Autonomicity – An Antidote for Complexity?


Link to publication record in Ulster University Research Portal

Published in:
Unknown Host Publication

Publication Status:
Published (in print/issue): 01/08/2005

DOI:
10.1109/CSBW.2005.28

Document Version
Publisher's PDF, also known as Version of record

General rights
Copyright for the publications made accessible via Ulster University's Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Ulster University's institutional repository that provides access to Ulster's research outputs. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact pure-support@ulster.ac.uk.
Autonomicity – An Antidote for Complexity?

Roy Sterritt
School of Computing and Mathematics
Faculty of Engineering
University of Ulster
Northern Ireland
r.sterritt@ulster.ac.uk

Mike Hinchey
NASA Goddard Space Flight Center
Software Engineering Laboratory
Greenbelt, MD 20771
USA
michael.g.hinchey@nasa.gov

Abstract

Autonomic Computing and other self-managing system initiatives, many strongly based on biological metaphors, are emerging as a significant new vision for the design and development of complex computer systems. They offer the promise of controlling complexity through the achievement of self governance (autonomy) and self management (autonomicity). We consider how complexity is exhibited in the computer industry as a whole, and how the situation is deteriorating, rather than improving. We consider how Autonomous and Autonomic Systems, with their biological inspiration, can provide a framework for tackling complexity and overcoming the problems of its (unavoidable) inherent existence in certain classes of systems.

1. Introduction

Computer systems are becoming increasingly demanding, challenging, and complex. Various visions and paradigms for the future of computing are emerging. Realization of these future visions, whether it is invisible, world, ubiquitous, pervasive, utility, grid, ambient intelligence, semantic web, etc., will require us to come to terms with complexity.

When one considers interdisciplinary fields such as Biology and Computer Science joining together in the guise of Computational Biology and Bioinformatics, most famously in the post-genomic era to harvest the fruits of the sequenced human genome, complexity sky rockets. Organizations that research and develop complex computer systems are facing additional market conditions such as demands for more functionality, ever decreasing time-to-market, domain-expert shortage and high employment costs, all of which leads to a need to utilize employees that are only semi-skilled.

This result is the need for self-managing systems and new development approaches that can deal with real-life complexity and uncertainty. The challenge is to produce practical methodologies and techniques for the development of such self-managing systems, so that they may be leveraged to deal with complexity.

2. The Challenge is Complexity

The world is becoming an ever-increasingly complex place. In terms of computer systems, this complexity has been confounded by the drive towards cheaper, faster and smaller hardware and functionally rich software. The infiltration of the computer into everyday life has made our reliance on it critical. As such, there is an increasing need throughout design, development and operation of computer systems to cope with this complexity and the inherent uncertainty within. There is an increasing need to change the way we view computing; there is a need to realign towards addressing computing in a complex world.

The IT industry is a marked success; within a half century it has grown to become a trillion dollar per year industry, obliterating barriers and setting records with astonishing regularity [1],[2]. Throughout this time the industry has had a single focus, namely to improve performance [3] which has resulted in some breathtaking statistics [4]:

- Performance/price ratio doubles around every 18 months,
- resulting in 100 fold increases per decade;
- Progress in the next 18 months will equal ALL previous progress;
- New storage = sum of all old storage, ever;
- New processing = sum of all old processing;
- Aggregate bandwidth doubles in 8 months.

This performance focus has resulted in the emergence of a small number of critical inherent behaviors in the way the industry operates when designing, developing and
deploying hardware, software, and systems [3], and in the tacit acceptance of a number of flawed assumptions:

- That humans can achieve perfection; that they avoid making mistakes during installation, upgrade, maintenance or repair.
- Software will eventually be bug free; the focus of companies has been to hire better programmers, and universities to train better software engineers in development life-cycle models.
- Hardware mean-time between failure (MTBF) is already very large—approximately 100 years—and will continue to increase.
- Maintenance costs are irrelevant compared to purchase price. The assumption that maintenance is a function of price as such cheaper helps keep estimates of maintenance costs artificially lower.

