

A. von Mayrhauser, J. Wang, M. C. Ohlsson and C. Wohlin, "Deriving a Fault Architecture from Defect History", Proceedings International Symposium on Software Reliability Engineering, pp. 295-303, Boca Raton, Florida, USA, 1999.

Deriving a Fault Architecture from Defect History

A. von Mayrhauser and J. Wang
Computer Science Department
Colorado State University
Fort Collins, CO 80523
USA
avm,wangj@cs.colostate.edu

M.C. Ohlsson and C. Wohlin
Dept. of Communication Systems
Lund Institute of Technology, Box 118
S-221 00 Lund
Sweden
magnuso,claes.wohlin@tts.lth.se

Abstract

As software systems mature, there is the danger that not only code decays, but software architecture as well. We adapt a reverse architecting technique to defect reports of a series of releases. Relationships among system components are identified based on whether they are involved in the same defect report, and for how many defect reports this occurs. There are degrees of fault-coupling between components depending on how often two components are involved in a defect fix. After these fault-coupling relationships between components are extracted, they are abstracted to the subsystem level. The resulting fault architecture figures show for each release what its most fault-prone relationships are. Comparing across releases shows whether some relationships between components are repeatedly fault prone, indicating an underlying systemic architecture problem. We illustrate our technique on a large commercial system consisting of over 800 KLOC of C, C++, and microcode.

1. Introduction

As systems change and evolve over a series of releases, not only the code, but also the architecture can decay. Early decay identification is desirable so that steps can be taken to prevent further degradation. The question is how to identify this decay and what to do to stop it.

Software architecture consists of a description of components and their relationships and interactions, both statically and behaviorally [20]. Thus decay can be spotted via defects related to relationships and interactions of the components. A software architecture decay analysis technique must identify and highlight problematic components and relationships, and elide component relationships that are not problematic. Consequently, we must identify fault prone components and their relationships as well as elide compo-

nents and relationships that are not problematic.

Identifying components and relationships can be done either through an existing, up-to-date software architecture document, or, in its absence, through reverse architecting techniques such as [4, 6, 7, 10, 12, 16, 24, 25]. This paper tries to deal with the latter situation: an obsolete or missing software architecture document and the need for some reverse architecture effort.

We propose to use the technique by Ohlsson et al. [18] to identify the most fault prone components across successive releases. This approach has also been used with simple coupling measures based on common code fixes as part of the same defect report [17]. Because the most fault prone components identified through this method still yielded too many relationships between components, further elision was necessary to focus on and highlight the worst architectural problems. It makes sense to solve those first.

Section 2 reports on existing work related to identifying (repeatedly) fault-prone components. It also summarizes existing classes of reverse architecting approaches. Few researchers have tried to combine the two [8, 9]. We preferred two steps rather than a combination, because we wanted to use the reverse architecting approach both for building a fault architecture and a reverse architecture. Section 3 details our approach. Section 4 reports on its application to a sizable embedded system across 4 releases. The results show identifiable persistent problems with a subset of the components and relationships between them, indicating systemic problems with the underlying architecture. Section 5 draws conclusions and points out further work.

2. Background

2.1. Fault-prone Components

It is important to know which software components are stable versus those which repeatedly need corrective main-

tenance, because of decay. Decaying components become worse as they evolve over releases. Software may decay due to adding new functionality with increasing complexity as a result of poor documentation of the system. Over time decay can become very costly. Therefore it is necessary to track the evolution of systems and to analyze causes for decay.

Ash et al. [2] provide mechanisms to track fault-prone components across releases. Schneidewind [19], Khosrotaar et al. [15] provide methods to predict whether a component will be fault-prone. [18, 17] combine prediction of fault-prone components with analysis of decay indicators. [17] is more interested in fault-prone component relationships rather than components alone. It ranks components based on the number of defects in which a component plays a role. The ranks and changes in ranks are used to classify components as green, yellow and red (GYR) over a series of releases. Corrective maintenance measures are analysed via Principal Components Analysis (PCA) [11]. This helps to track changes in the components over successive releases. [17] also uses box plots to visualize the corrective maintenance measures and to identify how they differ between releases.

2.2. Reverse Architecture

Reverse architecting is a specific type of reverse engineering. According to [22], a reverse engineering approach should consist of the following:

1. Extraction: This phase extracts information from source code, documentation, and documented system history (e. g. defect reports, change management data).
2. Abstraction: This phase abstracts the extracted information based on the objectives of the reverse engineering activity. Abstraction should distill the possibly very large amount of extracted information into a manageable set.
3. Presentation: This phase transforms abstracted data into a representation that is conducive to the user.

