Building Reliability in to IT Systems

Kutumba Velivela / Ramapantula Uday Shankar
Tata Consultancy Services

k.velivela@tcs.com/ ramapantula.shankar@tcs.com

Information Technology (IT) System reliability is in critical focus for government and business because of huge cost and reputation impacts. A well designed application will not only be failure-free but also will allow predicting failures so that preventive maintenance can take place. It will also have adequate resilience, capacity, security and data integrity.

IT reliability includes all parts of the system, including hardware, software, interfaces, support setup, operations and procedures. Due to the complexity in each of these areas, organisations are giving priority to developing end-to-end reliability-specific capabilities. These capabilities can delivered under the headings of: assessment, engineering, design, modelling, assurance and monitoring.

In this paper, we propose formal methods for developing reliability centre of excellence, with a customised maturity model, that will guarantee 5-9s availability to critical business functions. Positive effects of this approach, other than giving peace of mind to senior managers, include reduction in frequent re-design of applications, positive culture change within the organisation and increase in market share.

Keywords: IT Availability Management, Reliability, Centre of Excellence, Assessment, Engineering, Design, Modelling, Assurance, Monitoring, Metrics, Error Prevention, Fault Detection, Fault Removal, Service Level Agreement (SLA), Maintainability.

1. Introduction

“Availability Management is responsible for optimising and monitoring IT services so that they function reliably and without interruption, so as to comply with the SLAs, and all at a reasonable cost.”[ITIL OSIATIS]

Technology services failure has been making news headlines for last few years for causing extreme impacts in well established businesses and government departments. Payment / ATM failures, travel disruptions, medical operations cancellation, huge trading losses, reduced defence security, smart mobile blackout and unpaid wages headed top technology disasters in the last few years. Affected organisations include US Government, NHS, Walmart, Bank of England, M&S, Natwest, LBG, Stock exchanges, Airlines, Utilities and car manufacturers [Colin 2013] [Phil 2011] [Phil 2012]. A research summary on reasons for IT systems unavailability is included in Appendix A.

Technology Service Disruptions are costly...

- Recover from downtime 1.13hours to 27 hours
- Maximum Tolerant downtime 52.63minutes
- Average number major disaster events per year (not including medium and minor): 3.5

Figure 1 – Costs and other impacts of service disruptions
Even companies that had no major failures are hit by ever increasing hardware/software maintenance costs and delayed software deliveries. Operations are not able to cope when there is unexpected increases in faults. Essential services are being shut down with no prior notice due to communication failures and/or process failures. Backups and switching to redundant systems often does not work when needed.

Technology and IT systems reliability used to be the primary concern of installation designers and maintenance teams for more than 50 years. But, due to millions of dissatisfied customers, loss of data, fraud write-offs, regulatory fines and criminal/civil penalties, technology reliability has become a major concern for Business/IT account managers, Business analysts, IT Strategists / architects, Designers and Testers. This is more the case with safety critical, 24x7 web sites, systems software, embedded systems and other “high-availability must” applications.

This paper presents reliability-specific offerings organisations can adapt for preventing errors, detecting faults and removing them, maximising reliability and reducing the drastic impacts of failures. By using this paper as a roadmap, businesses can build IT Reliability skills which provide additional peace of mind to senior management.

2. Background

Reliability is an important, but hard to achieve, attribute of IT systems quality. These attributes are normally covered under non-functional requirements in the early stages of projects. Reliability analysis methods help identify critical components and to quantify their impact on the overall system reliability. Employing this sort of analysis early in the lifecycle saves large percentage of budget on maintenance and production support.

Hardware-specific reliability and related methods originated in Aerospace industry nearly 50 years ago and subsequently became ‘must-use’ in automotive, oil & gas and various other manufacturing industries. Arising from this appreciation of the importance of reliability and maintainability, a series of US defence standards (MIL-STDs) were introduced and implemented around 1960s. Subsequently the UK Ministry of Defence also introduced similar standards.

Reliability methods have successfully allowed hardware products to be built to satisfy high reliability requirements and the final product reliability to be evaluated with acceptable accuracy. In the recent years, many of these products have come to depend on software for their correct functioning, so the reliability of combined hardware + software components has become critically important. Even pure IT applications are dependent on hosting data centre, servers and other components to be reliable. Hence, software reliability has become important area of study for software engineers.

