Predicting and Learning the Performance of Configurable Software Systems

Lion Wagner

University of Stuttgart
Institute of Software Technology (ISTE)
70569 Stuttgart, Germany

Abstract. This is my abstract.

1 Introduction

Todays programs are mostly highly customizable. Whilst buying or downloading a program a user already decides on which version of program (s)he wants to have. On Installation there are mostly multiple Options reaching from 'Language' to a selection of software features. Once a program is installed there are usually multiple startup, customization and configuration options provided for a customer. But having such a large amount of options brings some problems into the development of modern applications. We need to have a stable development strategy that can offer the creation of multiple similar products with little redundant code or components.In practice software product lines are used for this case
Furthermore there is the problem of performance prediction. For this we introduce some Definitions. This paper will use the definition of *feature* and *configuration* as they are described in [**?**] "[...], where a feature is a stakeholder-visible behavior or characteristic of a program." and "a specific set of features, [is] called a configuration". The amount of possible configurations naturally lies in $\mathcal{O}(2^n)$. This scaling makes it hard to test each singular configuration for its performance or correctness. Especially if the configuration under test is unpredictably chosen by a user [**?**]. Nevertheless this paper will only look at the non-functional performance properties of an application. Consequently efficiently analyzing a product's performance is only partially possible and behavior beyond that has to be predicted. This opens up the challenge of efficiently and accurately predicting performance. Over time a lot of methods in different disciplines have been prosed.
This paper aims to give a short introduction into the importance of software product lines (SPL) in regards to performance engineering and an overview over solutions for predicting and learning the performance of a configurable software system. mention references TODO

2 Software Product Lines Performance Influence

TODO Intro tarditionell präprözessoren oder compiler optionen oder startup optionen

2.1 Introduction into Software Product Lines

As mentioned above modern software systems are often highly configurable and customizable. To be able to provide such a large amount of customization, software engineers adopted the concept of *mass customization* inside a product line. A technique mostly known from the automotive industry[6].[6] further describes how a general product-line functions: There is a *common platform*, which is in itself expandable and holds all basic functionalities. It also provides interfaces for adding on parts (features). Some parts are mandatory and some are optional. For each part there might be different versions or variations that are interchangeable.

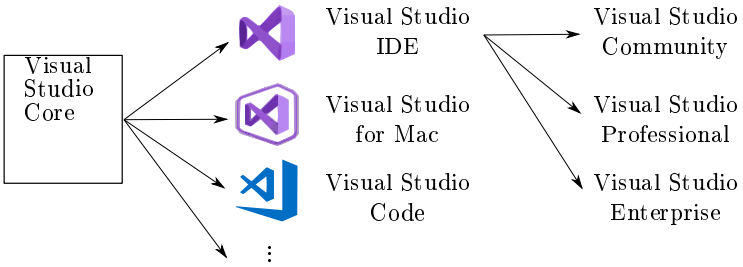

Fig. 1: Partial view of the Visual Studio Product Line [5]. (Assuming a Microsoft is using core libaries for its product. Which is likely since .Net Core runs on all major operating systems.)

Software product lines (SPL) inherit this behavior of reusing existing components to enhance the development process [2]. They are used to create a software product that is highly configurable and customizable. Products inside a SPL can even serve as a common platform for further customization. A more detailed look at the economical thought behind SPLs is given by Díaz-Herrera et al. [2]. An example of a SPL would be the Visual Studio product line from Microsoft as shown in Figure 1. The common platform are some core libraries that provide shared code to all products of this line. The products themselves add platform or usage specific features (parts) to the core to offer a broad variation of products. The Visual Studio IDE even splits up further into 3 products that all serve a different purpose, but use same code base. Note that figure 1 only displays static variations of the product line. Almost all versions are additionally dynamically expendable via plug-ins and customizable via options and preferences.