When made explicit is this way, it is obvious that these implicit behaviors are flawed and result in contributing factors to the complexity problem.

Within the last decade, problems have started to become more apparent. The industry that is used to rising metrics saw some key decreases. Figure 1 [1] highlights that key modern day systems—cell phones and the internet—have seen a decline in availability, changing the established trend of their counterparts.

![Figure 1 Systems Availability over the Decades](image)

When one also considers how dependant we have become on our systems and how much it costs for a single hour of downtime (e.g., in 2000, $6.5m for brokerage operations, $2.5m for credit card authorization and $¼m for eBay [3],[5]) it is obvious how much we increasingly require our complex systems to be dependable.

### 3. Dependability through Autonomicity

#### 3.1 Dependability

Dependability is defined as that property of a computer-based system that enables reliance to be placed on the service it delivers. That service is its behavior as perceived by other systems, or its human users [6].

Figure 2 [6]-[9] (which has been updated in [15]) depicts the concepts of dependability in terms of threats to, attributes of, and the means by which, dependability is attained.

The effectiveness of these four mechanisms has a substantial influence on the dependability of a computer-based system.

Randell describes dependability in terms of failures, faults and errors, arguing that they follow a “fundamental chain” [6], thus:

\[ \cdots \rightarrow \text{failure} \rightarrow \text{fault} \rightarrow \text{error} \rightarrow \text{failure} \rightarrow \text{fault} \rightarrow \cdots \]

More abstractly, this can be described by the sequence:

\[ \cdots \rightarrow \text{event} \rightarrow \text{cause} \rightarrow \text{state} \rightarrow \text{event} \rightarrow \text{cause} \rightarrow \cdots \]

For example, the failure of a system (event) occurs when a fault is encountered during its operation (cause), because of an error in its implementation (state). This might be attributed to a failure in the test process (event) because the relevant code was not exercised (cause), meaning that the test suite was incomplete (state).

These chains may, of course, be broken at any stage in the chain by effective fault means (fault prevention, fault tolerance, fault removal and fault forecasting, as shown in Figure 2).

Overall, the breadth of issues involved suggests the need for a holistic approach to designing dependable systems [36].
3.2 Autonomic Computing

Autonomic Computing, launched by IBM in 2001 [2], [10]-[13], is emerging as a significant new strategic and holistic approach to the design of computing systems. Two of IBM’s main objectives were to reduce the total cost of ownership of systems and to find better ways of managing their increasing complexity.

In addition to IBM, many major software and system vendors, such as HP, Sun, Cisco, and Microsoft, have established strategic initiatives to help create computer systems that manage themselves, concluding that this is the only viable long-term solution.

As the name implies, the influence for the new paradigm is the human body’s autonomic system, which regulates vital bodily functions such as the control of heart rate, the body’s temperature and blood flow—all without conscious effort.

The desire for automation and effective robust systems is not new; in fact this may be considered an aspect of systems and software engineering best practice. Similarly, the desires for system self-awareness, awareness of the external environment, and the ability to adapt, are also not new, being major goals of artificial intelligence (AI) research for many years. What may be considered new in Autonomic Computing is its overall breadth of vision and scope.

Research in Autonomic Computing is likely to see a greater collaboration between the AI and software engineering fields. Such collaboration has been motivated by increasing system complexity and a more demanding user community. For example, software engineers have used AI techniques to provide more sophisticated support for user interfaces, better and faster search techniques, and to help address soft issues in the development and operation of software. Similarly, the AI community has increasingly been looking to software engineering for disciplined methodologies to support the development of complex intelligent systems.

Consequently, Autonomic Computing is perhaps best considered a strategic refocus for the engineering of effective systems, rather than a revolutionary new approach [14], that said the overarching vision may be considered revolutionary.

