Quality-Driven Software Architecture Composition[1]

Hans de Bruin   Hans van Vliet

Vrije Universiteit Amsterdam

Mathematics and Computer Science Department

De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

e-mail: {hansdb,hans}@cs.vu.nl

---

[1] A shorter version of this paper will appear in the ECBS'02 conference proceedings.

Abstract

This paper discusses an approach for the top-down composition of software architectures. First, an architecture is derived that addresses functional requirements only. This architecture contains a number of variability points which are next filled in to address quality concerns. The quality requirements and associated architectural solution fragments are captured in a so-called Feature-Solution (FS) graph. The solution fragments captured in this graph are used to iteratively compose an architecture driven by quality requirements. Our versatile composition technique allows for pre- and post-refinements, refinements that involve multiple variability points, and functionality extensions. In addition, the usage of the FS graph supports Aspect-Oriented Programming (AOP) at the architecture level.

1   Introduction

The architecture of a software system captures early design decisions. These early design decisions reflect major quality concerns, including functionality. In order not to reinvent the wheel time and again, we would like to capture chunks of architectural knowledge explicitly, and use these chunks when deriving an architecture that fulfills some set of quality concerns. Our solution for capturing this knowledge is a Feature-Solution (FS) graph, which connects quality requirements with solution fragments at the architectural level (de Bruin and van Vliet, 2001). The present paper is concerned with systematically composing an architecture out of the solution fragments captured in the FS-graph.

The process we envisage for deriving the architecture is an iterative, quality-driven approach to software architecting (see for instance (Bosch, 2000)). The first step in this process is the derivation of a software architecture that meets the functional requirements set. This is called the reference architecture. Next, the attention focuses on non-functional requirements by iteratively applying known design solutions (e.g., architectural and design patterns) to refine the software architecture. Typically, this requires several iterations. These iterations might also involve backtracking steps because we usually have to deal with conflicting requirements.

At the heart of this iterative, quality-driven process for composing software architectures is the Feature-Solution (FS) graph. In this FS graph, requirements are connected with design solutions. On the basis of requirements specified in the feature space of the FS graph, solutions that are likely to meet these requirements are selected in the solution space. One way to look at the FS graph is that it embodies domain knowledge (expressed as requirements) along with design solutions for solving particular problems in that domain. The role of the FS graph in this iterative approach has been discussed extensively in (de Bruin and van Vliet, 2001).

In this paper we focus on a composition technique to systematically derive a software architecture by recursively applying design solutions to a given reference architecture. The reference architecture contains variability points at those places where the architecture is expected to be varied, for instance, to cater for non-functional requirements. The variations (e.g., design solutions in the form of patterns) in their turn may contain variability points as well. So in principle, the architecture can be refined indefinitely.

We may call our approach a top-down approach to software architecture composition, to contrast it with the (bottom-up) composition and (top-down) decomposition approaches that are well-known within software engineering, even though this may sound as a contradiction in terms. In a bottom-up composition, we start with a set of elementary components and aggregate these into higher level components using some glue mechanism. The abstraction level is raised each time we compose a component out of subcomponents. In a system decomposition, the opposite direction is followed.

We then start with a top level view of the system and recursively decompose it into manageable pieces. Thus, system decomposition lowers the abstraction level. In our approach, the starting point is a reference architecture, a description of basic functionality with variability points. The variability points are next used to plug-in existing design solutions that address certain quality concerns. These plug-ins in turn may contain variability points as well which are next used to further fine-tune the solution. This leads to a top-down approach of system composition.

The FS graph plays a key role in the top-down, iterative composition process. Because the FS graph is supposed to contain relevant domain knowledge for system construction, it is also used to explore design alternatives that have a system-wide impact. For instance, it is quite possible to change the interaction style of a system from user-centered (i.e., the user is in control) to system-centered (i.e., don't call the system, the system will call you). This can be seen as the architecture-level counterpart of aspect weaving as used in aspect-oriented programming languages (Kiczales et al., 1997).

Use Case Maps (UCM) (Buhr, 1998; Buhr and Casselman, 1996) are used in this paper as a vehicle to demonstrate the principles. UCM is a diagrammatic modeling technique to describe behavioral and, to a lesser extent structural, aspects of a system at a high level of abstraction. UCM provides stubs (i.e., the hooks or variability points) where the behavior of a system can be varied statically at construction time as well as dynamically at run time. The advantage of UCM is that it focuses on the larger, architectural issues, and its support of plug-ins and stubs, both static and dynamic. However, the use of UCM is not a prerequisite; the composition technique can also be applied using, for instance, UML.

The contributions of this paper are twofold. Firstly, we present a versatile composition technique for software architectures. This technique offers pre- and post refinement and functionality extension, supports multiple plug-ins in a design solution, and offers rules to enforce the well-formedness of refinements. Secondly, the usage of the FS graph supports a guided form of top-down composition and Aspect-Oriented Programming (AOP) at the architecture level.

The remainder of this paper is organized as follows. Section 2 gives a brief introduction to Use Case Maps. Section 3 explains the basics of our technique for top-down composition. In section 4, we illustrate the role of the FS graph in system composition. Section 5 illustrates the use of the FS-graph to realize global transformations to a software architecture. In section 6, the composition model is extended to cater for functionality extensions. Section 7 discusses how the concept of refinement can be formalized. Finally, section 8 discusses related work, and section 9 contains our conclusions.

2 A Brief Introduction to Use Case Maps

A Use Case Map (UCM) is a graphical notation for humans to use to understand the behavior of a system at a high level of abstraction. It is scenario-based, showing cause-effects by traveling over

paths through a system. UCMs do not have clearly defined semantics, their strong point is to show how things globally work.

The basic UCM notation is quite simple. It has three basic elements: responsibilities, paths and components. A simple UCM exemplifying these basic elements is shown in Figure 1. A path is executed as a result of the receipt of an external stimulus. Imagine an execution pointer is then placed on the start position. Next, the pointer is moved along the path thereby entering and leaving components, and touching responsibility points. A responsibility point represents a place where the state of a system is affected or interrogated. The effect of touching a responsibility point is not defined since the concept of state is not part of UCM. Typically, the effects are described in natural language. Finally, the end position is reached and the execution pointer is removed from the diagram.

A UCM is concurrency neutral, that is, a UCM does not prescribe the number of threads associated with a path. By the same token, nothing is said about the transfer of control or data when a pointer leaves one component and (re-)enters another one. The only thing that is guaranteed is the causal ordering of executing responsibility points along a path. This is not necessarily a strict temporal ordering. The execution of a responsibility point may overlap with the execution of subsequent responsibility points.
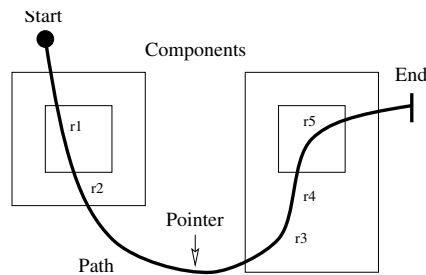

Figure 1: UCM basic elements.

A more realistic example is shown in Figure 2, depicting a distributed client-server system. Because the client communicates with the server over a network that can fail occasionally, a proxy server is included to provide transparent access to the real server. The proxy server is modeled as a stub for which two implementations are given: a transparent proxy server which passes the requests to and the replies from the server unaltered thereby denying the possibility of network failures, and a proxy server with a timeout facility with which failures are detected. The notation used in the figure is supposed to be self-explanatory.

