High-Availability and Standards – The Way to Go!

(Abstract)

Manfred Reitenspieß

Fujitsu Siemens Computers
Otto-Hahn-Ring 6
D-81739 München
manfred.reitenspiess@fujitsu-siemens.com

1 High-Availability – a Long Term Evolution

High-availability has been a long-term topic in computer science. There have always been applications with a need for extended, continuous operation. For example the NASA space program brought ideas about n-version programming, computer redundancy and state surveillance into real life applications ([SPC6x]) already in the 1960s. Advancement of computer technologies allowed Airbus Industries to introduce "drive by wire" in airplanes. These are just a few examples for the evolution of concepts to increase availability and reliability of systems and/or applications.

In this paper, we will not dive into the formal definition of terms such as dependability, availability or reliability. The interested reader is referred to widely available material such as in [Lap95]. For the following discussion, it is sufficient to use a relatively intuitive understanding of availability of a system or application – the system should do what it is supposed to do close to a 100% of the time.

Achieving such a high level of availability is a combination of a variety of technologies and methods. It is often their combination, which solves the reliability and/or availability requirements of the systems under implementation and maintenance. Particularly the maintenance aspect is often ignored and can introduce instability. During maintenance, new system components are introduced or defective components are replaced. The quality of the new components, but more importantly the quality of the replacement process implies the adequate use of availability techniques.

Redundancy

One of the most fundamental and widely used methods is the introduction of redundancy into systems. This approach has found wide industry acceptance. Even today, the use of redundant hardware is the most common approach to increase availability. It is now also applied for software objects such as processes or tasks. The concept of redundancy has also been applied to the development of software for high-availability systems. N-version programming is a typical example for such an approach. However, due to the limited applicability of such concepts and associated costs, they have not been widely adopted.

Programming Concepts

In dedicated application areas, security related programming concepts have been widely introduced. For example the use of transaction monitors, which support the programmer in assuring state consistency of a system, are a standard programming mechanism, mainly in the finance and other business related environments. Such concepts are now also transferred to the programming of Web applications ([WSR02]).

Programming starts with system design and the use of construction tools for the initial system structure set-up. The unified modeling language (UML) is getting widely accepted in the industry.

Quality Assurance Techniques

Last but not least, improved quality assurance (QA) technologies are applied in many application areas. QA includes the use of verification tools for the formal analysis of programs. Unfortunately, the formalization of a program and its formal analysis is still a very expensive process and often restricted to subsets of programming systems.

Another approach is the use of inspections and other methods for the analysis of designs and programs. Although not adhering to a formal analysis based on a mathematical model, such approaches are highly formalized to assure repeatability of results and the completeness of the analysis steps (dependent on method).

Such techniques, which also include the analysis of the design process, of the implementation and test process, are often used in highly critical areas such as the software development for airplanes.

Software validation in the sense of testing traditionally plays the most important part in industrial QA. This approach is heavily tool-based due to assure repeatability of results and their comparability. Also, the efforts for testing can become extremely high if not automated as far as possible (see [KR03] for a more detailed overview).

2 High-Availability – What Are the Issues

The number of availability technologies and methods are abundant and their importance has been widely accepted[1]. Nevertheless, the number of systems with a typical availability of more than 99,9% is still negligible compared to standard availability. According to an anonymous survey with 40+ Fujitsu Siemens Computers' customers, an availability of 98% is acceptable for 80% of them[2]. At the same time, all of the survey respondents see 99% and better as an achievable and relevant goal.

Some important factors to cross the gap between expected availability and its implementations are worthwhile mentioning.

Costs and Development Resources

It is (still) a challenging and expensive task to implement systems and applications supporting virtually continuous operation. Many reasons can be named:

- Quality assurance standards need to be raised. The degree of test coverage has to coincide with the availability expectations of the system or application. [5] gives an overview of the test efforts applied in telecommunications environments.

- Tools for implementation are not widely adopted. The implementation of systems supporting a high degree of reliability is not yet standard in software development. Requirements such as asynchronous events, exception handling (Ada), support for redundant structures need better acceptance in the programmer community.

Economic Pressure and New Requirements

However, economic pressures are growing to increase the overall system availability by factors. Main reason is the increased dependency of our society on the functioning of IT systems. Many business processes are dependent upon availability of the underlying IT infrastructure and the associated applications. Examples can be found in the banking or telecommunications industries, where a disruption of services can immediately lead to loss of revenue. The situation is aggravated by the introduction of interactive, web-based applications, the users can use any time they want and where the potential number of parallel users is only limited by the installed infrastructure.

[1] E.g. in the call for proposal for the 6th Framework Programme of the European community
[2] Unpublished material, can be requested from the author

In addition, changes in the value chain of network operators and their suppliers are now evolving, which increase the demand on system availability. E.g. the outsourcing of communications related processes of their corporate customers will be an important new stream of revenue for network operators. This offering will only be accepted if there is sufficient trust in the used infrastructure and services. Availability is an important trust component.

Standardization

A key aspect in the dissemination and exploitation of any technology is the support for standards. (De-facto) Standards were the gating factor for technical break-throughs such as GSM, the mobile telephone standard or XML, the mark-up language for web based applications. X/Open and POSIX interfaces form the basis for critical operating system functions to be used by application programmers.

High-availability interfaces (as specified by the Service Availability™ Forum) will only be accepted if they are used by software development companies world-wide. The vendors of computer platforms must support the interface standards on their systems. Customers must clearly state availability and reliability requirements in their project requirements.