The overall goal of Autonomic Computing is the creation of self-managing systems: these are proactive, robust, adaptable and easy to use. Such objectives are achieved through self-protecting, self-configuring, self-healing and self-optimizing activities, as indicated in Figure 3 [36].

To achieve these objectives a system must be both self-aware and environment-aware, meaning that it must have some concept of the current state of both itself and its operating environment. It must then self-monitor to recognize any change in that state that may require modification (self-adjusting) to meet its overall self-management goal. This means that a system must have knowledge of its available resources, its components, their desired performance characteristics, their current status, and the status of inter-connections with other systems. This self-monitoring and self-adjusting forms a feedback control loop between the managed component and the autonomic manager.

The ability to operate in a heterogeneous environment requires the use of open standards to understand and communicate with other systems. In effect, autonomic systems are proactive in their operation, hiding away much of the associated complexity from users.

Self-healing is concerned with ensuring effective recovery under fault conditions, without loss of data or
noticeable delays in processing, while identifying the fault and repairing it, when possible. Fault prediction techniques may also be used, leading to re-configuration to avoid the faults concerned, or to reduce the likelihood of their re-occurrence.

With self-optimization, the system seeks to optimize its operations in both proactive and reactive ways.

With self-protection, a system will defend itself from malicious attack and may also have to self-heal when problems are detected, or self-optimize to improve protection.

With self-configuring, the system may automatically install, configure, and integrate new software components seamlessly to meet defined business strategies.

IBM discusses the characteristics or “elements” of Autonomic Computing in more detail in its manifesto [10]. This is being expanded throughout the research community, as witnessed by the uptake of workshops and conferences on the topic (e.g., [22]-[27]).

3.3 Autonomic Computing and Dependability

Randell and colleagues [7]-[9] give two main reasons for their interest in and focus on the concepts and definitions of dependability, failures, errors, faults and tolerance. First, there is a need to clarify the subtleties involved. Secondly, and possibly more importantly, is a desire to avoid dependability concepts being reinvented in other research domains such as safety, survivability, trustworthiness, security, critical infrastructure protection, information survivability, etc. [6]. Often the associated research communities do not realize that they are dealing with different facets of the same concept, and are failing to build on existing research advances and insights [6].

This focus on concepts and definitions is also critical for Autonomic Computing. Research and development from many disciplines will be required and, as already mentioned, the successful integration of AI and software engineering, will be particularly important.

In the IBM manifesto for Autonomic Computing [10], success is linked to the use of open standards, open source code, and open technologies in general. Yet there is also a need for common concepts and indeed common or open definitions for researchers from the many disciplines that are needed to make Autonomic Computing a reality.

On first consideration, dependability and fault tolerance would appear to be specifically aligned to the self-healing facet of Autonomic Computing. Yet any system that is incorrectly or ineffectively configured and/or inefficiently optimized is likely to lead to failures in the future. Similarly, any system that is not adequately protected is vulnerable to malicious faults, whether from hackers or viruses. Thus, essentially all facets of Autonomic Computing are concerned with dependability [36].

Referring again to Randell’s fundamental chain:

… ➔ failure ➔ fault ➔ error ➔ …

and its abstract form:

… ➔ event ➔ cause ➔ state ➔ …

then each facet of Autonomic Computing (Figure 3) can be considered “states of undependability” or “states of dependability” according to how well they are addressed in a system.

3.3.1 States of Undependability

Faulty (unhealthy)
Ill-configured
Sub-optimal
Unprotected

That is, if any of these states exist within a system, they are liable to lead to subsequent errors; in turn, that may lead to subsequent faults and on to failure. Autonomic Computing, through self-healing, self-configuring, self-optimization and self-protection, will therefore increase dependability.

4. Towards Autonomicity in Complex Systems

This section takes a look at some exemplar complex areas to highlight the need for autonomicity.

