On the Criteria to be Used in Decomposing Systems into Aspects*

Kevin Sullivan[φ] William G. Griswold[λ] Yuanyuan Song[φ] Yuanfang Cai[φ]
Macneil Shonle[λ] Nishit Tewari[φ] Hridesh Rajan[φ]

[φ]Computer Science
University of Virginia
Charlottesville, VA 22903
{sullivan,ys8a,yc7a,nt6x,hr2j}@cs.virginia.edu

[λ]Computer Science & Engineering
UC San Diego
La Jolla, CA 92093-0114
{wgg,mshonle}@cs.ucsd.edu

ABSTRACT

With the growing popularity of aspect-oriented programming languages, like AspectJ, it is natural to ask how to best use such a language to achieve common modularity goals like comprehensibility, parallel development, and ease of change, among others. By means of a detailed case study, we compare a prevailing method called *obliviousness* with a method of our own that applies simple declarative *design rules* to the points in the code that the system designer deems to be of interest. The case study and a net-options-value analysis show several weaknesses in the oblivious approach, and that substantial benefits can be gained from using our design rules approach.

1. INTRODUCTION

Aspect-oriented programming (AOP) languages aim to improve the ability of designers to modularize concerns that cannot be modularized using procedural or object-oriented methods. Examples of so-called *crosscutting concerns* include tracing, logging, transactionality, caching, and resource pooling. The ability to better modularize these concerns is expected to ease development through improvements in comprehensibility, parallel development, reuse, and ease of change, among others. With such improvements should come reduced development costs, increased system dependability and adaptability, and ultimately better value for software producers and consumers.

The most prominent AOP model today is that embodied by the AspectJ language [2, 12]. AspectJ extends Java with several complementary mechanisms, notably *join points* (JPs), *pointcut descriptors* (PCDs), *advice*, and *aspects*. JPs are points in a program's execution, such as before a method is called or after it returns, that by definition of the programming language are subject to *advising.* Advising is the extension or overriding of the action at a join point by a CLOS-like [22, Ch. 28] *before, after,* or *around* anonymous method called *advice.* A PCD is a declarative expression that matches a set of JPs. An advice extends or overrides the action at each join point matched by a given PCD. Because a PCD can select JPs that span unrelated classes, advice behaviors are crosscutting in their effects, yet are locally specified. Advice, pointcuts, and ordinary data and method declarations are grouped into class-like modules called *aspects*.[1] At a conceptual level, aspects are intended to support modular expression of crosscutting concerns [13], although they admit other uses.

Thirty-three years ago, seeing the opportunities afforded by the new module composition mechanisms of procedural programming and separate compilation, David Parnas asked the question, by what criteria should systems be decomposed into modules [19]? Using the classic modularity objectives of comprehensibility, parallel development, and ease of change, he used a comparative analysis to argue that the prevailing criterion of modularizing a system according to the stages of processing in its flow chart performed poorly relative to his new approach, *information hiding*. Under the information hiding approach, "one begins with a list of difficult design decisions or design decisions that are likely to change. Each module is then designed to hide such a design decision from the others" [19, p. 1058]. The general idea was to impose stable, abstract interfaces to decouple design decisions that are expected to change. An example decision given by Parnas is the selection of a data representation, the hiding of which results in an abstract data type interface.

In this paper, seeing the opportunities afforded by the new mechanisms of aspect-oriented programming, we revisit Parnas's question, with a twist: by what criteria should systems be decomposed into aspects? The prevailing method of aspect-oriented design, popularly called *obliviousness*, advocates that the designers and developers of the base functionality[2] need not explicitly prepare code to be advised by aspects. Filman and Friedman put it as, "Just program like always, and we'll be able to add the aspects later" [7, p. 31]. This organizing principle has risen to the level of an architectural paradigm for aspect-oriented design: (1) modularize at the highest level by separating base and crosscutting concerns, (2) implement base concerns in an object-oriented style ignoring crosscutting concerns, (3) then implement the crosscutting concerns as a separate layer of aspect modules (See Section 2.1).