2.2 Performance Influences of Feature Binding in SPLs

The following section is based on Combining Static and Dynamic Feature Binding in Software Product Lines by Rosenmüller et al. [7].
When designing a SPL one has a choice for each feature to either bind it statically or dynamically. A statically bound feature is compiled directly into the binary of a program. Its always available to the environment when needed but uses up binary space to do so. When dynamically binding a feature it is not a part of the core binary, yet a part of an external *feature library*. Once such a feature is needed the core loads the according library from an external source. This leads to the core binary being smaller but the size needed for each feature is increased by some meta-data (size, location, format, etc. of the feature library). Multiple features can be combined into one feature library for optimization. Feature libraries are also called *binding units*. Note that [7] uses the *decorator pattern* for the realization of dynamic feature binding.
To describe the effects of binding types we use the terms *functional* and *compositional overhead*. A functional overhead originates form features (or binding units) that are loaded inside a program yet never used. In turn a compositional overhead is found as meta data that is needed to describe dynamic behavior.
An abstract example: A program core consisting out of 3 features and uses up 2MB of (hard-drive) memory. We extract 1 feature that is rarely used into a dynamic link library with the size of 1.1MB. Afterwards our core is only 1MB large. We can already conclude that our original program had a functional overhead of 1MB by subtracting the old and the new binary size. When using the extracted feature our program uses up 2.1MB of memory. So we can say that our compositional overhead for extracting the feature is 0.1MB. That is the data needed to describe the location, format, etc. of the feature library.

daten aus paper einfügen TODO

Fig 18 einbauen TODO Coming back to binding types we have the problem of finding an optimal combination of static and dynamic bindings. Rosenmüller et al. [7] suggest to prefer static binding if no dynamic extensibility is required and resources are not limited. A full static bound program has no compositional overhead and thus only requires additional (persistent and transient) memory. However programs that are more limited in their resources may need to be designed to minimize the overall overhead. This is basically the same task of finding an optimal configuration of a program. Techniques regarding the analysis and prediction will come up later in the paper reference einfügen. TODO The compositional overhead of binding units can be predicted partially since feature libaries have a constant size header (e.g. 5KB per Dynamic Link Library {DLL}). A general guideline is to avoid a large number and a large size of binding units. Overlapping binding units as well as dynamic splitting and merging of binding units are possible optimizations. Rosenmüller et al. [7] also mention that "[...] there is no optimal size for a binding unit, a domain expert can define binding units per application scenario.".

3 Measuring and Predicting the Performance of Highly Configurable Systems

As already established modern programs have a lot of customization options. With such a large amount of configuration options a developer needs some sort of mechanism to ensure his application has the required performance under most or all configurations. On one hand this is very important when it comes to contract terms and conditions with customers. A program should perform as good as the customer needs it to. Otherwise a developer may face fines. On the other hand performance prediction is helpful in finding performance problems or room for improvement. This either helps with the previous point of reaching a set performance target or to make the program more user friendly (less response time, smaller binary size, ...).
It is important to note that since we have such a large configuration space one can only talk about the average performance of either a program (considering all configurations) or the performance of a specific configuration(-family)/feature.
Why do we need to learn and predict the performance of a system? Berkley DB (C) is a database management program for embedded systems. It has 18 features and 2560 different configurations. In their paper "*Predicting Performance via Automated Feature-Interaction Detection*" Siegmund et al. [8] measured each configuration and it took 426h (=17,75d)[1]. This and other examples from the same paper show, that it is not practical to brute force measure each and every configuration possible.
By learning about the performance difference between multiple configurations it is possible accurately predict the behaviour of a program. This means that one does not need ot measure all possible configurations but instead a small sample size should be enought to predict a programs performance. This can be done in multiple ways. This paper will take a look at Automated Feature Interaction Detection, a statistic based approach and WHAT (machine/spectral learning).

3.1 Automated Feature Interaction Detection

Automated feature interaction detection (AFID) is the most straight forward approach to predicting the performance of a highly configurable system. It was developed by Siegmund et al. [8]. Unlike other methods it does not depend on machine learning but rather tires to directly identify the performance impact of each feature or a combination of features. This method reached a precision of up to 95% in the experiment conducted by Siegmund et al. [8].

Some **Formulars** are needed to descibe a Softwaresystem for AFID . The composition of using two (or more) units/features is denoted by $\cdot$. This composition is also called a configuration [3].

[1] These measurements were done on computers that are (from today's point of view) fairly slow [8, 10].

The interaction of two features is denoted by $a\#b$. By combining both we get a feature interaction:

$$a \times b = a\#b \cdot a \cdot b \tag{1}$$

This equation expresses, that when using both a and b we also need to consider their interaction $a\#b$. Note that either a or b can also be a configuration.
Further Sigmund uses an abstract performance function Π that is used to represent some performance value of a configuration:

$$\Pi(a \cdot b) = \Pi(a) + \Pi(b) \tag{2}$$
$$\Pi(a\#b) = \Pi(a \times b) - (\Pi(a) + \Pi(b)) \tag{3}$$
$$\Pi(a \times b) = \Pi(a\#b) + (\Pi(a) + \Pi(b)) \tag{4}$$