Objectives in why code is reverse architected drives what is extracted, how it is abstracted, and how it is presented. For example, if the objective is to reverse architect with the associated goal to re-engineer (let's say into an object oriented product), architecture extraction is likely based on identifying and abstracting implicit objects, abstract data types, and their instances. This is the case with [4, 10, 12, 24]. Alternatively, if it can be assumed that the code embodies certain architectural cliches, an associated reverse architecting approach would include their recognition. [7] describes an environment that uses recognizers that know about architectural cliches to produce different architectural views of the system.

Other ways to look at reverse architecting a system include using state machine information [9], or release history [8]. CAESAR [8] uses the release history for a system. It tries to capture logical dependencies instead of syntactic dependencies by analyzing common change patterns for components. This allows identification of dependencies that would not have been discovered through source code analysis. It requires data from many releases. This method could be seen as a combination of identification of problematic components and architectural recovery to identify architectural problems.

If we are interested in a high level fault architecture of the system, it is desirable not to extract too much information during phase 1, otherwise there is either too much information to abstract, or the information becomes overwhelming for large systems. This makes the simpler approaches more appealing. In this regard, we found Krikhaar's approach particularly attractive [16]. The approach consists of three steps:

1. defining and analyzing the import relation between files. [16] defines the import relation via `#include` statements in the source code. Each file is also assigned to a subsystem (creating a part-of relation). The import relation at the subsystem level is then derived as follows: if two files in different subsystems have an import relationship, the two subsystems to which the files belong have one as well.
2. analyzing the part-of hierarchy in more general terms (such as clustering, levels of subsystems). This includes defining the part-of relations at each level. These usually will be defined differently for each level. It also includes further definition of possible import relations and their abstractions.
3. analyzing use relations at the code level. Examples include call-called by relationships, definition versus use of global or shared variables, constants and structures. Analogous to the other steps, Krikhaar [16] also determines the part-of relation and abstracts use relations to higher levels of abstraction.

Within this general framework, there are many options to adapt it to a specific reverse architecting objective [6]. For example, Bowman et al. [3] also fits into his framework: the reverse architecting starts with identifying components as clusters of files. Import relations between components are defined through common authorship of files in the components (ownership relation). Use relationships are defined as calls-called-by relationships (dependency relation) of functions in components. Bowman et al. [3] also includes an evaluation of how well ownership and dependency relationships model the conceptual relationship.

We developed our adaptation based on the need to represent defect relationships between components and the ability to focus on the most problematic parts of the architecture by quickly filtering out information.

3. Approach

The first step is to identify the most problematic parts of the system. To identify the problematic components we apply GYR [18]. This identifies decaying components over successive releases. A component is considered fault-prone in a release if it is among the top 25% in terms of defect reports written against the component. This provides a manageable number of problematic components for further analysis. The threshold can be chosen subjectively based on available resources, quality, objectives of the analysis (most problematic versus all components that have problems).

For purposes of the case study we classified a component as problematic when it is in the most fault-prone quartile for at least one release. Components are defined as collections of files in the same directory. Thus the directory structure of the software can be used as the “part-of” relationship. Fault-prone components are illustrated as leaves in this fault-filtered directory structure. We denote this as a *Fault Component Directory Structure*. Subsystems are defined through the directory structure.

The next step is to develop the *Fault Architecture*. We adapted an existing reverse architecting technique [16] to identify the fault architecture of a system and to highlight both nature and magnitude of the architectural problem. Two or more components are related, if their files had to be changed in the same defect fix (i.e. in order to correct a defect, files in all these components needed to be changed). If too many relationships are identified it is necessary to set a threshold on the lowest number of fault relationships to include. This filter reduces the number of components and relationships to the most important. In our case, we decided to further investigate components with fault-relationships in the top 10%. This number could, for example, also be the top 5% or 25% depending on the objectives of the investigation and the number of relationships.

The fault relationship can be abstracted to the subsystem and system level: Two subsystems are related, if they contain components that are related. This represents Krikhaar’s “lift” operation [16]. To indicate the magnitude of the problem, we also report the number of defect reports associated with two components or subsystems. Changes in such patterns, or persistent fault relationships between components, across releases, is an indicator of systemic problems between components and thus architecture.

The result of this phase is a series of *Fault Architecture Diagrams*, one for each release. These results are also used to update the Fault Component Directory Structure as fol-

lows: components in bold face appear in the Fault Architecture Diagrams. These are components with fault relationships to others in at least one release. Bold components are also annotated with the release identifiers in which they were considered relationship fault-prone. Non-bold components are internally fault-prone in at least one release, but do not show fault relationships with other components.

