamlessly accessed from, and exchanged between, different remote locations, services, or SOKUs in order to achieve and preserve suitable performances.

6.5 Pervasiveness and Context Awareness of Services

In the future pervasive services will consist of self-contained entities capable of completing tasks on their own and able to discover other services and use them to compose a new higher level functionality. Based on this paradigm, service-orientation will entail high-level interoperability, as well as the smooth composition and automatic self-organisation of software. Furthermore, services should be able to organise their structure, behaviour, deployment and quality properties according to the available profiles and preferences, situation, and constraints. The distributed nature of this contextual information and its incompleteness, semantic variety and privacy-sensitivity make proper dealing with relevant context hard.

Semantic Web concepts are used to define intelligent web services that support automatic discovery, composition, invocation and interoperation. Some attempts have been made to integrate UML (Unified Modelling Language) and OWL (Web Ontology Language) based descriptions of software services to support self-descriptive service development. However, working solutions are still missing. Industrial and academic applications of SOAP (Simple Object Access Protocol), WSDL (Web Service Definition Language) and UDDI (Universal Description, Discovery and Integration) technologies have also highlighted their limitations in service development.

Existing software architectures are often modular, while service architecture or Service Oriented Architecture (SOA) means a new paradigm shift has emerged from the growing complexity of distributed software. It means a dynamic architecture, i.e. the structure and behaviour of software is changing at run-time as well as the location where the software is executed. The pervasive computing environment also brings new non-functional requirements related to interoperability, heterogeneity, mobility, and adaptability. This means that pervasive services have to be loosely coupled, and service composition shall be supported by service architectures that simplify complexity and allow dynamic service compositions, i.e. services shall be self-descriptive. Dynamic service compositions and dynamic binding techniques are required to enable dynamic architectures. Distributed self-aware software is required to manage the distributed service communication in a dynamic environment. New service protocols that also play an important role in loose coupling, self-organisation, and self-adaptation are also needed in order to guarantee the quality of services.

Standard based software technologies, i.e. reference service architectures and generic modelling and service technologies, like MDA (Model Driven Architecture) and Web Services provide the basis for the development of pervasive software. However, the context-aware co-ordination, which is one of the first steps towards self-organisation, is still missing in today's software development, and hinders the arrival of more automated, better and more efficient technologies.

6.6 Underpinning Semantic Technologies

The SOKU vision is predicated on machine-processable descriptions; e.g. of services, resources, people and policies. This enables the self-* behaviour. It also relies on an ability to work with knowledge from multiple sources. While 'a little semantics goes a long way', the full potential of this approach involves many research challenges, which have been articulated in the 'Semantic Grid' literature.

Service-oriented architectures will not scale and SOKUs will not be able to provide a solution without significant mechanisation of service discovery, service adaptation, negotiation, service composition, service invocation, and service monitoring; as well as data, protocol, and process mediation. In consequence, machine-processable semantics needs to be added to bring service orientation to its full potential. Only then can important subtasks be delegated to the computer and users can focus on the service they need. Therefore, bringing semantics into the core of modern computer science is the actual core of the NGG vision. Semantic description of services and information as enrichment of service-oriented architectures should be used to mechanise the common service layer that is needed to define domain and task-specific solutions based on the service paradigm. In order to provide this semantically enriched common service layer, it is necessary to define and add semantic

descriptions to data and information, service endpoints, collection of services and stateful resources, and to full-fledged service-oriented architectures. Questions that need to be addressed to achieve this are:

- How can semantics in computing really be used? How can services and data be described so that mechanisation of system composition can be achieved?
- How does an operational infrastructure make use of these semantic annotations of its services in practice and what impact does this have?
- How can scalable reasoning and formalisation methods on a scale of world-wide resource sharing be provided? Pragmatic approaches for the generation of specification, the usage of specification, and highly robust and efficient reasoning methods as well as restricted formal languages need to be developed.
- How do systems deal with semantic descriptions that are highly heterogeneous and dynamically changing? P2P network techniques for the resolution, alignments, as well as adaptation of meaning have to be developed.

In scientific terms, major achievements should include a better understanding of how semantics can become a major layer in modern computer engineering providing a new world wide operation system enabling resource sharing at a global scale. Whereas semantics has been studied in Artificial Intelligence for isolated tasks or tasks beyond the range of a computer (i.e. simulating or achieving human intelligence with a computer), it is necessary to realise it as an actual and central pillar of a new software architecture.