Is this straightforward method the best to be found? Aided by a case study of a modern object-oriented system, the HyperCast network overlay system [10], we address this question by comparing the oblivious method to a new one based on an extension of

*This research supported in part by NSF grant FCA-0429947.

[1]Named pointcuts can be declared in classes, but other AspectJ constructs are restricted to aspects.

[2]The term *base* in the AOP community refers to the non-aspect-oriented elements of the system. *Base code* more specifically refers to the actual program text, either as implementation or design.

information hiding called *design rules* [3]. We find that although the oblivious method can allow base code designers to ignore decisions made by aspect code designers, it does not decouple design decisions in the other direction. In particular, the freedom afforded to base code designers comes at considerable cost to aspect designers, who are thwarted by a lack of constraints on the code they need to advise. The results include unnecessary complexity in aspect code (especially pointcut descriptors); in some cases the inability to achieve desired results at all without forcing changes in the base code; and extreme sensitivity of the validity of aspect code to changes in base code. The challenges of the oblivious approach present opportunities to aspect language designers, in some cases extending the join point model [20], in others providing mechanisms to constrain it [1]. However, providing new language features is a costly proposition, fraught with risks such as semantic complications, runtime costs, and ultimately adoption.

Our proposed approach effectively decouples base and aspect design without incurring these costs. It is an instance of information-hiding in which design rules [3] are imposed as interfaces between aspects and base code, constraining their subsequent development so as to decouple them. Our design rules govern how base code creates join points and how aspects use them to ensure that specific join points are exposed in a way that enables the integration of separately implemented aspect modules. Concretely, our design rules specify (1) the types of join points used to expose specified behavior (e.g., field access versus method call), (2) naming conventions for join points, and (3) constraints on base code and aspect behaviors. The first guarantees that join points needed by aspects are visible. The first two together permit aspect designers to write PCDs that do not have to change when the base code evolves. The last ensures that only intended effects occur across join points, between base code and aspects.

These rules are documented in interface specifications that base code designers are constrained to "implement," and that aspect designers are licensed to depend upon. Once the interfaces are defined aspect and base code can be developed independently and concurrently. They are symmetrically "oblivious" to each others' design decisions. What they cannot be oblivious to is the presence of the design rule interfaces. We find that formulating design rules in an application-centric manner—for example, in terms of interesting states of applications rather than, say, logging—improves separation and resilience of the design.

Our method does not require any auxiliary code to be added to base code (e.g., to signal events), or demand references from base code to aspects or other external code modules. Nor does it involve the introduction of any new programming mechanisms or semantics, although we discuss how the checking of design rules could be valuable in some cases. In particular, our approach does not constrain the JPM model, nor does it demand one with any particular characteristics; and our approach does not impose any prior constraint on the exposure of join points by a program under the JPM of the given language.

The following section provides some background on obliviousness, after which we introduce the HyperCast case study, including its design considerations and crosscutting concerns. Next we consider designing, developing, and extending HyperCast through the oblivious aspect-oriented design approach. We then repeat the exercise with the information-hiding design-rules approach. Finally, we compare the two approaches, first qualitatively—from the designer's perspective—and then quantitatively using an economic analysis based on options theory [3]. In the process, the analysis shows that both are better than the plain OO design of HyperCast.

2. BACKGROUND

2.1 Obliviousness

Obliviousness as a distinguishing characteristic of AOP was first proposed by Filman and Friedman [6, p. 2]:

> *AOP can be understood as the desire to make quantified statements about the behavior of programs, and to have these quantifications hold over programs written by oblivious programmers.*

The above-cited paper, whose purpose is to justify this definition, has at least 50 citations in the literature. The vast majority use this paper's definition of AOP, while a few distinguish their work from this definition.