To investigate further the nature of continued problems between components, we also aggregate these diagrams into a *Cumulative Release Diagram*. The nodes represent components that occur in at least one Fault Architecture Diagram. Two nodes are related (i.e. have an arc between them), if there is a relationship between corresponding nodes in at least one Fault Architecture Diagram. The arcs are annotated as follows: if Fault Architecture Diagrams for releases n and m show an arc between components c_i and c_j , then the Cumulative Release Diagram’s arc between c_i and c_j is annotated with $T_{n,m}$ (for release transition n to m). This highlights repeatedly problematic relationships in the Fault Architecture.

4. Case Study

4.1. Environment

We applied this technique to a large embedded mass storage system of about 800 KLOC of C, C++, and microcode in 130 software components. Each component contains a number of files. We studied four releases. The data is based on defect/fix reports or source change notices (SCN). Every report indicates a problem that had to be corrected.

4.2. Identification of fault-prone components

To extract a subset of the components we decided to identify the top 25% of the most fault-prone components in each release¹. This focuses on the fault-prone relationships between components. In case of ties in rank that would cause more than 25% of the fault-prone components to be included, the smaller set was chosen. Table 1 shows to which degree components were repeatedly fault-prone. For example, 68 components were never identified as fault-prone while 3 were identified as fault-prone in all four releases. Fault-prone components in at least one release are included in the Fault Component Directory Structure (see Figure 1). We refer to a component as a collection of files in the same directory.

¹Extraction of relations for all 130 components would have provided 5400 relationships for Release 1 and 4800, 3000 and 1000 for the following releases. This is an unmanageable number from our point of view and would only obscure the most pressing problems.

Times fault-prone	0	1	2	3	4
Number of components	68	29	17	13	3

Table 1. Number of Releases in which Components were Fault-Prone

4.3. Analysis of fault relationships

This analysis extracts the fault architecture of the system. Two fault-prone components have a fault relationship between them if they each contain files that had to be fixed as part of corrective maintenance for the same defect report. All data is stored in a database which contains records with information for every defect report, including the files that had to be changed. The component fault-relationships were extracted with SQL scripts. The first half of Table 2 shows the results of this analysis. Column 2 identifies the number of components to be included in the fault architecture for each release. Column 3 states the number of relationships (arcs) between components. Column 4 lists a range for the strength of these relationships (i.e. the number of times related components shared a defect report). The number of relationships between components is very large making it difficult to get an overview of the most problematic component relationships. Therefore we decided to focus on the top 10% of fault-prone component relationships. The results from this second reduction are in the second half of Table 2. Note that this filtering step also reduced the number of components (they are omitted if the filter eliminated all arcs related to them).

Based on these results we also updated the Fault Component Directory Structure in Figure 1, marking components with fault-prone relationships to other components in bold. Bold components are annotated with the release identifiers in which they were considered relationship fault-prone. For example, *sd52* and *sd53* were relationship fault-prone in Release 2 and Release 3. The components not marked bold are fault-prone components with internal problems instead of fault-prone relationships to other components.

Figure 2 shows the Fault Architecture Diagram for Release 1. Nodes represent components. Arcs between two vertices show that components are fault-prone in their relationship. The weights on the arcs indicate the number of times this happened. Note that Figure 2 includes subdirectories at different levels of the directory structure. This was necessary when they included (changed) files. For example, both *A/c/sd21* and */A/c/cc/sd23* contain files.

4.4. Fault Architecture Diagrams

Figure 3 was constructed from the Component Level Diagram by lifting the component level relationships to the next higher level in the directory structure. For example, the components *A/c/cc/sd23*, *A/c/cc/sd24* are aggregated to *A/c/cc*. The relationship arcs of these components (to *A/b* and *A/e*) are aggregated as well. Thus the strength of the aggregated relationship between *A/e* and *A/c/cc* became the sum of the component relationships: $142 + 119 = 261$. The results from this operation are presented in Figure 3. Table 3 shows the number of components, the number of fault-relationships and their strength (as a range) for all four releases after the lift operation.

	Components	Num. relations	Strength of rel.
Re. 1	7	12	92-1391
Re. 2	7	11	45-942
Re. 3	3	3	132-562
Re. 4	8	14	13-206

Table 3. Lift Operation

Figures 3-6 show the aggregated Fault Architecture Diagrams for all four releases. The thickest arcs indicate at least 500 problems involving both components, while medium thick arcs represent between 200 and 500 problems. Thin arcs indicate 50 and 200 problems, and thin dashed arcs represent less than 50 problems.

In Release 1 most of the problems centered around subsystem *System/A* and its components. *System/A/b* and *System/A/e* are the central fault-relationship nodes indicating faults shared with many components within *System/A*. *System/A/b* was very fault-prone as a single entity.