It is interesting to note that many things are unspecified in UCMs, but the intended meaning is suggested strongly. For instance, distribution aspects (e.g., connection mechanisms and the amount of concurrency in a component) are not dealt with. However, the client, the server and the proxy server

3

are distinct components that are connected by a network, which is also modeled as a component. By using these names, it is natural to assume that the components are distributed over a number of computer systems. But again, it is not specified, it is all in the eye of the beholder.
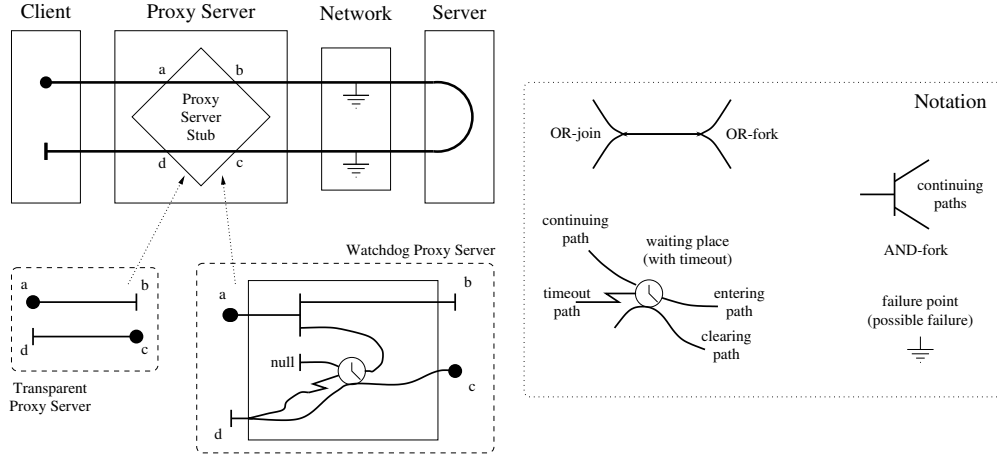

Figure 2: Distributed Client-Server UCM.

3   The Basic Composition Technique

In (de Bruin and van Vliet, 2001), we described a very simple, but effective way of system composition through successive refinements. The basic idea is to provide a UCM with stubs in which plug-ins can be placed. A plug-in may contain a stub as well, so a plug-in can be placed in a plug-in, and so on. This approach was used, amongst others, to secure the communication between WEB-browsers and a WEB-based system by first adding encryption/decryption components and next a firewall component.

This idea works out fine for local refinements, that is, refinements such as encryption and firewall that apply to *one* variability point (UCM stub) only. It falls short when we have to deal with refinements that should be applied to multiple variability points. An example of the latter type of refinement is the use of the Observer design pattern, which affects both a component in the role of subject and one or more components in the role of observers. The solution presented below can handle refinement of multiple variability points.

The basic concepts of refinement are explained using the example shown in Figure 3 and Figure 4. This first example again deals with encryption/decryption and firewall refinement. This time, though, we model encryption and decryption as one refinement with two variability points, which gives greater flexibility than the solution presented in (de Bruin and van Vliet, 2001). The example involves two components, Peer1 and Peer2. Component Peer1 communicates with component Peer2. Peer1 provides a socket for sending data, whereas Peer2 provides a socket for receiving data. The communi-

4

cation between the peers is refined by first adding encryption and decryption components in Peer1 and Peer2, respectively, and next a firewall component in Peer2. The meaning of the UCM elements used for refinement is given in Table 1. Notice that we have two options for firewall placement in this example. Either the firewall component is placed before the Decryptor component, or it is placed after it, depending on whether the firewall is placed in the Pre-stub or in the Post-stub of the Rcv (Receiver) socket in Peer2, respectively.

A basic UCM modeling peer to peer communication in which each peer provides a socket for refinement:
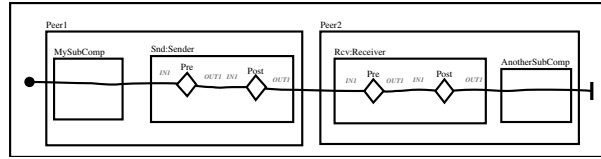
A refinement containing a plug-in for encryption and decryption:
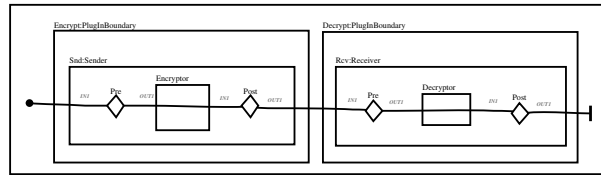
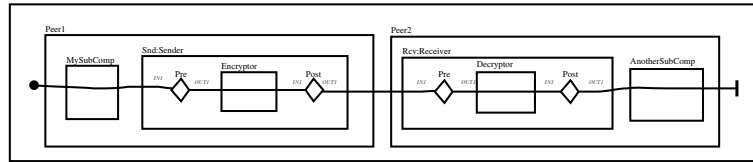The result of refining the basic UCM with encryption and decryption:

Figure 3: Refinement basics.

Refinements have to obey certain rules. These rules enforce the well-formedness of the transformation between successive refinements of architectures. These rules are easily understood by viewing a refinement and the UCM to which it is applied as patterns that should match. The rules for refinement are as follows:

Type matching. Sockets and plug-ins have names and type identifiers, specified as a Name:Type pair. Types are organized in type hierarchies supporting single as well as multiple inheritance. A plug-in can be placed in a stub of a socket only if the plug-in and socket have matching types. That is, the type of a plug-in must be identical to or be a subtype of the stub's type.

Plug-in matching. Three kinds of plug-ins are recognized:

A refinement containing a firewall plug-in:
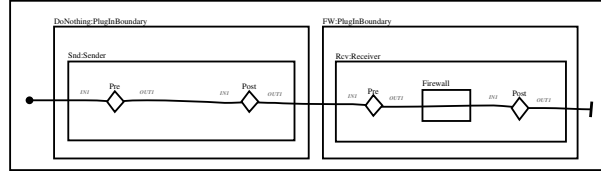
The result of placing the firewall refinement in the Pre-stub:
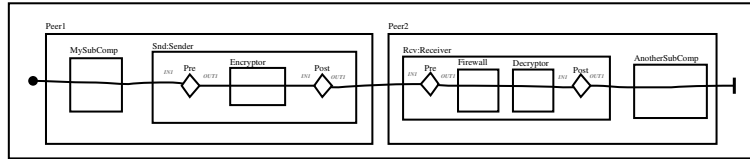
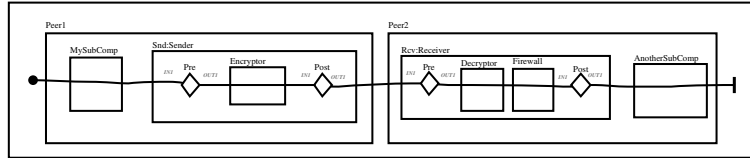The result of placing the firewall refinement in the Post-stub:

Figure 4: Refinement basics (continued).

Mandatory. All mandatory plug-ins in a refinement must be placeable in sockets with matching types.

Optional. Only if a socket is provided for an optional plug-in, the plug-in will be placed in the socket, otherwise it is discarded.

Extension. In contrast to an optional plug-in, no socket needs to be provided, i.e., an extension is never discarded.

Having optional and extension plug-ins allows us to capture broadly applicable design patterns in quite general UCM refinements that can next be applied in a variety of circumstances. The plug-in kind is specified syntactically as Name:Type (Plug-in kind)$_{optional}$. If no plug-in kind is specified, the plug-in is qualified as mandatory.

Boundary matching. All plug-ins enclosed in a plug-in boundary must be placed in one component.