Quantification needs to be understood before obliviousness can be discussed. One property that makes AspectJ unique is the ability to name sets of join points declaratively. For example, the pointcut designator `call(* *State(..))` refers to all calls to methods with names ending in `State` (and having any parameter list). Not only does this save programming effort, but the pointcut, to the extent that it exploits a naming convention, automatically matches join points of newly programmed methods following the naming convention. Some have referred to this extension property of properly quantified aspects as *shyness* [18]. In the terminology of obliviousness, such aspects are oblivious of base code.

Quantification's link to obliviousness is that more expressiveness in the quantification language—for example extending the join point model and pointcut language of AspectJ—provides for more obliviousness. That is, more power in the hands of aspect designers means that less help is required from base code developers to add aspects to a system. As often is the case, greater power can be a double-edged sword, as discussed in Section 4. The desire for greater expressiveness in the quantification language and its attendant difficulties have been a driver for aspect-oriented programming language researchers.

Many variants and degrees of obliviousness can be found in the literature, each of which carries different implications. We list the definitions most relevant to the current paper in an approximate hierarchy from weakest to strongest, and give them appropriate names:

Language-level obliviousness is what is allowed when advising constructs are introduced to a programming language. Filman and Friedman provide an apt definition, arguing that "...the distinguihing characteristic of aspect-oriented programming (AOP) languages is that they allow quantified programmatic assertions over programs that lack local notation indicating the invocation of these assertions" [7, p. 21]. Specifically, the language enables the base code developer to write code without needing to use callback hooks or macros.

Feature obliviousness is when the base code developer is unaware of the features that aspects implement. The base code designer may prepare the code for aspects, for example using callback hooks, thus sacrificing language-level obliviousness. The approach proposed in this paper seeks what amounts to a combination of language-level obliviousness and feature obliviousness by establishing rules such as syntactic conventions and naming conventions. This permits aspects to be more "base code shy" and hence do not have to change when the base code evolves.

From an information-hiding perspective [19], feature obliviousness matches the classic notion of obliviousness: services are unaware of their clients, but the services are obligated to provide the service if its preconditions are met by the client.

Designer obliviousness is when the designers of the base code can be oblivious to the existence of aspects; in particular, not de-

signing any differently than they normally would. As Filman and Friedman say "For true AOP, we want our system to work with oblivious programmers—ones who don't have to expend any additional effort to make the AOP mechanism work" [6, p. 2]. In addressing the properties of AOP that aid decoupling (i.e., separation of concerns), Erad, Filman, and Bader note [5, p. 31]:

> *This includes obliviousness, whether the writer of the main code be aware that aspects will be applied to it; intimacy, what the programmer has to do to prepare code for aspects; and globality versus locality, whether aspects apply to the program as a whole or only parts of it.*

Intimacy, as used here, is at the other end of the spectrum from obliviousness—the oblivious base code developer would not *prepare code for aspects.* To satisfy designer obliviousness, however, the developer has to meet the stronger criterion of not even needing to be *aware that aspects will be applied to it*. This would generally rule out communication with aspect designers.

The adoption of this definition of AOP has reached the popular press. In Laddad's widely acclaimed book on programming with AspectJ, he says of AspectJ, "AOP modularizes the individual aspects and makes core modules oblivious to the aspects. Adding a new functionality is now a matter of including a new aspect and requires no change to the core modules" [14, p. 28].

Pure obliviousness is the limit where perfect obliviousness and shyness are achieved, allowing for total, symmetric separation of concerns. As Filman and Friedman laid it out as the ultimate goal to be sought by AOP researchers [7, p. 31]:

> *It's a really nice bumper sticker to say "Just program like always, and we'll be able to add the aspects later." (And change policies later, and we'll painlessly transform the code for that, too.)*[3]

Such a degree of obliviousness is viewed as a far-off ideal by others [4]. In the presence of "harmful aspects" [11], however, complete obliviousness might not be desirable.