Even though still maturing, reliability methods have been adapted either as standard or as a best practice by a few large organizations. It is regulatory in some of these for IT systems to be certified to meet prior specified availability and reliability requirements. Many other organisations are yet to lap up this standard and reap the rich rewards through focusing on the availability criteria that all critical IT development processes must comply with.

3. Reliability Engineering = 5-9s or 99.999% Availability

In order to meet raising customer expectation for having quality software running 24X7, often defined as 5-9s in requirements, there is a need for a fundamental shift in the way IT applications are developed and maintained. Detailed hardware and software reliability requirements required to be documented and special focus is to be given to meet these from design to implementation stages. Reliability skills proposed in this paper will offer a comprehensive approach for addressing all IT reliability-related issues including capacity, redundancy, data integrity, security and maintainability.

Critical applications developed without a proper reliability approach would lead to frequent partial re-designs or full re-development because they become cumbersome to maintain. A well designed application will either be failure-free or will allow predicting failures so that preventive maintenance can take place. If a failure has safety or environmental impact, it must be preventively maintainable, preferably before it starts disrupting production.

New reliability-specific capabilities will help businesses substantially shift from reacting to failures when they happen to pro-actively manage them through approaches like Reliability Centred Analysis and Design (RCDA), covered in more detail in next section.

Setting up a separate reliability Centre-of-Excellence (COE) will not only help directly enhance business image and customer satisfaction, but also indirectly contribute to increase in market share and cost-savings. Developers who have applied these methods have described them as “unique, powerful, thorough, methodical, and focused.” The skills developed
highly correlated with attaining best-in-class levels 4 and 5 of Capability Maturity Model. Based on multiple projects experience, when done properly, Software Reliability Engineering only adds approximate maximum of 2-3% to project cost.

- 3.1 Reliability-Specific Capabilities

Reliability COE focusses on related business issues and help them efficiently meet their expectations. A combination of offerings can be provided under major headings of:

- Reliability Engineering,
- Reliability Assessment,
- Reliability Modelling,
- Reliability Centred Design Analysis (RCDA),
- Software Reliability Acceptance Testing,
- Reliability Analysis and Monitoring using specialist tools.

The methods used under the above heading are fundamentally similar but reliability offerings often have to be customised depending on different stages of development. For example, a reliability assessment offering will apply mainly to existing applications and will need some modelling, use of tools and some testing. Similarly, a reliability engineering offering applies to new or being redesigned applications and will need some assessment, modelling, RCDA, testing and tools use.

3.1.1 Reliability Engineering

Reliability Engineering involves defining reliability objectives and adapting required fault prevention, fault removal and failure forecasting modelling techniques to meet the defined objectives all through the development lifecycle. The emphasis is on quantifying availability by planning and guiding software development, test and build processes to meet the target service levels. A collaborative culture change is needed in solution architecture, application development, service delivery, operational and maintenance teams to implement this approach.

Fault prevention during build requires better development and test methods that will reduce error occurrences. Smart error handling and debugging techniques are to be adapted during design and test reviews so that faults are removed at the earliest possible time. By modelling occurrences of failures and using statistical methods to predict and estimate reliability of IT systems, more focus can be given to high risk components and Single Points of Failure (SPOFs). Refer to Figure 2 for a representation of engineering components.

Reliability engineering is a continuous process as the analysis may have to be repeated as more IT system releases are delivered. On-going improvements in fault tolerant and defensive programming techniques will be required to meet business expected targets for reliability.

![Image of reliability components](image)

**Figure 2 – Reliability Engineering Components**

3.1.1.1 Reliability Engineering Techniques

Popular Hardware techniques include redundancy, load-sharing, synchronisation, mirroring and reconciliation at different architecture tiers. Some of the software techniques include Modularity for Fault Containment, Programming for Failures, Defensive Programming, N-Version Programming, Auditors, and Transactions to clean up state after failure.