Structure matching. If a UCM path (indicated by a straight or curved line) connects two components, that means that those components communicate with each other, in some way or another. There then exists a structural relationship between those components. There are two ways to specify further details about this relationship: the communication may be one-to-one or one-

6

Element	Meaning
Socket and Stubs	A socket is a placeholder that provides at most two (UCM) stubs in which a (UCM) plug-in can be placed. One stub is called Pre, the other one Post. The intention is that the Pre-stub provides the ability for pre-processing. Likewise, the Post-stub offers the possibility for post-processing. Stubs are depicted as diamonds.
Plug-In	A plug-in can be placed in either the Pre-stub or the Post-stub of a socket. It may itself contain a Pre-stub and Post-stub, thereby creating the opportunity to recursively plug in plug-ins. If a plug-in is placed in the Pre-stub, the Post-stub of the plug-in is discarded, and vice-versa for Post-stub placement.
Plug-In-Boundary	A plug-in-boundary encloses a set of plug-ins that should be placed in a single component. It is used as a mechanism to ensure that plug-ins that belong together are not scattered over a number of components. A plug-in boundary is depicted as a rectangle surrounding its plug-ins.
Refinement	A refinement is a combination of plug-ins enclosed in one or more plug-in boundaries that can be placed in stubs provided by sockets.

Table 1: The meaning of UCM refinement elements.

to-many, and the communication may be direct or indirect. Since components reside within plug-in boundaries, we may also say that these relationships exist between plug-in boundaries.

The structural relations between plug-in boundaries in a refinement must correspond with the structural relations provided by sockets.

A path to a stack of components denotes a one-to-many relation. For instance, in the Observer design pattern, a stack of observers may be connected to a single subject. A structural one-to-many relation in a refinement matches a one-to-one relation in the target UCM, but not vice versa.

A distinction is further made between indirectly and directly connected plug-ins in a refinement. In case of indirectly connected plug-ins (as indicated by a UCM end symbol attached to a start symbol), the corresponding sockets need not necessarily have a direct connection. All that is required is that there is a UCM path from one socket to the other one, i.e., the path may go through an arbitrary number of intermediate components and sockets. The latter is not permitted for direct connections.

To illustrate the finer details of the refinement process, consider again the example of peer to peer communication. In the example in Figure 3 and 4, we have abstracted away from how the peers communicate. One way of realizing this is by deploying the Observer design pattern, which can be

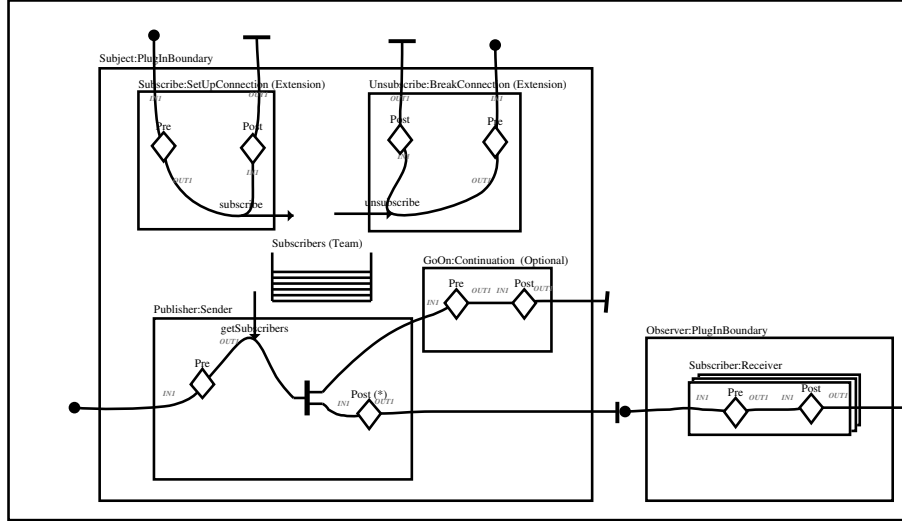captured in a UCM refinement as shown in Figure 5.

Figure 5: Observer design pattern captured in a UCM refinement.

Figure 6 shows how the Observer refinement can be plugged in the peer to peer UCM with encryption/decryption and firewall components. Firstly, notice that the types specify where plug-ins are placed. The plug-in of type Sender fits into Peer1, while the plug-in of type Receiver fits into Peer2. Both these plug-ins are mandatory. There is an optional continuation plug-in in the Observer design pattern depicted in Figure 5 for which no socket is provided, so that one is discarded in the refinement. There are two extension plug-ins which are placed into Peer1 (because they are contained within the plug-in boundary that also contains the Publisher:Sender plug-in), which now contains functionality for subscribing and unsubscribing. The structural relationship between the two plug-in boundaries of the Observer refinement are one-to-many and indirect, so it fits the one-to-one direct relationship between the two peers in Figure 3 and 4.

A few remarks are in order here. Firstly, by using a specific solution captured in a UCM refinement, more and more details are added to the basic UCM we started with in the first place. We call this process abstraction lowering, which is discussed further in the next section. Conversely, the Observer design pattern is captured in a general UCM refinement that can be used in a variety of circumstances. Secondly, a flexible refinement mechanism is provided in the form of Pre-stubs and Post-stubs. In this particular case, the multicast to interested observers can be positioned either before or after the encryption component (in Figure 6, we have chosen to place it before the Encryptor).
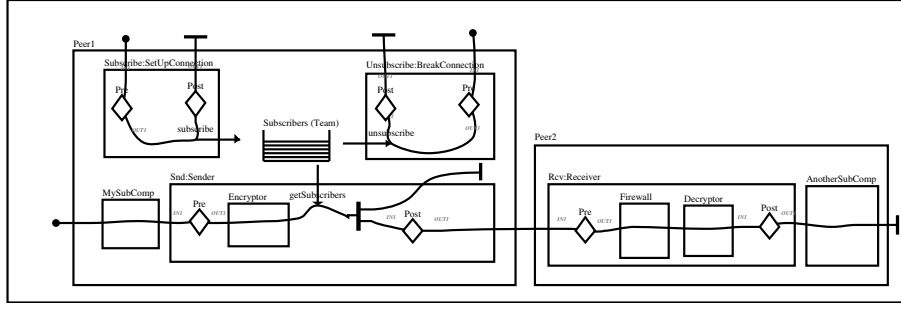
8

Figure 6: Lowering the abstraction level by deploying the Observer pattern.

4 Abstraction lowering and Refinement

By starting at the top-level system structure, each refinement in the form of plug-in placement lowers the abstraction level. An example was given in the previous section in which the Observer pattern refinement resulted in the extension of the peer to peer system with subscription management. In our approach, the architectural knowledge of these refinements and the quality concerns they address are captured in a Feature-Solution (FS) graph.

As an example, consider again peer to peer communication. A FS graph that details how the basic peer to peer UCM can be refined is given in Figure 7. Two spaces are recognized in the FS graph. The Feature (F) space contains the requirements, whereas the Solution (S) space contains solutions addressing these requirements in the form of plug-ins that can be placed in the basic peer to peer UCM. Features as well as solutions are decomposed in AND-(EX)OR decomposition trees. An AND decomposition of a node in either the feature or the solution space means that all its constituents must be available, an OR requires an arbitrary (≥ 0) number of constituents, and an EXOR requires precisely one constituent. The key idea is that a feature in the F-space may select a solution in the S-space as defined by directed selection links between nodes (indicated by a solid line). It is also possible to explicitly rule out a particular solution. This is done by connecting a feature to a solution with a *negative* selection link (indicated by a dashed line).