The research on the Semantic Web, with its focus on defining semantics of data, already has a research record of nearly ten years. Therefore, the description of data via semantics is already better understood; however, the scalable semantic description of services is still in its infancy. Since the approach is aiming for an integration of resources and services on a global scale, the semantic means must provide such coverage, too. Therefore, part of the approach that has to be followed is to co-develop standards at an international scale in cooperation with existing initiatives at Standards Development Organisations.

The lifecycle of the knowledge in this new very large scale and dynamic context, where knowledge both describes services and is processed by them, will have quite different characteristics to current systems. Established knowledge management and knowledge engineering approaches require revisiting in the SOKU context to identify new techniques and practices. For example, techniques are required for reasoning when there is incomplete and uncertain knowledge on a large scale, and there is a requirement for comprehensive handling of provenance. As well as addressing knowledge delivery in a human-centred way, there are significant opportunities for acquiring knowledge through the large scale interaction of users with future systems, emphasising the relevance of contemporary approaches such as ‘social tagging’, folksomies and statistical approaches.

6.7 Human Factors and Societal Issues

The match between the user requirements and the new technologies is essential for successful uptake. There are many classes of users within the value chain, including systems designers, service providers, solution providers, deployers, operators and maintainers, each with specific requirements. SOKU challenges the traditional notion of applications and these issues may entail new methodologies for evaluating and designing service-oriented and interactive systems.

The intersection between the physical world and the digital world of Next Generation Grids and SOKU requires an understanding of user interaction from a semantics perspective – for example capturing semantic annotations at the point of intersection, explicitly or as incidental knowledge capture. This intersection provides a real basis for ambient intelligence and extends the human-computer interaction issues associated with traditional interfaces and devices to address also the ‘intelligent’ behaviour of what can be regarded as an adaptive information system.

End-users will require personalisation techniques that allow the services and infrastructure to be tailored to their individual needs, including the creation of information and knowledge as well as its delivery. Interactions of individuals and groups with SOKU will demand new techniques to understand and support communities (virtual organisations, or social networks) that combine both human and computational agents and may vary from ad hoc to highly managed.

The socio-economic aspects also demand investigation. Issues of ethics, privacy, liability, risk and responsibility will arise which will impact future public policies. Economic and legal issues will emerge from combining open shared systems with other systems, and new forms of business models will emerge. These studies will require interdisciplinary research.

6.8 NGG2 Topics as yet unaddressed

The NGG2 report [NGG2] proposed a range of topics that required addressing through funded projects within FP6. It appears likely that not all of these topics will be addressed in projects that are to be funded. Thus, there are a set of topics remaining from [NGG2] that require attention in the early calls of FP7. The scenarios presented above reinforce the Next Generation Grids requirements identified in the NGG2 report.

Perhaps most significant is fundamental research on network-centric operating systems that are scalable (via components) from RFID scale devices through palmtops, laptops, desktops, servers to supercomputers. Network-centric operating systems support properties and provide functionalities that are usually addressed at middleware level to enable seamless integration and management of distributed resources while providing a uniform interface to applications and services. In a world where computing and knowledge capabilities are "escaping from the box" to pervade our everyday lives, there is a challenge to master immense power in the same way that Operating Systems did in the past forty years for the capabilities "inside-the-box".

7. Conclusions

The NGG3 expert group has further substantiated and widened the Next Generation Grid vision and research priorities, based on the notion of Service Oriented Knowledge Utilities.

The main results of the NGG3 expert group are:

- The articulation of the NGG and SOKU vision, which is significantly ahead of established Grid visions and constitutes an advanced e-infrastructure, enabling the development of new products and value-added services, thus improving European competitiveness and growth;
- The identification of research topics to enable the NGG and SOKU vision. The development of a coherent programme of research activities is considered essential to consolidate and further strengthen European positioning in the ICT market sector.

The vision represents a longer-term agenda and we anticipate multiple steps towards it, with commercial developments possible at each stage.

The group concludes that, in order to realise the compelling benefits of NGG and SOKU for European industry and society, future research beyond FP6 should include:

- (a) Lifecycle management: on-the-fly service creation and deployment; robust, efficient and semantically aware discovery of services; composition of services; management of functional and non-functional properties and requirements; and support for multiple “economy models” for the grid.
- (b) Trust and security: ad hoc and managed virtual organisations of digital and physical entities; policy and business practice; service-level agreements; authentication and authorisation in a multi-domain environment in which entities have multiple identities and multiple roles.
- (c) Adaptability, dependability and scalability: self-* systems; peer-to-peer; scalability.
- (d) Raising the level of abstraction: higher level programming models and tools; new or improved management abstractions; better operating systems capable of managing more complex resources and requirements from application, service and system contexts; abstract/virtual service containers; compact data formats.
- (e) Pervasiveness and context awareness of services: high-level interoperability, smooth composition and automatic self-organisation of software with structure and behaviour changing at run-time; non-functional requirements related to interoperability, heterogeneity, mobility, and adaptability.
- (f) Underpinning semantic technologies: mechanisation of composition; scalable reasoning and formalisation; heterogeneous and dynamic semantic descriptions; lifecycle of knowledge; collaboration and sharing.
- (g) Human factors and societal issues: user requirements and evaluation; intersection between the physical world and the digital; personalisation techniques; issues of collaboration and community; socio-economic aspects.