Following that the performance of a program $P = a \times b \times c$ can be written down as

$$\Pi(P) = \Pi(a) + \Pi(b) + \Pi(c) + \Pi(a\#b) + \Pi(a\#c) + \Pi(b\#c) + \Pi(a\#b\#c). \tag{5}$$

The Problem with the equations 2-5 is that they assumes that we can measure the performance of a feature in isolation. This is in general not possible [8]. Also we are still in the space of $\mathcal{O}(2^n)$ of possible configurations that we need to measure. To reduce this Sigmund et al. uses a interaction *delta*.

$$\begin{aligned} \Delta a_C &= \Pi(C \times a) - \Pi(C) \\ &= \Pi(a\#C) + \Pi(a) \end{aligned} \tag{6}$$

Where C is a base configuration. This formula describes how the performance influence (Δ) of a on a configuration C can be calculated. Its either the performance difference between using C with and without a, or the performance influence of a itself plus the influence of the interaction between C and a.
As a general approach to reduce its search space AFID looks at:

$$\Delta a_{min} = \Pi(a \times min(a)) - \Pi(min(a)) \tag{7}$$

and

$$\Delta a_{max} = \Pi(a \times max(a)) - \Pi(max(a)) \tag{8}$$

Where $min(a)$ is a valid minimal configuration not containing a but to which a can be added to create another valid configuration. $max(a)$ is a valid maximal configuration not containing a but to which a can be added to create another valid configuration.

For **AFID** one first needs to define when a feature is interacting. For this Siegmund et al. [8] use the definition of

$$a \text{ interacts} \Leftrightarrow \exists C, D | C \neq D \wedge \Delta a_C \neq \Delta a_D. \tag{9}$$

$C = min(a)$ and $D = max(a)$ are chosen to find interacting features and to reduce the search space for C or D from $\mathcal{O}(2^n)$ to $\mathcal{O}(n)$. By measuring $\Delta a_{min(a)} = \Delta a_C$ and $\Delta a_{max(a)} = \Delta a_D$ for each feature some first information about their behavior can be obtained. If both values for a feature a are similar it does not interact with the features of $max(a) \backslash min(a)$. Otherwise a is marked as interacting. In both cases it can still interact with the features of $min(a)$. In total 4 measurements per feature are required ($\Pi(a \times min(a))$, $\Pi(min(a))$, $\Pi(a \times max(a))$, $\Pi(max(a))$)[8].
Since most of the interacting features are known by now one can look for the groups of features whose interaction does have an influence on performance. Again the problem arises that there is an exponential number of possible combinations. Three heuristics are used to simplify the finding of these groups.

Pair-Wise Heuristik (PW): Most groups of interacting features appear in the size of two[8, 4]. So it makes sense to look for pair interaction first.

Higher-Order Interactions Heuristic (HO): Siegmund et al. [8] only look at higher order interactions of the rank of three. More on this later.

Hot-Spot Features (HS): Based on [1, 9] Siegmund et al. [8] assume that hot spot features exist. "[...] There are usually a few features that interact with many features and there are many features that interact only with few features.", these features are the hot spot features.

Using a SAT-Solver an implication graph as seen in Figure 2 is generated. Each implication chain in this tree should have at least one interacting feature. When analysing the tree each chain is walked from the top down. The three heuristics will be applied in the order of PW $\rightarrow$ HO $\rightarrow$HS.

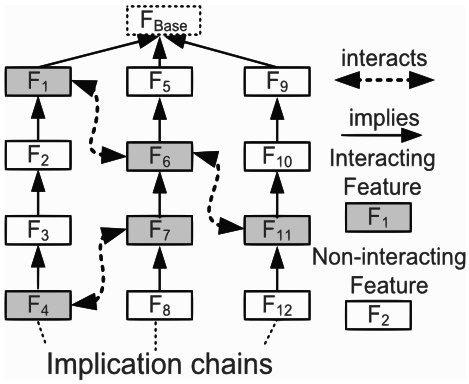

Fig. 2: Implication tree example found in [8]

First the influence of every feature on another chain is measured (PW-heuristic). In the example of Figure 2 the interactions would be measured in this order:"$F1\#F6, F1\#F7, F4\#F6, F4\#F7, F6\#F11, F7\#F11, F1\#F11, F4\#F11$"[8]. If an interaction impact $\Delta a\#b_C$ exceeds a threshold it is recorded.