In Release 2 some of the fault relationships persist, for example, *System/A/b* and *System/A/e* have relationships to *System/A/c* and *System/A/d*. *System/A/b* still has some internal problems. In addition, there are three new fault-prone subsystems, *System/B*, *System/Comm1* and *System/Comm2*. *System/B* only has internal problems while the other two have faults in common with subsystems *System/A/d* and *System/A/e*. *System/A/d* plays a central role in this release with regards to fault relationships. One positive development in this release is the decreasing magnitude of the fault relationships.

Release 3 does not seem to be as problematic as the other releases. Only three components have fault-prone relationships. One of them is new, *System/C*, but the fault relationships represent problems local to subsystem *System/C*. *System/B* has both internal problems and

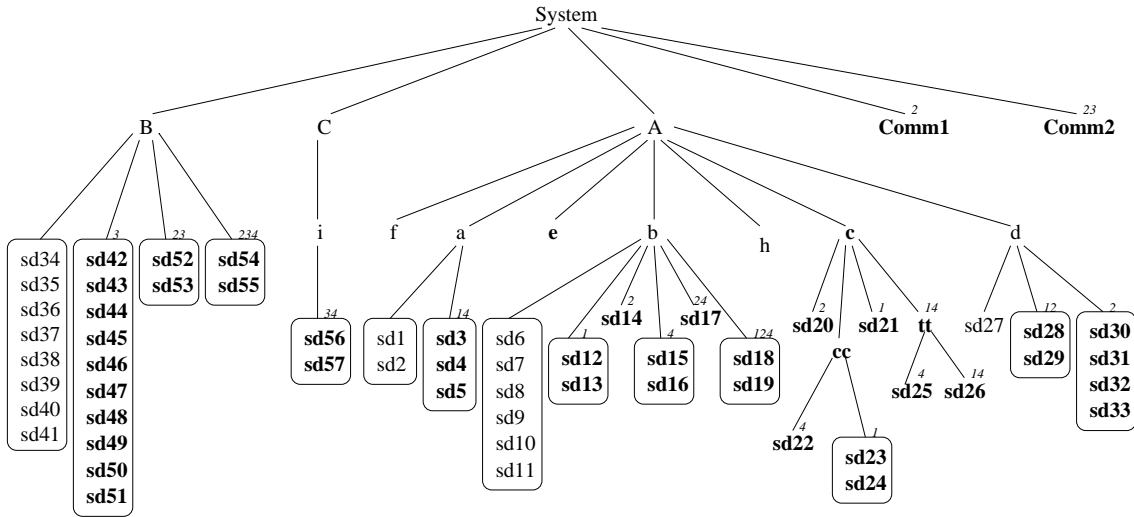


Figure 1. Fault Component Directory Structure

	Without Fault Prone Filter			With Fault Prone Filter		
	Components	Num. relations	Strength of rel.	Components	Num. relations	Strength of rel.
Release 1	29	300	1-681	16	29	70-681
Release 2	32	250	1-330	18	27	42-330
Release 3	25	50	1-200	17	22	10-200
Release 4	29	100	1-280	18	29	11-128

Table 2. Relationship Information

problems in common with System/Comm2.

In the last release the local problems in System/B and System/C persist even though there are fewer than in previous releases. Release 4 is very similar to Release 1. In both releases, relationship problems center around System/A/c which has fault-relationships with subsystems System/A/c/cc, System/A/c/tt and System/A/e. The main difference between the two releases is that System/A/d is not fault-prone in the later release and that the number of fault relationships has decreased.

4.5. Cumulative Release Analysis

Looking across the releases we can see some trends. Three releases show a large number of problems between subsystems System/A/b and System/A/e. Persistent problems for Release 1 and Release 2 involve subsystems System/A/b, System/A/c, System/A/d, and System/A/e. Even in the fourth release there are problems in the relationships between subsystems System/A/b, System/A/c, and System/A/e. We

can also see some problems resurface in System/A/a. This indicates a systemic problem that is not going away, and is a strong indicator of architectural problems in the relationship between these components. It could also be accompanied by code decay due to repeated problem fixes. A positive indication is the decreasing number of problems between System/A/a and System/A/e.