4.1 Telecommunications Systems

As the size and complexity of networks and communications continue to grow, there is a heightened need to develop new techniques capable of achieving a level of service with successful operations upon which users can place even more reliance. Key emerging strategies for meeting this demand are “autonomic networks” and “autonomic communications”, concepts similar to Autonomic Computing, while specific to the communications field.

The Autonomic initiatives are about much more than faults and self-healing, yet this is a critical area to address considering that it has been estimated that companies spend 33% to 50% of their total cost of ownership recovering from or preparing against failures.
All properties within self-management can also be related to a fault focus.

The Internet, with its vast infrastructure supporting millions of interconnected computers is perhaps the most significant development. The complexity of networks has grown in various ways [29]. As user demands and expectations become more varied and complex, so too do the networks themselves.

Data, voice, image, and other information now travels under the control of different protocols through numerous physical devices manufactured and operated by different vendors. It is expected that the trend towards increasing complexity will continue, due to several factors such as the increasing complexity of individual network elements, the need for sophisticated network and communication services, and the heterogeneity of connected equipment [30]. The promise of Autonomic Networks, networks that manage themselves, will substantially abate this complexity crisis.

Survivable Network Architectures demonstrate some autonomic behavior in the physical layer of the telecommunications networks, yet this is just the beginning of the autonomic vision: zero touch, self-sensing, context-aware, dynamic, self-programming and evolvable networks. To create Autonomic Networks will require the cooperation of the industry to develop open standards to evolve from the current network elements (NEs) to autonomic network elements (ANEs). From a Telco’s perspective, the physical layer tends to be outside their immediate design control as the NEs are supplied by third party vendors.

Telcos offer communications and services across a large variety of technologies. Each technology within the network: SDH (SONET in USA), PDH, ATM, IP, and so on, all have their own specific domain technology fault managers. SDH frames may be carrying ATM frames which may be carrying IP and so on. As such, at the physical layer, Autonomic Networks may resolve their own management issues, but these may have affected the traffic/service they are carrying. This can only be determined at a higher layer.

Essentially, due to complexity, the situation has arisen that a large number of uncorrelated alarm event messages may reside on a network at any one time. One estimate concerning BT’s UK network was that 95% of all alarm events raised remain uncorrelated, amounting to tens of thousands of alarm events being active at any one time.

Over time this amounts to a substantial load of data. Another concern is that these problems with root cause analysis are preventing the development of further autonemics particularly in self-healing, and with increasing mean-time to human intervention.

Autonomic Networks in themselves will not be an easy goal to achieve, yet the longer term goal of Autonomic Communications is much more than this, having commonality with Ubiquitous and Pervasive Computing, a vision of communications services anytime, anyplace, from any device, adapting to the user’s current needs and situation. Effective problem determination in the networks will assist in enabling other autonemics to advance.

The introduction of autonomic principles requires the monitoring of individual system components through sensors and the ability of those components to respond to requests through effectors. Monitoring will typically involve the correlation of several related pieces of information. Correlation is important in both self-assessment (self-awareness) and in the assessment of a component’s operating environment (environment awareness). This helps in deciding when action is required and what should be done.

By analogy with the human autonomic nervous system, event messages are similar to the electric pulses that travel along nerves. When a fault occurs in an SDH network, a series of triggered events are usually reported to the element controller (manager). The behavior of the alarms is often so complex that it appears non-deterministic [31], making it very difficult to isolate the true cause of the fault [32]. Yet at this level this is one of the primary goals of Autonomic Networks.

Currently, the skill of the operator is central to identifying faults. So, although automation prevents the immediate loss of traffic and preserves the general function of the system (as in the SNA), intervention is necessary to determine and resolve problems that arise. The promise of autonomic networks would bring about a significant reduction in the role of the operator.

IBM concurs with this assessment that root cause analysis in complex systems is key to achieving autonemics. In their white paper “Autonomic problem determination: A first step towards self-healing computing systems” [33] they state that, in effect, complexity in problem determination is diluting the effectiveness of computing in the corporate environment. The same can be said for communications and networks.