3.1.2 Reliability Assessment

Reliability Assessment can be conducted on multi-location systems, single data centres, services, servers and/or component levels. Diagram below shows three popular assessment methods and how they can be implemented together in a continuous improvement scenario. Each of the approaches can be implemented on their own as one-off exercises depending on the life cycle stage the IT system is in.
Architecture-based reliability analysis focuses on understanding relationships among system components and their influence on system reliability. This is based on the process of identifying critical components/interfaces and concentrating more on the potential problem areas and SPOFs. It assumes reliability and availability of IT systems is proportionate to corresponding measurements of its reusable hardware/software components. For example,

![Figure 3 – Reliability Assessment Methods](image)

Metric based Reliability analysis is based on the static analysis of the hardware/software complexity and maturity of the design and development process and conditions. This approach is particularly useful when there is no failure data is available, for example, when the new IT system is still in design stages. IEEE had developed a standard IEEE Std. 982.2 (1988) and a few other product metrics are available to support reliability assessors in achieving optimum reliability levels in software products. Similar vendor supplied reliability data available for hardware components and third-party components used.

The black box approach ignores information about the internal structure of the application and relationships among system components. It is based on collecting failure data during testing and/or operation and using such data to predict/estimate when the next failure occurs. Black-box reliability analysis evaluates how reliability improves during testing and varies after delivery. As pointed out in Appendix A, not adapting best practices in long-term monitoring of relevant components is one of the major reasons for IT unavailability.

A combination of these methods will be required for IT systems that require high levels of reliability.

### 3.1.3 Reliability Modelling

Over 200 models have been developed to help IT Project Managers to deliver reliable software on-time and within budget. A good practical modelling exercise can be used to initiate enhancements that improve reliability from early development phase. Based on predictive analytics concepts, different models are used depending on the type of analysis needed:

- Predict reliability at some future time based on past historical data even during design stages,
- Estimate reliability at some present or future time based on data collected from current tests,
- Estimate the number of errors remaining in a partially tested software and guide the test manager as to when to stop testing.

Like performance models, no single reliability model can be used in every situation because they are based on a number of assumptions, parameters, mathematical calculation and probabilities.

The modelling field is fast maturing and carefully chosen models can be applied in practical situations and give meaningful results.

### 3.1.4 Reliability Centred Design and Analysis (RCDA)

Reliability should be designed-in at the IT strategy level and a formalized RCDA methodology is needed to reduce the probability and consequence of failure. Various statistics have been published that prove large % of failures can be prevented by making needed changes at design stage. Successfully implemented RCDA can result in an improved productivity and reduced maintenance costs.

The focus of RCDA all through the life cycle is to ensure services are available whenever business users need it. For that to happen IT capacity has to be aligned to business needs, sufficient redundancy is built-in such that critical services still run during significant failures and data integrity/confidentiality is maintained at all times. Below is a high level flow diagram that shows a sequence of basic steps to be followed as part of RCDA:
3.1.4.1 Load-balancing and Failover

Reliable IT systems should be housed in a highly secure and resilient data centres and the solutions should be built around a redundant architecture able to ensure hardware, network, databases, and power availability as needed. Latest active/active failover, recovery and continuity mechanisms to be considered to help meet the high business availability requirements.

However, IT architects need to be careful while employing complex redundant solutions as they can often be the sources for major failures. Some of the latest major business IT failures are due to incorrect setup or inadequate testing of complex redundancy and backup solutions.

3.1.4.2 Other Design Factors

Business will not accept IT systems just because they are available 24x7. Reliable IT systems must meet various business specified requirements including performance, capacity to match business growth, security, data integrity and on-going maintainability.

[Evan 2003] identified Top-20 Key High Availability Design Principles that range from removing Single Point of Failures to keeping things simple. This kind of analysis will guide reliability designers and architects in developing customised best practices.

3.1.5 Reliability Acceptance Testing

Like all other non-functional requirements, reliability and availability for IT systems need good validation and verification phases. However, traditional software development and testing often focus on the success scenarios whereas reliability-specific testing focuses on things that can go wrong. New testing methods focus on failure modes related to timing, sequence, faulty data, memory management, algorithms, I/O, DB issues, schedule, execution and tools.