2.2 Design Structure

The analysis in this paper depends heavily on the notion of the *dependence* between elements in a software design, and especially on the *structure* of dependences in a given design. We say that an element B *depends upon* an element A iff B makes assumptions about A that, if invalid, can cause B not to achieve its intended properties. Invalid assumptions can arise for many reasons, such as: A's developer misimplements A's specification; A's specification changes but A's developer is not aware of the change; B's developer misunderstands A's specification; B is unaware of changes in A on which B depends.

These design dependences are de facto dependences among the developers involved in the system design and evolution task. The depends-upon relation is critical for our analysis because changes to any property of A on which B depends create possible obligations for B to change. The structure of dependences on design elements largely determines the dynamics of the design task.

We say that a design or a part of a design is *modular* if its elements do not depend on each other (although they may depend on external contracts, or *design rules*, that serve to decouple them). The corresponding work in realizing the design is parallel in that these elements can be developed or changed independently, modulo their conformance to any prevailing contracts. A design and its corresponding implementation process are said to be *sequential* if elements are linked in a chain of dependences. In such a structure, upstream design elements are resolved before downstream decisions that depend on them. A design and its implementation process are called *coupled* if the elements depend on each other in some cyclic relation.

It's common for different dependence structures to prevail in different stages of the software process. In a waterfall process, for example, the relationship between the design-specification process and the implementation-design process is sequential, but the implementation process itself is modular: the code modules can be developed independently as long as they conform to the prevailing rules (e.g., the syntax, behavioral, performance, and dependability specifications).

Following Baldwin and Clark [3] and the subsequent work of Sullivan et al. [24], we represent design dependence structures using Design Structure Matrices (DSMs) [23]. DSMs present the dependence structures of designs and and of their corresponding development and evolution process in a graphical form.

Figures 1, 2, and 4 in the following sections are examples. The rows and columns of a DSM are labeled with *design variables*. These are the design elements or dimensions for which the designers must make design decisions: e.g., the selection of an algorithm, the naming of a class, the formulation of an interface, the choice of minimum acceptable reliability. Cells in a DSM are marked to represent dependences among these decisions. For a given row (e.g., a design variable A), the marks in that row show which other design variables A depends upon. The choices made for, or changes to, those design variables influence the best choice for A. In Figure 1, for example, variable 28, socket_impl, depends on 3, 14, 29, and 30, socket_spec, socket_interface, exception_logging_impl and non_exception_logging_impl. For a given column, for B, the marks show which variables depend upon B.

Carefully ordering and clustering the rows and columns of a DSM can reveal large- and small-scale dependence structures in patterns of marked cells. Consider as an example Figure 1, which we discuss in detail in the next section. At the outermost level of aggregation, the upper left part of the DSM represents abstract design concerns (i.e., the software specification), and the lower right, the code base that implements them. The code base comprises both interface and implementation elements. The marks below the concerns and to the left of the code base indicate a sequential structure: the code base depends on the abstract concerns. Similarly, within the code base, the implementation modules depend on the interfaces. The block diagonal structure of the implementation has a modular structure. The absence of off-diagonal marks indicates no dependences between the nested smaller boxes (each representing the interdependent design decisions within an implementation module). The dense marks in each implementation module reflect a coupled structure.

An *interface* is a statement of the properties of an element B upon which other elements, such as A, are permitted to depend upon, and that B's designer is obliged to produce in B. An interface thus comprises a set of design rules to which B is subject and which A may assume to be followed. One of the central operations in a design activity is to create interfaces to decouple otherwise coupled design elements and corresponding processes. Both elements come to depend on an interface, but are made independent of each other. This property is what would commonly be called a modular design. However, by Parnas's criterion, a design exhibits information hiding modularity only if the interfaces themselves are relatively immune to change. By modeling the external forces of change on a system as a special class of design variables that we call *environment variables*, it is possible to literally see whether the

[3] A similar quotation can be found in their 2000 paper [6, p. 6].

design is modular in this sense [24]. The interface cluster should not depend on the environment cluster; only implementation variables should depend on the environment variables. In Figure 1 the environment variables are modeled as elements of the system requirements specification (or concerns) that the Hypercast designers believe will change.