In the example, we focus on non-functional requirements, in particular flexibility, security and performance requirements. If a high flexibility level is desired, the FS graph dictates that we should use the Observer design pattern, because of its properties of reducing the coupling between peers and supporting multiple observers. On the other hand, if we want high performance, the FS graph selects a direct invocation style in the form of (remote) procedure calls. It is interesting to observe that a high level of flexibility and a high level of performance cannot be obtained simultaneously since these requirements select solutions that rule out each other, as implied by the EXOR decomposition of the communication node. Thus, a FS graph contains trade-off information as well. Typically, several design process
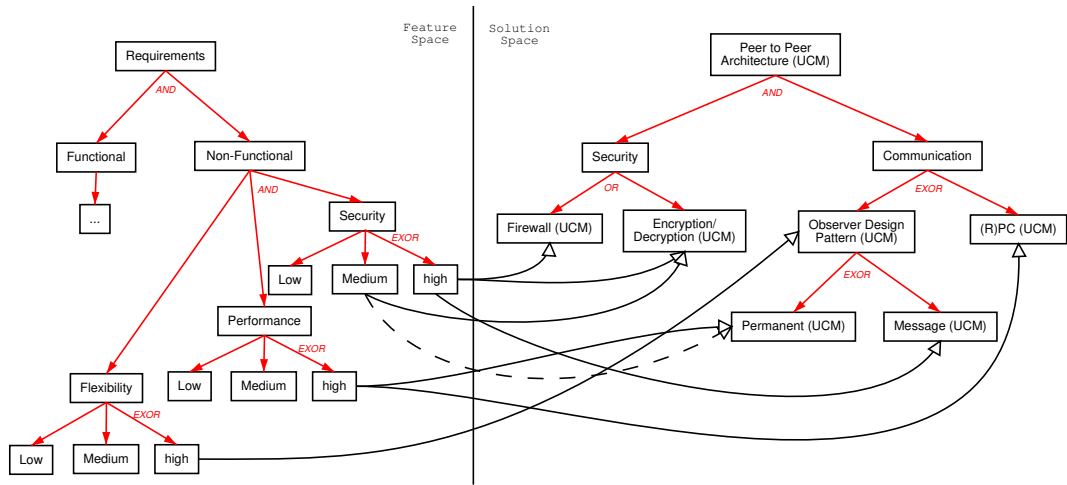
9

Figure 7: Feature-Solution graph for peer to peer communication.

cycles are required to arrive at an architectural design that satisfies all non-functional requirements. This quality-driven, iterative approach to architecture generation and evaluation is described in more detail in (de Bruin and van Vliet, 2001).

As discussed before, the abstraction level is lowered each time a refinement step is performed, which generally requires that more and more detail must be added to the general solution we started with. For example, in the case of the application of the Observer design pattern, details must be filled in as to when and how to subscribe and unsubscribe observers. As can be deduced from Figure 7, the FS graph also contains information about when and where to apply abstraction lowering. Suppose we have opted for a high flexibility solution, which resulted in selecting the Observer design pattern. The Observer design pattern node in the FS graph is EXOR decomposed in a Permanent and a Message solution, which stands for setting up a connection permanently or setting up and breaking down a connection for each message, respectively. In case of high performance requirements, a permanent connection is selected, whereas for high security a connection is established on a message basis. Notice that if a medium level of security is chosen, a negative selection link rules out a permanent connection explicitly, and because of the EXOR decomposition relation, this implies the selection of the message solution. In many cases, especially when the required flexibility, security and performance levels are set to low, no particular solution is favored. To put it differently, the FS graph contains degrees of freedom that can be used to explore design alternatives in an iterative design process.

The result of abstraction lowering for a permanent connection is shown in Figure 8. The plug-in for setting up a permanent connection has two plug-in boundaries. One matches the Sub-

scribe:SetUpConnection socket of Peer1, the other matches the Subscriber:Receiver socket of Peer2.
Setting up a permanent connection requires some initialization, which is modeled as an extension in
the plug-in. When applying the refinement, this extension is added to the result. Once the connection
is established, no additional functionality is needed for the actual communication. This is modeled as
an 'empty' plug-in of type Subscriber:Receiver. This empty plug-in is needed to determine where the
plug-in is to be applied. When applied, it leaves the existing structure intact.

A plug-in for setting up a permanent connection:
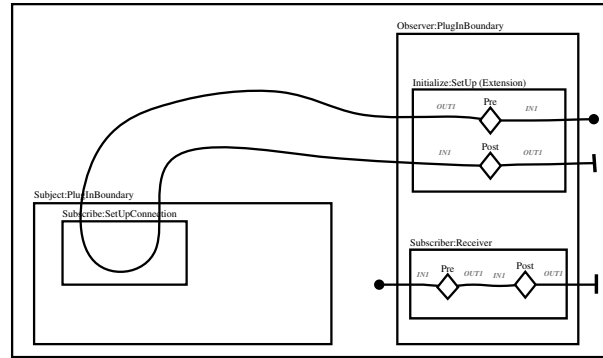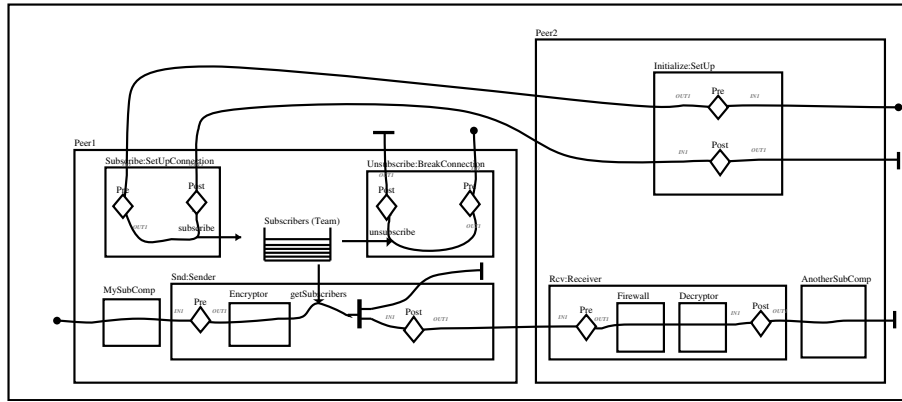

The result of applying the plug-in:

Figure 8: Filling in the details for a permanent connection.

In conclusion, the solution part of the FS-graph contains a collection of UCM plug-ins. These plug-ins
are used to iteratively refine an architectural solution, guided by the quality requirements as expressed
in the feature part of that same graph. Each time we apply such a refinement, the abstraction level of
the architecture is lowered.

5 Crosscutting Concerns

By means of a larger example, we now show that a FS graph cannot only be used to refine a particular solution locally, but also globally, at the system level. That is, we can handle aspects, such as performance and security, that in general require the adaptation of many components rather than one or a few. This type of refinement may be called Aspect-Oriented Programming (AOP) at the architectural level.

The example we use to illustrate this type of global refinement is about clients ordering goods by a retailer. A typical but simplified scenario for ordering is as follows. First, a client fills in a form and sends this form electronically to the retailer. Next, the retailer checks its database to see whether the goods are in stock. Finally, it sends a report to the client stating the price and when the goods are expected to be shipped.

Now assume this basic scenario is extended with a brokering facility. If certain goods are not supported by the retailer, he tries to obtain them from wholesalers. Since it may take some time to get answers from wholesalers (especially if they do not provide on-line services), a client-driven approach as in the basic ordering scenario is no longer applicable. Instead, the retailer takes the initiative to inform the client about the status of an order after all outstanding requests to wholesalers have been collected. Thus, the communication between a client and a retailer has changed from client-driven to a mixture of client-driven and retailer-driven communication.