Figure 6 – Example Assurance Team Structure

Some of the methods that guide these tests are Reliability Block Diagrams (RBDs), Failure Mode Effect Analysis (FMEA), Fault Tree Analysis, Defect Classification, Operational Profiles and error handling/reporting functions. These methods help testers develop reliability-specific test cases during integration, user acceptance, non-functional, regression and deployment test phases.

Some sectors need their IT systems to be certified along with hardware components and they need reliability based acceptance criteria to be defined and met before releasing any changes into production. Given a component of IT system advertised as having a failure rate, Assurance team can analyse if it meets that failure rate to a specific level of confidence.
3.1.6 Reliability Monitoring and Analysis using Specialist Tools

Reliability is measured by counting the number of operational failures and their effect on IT systems at the time of failure and afterwards. A long-term measurement program is required to assess the reliability of critical systems. Some of the well-known software reliability metrics that can be used include Probability of Failure on Demand (POFOD), Rate of Fault Occurrence (ROCOF), Mean Time to Failure (MTTF), Mean Time Between Failure (MTTR), and Mean Time to Repair (MTTR).

Most of the analysis mentioned above can be performed by using office tools by an experienced analyst. However, a few specialized tools and workbenches available that will help in completing different types of analysis including reliability modelling and estimation/prediction. Partial list of these tools is available in references [Kishor 2013, Goel 1985]. Prediction/estimation using these tools need good understanding of analytics methods and basic probability theory.

Reliability specialist team has to master the tool related skills before recommending any of them to the customer area. Often tools related skills result in continuous source of budgets / revenue for the CoE for prolonged periods of time.

4. How to Setup Reliability CoE?

There is no one fixed method for setting up Reliability CoE and, whichever way, it is not going to be simple journey. Constructing any niche team requires commitment, hard work and support from all stakeholders. Below sample model shows, some of the factors that will bring maturity to the CoE organisation.

The maturity model similar to the above can be used as a basis for a ‘CoE development plan of action’ and as a means of tracking progress against targets.

The model above shows sample 9 central and 4 interface headings. In a real model, these headings are to be chosen in consultation with senior management and other stakeholders.

4.1 Strategy

Most organisations prefer to start with small steps when it comes to new CoEs and customise the approach as the concept catches on with more partners and customers. Here are a list of generic steps that can be followed:

- Consult with industry sponsors and outside partners,
- Appoint a talent leadership with high-level of business knowledge,
- Establish vision for the reliability practices,
- Identify software reliability champions internally and customer areas,
- Define organisation structure and secure funding,
- Start building a knowledge repository and sharing mechanisms,
- Develop action plan for each of the areas mentioned in the maturity model,
- Develop strict metrics for each area mentioned in maturity model,
- Evaluate, select, and mandate vendor products and standards,
- Collaborate with other IT consultancy areas to create reusable assets,
- Setup review and approval mechanism for deliverables,
- Seek feedback and use it for continuous improvement,
- Encourage innovation and allow challenging status-quo,
Customise to fit different customer cultures.

4.2 Processes

IT processes often constrained by resources, backlog of projects, governance processes and controls, and lack of focus on security and maintainability, fail to deliver any of the set objectives. Other than some of the generic processes like project management, software engineering, and marketing, Reliability CoE need the following for quick delivery of set availability objectives:

- an agile assessing, modelling, testing and measurement process for reliability,
- techniques that focus on error prevention, fault detection and removal,
- process to adapt for real time, online/web, batch applications,
- an early defect/ SPOFs detection framework supported by comprehensive error handling process,
- knowledge repository and reliability governance program,
- adaption programs to find better ways of working with partners, vendors and governmental departments,
- processes to identify areas which will need less effort but likely to have bigger outcome,
- review process with the aim of continuous improvement.

4.3 Technologies

Reliability technologies are fast evolving but currently there are no uniformly recognised and matured ones.

Most companies have their own selection of products and methods that fall in their own comfort zone. That means, a thorough assessment with customer engagement and a Proof-of-Concept is needed before adapting these technologies in customer areas. Below diagram shows where customer engagement and POC fits in an IT technology lifecycle.