3. THE HYPERCAST CASE STUDY

This paper centers around a comparative analysis of three designs for HyperCast, a scalable, self-organizing overlay network system [10, 15, 16]. The three designs are (1) its actual OO design, (2) an oblivious AO design that we produced by moving scattered code into aspects using oblivious design, and (3) one based on the design rules approach. HyperCast is a real system developed independently of the analysis reported in this paper. It includes scattered code fragments for classic crosscutting concerns, including logging. The placement of each fragment indicates objectively and precisely where aspects need to advise in order to factor concern-related code into aspects using an oblivious approach.

3.1 What HyperCast Does

The key abstraction provided by HyperCast is the *overlay socket*. An overlay socket supports point-to-point and multicast communication in overlay networks. HyperCast integrates overlay sockets, viewed as nodes, into networks in a decentralized manner. It also offers network services including naming, reliable transport and network management.

Key concerns in the design of HyperCast include the following: *Socket*—the design of the overlay socket API; *Protocol*—the protocols for maintaining various network topologies; *Monitor*—a HyperCast network management capability; *Service*—network mechanisms for end-to-end service; *Adapter*—a layer that virtualizes underlying networks; *Logging*—a mechanism to record selected events. These concerns map to loosely coupled classes in the implementation.

There are several areas in which scattering and tangling of code is evident. In this case study we address two. The first is logging. A careful study of the logging code in HyperCast revealed three sub-concerns: logging of informational messages, of raised exceptions, and of errors that do not raise exceptions. Second, several HyperCast modules use implicit invocation to notify clients of key events. The protocol module, for example, announces events for some transitions in the state machine that implements the self-organizing behavior of HyperCast. The Service module announces end-to-end service-related events. We thus inferred the following additional concerns: *Information logging, Exception logging, Non-exception error logging, State machine events,* and *Service events.*

3.2 The Design Structure of HyperCast

Each concern leads to specification and corresponding implementation decisions, which we model as design variables. For example, one must specify a set of supported underlay networks. This information is passed to the Adapter developer who then produces an implementation. Figure 1 presents a DSM showing that design structure in terms of these design variables. We view the specifications as environment parameters. They are grouped as the first major module, in the block on the upper left. We distinguish the crosscutting concerns from non-crosscutting concerns. Implementation decisions are grouped in the lower right.

In terms of these large blocks, we have a classic sequential (lower triangular) structure: Specification precedes implementation. The implementation block, by contrast, is classically modular (block diagonal). Once the specifications are fixed, the implementations

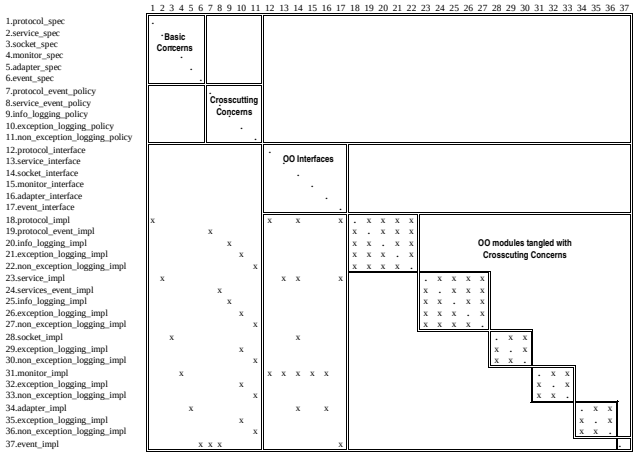

Figure 1: Basic Object-Oriented Design of HyperCast.

can be developed and changed independently. However, each implementation module exhibits serious scattering and tangling internally due to the influences of the crosscutting concerns. Tangling is seen by reading rows. For example, the protocol implementation module depends on the protocol specification but also on that of protocol events and each of the logging concerns. Scattering is seen by reading columns. Changes in the exception logging policy (parameter 10), for example, would likely effect every implementation module in the system. AOP is meant to enable improved modularity.