This research will be driven by the development of novel applications that are wealth-creating and / or improve the quality of life, particularly in the e-business domain, but also in e-health, e-environment, e-culture, e-science and e-government. It also requires work on the computer science issues that underpin these research areas.

Each of these areas should be accomplished in two phases: preliminary investigations, prototypes and demonstrators should be requested first; full developments and deployments should be pursued in subsequent FP7 IST calls.

The development of a coherent programme of research activities in the above areas is considered essential to consolidate and further strengthen the European excellence in Grid technologies and in other related areas such as software and service technologies and applications, a fundamental pre-requisite to boost industrial competitiveness in the European ICT sector. In view of its broad scope and cross-disciplinary character, this research should be developed through different types of funding interventions in order to build a critical mass at the pan-European level.

References

- [NGG1] “Next Generation Grids, European Grid Research 2005-2010”
ftp://ftp.cordis.lu/pub/ist/docs/ngg_eg_final.pdf
- [NGG2] Expert Group Final Report
ftp://ftp.cordis.lu/pub/ist/docs/ngg2_eg_final.pdf
- [NESSI] Networked European Software and Services Initiative
<http://www.nessi-europe.com>
- [COM(2005) 119] Proposals for a Seventh Framework Programme (FP7) for research, 2007-2013, COM(2005) 119
- [COM(2005) 440] Proposal for a Council Decision concerning the Specific Programme "Cooperation" implementing the Seventh Framework Programme (2007-2013) of the European Community for research, technological development and demonstration activities COM(2005) 440
- [SGSD] Software, Grid, Security and Dependability - Report of the workshop held in Brussels on 22nd June 2005
http://www.cordis.lu/ist/trust-security/towards_fp7.htm
- [i2010] A European Information Society for growth and employment
<http://europa.eu.int/i2010>
- [AMI] Ambient Intelligence: from vision to reality
ftp://ftp.cordis.lu/pub/ist/docs/istag-ist2003_consolidated_report.pdf
- [GC] Global Computing initiative
<http://www.cordis.lu/ist/fet/gc.htm>

Appendix 1 – Market/SWOT Analysis for Next Generation Grids

Market evolution

Today's trend in the IT market in shifting revenues from the sales of products towards the provision of on-demand services is expected to continue, driven by the increasing need of cutting costs of operations management and making business processes more effective. Several large IT multinationals have developed or adapted their concepts and strategies towards the emerging paradigm of providing IT services as a set of utility services, in a fashion similar to traditional utilities. Recent market forecasts made by highly renowned business analysts stated that the worldwide market for IT services is expected to increase considerably in the future; in particular, process management services will grow fastest, as demand for outsourcing IT management and applications rises. Grid technologies have the potential to drive the market evolution of the IT industrial sector "toward IT services".

In parallel with this trend, IT infrastructures are experiencing a fundamental paradigm shift in moving away from scalable client/server approaches towards decentralised, scale-free, service-oriented approaches. They may consist of millions of heterogeneous networked components and billions of "non-functional" dependencies. The unprecedented level of complexity, instability and pervasiveness reached by today's IT systems very often leads to situation where local or component failures can be propagated without control across the IT infrastructure and degenerate into global crashes or misbehaviors.

When confronted with such a level of complexity, the traditional computing models show limitations as they lack the capabilities, constructs and associated semantics necessary to express emergent and non-functional properties and behaviors and ensure fundamental properties such as composability, consistency, completeness. Furthermore, the existing implementation models (mostly based on the inter-leaving of several middleware layers) offer very little support for managing, adapting, and reacting to complex contextual changes and failures.

Clearly, new theoretical foundations are required to allow each component to react to changing circumstances dynamically and adapt to failures without compromising the system function or performance as a whole. Key research issues are, for example the development of new flexible and adaptable architectures; new software engineering methodologies, languages and tools, new approaches to operational semantics, new computational models to express emergent and non-functional behaviours; new approaches to resource virtualisation, interoperability and integration; new models for trust and security; complemented by horizontal issues such as economic models for IT service provision, societal and commercial acceptance.