A frequently used architecture for systems like the ordering system is a Client-Server (CS) architecture. In figure 9, a reference architecture is shown in which the basic scenario of ordering goods is contained. Both the client and the server component provide sockets for further refinement.
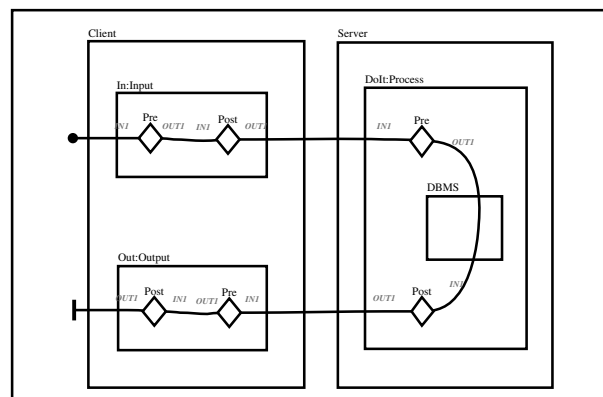

Figure 9: Client-Server reference architecture.

The next step is to extend the system with broker functionality. This is shown in Figure 10. The functionality that is expressed in the brokering UCM is that, first, the database is consulted. If certain

goods cannot be delivered by the retailer (i.e., the server), the broker sends a request to a number of wholesalers. Simultaneously, the broker sends a message to the client to inform that brokering is in progress. After all brokering requests have been collected, the database is updated and, finally, a status report is sent to the client. Whereas the basic scenario of ordering goods without brokering is client-driven, the reverse holds for sending a status report after the results of the brokering process have been sorted out. In the latter case, the system takes the initiative. As a consequence, the client has to be adapted to support this system driven approach. For instance, the client can be equipped with an E-mail component or, alternatively, it may contain some kind of message broker. The socket Notify:Messenger contained in the client component acts as a placeholder for system-driven message delivery.

A plug-in for brokering:
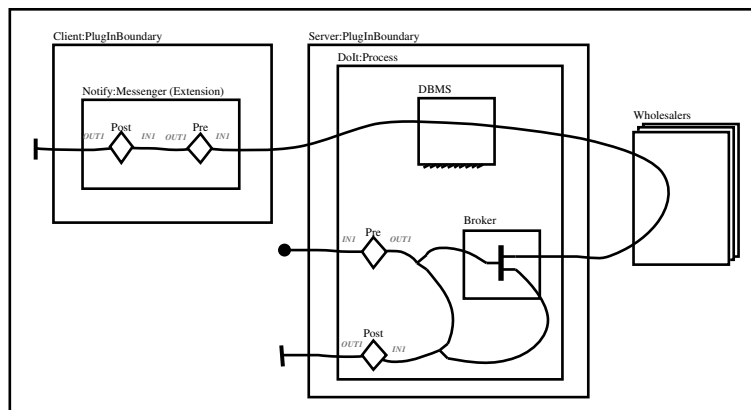
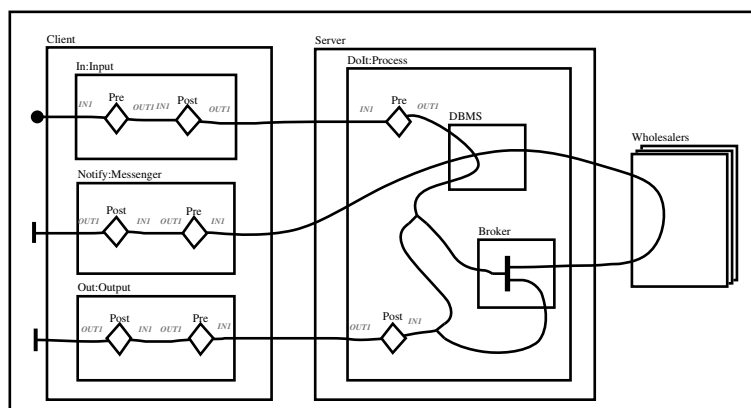The resulting Client-Server system:

Figure 10: Extending the Client-Server system with broker functionality.

The point is that, because we extend the server with broker functionality, we have to extend the client as well. So the effect of this extension is not localized. Rather, the extension has a system-wide

impact. This we may call AOP at the architectural level. The knowledge of how and where the system has to be extended or adapted can again be captured in a FS graph, as is illustrated in Figure 11. Observe the effect of requiring broker functionality. The selection links stemming from the brokering functional requirement not only select the broker component to be placed in the server, but also a message delivery component for placement in the client. Effectively, the knowledge captured in the FS graph ensures that all refinements and extensions needed to satisfy a particular requirement are effectuated.
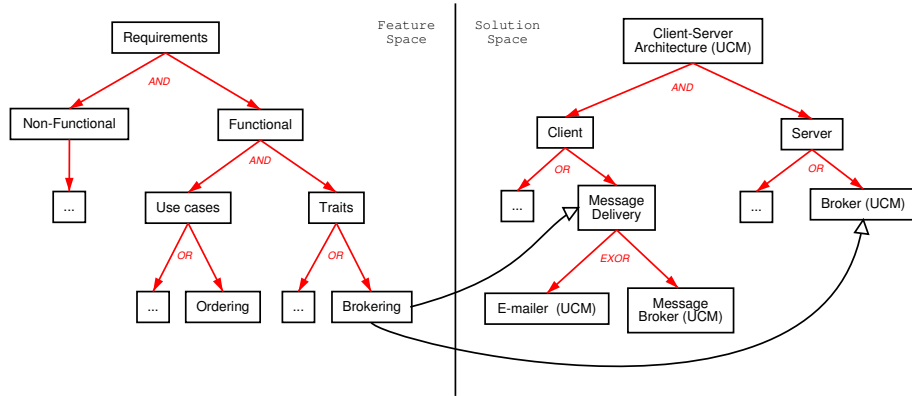

Figure 11: Feature-Solution graph for the Client-Server system.

6 Functionality Extensions

The examples given in the previous sections all used a reference architecture that contains basic functionality and variability points in order to offer the possibility for refinement to satisfy non-functional requirements. This approach is restricted in the sense that we can only refine a given functionality as expressed in the reference architecture. Obviously, this restriction is too tight. In reality, systems evolve over time; obsolete functionality will be replaced and new functionality will be added. We therefore need to support the concept of functionality extension.

Consider again the example of ordering goods by a retailer. Rather than starting with a reference architecture in which a database is consulted, we start with a Client-Server (CS) architecture that does nothing at all, but is gradually extended with new functionality. For instance, ordering and broker functionality can be added first, in much the same way as in the example given before. Next, functionality can be added to deal with, for instance, payment of the goods.

In figure 12, the CS architecture is captured in a UCM that has been stripped down to its absolute minimum. It consists of small pieces of functionality in the form of UCM paths that can be connected

to form larger pieces of functionality. Such UCM paths may start and end with named start and end points (specified as Name:Type). The idea is that a named end point can be connected to a named start point if they have identical types (see Table 2 for the meaning of the UCM elements that are used to express extensions). A UCM path can be cloned in order to add new functionality. This is done by marking that path in a refinement with the keyword "Clone". When a refinement is applied to a target UCM, the UCM paths marked as cloned are not only used in the actual substitution of plug-ins in stubs, but a copy (clone) thereof is retained in the result as well. This latter path can later be used to add further functionality.
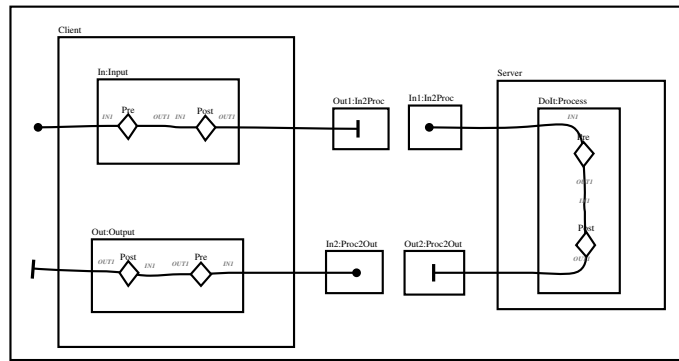

Figure 12: Stripped CS system.