3.3 Comparative Methodology

We compare the actual HyperCast design with two AO alternatives: an oblivious design and one produced using the design rule method. We produced and tested both alternatives by refactoring HyperCast. The oblivious design was obtained by assuming that, had the developers been able to ignore crosscutting concerns, they quite plausibly would have written the same code, just leaving out the code for the crosscutting concerns. Aspects then advise the code in precisely the places where scattered fragments appear in the original design.

It was easy to find scattered code for the concerns of interest: they made explicit calls to methods for logging or event notification. We refactored the code to move them (or functional equivalents) into aspects using AspectJ 1.2 [2]. The aspects had to advise the new base code with PCDs that caused the given fragments, now in the form of advice bodies, to be executed at those points from which they had been removed.

The design rules approach is different: Here we asked the question, what constraints on the code would shape it to make it relatively easy to write the aspects at hand, as well as support future aspects? We thus not only moved fragments from the original code to aspects, but refactored the original code in ways dictated by the design rule interfaces that we designed. In the following sections, we compare the results in order to gain insights into the properties of the respective design methods and resulting designs.

The phenomena we sought to understand include the difficulty of writing the aspects in the first place, their sensitivity to changes in the base code, and the value of the oblivious approach relative to the design rule interfaces method. For our design rule method, we also ask how intrusive it is into the developer's practice and the

resulting code.

4. THE OBLIVIOUS APPROACH

In this section, through study of HyperCast, we explore benefits and costs of a commitment to the notion of oblivious design. Following the method outlined in the previous section, we refactored HyperCast into an oblivious aspect-oriented design. We assume—reasonably, we believe—that the developers could plausibly have written essentially the same OO code, just leaving out the scattered parts that implement crosscutting concerns. We removed the scattered fragments and localized them (or functional equivalents) in new aspect modules: *ao_protocol_events*, *ao_service_events*, *ao_info_logging*, *ao_exception_logging*, and *ao_non_exception_logging*. With implicit invocation now replaced by aspect-oriented advising, we also eliminated the whole OO *events* module.

In writing aspects that would result in these fragments being woven back into the base code at the places from which it had been removed, we encountered a number of issues, which we describe below. We give code for each example in its original form to convey what behavior is required and where.

We studied the HyperCast code to try to characterize the join points that we'd have to advise to weave the extracted code fragments back into the system. We found out that the context of these join points was in one of the following four categories.

1. Private join points. In many cases, fragments had to be woven where there were no public join points (e.g., calls to public methods).In some cases, code had to be woven into nested switch or if statements, but join points were available, such as setting of a data member. To identify these join points, we often used the `set` and `withincode` pointcut designators. Below is an example.

```
pointcut HCLogicalAddrChanged(HC_Node node):
 set(I_LogicalAddress HC_Node.MyLogicalAddress)
 && (withincode(void HC_Node.
       messageArrivedFromAdapter(I_Message))
 || withincode(void HC_Node.timerExpired(Object))
 || withincode(void HC_Node.resetNeighborhood()))
 && target(node);
```

The pointcut works, but tightly couples the aspect to hidden details that the base code developer is free to change. If the developer changes the field name, MyLogicalAddress, the aspect is will not compile, and the aspect must be rewritten. In other words, the base code change is non-modular. Unless the base code developer can wait for the aspect developer to discover that the application no longer compiles, the base code developer will need to either make the change herself or notify the aspect developer. Here we highlight either the lack of obliviousness of the base code developer or the potential for high coordination costs, and chaotic, continual introduction of incompatibilities (bugs) into code.