It is therefore of overall importance to work with industry and academia in the specification of the interfaces, in the establishment of a convincing value proposition, and in the creating of market pull for high-availability and reliability.

Relationship to Security

The Internet has shown how closely related security requirements and dependability requirements are. Denial of service attacks and virus attacks are seen as security threats, but reduce the overall availability of the system. On the other hand, the availability of security tools such as firewalls and security gateways is crucial for the security of the system.

This close relationship has been addressed by the German Computer Society when a new security division was founded in 2002 to combine both aspects of dependability. Synergies in technologies and value proposition are expected when addressing both aspects in a joint group.

3 High-Availability – Trends and Standards in the Telecommunications Industry

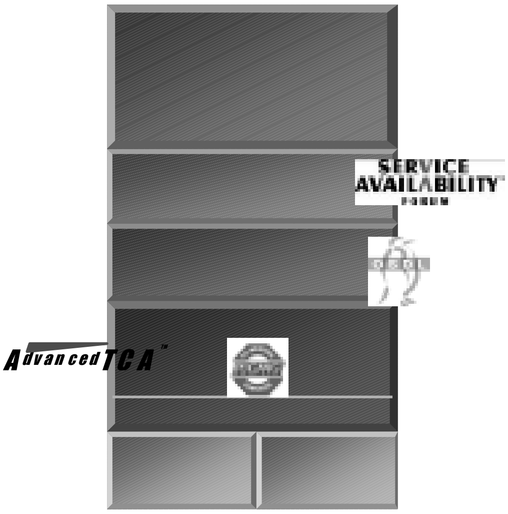

Figure 1: standardized layers of high-availability telecommunications platforms

Today, the dependability of the global communication infrastructure is more important than ever. As new technologies emerge to power new services, users quickly become dependent on those services to conduct their personal and professional lives. With this development, comes the challenge of accommodating growth and emerging technologies while maintaining uninterrupted availability and dependability. Users expect new, innovative services to be delivered on demand and without interruption. Service providers are expected to rapidly deploy new services and vouch for their reliability to compete successfully for users and meet customer expectations. This means that communications equipment must incorporate the highest possible levels of availability and dependability while balancing the constraints of short development cycles and increasing pressure to reduce development costs. The communications industry recognizes that the most urgent part of an effective solution is the broad adoption of open standards.

The Service Availability™ Forum was formed to address this situation by developing standard interfaces necessary to enable the delivery of highly available carrier-grade systems with off-the-shelf hardware platforms, middleware and service applications (see [SAF], [JHMP03]).

The Benefits of Service Availability™ Specifications

The goal of the Service Availability™ Forum is to develop a framework and specifications for service availability to enable carrier-grade systems to be developed using off-the-shelf computing solutions, and which meet user expectations for availability and dependability. Interface standardization eliminates the need for network equipment manufacturers to develop applications from scratch and port them to each new hardware platform. Service continuity is achievable only with the cooperation of all elements in the stack – hardware, middleware, and applications software. One of the primary motivations for creating the SAForum Application Interface Specification is, so application software can be developed independent of a specific hardware or system platform.

The Service Availability™ Forum Interfaces

To-date the Service Availability™ Forum has developed two interface specifications for carrier-grade platform and middleware applications that offer an open standard for the unification of carrier-grade network equipment elements. These standardized specifications provide the ability to vertically integrate applications and systems without resorting to customization of the application for a particular platform. The Service Availability™ interfaces specify the information that flows between software entities and the semantics of that information. The implementations on either side of the interfaces are not addressed. The Service Availability™ Forum intends to create an "ecosystem" which fosters rapid development of platforms, Service Availability™ compliant high-availability middleware and application software that can be used in a building block architecture while delivering highly available user services.

Two interfaces are specified by the Service Availability™ Forum (see Figure 1):

- The Application Interface Specification (AIS) - a programming interface between customer applications and high availability middleware.

- The Hardware Platform Interface (HPI) – a programming interface between the high availability and system level middleware and platform components.

Literature

[SPC6x] "Computers in Spaceflight: The NASA Experience - Chapter Two - Computers On Board The Apollo Spacecraft - The Apollo guidance computer: Software". www.hq.nasa.gov/office/pao/History/computers/Ch2-6.html.

[Lap95] Jean-Claude Laprie: "Dependability - Its Attributes, Impairments and Means". In (B.Randell, J.-C.Laprie, H.Kopetz, B.Littlewood Edts): Predictably Dependable Computing Systems,. Springer, Berlin Heidelberg New York. 1995. ISBN: 3-540-59334-9.

[WSR02]Web Services Reliable Messaging, http://www.oasisopen.org/committees/download.php/1461/WS-ReliabilityV1.0Public.zip. Web Services Transactions Specification, August 2002. http://msdn.microsoft.com/library/default.asp?url=/library/enus/dnglobspec/html/ws-transaction.asp.

[KR03] B.Kellerer, M.Reitenspiess: High-Availability Middleware - Design Principles, Quality Requirements and Test Methods. In Proceedings of the 7th World Conference on Integrated Design and Process Technology, Austin, Texas, December 2003. IDPT Vol. 2, 2003. ISSN No. 1090-9389. . www.sdpsnet.org.

[SAF] www.saforum.org

[JMHP03] Timo Jokiaho, Fred Herrmann, Dave Penkler, Louise Moser: The Service Availability™ Forum Application Interface Specification (AIS 1.0). Board and Solutions Magazine. http://www.saforum.org/press/releases/Board_and_Solutions_June_2003.pdf