Secondly, the higher order interaction heuristic is applied. Higher order interactions can be relatively easily found by looking hat the results of the PW-Heuristik. Three features that interact pair-wise are likely to interact in a third order interaction. For

example, looking at features a, b and c- If $\Delta a\#b_{C1}$ and $\Delta b\#c_{C2}$ have been recorded $\{a\#b, b\#c, a\#c\}$ all have to be non zero to find a third order interaction. Interactions with and order higher than three are not considered to prevent too many measurements.

Lastly Hot-Spot features are detected (HS-heuristic). This is done by counting the interactions per feature. If the number of interactions of a feature is above a certain threshold (e.g. the arithmetic mean) it is categorized as a Hot-Spot feature. Based on the hotspot features further third order interactions are explored. Again higher order interactions are not considered to prevent too many measurements.

After applying the three heuristics all detected interacting features or feature combinations are assigned a Δ to represent their performance influence on the program.

Siegmund et al. [8] tested AFID on six different SPLs (Berkely DB C,Berkely DB Java, Apache, SQLite, LLVM, x264). Each program was tested under four approaches: Feature-Wise, Pair-Wise, Higher-Order, Hot-Spot (in this order). Each approach also used the data found by the previous one. Accordingly the results get better the more heuristics are used as seen in Table 1. Using only the FW approach means that interactions (and the heuristics) are not considered, yet the accuracy is already at about 80% on average. A significant improvement can be made by using the PW heuristic. It uses on average 8.5 times more measurements than the FW approach but improves the accuracy to 91%. Using the HO or HS approach improves the accuracy further by about 2-4%. However for Apache using the HO over the PW approach even deteriorated the average result by 3.9% and doubled the standard variation. As already mentioned using the HS approach gives the best accuracy this is true for all 6 tested applications. Siegmund et al. [8] also notes that analysing SQLite only needed about 0.1% of all possible configurations. This hints to the good scalability of AFID.

Table 1: Results of average accurcy found by Siegmund et al. [8]

Approach	avg. Accuracy
FW	79.7%
PW	91%
HO	93.7%
HS	95.4%

4 Measuring Option and Problems

5 Prediction

References

[1] S. Apel and D. Beyer. Feature cohesion in software product lines: An exploratory study. In *Proceedings of the 33rd International Conference on Software Engineering*, ICSE '11, pages 421–430, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0445-0. doi: 10.1145/1985793.1985851. URL `http://doi.acm.org/10.1145/1985793.1985851`.

[2] J. L. Díaz-Herrera, P. Knauber, and G. Succi. Issues and models in software product lines. *International Journal of Software Engineering and Knowledge Engineering*, 10(4):527–539, 2000. doi: 10.1142/S0218194000000286. URL `https://doi.org/10.1142/S0218194000000286`.

[3] J. Guo, K. Czarnecki, S. Apel, N. Siegmund, and A. Wasowski. Variability-aware performance prediction: A statistical learning approach. In *Automated Software Engineering (ASE), 2013 IEEE/ACM 28th International Conference on*, pages 301–311. IEEE Press, 2013. ISBN 978-1-4799-0215-6. doi: http://dx.doi.org/10.1109/ASE.2013.6693089.

[4] J. Liebig, S. Apel, C. Lengauer, C. Kästner, and M. Schulze. An analysis of the variability in forty preprocessor-based software product lines, 2010. URL `http://doi.acm.org/10.1145/1806799.1806819`.

[5] Microsoft. `https://visualstudio.microsoft.com/de/products/`, Accessed: 10.05.2019.

[6] K. Pohl, G. Böckle, and F. J. van der Linden. *Software Product Line Engineering: Foundations, Principles and Techniques.* Springer-Verlag Berlin Heidelberg, Tiergartenstraße 17, 69121 Heidelberg, 1 edition, 2005. ISBN 3540243720. doi: 10.1007/3-540-28901-1.

[7] M. Rosenmüller, N. Siegmund, G. Saake, and S. Apel. Combining static and dynamic feature binding in software product lines. technical report 13, 2009.

[8] N. Siegmund, S. S. Kolesnikov, C. Kästner, S. Apel, D. Batory, M. Rosenmüller, and G. Saake. Predicting performance via automated feature-interaction detection. In *In Proc. of ICSE*, pages 167–177. IEEE, 2012.

[9] C. Taube-Schock, R. Walker, and I. Witten. Can we avoid high coupling? pages 204–228, 07 2011. doi: 10.1007/978-3-642-22655-7_10.

[10] techpowerup. `https://www.techpowerup.com/cpudb/`, Accessed: 17.05.2019.