2. State–point separation. In many cases, the setting of a variable of interest and the join point at which weaving is needed are separated, and the given variable is not accessible to advice through the AspectJ join point model. For example, we need to access an IP address error at certain place in order to construct a log message. The required value is stored in the local variable `addrStr`, which advice cannot access. It is computed earlier by an inlined block of code, which is neither governed by the same nesting of if and switch statements as the logging code, nor solely reserved for use by the logging concern:

```
String addrStr;
InetAddress ipAddr
if (AddrString != null) {
  addrStr = AddrString;
}
else {
  String addrType = config.getStringProperty
         (PROPERTY_NAME_PREFIX + ".PhyAddr",
          "INETV4AndTwoPorts");
  addrStr = config.getStringProperty
  (PROPERTY_NAME_PREFIX + "." + addrType +
  ".Address",":0:0");
}
String[] paFields = addrStr.split(":");
...
for (int i = 0; i < paFields.length; i++) {
   paStr[i] = paFields[i].trim();
}
if (paFields.length > 3) {
  //logging code
  config.err("String" + addrStr + "
   has wrong format for a physical address.");
}
```

There are a few possible ways to capture the IP address. First, a two-stage advising sequence could be programmed, in which one advice advises calls to `config.getStringProperty` to save the IP address, and a separate advice advises the logging join point. Not only is this possibly computationally costly, but its complexity is sensitive to both whether the local variable computation dominates the logging statement and whether such a sequence could be nested (which would require a stack to capture all the relevant state). Two, the aspect developer could perhaps write a method in the aspect to compute the IP address from scratch. This would introduce unwanted scattering of IP address computations: in essence, base code is being copied into the aspect code. Finally, stepping outside the options available to the developer, the join point model of AspectJ could be extended for access to local variables. However, this approach would only exacerbate the coupling problems observed in category 1.

3. Inaccessible join point. This category includes join points within nested switch and if statements, where there is no proxy join point to advise. The check-and-branch sequence alone defines the join point. Here is an example:

```
switch(MyState) {
 case WaitforACK: {
  switch(e.getType()) {
    case FULL_E2E_ACK: {
      processAck(msg.getSourceAddress());
      if(ACKExpected.isEmpty()) {
        MyState = Done;
        MStore.setTimer(...);
        /* Notification concern - removed
        if(mylogicaladdress==root) {
          notifyApplication();
        }
        */
     }
    }
    break;
    case {
      ...
}
```

Here `notifyApplication()` notifies the application of certain events. We want to replace the use of event notification with advising. The AspectJ JPM does not provide visibility into branches taken by a program. The solution of recreating these conditionals in the advice body at best only scatters the concern, would make advice complex and hard to understand, and might not always be feasible.

4. Quantification failure. Many join points that have to be advised in the same way cannot be captured by a quantified PCD, e.g., using wild-card notations. A separate PCD is required for each join

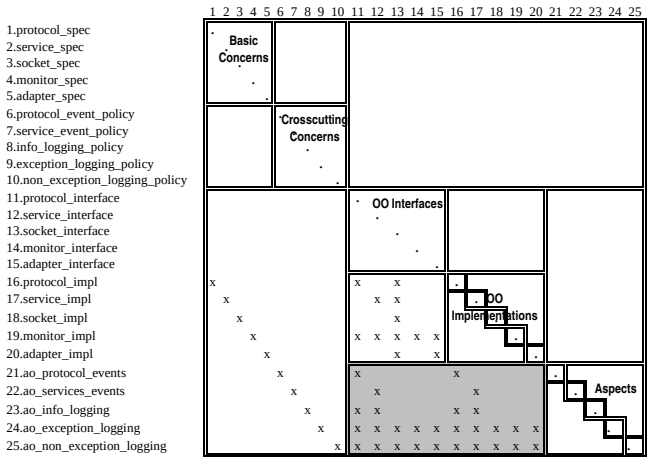

Figure 2: Obliviousness Aspect-Oriented Design of HyperCast.