The Cumulative Release Diagram (see Figure 7) illustrates persistent problems. This diagram aggregates relationships between releases. The arc annotations in the diagram describe which fault relationships persisted across releases. For example, System/B had internal problems carried from Release 2 to Release 3, and from Release 3 to Release 4 and is therefore annotated with T23 and T34. Key fault relationship “drivers” are subsystems System/A/b and System/A/e. They have fault relationships with most other subsystems in the upper portion of the diagram. The problems internal to System/B and System/C also are persistent. The only two components without any transitions are System/Comm1 and System/Comm2. The reason is that these components only contain files and no sub-

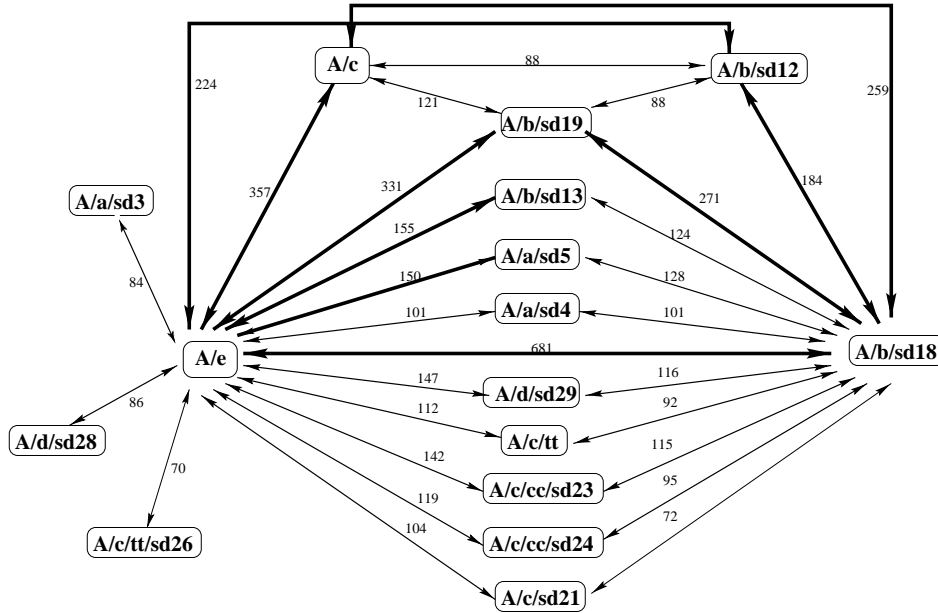


Figure 2. Fault Architecture Component Level Diagram for Release 1

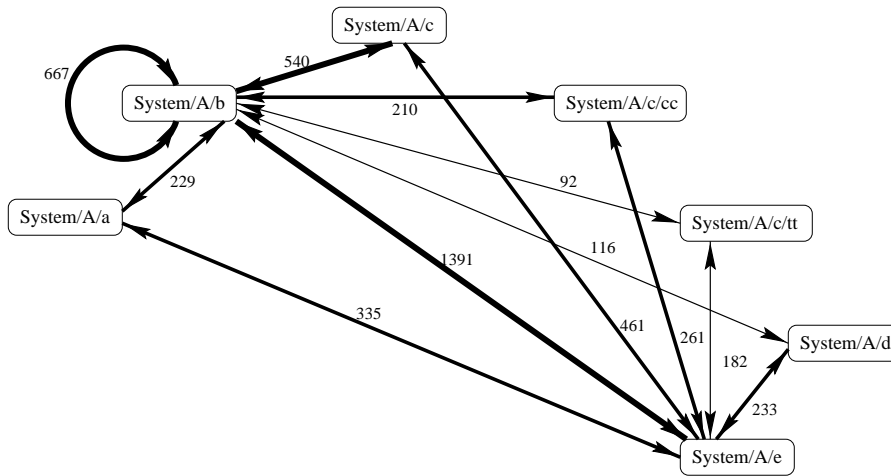


Figure 3. Fault Architecture Release 1

directories. Thus there are no relationships between components.

T14 represents an interesting phenomenon. Arcs annotated with T14 reflect relationships that were present in Release 1 and Release 4, but not in Release 2 and Release 3. One interpretation could be that the underlying architectural problems were never fully solved in Release 1 and, with addition of new features, reappeared in Release 4. We recommend to look at System/A, especially System/A/b and System/A/e in conjunction with their fault-related components which seem to be key to the biggest problems rela-

tive to multiple components.