4.3.1 Tools

A few suites of tools/workbenches available that will support reliability analyst in documenting Reliability Block Diagrams(RBDs), Fault Tree Analysis, Markov Modelling, Failure Mode and Effect Analysis (FMEA), Root cause Analysis, Weibull Analysis, Availability Simulation, Reliability Centred Maintenance and Life-cycle Cost Analysis.

The focus of these tools is mostly hardware reliability but recently they have been adapted for IT infrastructure, software and process components. A few software specific tools available that help in Software Reliability Modeling, Statistical Modeling and Estimation, Software Reliability Prediction Tools[Kishor 2013], [Allen 1999].

4.4 People

Supply of people with proven and practical reliability analysis experience is very limited. Because of this companies need to find people with partially available skills and have to train them in the rest of the areas. Below shows a good proportion of skills needed in a reliability CoE.

Other than the generic roles like project manager, business analyst, architect, operational analyst, a
few companies recruit specialist Reliability Managers and Reliability Analysts. Sample position descriptions for these roles is provided in Appendix B.

In general, staff with 6-10 years of experience in 3-4 areas of the below list could be trained into the specialised reliability roles.

- Capacity Management,
- Service Level Management,
- Configuration Management,
- Change Management,
- Test/Release Management,
- Incident Management,
- Production Support and Operations,
- Maintenance Management,
- Product Life Cycle Management,
- Vendor Management,
- Resilience and Disaster Recovery,
- Supply chain Management,
- Asset Management.

When the focus is on a particular IT application, participation from SMEs in the areas of business functions, hardware, network, process, security, software, tools, data, operations, and maintenance would be needed.

5. Conclusion

IT organisations must focus on what is going on in business areas and customise to help them efficiently meet their requirements for systems availability and reliability. Good set of reliability practices can halve the re-active fixes needed for the IT systems. The earlier they are adapted in the lifecycle the better the savings for businesses. Based on experience, up to 30% productivity gains and roughly same percentage in reduction in maintenance costs is predicted to be achievable through these practices.

Reliability is one of the characteristic of IT systems and, with systematic approach, it is possible to meet business requirements with smaller cost and minimum disruption. Implementation of any chosen reliability methods will succeed with seamless integration with current SDLC, Agile and Transformation methodologies. Marketed properly, Reliability Capabilities has good potential for generating regular income and on-going project work for commercial organisations.

Setting up separate reliability excellence team in specialist IT departments would require broader effort and participation from the strategy, architecture, assurance, tools and industry vertical solution teams. Developing a system for proper data capture, its interpretation and taking action to reflect in terms of KPIs like reliability and availability, identifying critical failure areas is the key.

Setting up the reliability CoE will not only help in giving reliability the priority it needs but also enhance organisation image and improve customer satisfaction, greatly reducing the risk of angry customers. In the long term, best reliability practices will result in positive culture change within the team as well as increased market share.

6. References

[ITIL OSIATIS] http://itil.osiatis.es/ITIL_course
/it_service_management/availability_management/
overview_availability_management/overview_availability_management.php


technology-disasters-of-2013.html - By Colin Armitage, chief executive, Original Software


[Phil 2012] http://www.business_computing_world.co.uk/top-10-software-failures-of-2012 - By Phil Codd, Managing Director, SQS.


[JBS 2013] http://www.jbs.cam.ac.uk/media/2013/research-by--cambridge--mbas-for-
tech-firm--undo-finds--software--bugs--cost-the-industry-316-billion-a-year/


Acknowledgement

Authors are grateful to Girish Chaudhari, Peter Andrew, Carl Borthwick and Jonathan Wright who reviewed the material when it was first prepared for an internal team discussion. We would also like to thank Prajakta Vijay Bhatt and the anonymous CMG referees for their comments which have helped make this paper better.
Appendix A: Surveyed Reasons for Unavailability

A survey among a few academic availability experts in 2010 ranked reasons for unavailability of enterprise IT systems [Ulrik 2010]. They identified the lack of best practices in the following areas are the causes:

- Monitoring of the relevant components
- Requirements and procurement
- Operations
- Avoidance of network failures, internal application failures, and external services that fail
- Network redundancy
- Technical solution of backup, and process solution of backup
- Physical location,
- Infrastructure redundancy
- Storage architecture redundancy
- Change Control

[Evan 2003] identified investment in the following areas will help improve the availability of IT systems.