SWOT analysis

The NGG expert group carried out a SWOT (Strengths, Weaknesses, Opportunities and Threats) analysis of European Grid capabilities and research. The SWOT analysis highlights the strategic importance of Grids both as primary ICT technology on which Europe can build its strengths, and as a key enabling tool for boosting productivity, competitiveness and growth of European economy. The confluence with SOKU is evident from this analysis, in particular with respect to data, Semantic Grid and the shift towards IT services.

The Commission has already sought advice on vision and research challenges in several ICT-related fields. For example, the IST Advisory Group Report of June 2002 has produced an extensive report in the area of software technologies, embedded and distributed systems. Many considerations identified in this report, such as the for example the effects of OS vendor monopolies and skills shortages, obviously apply to Grids software development as well. The SWOT analysis focuses mainly on those issues which have direct impact on Grids technology, or where Grid technologies may have a wider impact on ICT policies.

The complete SWOT analysis of European ICT capabilities and research as they relate to Grids technologies is described in Table 1 and Table 2 below.

The main conclusions of the SWOT analysis can be summarised as follow:

- Ontologies and Semantic Web technologies will be crucial to provide scalable support for complex, heterogeneous Grids middleware and applications.
- The strengths of the European telecommunications industry and the diversity of its market for electronic control systems have given Europe a leading position in the areas of mobile and embedded technology. This is of particular relevance for the realisation of the vision of a Grid as a pervasive, user-centered utility.
- The weakness in hardware and primary software products (e.g. commodity processors, server and desktop Operating systems, Programming Languages, etc.) may hamper the development of a European leadership in Grids Technologies.
- The convergence between Grids and Web Services provides a significant opportunity to move to a model of software development and service provision where the market dominance of particular OS vendors is no longer a major economic issue.
- The distinctive European vision of a Grids environment that operates from the level of devices to supercomputers, to serve communities ranging from individuals to whole industries, including data, information and knowledge and emphasising resilience and scalability could have a significant economic and social impact far beyond the scope of existing compute and data Grids. This should be contrasted with the North American Grid vision of programmer-level meta-computing.
- It is vital that any European vision for the evolution of Grids is accompanied by a clear representation of that vision to the key standards bodies and technology providers worldwide.

Table 1: SWOT analysis (Strengths and Weaknesses)

Strengths	Weaknesses
<p>Ontologies and web semantics: EU member states have a strong presence in the research and business communities working on the theory, tools and applications of ontologies and the Semantic Web. These technologies will be crucial if Grids are to provide scalable support for complex, heterogeneous applications.</p> <p>Collaborative approach to data: Member states benefit from national and European programmes for the collection and curation of research data and support for networking between researchers. This has led to a significant number of world-leading data collections and an open, collaborative approach to knowledge sharing, which is already forming a valuable underpinning for Grid based research. European research activities in the area of knowledge Grids represents a valuable contribution that can enforce the EU technology in data and knowledge management.</p> <p>Mobile devices and embedded systems: The strength of the European telecommunications industry and the diversity of its market for electronic control systems has given Europe a leading position in some areas of mobile and embedded technology. This is of particular relevance to the vision of Grids as a pervasive, user-centred utility.</p> <p>Grids Middleware: Europe has established a strong position on higher-level Grids middleware research. Most European developments have their particular strength in offering high-level services and tools for building applications, targeting the special needs of user and service provider communities such as industrial engineering, business applications, and application services provision. In this research area of higher level Grids services, European players seem to have gained a worldwide leading role.</p>	<p>Hardware: Europe lacks research and commercial leadership in desktop and server scale computer system design. This may hamper development of efficient, scalable Grids components in favour of more traditional server-based architectures.</p> <p>Operating Systems: Historically, support for adopted middleware technology has eventually migrated to the operating system level. Middleware and Grids technology choices made by the major OS vendors may therefore have a disproportionate effect on the uptake of particular Grids technologies.</p> <p>Programming Languages: A significant proportion of Grids and web service research and application development is dependent on language platforms which lie outside the control of European user communities. Despite excellent European academic R&D there is little pull-through to wealth creation. There is a problem overcoming the inertia of existing widely-used languages.</p>

<p>Semantic Grids: In Europe there is a long-standing expertise in Semantic Grids, agent and cognition technologies, which could be better capitalised and integrated towards Grids developments.</p> <p>Open Source: Europe has a strong Open Source community, including SME players with international visibility and market, covering the whole value chain.</p>	
--	--