One of the major differences that the vision of autonemics brings to the existing efforts aimed at advanced automation (often including AI research) is the situated aspect — the goal being to deal with the problem as locally as possible, and within the context of the situation.

Although complex telecoms systems have automated fail-over by having alarm event messages passed off to an element manager, the vision of Autonomic Computing is to have each component with its own manager.
4.2 Space Flight Systems

Complexity in Space Systems has been well documented [18],[19],[21]. New paradigms in spacecraft design are leading to radical changes in the way NASA designs spacecraft operations [16]. Increasing constraints on resources, and greater focus on the cost of operations, has led NASA, and other agencies, to use adaptive operations and move towards almost total onboard autonomy in certain classes of mission operations [17],[18]. NASA missions, particularly those to deep space, where manned craft cannot currently being sent, are considering the use of almost wholly autonomous decision-making to overcome the unacceptable time lag between a craft encountering new situations and the round-trip delay (of upwards of 40 (Earth) minutes) in obtaining responses and guidance from mission control.

More and more NASA missions will, and must, incorporate autonomy as well as autonomy [19],[20], and the Autonomic Computing initiative has been identified as having potential to contribute to NASA’s goals of autonomy and cost reduction in future space exploration missions [18],[19],[20].

ANTS, Autonomous Nano-Technology Swarm, is a NASA concept mission that will launch sometime between 2020 and 2030 (”any day now” in terms of NASA missions). The mission is viewed as a prototype for how many future unmanned missions will be developed and how future space exploration will exploit autonomous and autonomic behavior. The mission will involve the launch of 1000 picoclass spacecraft swarm from a stationary factory ship, on which the spacecraft will be assembled. The spacecraft will explore the asteroid belt from close-up, giving scientists data that heretofore has not been available.

As much as 60% to 70% of the spacecraft will be lost on first launch as they enter the asteroid belt. The surviving craft will work as a swarm, forming smaller groupings of worker craft (each containing a unique instrument for data gathering), a coordinating ruler, that will use the data it receives from workers to determine which asteroids are of interest and to issue instructions to the workers and act as a coordinator, and messenger craft which will coordinate communications between the swarm and between the swarm and ground control. Communications with Earth will be limited to the download of science data and status information, and requests for additional craft to be launched from earth as necessary.

A current project (FAST) is studying advanced technologies for the verification of this incredibly complex mission; the reader is directed to [18],[20] for a more detailed exposition of the ANTS mission and the FAST (Formal Approaches to Swarm Technologies) project. Formal approaches to verification of such complex autonomic systems are essential, as all possible behavior cannot possibly be determined in advance, and no a priori testing plan is likely to be realistic.

4.3 Towards an Autonomic Grid

Virtualization of resources such as machine, memory, storage, and I/O are enabling virtual, collaborative organizations sharing applications and data in an open heterogeneous environment. This empowers the organization yet also creates a more complex infrastructure to manage.

A grid infrastructure promises seamless access to computational and storage resources, and offers the possibility of cheap, ubiquitous distributed computing. Grid technology is beginning to have a fundamental impact on the economy by creating new areas, such as e-Science, e-Government and e-Health, new business opportunities, such as computational and data storage services, and changing business models, such as greater organizational and service devolution [37],[38]. The Grid is a very active area of research and development; with the number of academic grids jumping six-fold in the last year [41].

Historically, the Grid arose out of a need to perform massive computation, the current direction demonstrates the potential to change the structure of electronic service provision and create a new grid service economy. The success of the Grid will be founded on the development of new grid-enabled software systems and the evolution of legacy systems to grid-enabled systems. There are many middleware frameworks for distributed computing, many modeling techniques for software artifacts, and many development processes for controlling the creation of new software systems and managing the evolution of existing software systems.