Element	Meaning
Start and End Points	Start and end points are used to delimit UCM paths that represent a small piece of functionality. The idea is that a larger piece of functionality can be composed out of such small pieces. Start and end points are depicted as UCM start and end points, but are enclosed in a rectangle.
	A start and/or end point may be marked as "Clone" in a refinement to indicate that the corresponding UCM path in a target UCM is cloned first before substitution takes place.

Table 2: The meaning of UCM extension elements.

To illustrate functionality extension, the CS architecture is first transformed to support the scenario of ordering goods without brokering (see Figure 13). Observe that all three paths in the DBMS refinement are marked as "Clone". This means that in the resulting UCM, clones of these UCM paths appear that serve as placeholders to cater again for supporting functionality extension. In this way, the functionality is not fixed beforehand. We keep the option to add new functionality to the system over and over again.

In Figure 14, a refinement is shown for brokering. This plug-in should be contrasted with the broker
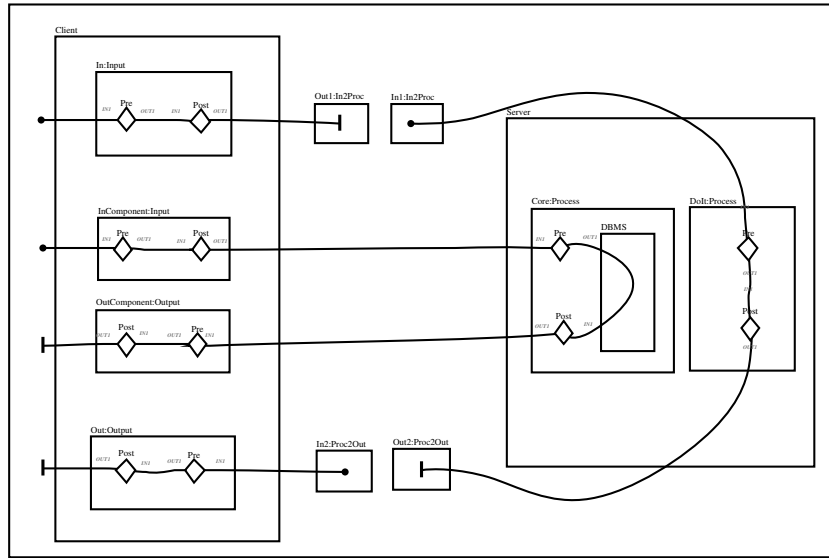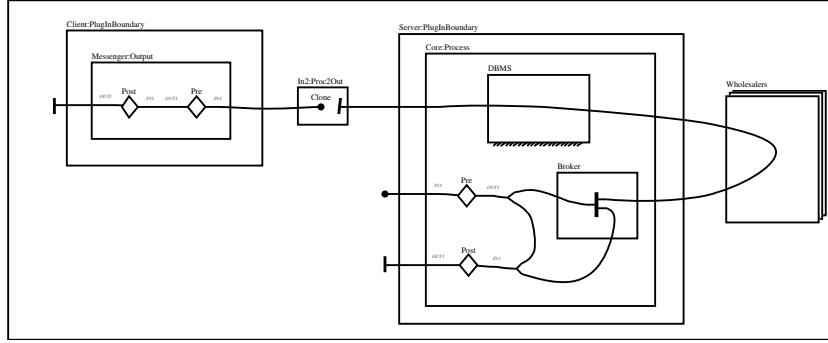
DBMS refinement:
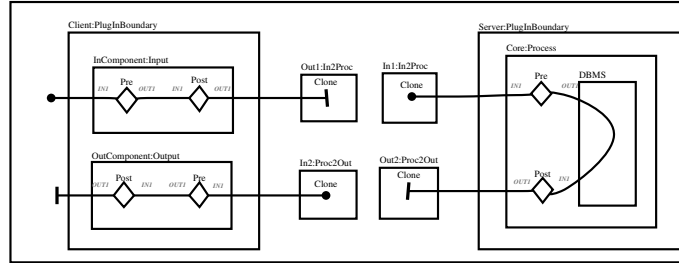
Result:

Figure 13: DBMS functionality extension.

plug-in shown in Figure 10. The two plug-ins are identical except for the way a new Messenger socket is introduced. In the case of functionality extension, the Messenger:Output plug-in is substituted in the Out:Output socket in the CS architecture, but only after the Out:Output socket has been cloned. Again, by cloning a UCM path we keep the option for functionality extension. Notice that the broker refinement can be seen as both a refinement and an extension. The actual brokering process refines the core functionality of the server, whereas the messenger plug-in extends the functionality of the client.

The functionality of the CS can be further extended with, for instance, facilities for credit-card payment. The client can decide to accept the offer by sending back an acknowledgement to the retailer along with his credit-card information. The retailer then checks the validity of the credit-card information, and if all went well, the retailer confirms shipping the goods, otherwise the transaction is cancelled. This functionality can be defined along the same lines as discussed before.

To summarize, the concept of functionality extension provides a mechanism to add new functionality

Broker refinement:

Result:
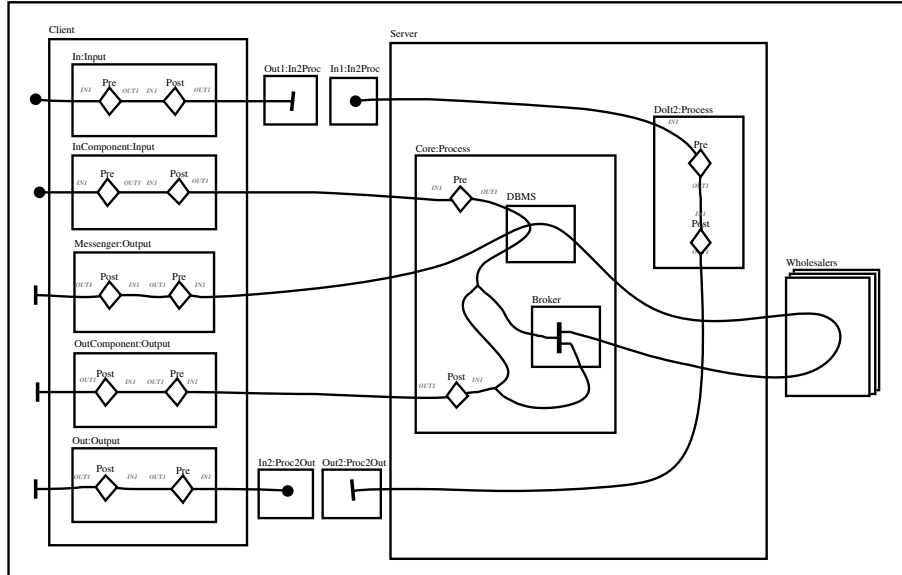
Figure 14: Broker plug-in.

to the system at any one time. New functionality can then be refined using the refinement principles described before in order to address non-functional requirements. These techniques allow us to extend and refine a basic reference architecture indefinitely.

7 Formalizing the Composition Model

The refinement rules presented so far are syntactic rules. They are used to ensure that a refinement consisting of a set of plug-ins are compatible with a set of stubs as far as type and structure is concerned. This is a necessary condition, but it is not sufficient for ensuring that the refined composition meets its initial behavioral specification. Thus, besides syntactic subtyping rules we need a stronger mechanism

that captures the concept of behavioral subtyping.