5. Conclusions and Further Work

This paper reported a study to evaluate the usefulness of using defect reports for building fault architectures. Defect reports are easily available and could therefore be used to identify the problematic parts of a system. We have applied a technique that identifies the most fault-prone relationships between components and subsystems in a number of releases. We created a Fault Component Directory Struc-

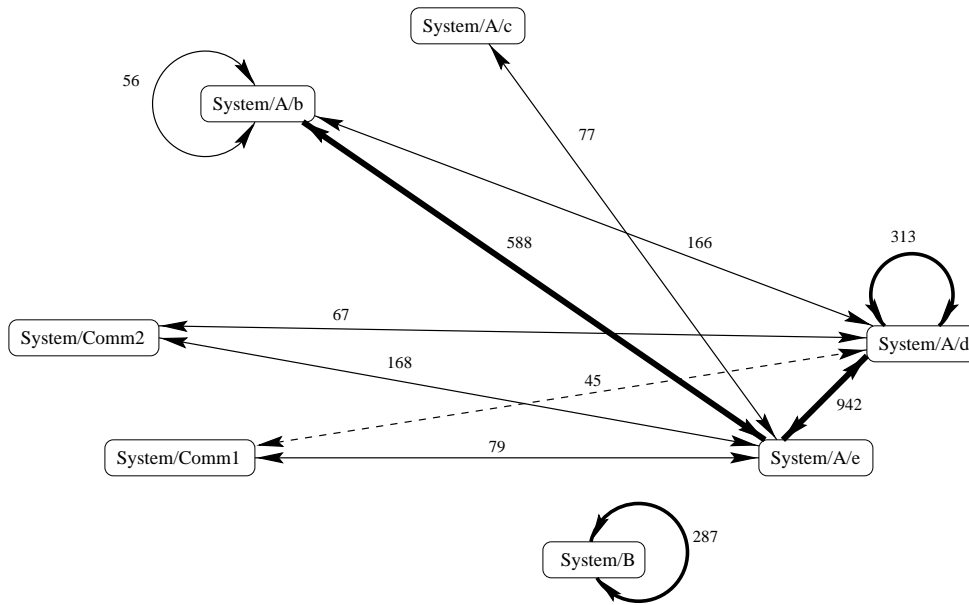


Figure 4. Fault Architecture Release 2

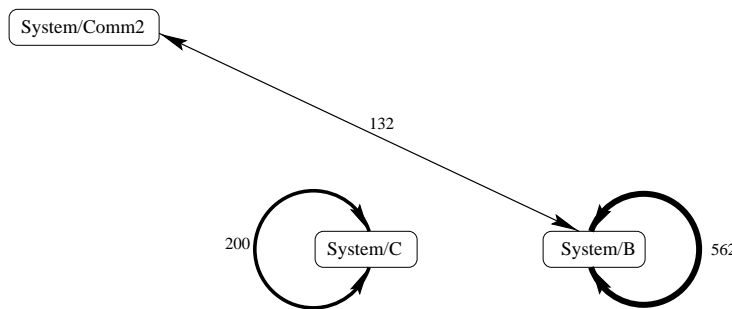


Figure 5. Fault Architecture Release 3

ture and investigated the fault-driven relationships between components. Fault Architecture Diagrams show fault-prone relationships at several levels. Finally, a Cumulative Release Diagram is created to track problematic relationships across releases.

This simple technique visualizes problems due to architecture fairly well. We were able to clearly identify for every release what the most problematic component relationships are. The most problematic stayed that way over more than one release or reappeared. Even with improvement efforts in successive releases, the core problems in the architecture, while mitigated, never disappeared completely. Lesser problems, as for example the difficulties with the System/Comm1 and System/Comm2 subsystems in Figure 4, may be due to premature release, rather than more deeply rooted architectural problems. In the latter case, they will disappear from the fault architecture. We see the key advantage to providing a fault architecture in drawing attention to the most pressing faulty component relationships. This

identifies which relationships should be scrutinized whether they require corrective maintenance or re-architecting. The most central problems seem to revolve around interactions related to System/A/b and System/A/e. These subsystems have relationships to many other components in System/A and should therefore be analyzed in more depth.

Further, root cause analysis would benefit from counting other indicators, but that depends on their availability. In [17] the same system was analyzed using Principal Components Analysis on detailed measures related to code changes in (shared) files. They derived measures for impact of change to components and to related components. The most fault-prone show more decay. Our fault architecture concentrates more specifically on the relationships between fault-prone components in terms of the magnitude of the problems in which they are involved. The fault architecture could also be used in conjunction with the analysis in [17] to further investigate architectural problems. The fault architecture identifies which components and compo-