A fundamental challenge is creating correct, robust, flexible and cost-effective grid-enabled software [39]. The Grid aims to be self-configuring, self-tuning and self-healing, similar to the goals of Autonomic Computing [40]. Its aim to fulfill the vision of Corbato’s Multics [42]– like a utility company, a massive resource to which a customer gives his or her computational and storage needs [40]. As such, Autonomic Computing will be required to provide some of the answers to achieve this vision.

5. Discussion

There is a need to establish standards and mechanisms in order for Autonomic Computing to work.
For instance, it is possible to develop a self-healing tool with a control loop that constantly monitors the applications (processes) running on your laptop [35]. If any of these applications (processes) should “hang”, the autonomic tool can restart that process. Yet there is no means to inform the process where to restart (unless it is designed to do so)—effectively it’s a process being started from fresh with any previous state lost, unless the process’s application itself handles this. There is a need for standards for autonomic signals and communications to take place not only at this level—from autonomic manager to processes running on the managed component—but also from autonomic manager to autonomic manager. Allowing standard “autonomic signal” routes into processes would raise security issues, yet this will need to be part of the self-protection autonomic property.

This implies that all processes effectively need to be designed with autonomicity and self-managing capabilities in mind (not only from within, but taking direction from the external environment). This not only raises issues of standards to achieve this but raises questions as to whether current design and development approaches meet the needs for developing autonomicity, and handling human error due to complexity. The realization of self-managing systems, which will still be complex to design, may only move the human error aspect from the administrator (who had been manually managing the running systems) to the designer.

The telecommunications domain was discussed as an exemplar as its alternative evolution may place it much further down this path than the computer industry. That is, its systems have a management layer, with standards allowing heterogeneous elements to communicate management information. Consider how often our phones go down compared to our PCs or Internet connections. Yet the design of the management layer has created a complex system in itself, where, it has been claimed, 95% of the event messages under fault conditions cannot be automatically correlated, and this has created a bottleneck for further advanced automation.

It is essential that the emerging Autonomic research community find a way forward to deal with this hard problem of root cause analysis from the start and avoids this situation.

The NASA example illustrates a complex system that cannot be managed from Earth due to bandwidth limits and time delays. Moreover, it is a complex system where decisions need to be made with real-time constraints. Even without bandwidth issues, the system would likely be too complex to be managed in real-time by human beings.

Fully autonomous behavior is realistically the only alternative. But, in order for this to be successful, the mission must embody the autonomic properties of self-healing, self-protecting, self-configuring and self-optimizing. In short: the mission must be self-managing.

6. Conclusion

Autonomic computing is an emerging holistic approach to computer system development that aims to cope with complexity and bring a new level of automation and dependability to systems through self-healing, self-optimizing, self-configuring and self-protection functions.

To illustrate that autonomicity may assist with coping with complexity, examples from research in telecommunications and space systems were discussed.

While Autonomic Computing may not be a panacea for complex computer system, it clearly does have a role to play in overcoming complexity, and offers a promising antidote to some of the problems of complex systems.

Open standards and technologies are required for Autonomic Computing to reach its goals. The challenges of addressing these issues must be taken up by the wider computing community. A remit needs to be established to encourage interaction and research among researchers and developers in industry, government, and academia, in determining standards, techniques, development processes and mechanisms that can be exploited in creating self-managing systems.

As Bioinformatics produces even more complex applications, the Autonomic Computing initiative will have a role to play in ensuring the emergence of better, more reliable, and effective systems.

Acknowledgements

The development of this paper was supported at University of Ulster by the Computer Science Research Institute (CSRI) and by the Centre for Software Process Technologies (CSPT) which is funded by Invest NI through the Centres of Excellence Programme, under the EU Peace II initiative.

Part of this work has been supported by the NASA Office of Systems and Mission Assurance (OSMA) through its Software Assurance Research Program (SARP) project, Formal Approaches to Swarm Technologies (FAST), and by NASA Goddard Space Flight Center, Software Engineering Laboratory (Code 581).
References

[2] P. Horn, Invited Talk to the National Academy of Engineering at Harvard University, March 8, 2001