- Good systems and admin procedures
- Reliable backups
- Disk and volume management
- Networking
- Local Environment
- Client Management
- Services and application
- Failovers
- Replication

Even though these studies does not apply in all cases, they provide useful guidelines for architects and designers of IT systems. This paper proposes more structured approach for availability management that applies to most business organisations.
Appendix B – Sample Reliability Engineer Position Descriptions

Senior Reliability Engineer – Technical IT Infrastructure

- This position is located within xxx Team in Reliability, Maintainability and Testability Support Discipline,
- xxx team has a role of increasing the availability and reducing the through life cost of ownership of IT systems for customers.

Main responsibilities:

- Own end-to-end availability and performance of customer critical services from infrastructure point of view,
- Ensure five 9s reliable experience for IT systems users located in UK and abroad,
- Liaison with customer teams and other partners to obtain Reliability data,
- Analyse, Model and interpret arising data to forecast the reliability of customer IT systems,
- Utilisation of reliability data to produce analysis and system performance reports for customers,
- Capable of technical deep-dives into code, networking, operating systems and storage problem areas,
- Respond to and resolve emergent service problems to prevent problem recurrence,
- Liaising with Design, Support, Maintenance, Procurement and Commercial functions to identify suitable recommendations for improvements,
- Understanding and interpreting IT maintenance and support information to identify root causes of IT failure,
- Attendance at customer high-level service reviews and support root cause analysis,
- Detailed IT systems analysis to support releases in different production environments,
- Representing xxx team in internal and external customer meetings,
- Participate in service capacity planning, demand forecasting, software performance analysis and system tuning activities

Minimum qualifications

- BS degree in Computer Science or related field or equivalent practical experience,
- Proven experience in similar role in a commercial organisation, using formal reliability tools and procedures,
- Good understanding of reliability, maintainability and testability practices

Preferred qualifications

- MS degree in Computer Science or related field,
- Experience with different M/F, servers, desktop systems administration and logistics,
- Expertise in data structures, algorithms and basic statistical probability theory,
- Expertise in analysing and troubleshooting large-scale distributed systems,
- Knowledge of network analysis, performance and application issues using standard tools: BMC Patrol, Teamquest or similar,
- Experience in a high-volume or critical production service environment,
- Sound understanding of understanding of IT life-cycle management and maturity gates,
- Strong leadership, communication, report writing, and presentation skills.
Senior Reliability Engineer – Software Engineering

- This position is located within the xxx Team in Reliability, Maintainability and Testability Support Discipline,
- xxx team has a role of increasing the availability and reducing the through life cost of ownership of IT systems for customers.

Main responsibilities:

- Own end-to-end availability and performance of customer critical services from software design point of view,
- Manage availability, latency, scalability and efficiency of customer services by engineering reliability into software and systems,
- Review and influence ongoing design, architecture, standards and methods for operating services and systems,
- Work in conjunction with software engineers, systems administrators, network engineers and hardware teams to derive detailed reliability requirements,
- Identify metrics and drive initiatives to improve the quality of design processes,
- Understanding of fault prevention, fault removal, fault tolerance & defensive programming design techniques,
- Liaison with customer teams and other partners to build five 9s reliability into software delivery procedures,
- Capable of technical deep-dives into code, networking, operating systems and storage design problem areas,
- Attendance at customer high-level IT design reviews,
- Representing xxx team in internal and external customer meetings,
- Participate in capacity planning, demand forecasting, software performance analysis and system tuning activities.

Minimum qualifications

- BS degree in Computer Science or related field or equivalent practical experience,
- Proven experience in similar role in a commercial organisation, using formal reliability tools and procedures,
- Good understanding of reliability, maintainability and testability practices.

Preferred qualifications

- MS degree in Computer Science or related field,
- Expertise in complexity analysis and basic statistical probability theory,
- Expertise in designing end-to-end large-scale distributed systems with full resilience,
- Experience in end-to-end infrastructure, data, applications, security and service design,
- Experience in a high-volume or critical production service environment,
- Sound understanding of understanding of IT life-cycle management and maturity gates,
- Strong leadership, communication, report writing, and presentation skills.