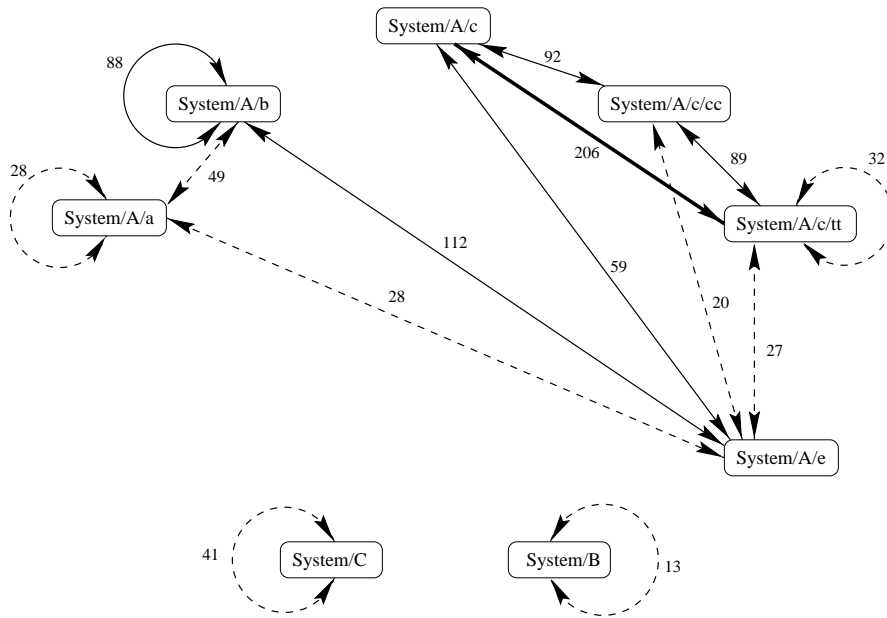


Figure 6. Fault Architecture Release 4

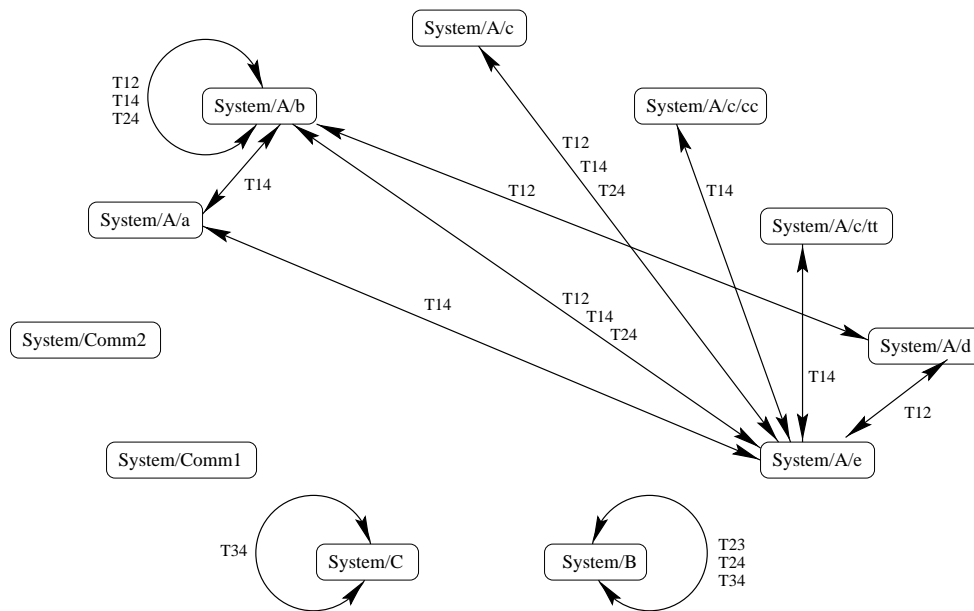


Figure 7. Cumulative Release Diagram

relationships should be analyzed further through Principal Components Analysis or box plot trends.

We would also like to apply other techniques like [8] to the defect analysis reports and compare the results. Adapting a reverse architecture technique like [16] has the advantage that it can be used to identify both the existing module

architecture as well as its fault related parts.

References

- [1] R. Allen and D. Garlan, "Formalizing Architectural Connection", *Procs. International Conference on Software Engi-*