Table 2: SWOT analysis (Opportunities and Threats)

Opportunities	Threats
<p>Paradigm shift toward IT services: Today's trend in the IT market in shifting revenues from the sales of products towards the provision of on-demand services creates unprecedented opportunities to develop a European competitiveness in IT services, which could have a catalysing, positive effect on adjacent sectors both upstream and downstream the user-supplier value chain.</p> <p>Operating system virtualisation: The adoption of Web Service technology by the major OS vendors allows the development of distributed applications that are independent of the underlying operating system and language technologies. The convergence between Grids and Web Services therefore provides a significant opportunity to move to a model of software development and service provision where the market dominance of particular OS vendors is no longer a major economic issue.</p> <p>Existing infrastructure for collaboration: FP6 already has instruments in place to allow for the effective exploitation of Europe's diverse Grids-related research community, and can rapidly build collaborative projects to allow researchers and businesses to exploit Grids applications.</p>	<p>Dependency on development tool support: Support for interoperable messaging protocols (such as SOAP) depends on the tools provided by the various language and OS platform owners. While at the moment there is agreement on the overall direction of Grids middleware and Web Service evolution, disputes or changes in policy over supported technologies could have a rapid impact on the ability of Grids developers and service providers to support particular language and operating system combinations.</p> <p>Standards evolution: As Grids technologies mature and become more complex, the adoption of standards (official or de facto) will be a requirement for sustained development of Grids, and only applications compatible with those standards will gain widespread adoption. It is vital that any European vision for the evolution of Grids is accompanied by a clear representation of that vision to the key standards bodies and technology providers worldwide.</p> <p>Non-acceptance and Lack of Use: Industry may not accept the developed / developing European Grids Foundations / Middleware leading to a non-interoperable environment thus reducing the potential market size and the advancement of the knowledge society.</p>

Service Model for Industry: The Grid today is used most heavily in particle physics, environmental science, life science applications, genomic research, protein folding, and medical applications, in advanced engineering R&D, in chemistry and materials science. It is expected that business, like finance or media, and many industries, such as aerospace, automotive, or entertainment will seize the opportunity to use existing IT resources more efficiently. Grids offer a new range of service models for the IT service industry.

Standardisation for European Advantage: a window of opportunity exists for a pan-European effort to develop the Grids foundations / middleware and have a workforce familiar with Grids so that Europe (business, industry, science, healthcare, environment, culture, education...) may have an interoperating advantage.

Next Generation Grids: The distinctive European vision of Grids operating from the level of devices to supercomputers, to serve communities ranging from individuals to whole industries, could have a significant economic and social impact far beyond the scope of existing compute and data Grids.

Appendix 2 – NGG3 Participant List

Expert	Country	Affiliation
J.-P. Banatre	FR	INRIA & University of Rennes
S. Campadello	IT	NOKIA
M. Danelutto	IT	University of Pisa
S. De Panfilis	IT	Engineering Ingegneria Informatica S.p.A.
D. De Roure	UK	University of Southampton
S. Druais	FR	Thales
J. Easton	UK	IBM UK
M. Fehse	DE	T-Systems
D. Fensel	AT	Leopold-Franzens-Universität Innsbruck
I. Fikouras	DE	ERICSSON
M. Fisher	UK	BT Group
A. Fuggetta	IT	Politecnico di Milano
W. Gerteis	DE	SAP
C. Goble	UK	The University of Manchester
Y. Guo	UK	Inforsense Ltd
J. Hierro	ES	Telefónica I+D
K. Jeffery	UK	Rutherford Appleton Laboratory
T. Kielmann	NL	Vrije Universiteit, Amsterdam
D. Laforenza	IT	ISTI-CNR
P. McCallum	UK	University of Cambridge
B. Neidecker-Lutz	DE	SAP
T. Priol	FR	INRIA
A. Reinefeld	DE	Konrad-Zuse-Zentrum für Informationstechnik
A. Reuter	DE	European Media Lab GmbH
M. Riguidel	FR	ENST
H. Saikkonen	FI	NOKIA
J. Sairamesh	USA	IBM
D. Snelling	UK	Fujitsu Laboratories of Europe
C. Thole	DE	Fraunhofer
T.A. Varvarigou	GR	National Technical University of Athens
W. Waterfeld	DE	Software AG

Chair: K. Jeffery; Editor-in-Chief: D. De Roure

The work of the Next Generation Grids expert group was supported by the European Commission, Directorate-General Information Society and Media, Unit F2 “Grid Technologies”: F. Accordino, W. Boch, M. Lemke