Formalizing behavioral refinement is difficult to achieve with UCM's alone. In particular, responsibility points are not defined formally and neither can we give precise component interaction specifications. This should not be seen as an omission of UCM. UCM simply serves another purpose, namely to illustrate the big picture. To overcome the lack of formality, we have extended UCM with BCOOPL interface descriptions to formalize the external behavior of a component. A BCOOPL interface describes the operations that a component should support. It is specified as an augmented regular expression over operations, not only specifying the signature of the operation, but also detailing when and by whom a specific operation may be invoked (de Bruin, 2000a). Together with internal behavior descriptions in the form of UCMs, the behavior of a component is formally captured at a high level of abstraction (de Bruin, 1999). In (de Bruin, 2000b), we describe how components described by BCOOPL's interfaces and UCMs can be translated into BCOOPL code and how the resulting specification can then be subjected to a formal analysis. Effectively, BCOOPL interfaces and UCMs establish cause-effect relations not only inside a component, but on the system level as well. The temporal orderings between events (e.g., sending and receiving messages) can be analyzed. For instance, we have shown how architectural mismatch can be detected, that is, components that work perfectly well in isolation may fail when they are brought together (Garlan et al., 1995).

BCOOPL interfaces can be refined following co- and contra-variance rules. With this, we arrive at additional matching rules to include behavioral subtyping:

- The external behavior of a component specified as a BCOOPL interface may be refined using BCOOPL's subtyping rules.

- The scenarios as manifested by UCM paths may be extended with new ones, but the existing scenarios may not be changed.

These rules ensure that the basic functionality remains the same as far as the external behavior of components is concerned. That is, a refined component still reacts to the same events and still emits the same events, but the set of events may have been extended with new ones. For instance, the refinement of using the Observer pattern results in an extension of the component's interfaces (to deal with notifications and subscription management) and in the extensions of UCM paths (to subscribe and unsubscribe). However, the basic functionality is not changed.

UCM components do not have a formal notion of state. (BCOOPL interfaces, however, do introduce an abstract state in the sense that it describes when and by whom operations may be invoked.) UCM's responsibility points provide the means to address this issue. There are no principal reasons for *not* formalizing responsibility points, and one way to do this is by using a language akin to UML's Object

Constraint Language (OCL). In this way, we can give precise semantics of operations by means of pre- and post-conditions, and hence we can state additional criteria for refinement. It also provides a road to executable specifications, that is, the application of a sequence of refinement steps ultimately leading to executable code.

8   Related Work

Architecture-Based Architectural Styles (ABASs) are proposed in (Klein et al., 1999) as a means to capture structural and behavioral aspects of (partial) design solutions together with their quality properties. An ABAS combines an architectural style with certain quality attributes. We do essentially the same in our FS-graph, but we have a stronger focus on a construction/generation-oriented representation of the architectural knowledge.

A Design Rationale (DR) (Buckingham Shum and Hammond, 1994; Moran and Carroll, 1994) is a representation of the reasoning behind the design of an artefact. It is concerned with capturing why designers have made the decisions they have made. A well-known approach to representing DR is Design Space Analysis, whose notation is called QOC (Questions, Options and Criteria) (MacLean et al., 1991). Questions in QOC are key design issues, and Options are possible answers to the Questions. The solution part of our FS-graph contains both Questions and Options. Criteria in QOC are used to choose between Options, and thus resemble the requirements as captured in the feature part of the FS-graph. DR is most often used as a tool in the design process, especially the user-interface design process, to augment design reasoning, and to help in formulating and communicating arguments.

In feature-oriented domain analysis (FODA) (Kang et al., 1990) and variants thereof, a family or product line is represented in a feature tree. Features can be mandatory, alternative, or optional. A specific product is then composed by choosing a set of alternative and optional features; these express variabilities within the product line. The feature tree thus resembles the feature part of our FS-graph. Usually though, features of a product line are units of functionality rather than solutions to different quality concerns.

In goal-oriented requirements engineering, the relation between goals and requirements is represented explicitly (Lamsweerde, 2000; Mylopoulos et al., 2001). A natural continuation of this line of thought is to connect requirements with high-level design (i.e., architecture). This is done in GRL-UCM (Liu and Yu, 2001). The work on Goal-Oriented Language (GRL) in combination with UCM (GRL-UCM) has a lot of similarities with the concepts presented in this paper. As in our approach, a link is established between requirements (AND-OR decomposition, soft goals, and tasks) and design solution in the form of UCM scenarios. Also a notion of refinement with UCM stubs and plug-ins is supported.

Our approach differs in the following respects:

- The explicit distinction between feature and solution space and the connections between them. This opens the way to recursively apply this idea. For instance, we can establish linkages between requirements (F-space) and software architecture (S-space). By the same token, we can establish linkages between architecture and detailed design solutions. But now, the software architecture level is the F-space w.r.t. to the detailed design level (S-space). The detailed design space in its turn can act as a F-space for the implementation (S-space).

- The S-space may contain nodes that are not connected to nodes in the F-space. This can be seen as degrees of freedom with which design alternatives can be explored.

- The idea of refinement can be applied recursively (i.e., plug-ins provide stubs). The GRL-UCM supports, on the other hand, one level of refinement only.

- We support a more advanced refinement mechanism in the form of multiple variability points. This allows us to use refine patterns that are not necessarily restricted to one, local point. The Observer pattern is a typical example to support this point.

The concept of refinement has been applied for many years, especially in the field of formal specifications, e.g., VDM and Z. A well-known approach to refinement in a component-oriented context is Catalysis (D'Souza and Wills, 1998). The main idea in Catalysis is the concept of collaboration, which is a set of related actions involving multiple objects and resulting in a goal. A collaboration can be seen as a formalization of a use case. It is specified formally in terms of pre- and post-condition, and invariants using UML's Object Constraint Language (OCL). A collaboration is first described at a high level of abstraction. This description serves as a formal specification to be refined in subsequent steps. Although refinement is a key concept in Catalysis, it does not support the notion of variability points. That is, no clues are given whatsoever on where and how to refine. In our approach, refinement is done automatically on the basis of type and structure matching and the content of the FS graph. Also, no attention is paid to non-functional quality attributes, like flexibility, security and performance. The main emphasis in Catalysis is on behavioral refinement.

Surprisingly, refinement is not a main issue in most Architectural Description Languages (ADL) (see (Medvidovic and Taylor, 2000) for a classification framework and a survey of ADLs). Most ADLs favor bottom-up composition in the sense that components are treated as black-box, building blocks. SADL (Moriconi et al., 1995) and Rapide (Luckhalm and Vera, 1995) are noticeable exceptions.

SADL supports a method for stepwise refinement of an abstract architecture into a lower level architecture. A refinement pattern is applied in each refinement step and maps the architectural style (e.g., dataflow, pipes and filters, implicit invocation) of the abstract architecture to the architectural style of

the refined architecture. These refinement patterns are defined and proved correct independent of a particular architecture. SADL mappings have a rather global nature, whereas in our approach we can handle local as well as global refinements guided by the FS graph.

Rapide is a concurrent event-based simulation language for defining and simulating the behavior of architectures. The underlying semantical model of Rapide is based on partially ordered event sets (posets). An abstract and a refined architecture can be related by mapping concrete events to abstract events. These mappings provide the means for comparative simulations of architectures at different abstraction levels to check whether a refined architecture is in conformance with the abstract architecture. It could be interesting to explore this idea in our approach, after all, UCMs do imply ordering of events too. However, this is a subject that has currently not our focus of attention.

The idea of Pre- and Post-stubs has been derived from the notion of composition filters as used in the programming language SINA (Aksit et al., 1996). The composition filter model consists of input and output filters that surround an object and affect the messages sent to and received by that object. Composition filters can be seen as objects in the role of proxies that perform additional pre- and post-processing.

9 Concluding Remarks

We have discussed an approach for the top-down composition of software architectures. It is centered around a Feature-Solution (FS) graph in which requirements are linked to design solutions and alternatives. The design solutions are expressed in UCM, which provides an intuitive way of showing how things work globally. UCM is not a necessary prerequisite for our approach, though. The idea of supporting multiple variability points and plug-ins in combination with a FS graph can also be used with modeling techniques like UML, or even with a text-based specification language.