- neering, ICSE'94, May 1994, Sorrento, Italy, pp. 71-80.
- [2] D. Ash, J. Alderete, P.W. Oman and B. Lowther, "Using Software Models to Track Code Health", *Procs. International Conference on Software Maintenance*, ICSM'94, September 1994, Vitoria, British Columbia, Canada, pp. 154-160.
- [3] I.T. Bowman and R.C. Holt, "Software Architecture Recovery Using Conway's Law", *Procs. CASCON'98*, November-December 1998, Mississauga, Ontario, Canada, pp. 123-133.
- [4] G. Canfora, A. Cimitile, M. Munro and C.J. Taylor, "Extracting Abstract Data Types from C Programs: a Case Study", *Procs. International Conference on Software Maintenance*, ICSM'93, September 1993, pp. 200-209.
- [5] E. Chikofsky and J. Cross, "Reverse Engineering and Design Recovery: A Taxonomy", *IEEE Software*, Jan. 1990, pp. 13-17.
- [6] L. Feijs, R. Krikhaar and R. van Ommering, "A Relational Approach to Software Architecture Analysis", *Software Practice and Experience*, vol. 28, 4(April 1998), pp. 371-400.
- [7] R. Fiutem, P. Tonella, G. Antonioli and E. Merlo, "A Cliche-based Environment to Support Architectural Reverse Engineering" *Proc. International Conference on Software Maintenance*, ICSM'96, Monterey, CA, pp. 319-328.
- [8] H. Gall, K. Hajek and M. Jazayeri, "Detection of Logical Coupling Based on Product Release History", *Procs. International Conference on Software Maintenance*, November 1998, Bethesda, MD, pp. 190-198.
- [9] H. Gall, M. Jazayeri, R. Kloesch, W. Lugmayr and G. Trausmuth, "Architecture Recovery in ARES", *Proceedings of the Second International Software Architecture Workshop (ISAW-2)*, San Francisco, CA, Oct. 1996, ACM Press, pp. 111-115.
- [10] J.-F. Girard and R. Koschke, "Finding Components in a Hierarchy of Modules: A Step Towards Architectural Understanding", *Procs. International Conference on Software Maintenance*, September 1997, Bari Italy, pp. 58-65.
- [11] R.L. Gorsuch, "Factor Analysis", 2:nd edition, Laurence Erlbaum Associates, Hillsdale, New Jersey, 1983.
- [12] D.R. Harris, A.S. Yeh and H.B. Reubenstein, "Recognizers for Extracting Architectural Features from Source Code", *Procs. Working Conference on Reverse Engineering 1995*.
- [13] D. Jerding and S. Rugaber, "Using Visualization for Architectural Localization and Extraction", *Procs. Working Conference on Reverse Engineering*, WCRE'97, October 1997, Amsterdam, Netherlands, pp. 56-65.
- [14] R. Kazman, L. Bass, G. Abowd and M. Webb, "SAAM: A Method for Analyzing the Properties of Software Architectures", *Procs. International Conference on Software Engineering*, ICSE'94, May 1994, Sorrento, Italy, pp. 81-90.
- [15] T.M. Khoshgoftaar and R.M. Szabo, "Improving Code Churn Predictions During the System Test and Maintenance Phases", *Procs. International Conference on Software Maintenance*, ICSM'94, September 1994, Vitoria, British Columbia, Canada, pp. 58-66.
- [16] R. L. Krikhaar, "Reverse Architecting Approach for Complex Systems", *Procs. International Conference on Software Maintenance 1997*, September 1997, Bari Italy, pp. 1-11.
- [17] M.C. Ohlsson, A. von Mayrhauser, B. McGuire and C. Wohlin, "Code Decay Analysis of Legacy Software through Successive Releases", *Procs. IEEE Aerospace Conference*, March 1999.
- [18] M.C. Ohlsson and C. Wohlin, "Identification of green, Yellow and Red Legacy Components", *Procs. International Conference on Software Maintenance*, ICSM'98, November 1998, Bethesda, Washinton D.C, pp. 6-15.
- [19] N.F. Schneidewind, "Software Metrics Model for Quality Control", *Procs. International Symposium of Software Metrics*, Metrics'97, November 1997, Albuquerque, New Mexico, pp. 127-136.
- [20] M. Shaw and D. Garlan, "Software Architecture: Perspectives on an Emerging Discipline", Prentice-Hall, Upper Saddle River, NJ, 1996.
- [21] D. Soni, R. Nord and C. Hofmeister, "Software Architecture in Industrial Applications", *Procs. 17th ICSE*, April 1995.
- [22] S.R. Tilley, K. Wong, M.-A.D. Storey, and H. Muller, "Programmable Reverse Engineering", *International Journal of Software Engineering and Knowledge Engineering*, Vol. 4, No. 4, December 1994, pp. 501-520.
- [23] V. Tzerpos and R.C. Holt, "The Orphan Adoption Problem in Architecture Maintenance", *Procs. Working Conference on Reverse Engineering*, WCRE'97, October 1997, Amsterdam, Netherlands, pp. 76-82.
- [24] A.S. Yeh, D.R. Harris and H.B. Reubenstein, "Recovering Abstract Datatypes and Object Instances from a Conventional Procedural Language", *Second Working Conference on Reverse Engineering*, July 1995, pp. 227-236.
- [25] E.J. Younger, Z. Luo, K.H. Bennett and T.M. Bull, "Reverse Engineering Concurrent Programs using Formal Modeling and Analysis", *International Conference on Software Maintenance 1996*, Nov. 1996, Monterey, CA, pp. 255-264.