Our FS-graph serves two purposes. Firstly, it contains knowledge for solving problems in a particular domain. Thus viewed, the graph can be used to guide an iterative, quality-driven process to software architecture generation. In this iterative process, the abstraction level is successively lowered because of refinements made. Secondly, the FS graph can be used to make system-wide adaptations. This ensures a consistent, system-wide application of refinements. These system-wide refinements resemble aspect-oriented programming, albeit applied at a much higher level of abstraction.

We are currently using the FS-graph to codify the architectural knowledge from various real-life architecture projects. Our experiences indicate some promising areas of further research:

- The version of the FS-graph discussed in this paper only contains product knowledge. We are currently investigating the incorporation of process knowledge in the FS-graph as well. When

selected, such process nodes then guide the the architect in finding appropriate design solutions. The process steps encoded in process nodes can be based on heuristics, for instance, in the form of production rules (if ≪a particular situation X is encountered≫ then ≪select solution Y≫). Alternatively, process nodes may alert the designer that certain design choices have to be made. These small process steps can be embedded in the larger process cycles defined in a quality-driven, iterative approach. This leads to a content-driven refinement process.

- The relationships currently captured in the FS-graph are rather strong: "when high flexibility is required, select the Observer pattern". Actual relationships between requirements and solution chunks are often of a much weaker kind: "when high flexibility is required, we suggest to apply the Observer pattern". To represent these weaker links, the FS-graph will have to be enriched with additional kinds of relationships which, in turn, require additional resolution mechanisms to get from (quality) requirements to acceptable architectural solutions.

Finally, we have started the QUASAR project, which stands for QUAlity-driven Software ARchitecture. The goal is to generate software architectures and implementations using the techniques described in this paper. The starting point is the FS-graph in which domain and application specific knowledge is captured. The FS-graph is then used to (semi-)automatically generate and evaluate candidate architectures. Promising architectures are stepwise refined until we end with executable specifications, that is, code.

References

Aksit, M., Tekinerdogan, B., and Bergmans, L. (1996). Achieving adaptability through seperation and compositin of concerns. In *Proceedings of ECOOP Workshop on Adaptability in Object-Oriented Software Development*, pages 12–23, Linz, Austria.

Bosch, J. (2000). *Design and Use of Software Architectures: Adopting and Evolving a Product-Line Approach*. Addison-Wesley.

Buckingham Shum, S. and Hammond, N. (1994). Argumentation-based design rationale: What use at what cost? *International Journal of Man-Machine Studies*, 40(4):603–652.

Buhr, R. (1998). Use Case Maps as architecture entities for complex systems. *IEEE Transactions on Software Engineering*, 24(12):1131–1155.

Buhr, R. and Casselman, R. (1996). *Use CASE Maps for Object-Oriented Systems*. Prentice Hall, Upper Saddle River, New Jersey.

de Bruin, H. (1999). A grey-box approach to component composition. In Czarnecki, K. and Eisenecker, U. W., editors, *Proceedings of the First Symposium on Generative and Component-Based Software Engineering*

(GCSE'99), Erfurt, Germany, volume 1799 of *Lecture Notes in Computer Science (LNCS)*, pages 195–209, Berlin, Germany. Springer-Verlag.

de Bruin, H. (2000a). BCOOPL: Basic Concurrent Object-Oriented Programming Language. *Software Practice & Experience*, 30(8):849–894.

de Bruin, H. (2000b). Scenario-based analysis of component compositions. In Butler, G. and Jarzabek, S., editors, *Proceedings of the Second Symposium on Generative and Component-Based Software Engineering (GCSE'2000), Erfurt, Germany*, volume 2177 of *Lecture Notes in Computer Science (LNCS)*, pages 1–18, Berlin, Germany. Springer-Verlag.

de Bruin, H. and van Vliet, H. (2001). Scenario-based generation and evaluation of software architectures. In Bosch, J., editor, *Proceedings of the Third Symposium on Generative and Component-Based Software Engineering (GCSE'2001), Erfurt, Germany*, volume 2186 of *Lecture Notes in Computer Science (LNCS)*, pages 128–139, Berlin, Germany. Springer-Verlag.

D'Souza, D. F. and Wills, A. C. (1998). *Objects, Components, and Frameworks with UML: The Catalysis Approach*. Object Technology Series. Addison-Wesley, Reading, Massachusetts.

Garlan, D., Allen, R., and Ockerbloom, J. (1995). Architectural mismatch: Why reuse is so hard. *IEEE Software*, 12(6):17–26. Carnegie Mellon University.

Kang, K., Cohen, S., Hess, J., Novak, W., and Peterson, S. (1990). Feature-Oriented Domain Analysis (FODA) Feasibility Study. Technical report, Software Engineering Institute.

Kiczales, G., Lamping, J., Mendhekar, A., Maeda, C., Lopes, C. V., Loingtier, J.-M., and Irwin, J. (1997). Aspect-oriented programming. In Askit, M. and Matsuoka, M., editors, *Proceedings of 11th European Conference on Object-Oriented Programming (ECOOP'97), Finland*, volume 1241 of *Lecture Notes in Computer Science (LNCS)*, pages 220–242, Berlin, Germany. Springer-Verlag.

Klein, M., Kazman, R., Bass, L., Carriere, J., Barbacci, M., and Lipson, H. (1999). Attribute-based architectural styles. In Donohue, P., editor, *Software Architecture*, pages 225–244. Kluwer Academic Publishers.

Lamsweerde, A. v. (2000). Requirements engineering in the year 00: A research perspective. In *Conference Proceedings ICSE'00*, pages 5–19, Limerick, Ireland. ACM.

Liu, L. and Yu, E. (2001). From requirements to architectural design: Using goals and scenarios. In *ICSE'2001 Workshop 9, From Software Requirements to Architectures (STRAW'2001)*, pages 22–30, Toronto, Ontario, Canada. ACM.

Luckhalm, D. and Vera, J. (1995). An event-based architectural description language. *IEEE Transactions on Software Engineering*, 21(9):717–734.

MacLean, A., Young, R. M., Bellotti, V. M., and Moran, T. P. (1991). Questions, options and criteria: Elements of design space analysis. *Human-Computer Interaction*, 6(3 & 4):201–250.

Medvidovic, N. and Taylor, R. N. (2000). A classification and comparison framework for software architecture description languages. *IEEE Transactions on Software Engineering*, 26(1):70–93.

Moran, T. and Carroll, J., editors (1994). *Design Rationale: Concepts, Techniques, and Use*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Moriconi, M., Qian, X., and Riemenschneider, R. (1995). Correct architecture refinement. *IEEE Transactions on Software Engineering*, 21(4):356–372.

Mylopoulos, J., Chung, L., Liao, S., Wang, H., and Yu, E. (2001). Exploring alternatives during requirements analysis. *IEEE Software*, 18(1):92–96.

Biographies

Hans de Bruin is a lecturer and a researcher at the Vrije Universiteit. He recived his masters degree from the Technical University of Delft in 1987 and his PhD from the Erasmus University Rotterdam in 1995. Before returning to the university in 1998, he worked as a software engineer and consultant in the industry. Currently, he teaches software engineering at the Vrije Universiteit. His research interests include software engineering, in particular software architecture and component-based development, and programming language design.

Hans van Vliet is a Professor in Software Engineering at the Vrije Universiteit. His research interests include software architecture and software measurement. Before joining the Vrije Universiteit, he worked as a researcher at the Centrum voor Wiskunde en Informatica (Amsterdam) and he spent a year as a visiting researcher at the IBM Almaden Research Center in San Jose, California. Hans has an M.Sc. in Computer Science from the Vrije Universiteit and a Ph.D. in Computer Science form the University of Amsterdam.