point. There were about 180 places in the base code where logging was required. Most of the join points do not follow a common pattern. Not only is there a lack of meaningful naming conventions across the set of join points, but also variation in syntax: method calls, field setting, etc. One failure mode here is that creating many PCDs is costly. More seriously, extensions to the base code that should be logged will require the logging pointcut to be updated to capture the new logging points. If the pointcut is not updated, it will silently malfunction, as the non-advising of a join point does not manifest a syntax or type error that can be reported at compile time. In this case, then, it is imperative that the base code developer communicate the changes to the aspect developer.

The common theme that runs through all of these problems is that the oblivious approach might simplify the task of the base code designer, but it can significantly complicate overall system development because there is no agreement between the base and aspect code developers about how their respective parts will be integrated. Rather, the base code developer proceeds to blithely implement "as he would anyway," and the resulting design decisions then dictate the conditions to which the aspect developers must conform.

In a nutshell, the oblivious design process is a sequential process. Specifications are provided for base and aspect code modules. Then the base code is written. Finally the aspect code is written under the constraints imposed by both the external specifications and the coding decisions made by the base developer.

Figure 2, which presents a DSM for the oblivious AO version of HyperCast, highlights this design structure. The base code modules no longer depend on the crosscutting concerns, but the aspect modules do strongly depend on the base implementation modules. The cell for (row 21, column 16), for example, indicates that *ao_protocol_event* depends on *protocol_impl*. This dependence arises because the PCD of the aspect module depends on the form of the join points that signal protocol events in the base code, entirely under the control of an oblivious base code designer.

5. THE DESIGN RULES APPROACH

The oblivious design process assumes that it's bad for base developers to be aware of aspects, and that intimacy—the explicit preparation of code for advising by aspects—is even worse. Why? For one, the comprehensibility of the base code could be compromised: it might not manifest a design that matches the designer's conceptions of the base functionality. Two, the base code could contain tangled concern code, compromising ease of change. Similarly, explicit measures in the base code to anticipate certain extensions might effectively exclude others. Three, parallel development could be compromised, in that the base code designers would have to coordinate with the aspect developers on the proper way to prepare the base code for aspects.

Yet, as the previous section highlights, obliviousness in the design of the base code can cause numerous problems. The question, then, is whether a new method can be devised that realizes the benefits of AO design and yet minimizes or eliminates these problems. The first two issues could be addressed with a method that molds base code in ways that help aspect designers but without the need for any auxiliary code, such as event notifications, method calls, or tags. The OO design would be uncompromised, and there would be no tangling. The third problem in fact highlights an opportunity: the freedom afforded to base code designers by obliviousness delays aspect design because it flows precisely from the sequential nature of the design process. In the oblivious approach, pointcuts cannot be written and advice parameter lists cannot be formulated until the base code is written. A short design phase that establishes symmetric separation of concerns between base designers and aspect designers could actually increase overall parallel development.

The design rules methodology provides the basis for such a method. An essential idea is that for each crosscutting concern, a crosscutting design rule interface is established to decouple the base design and the aspect design. The constraints imposed by a design rule govern three things: (1) which execution phenomena must be exposed as join points, (2) how they are exposed through the JPM of the given language (e.g., in AspectJ, rules could govern syntax, name, and stack shapes), (3) constraints on behavior across join points (e.g., pre- and post-conditions for the execution of advice compositions). These elements ensure two important properties. One, the PCDs required by aspects can be constructed and will not have to change when the base code is evolved. Two, the state at a join point is the actual behavioral point in the execution of interest to the aspect, and that the system state after advice has returned is not compromised. Our method is thus an instance of the information hiding approach, but for crosscutting join-point-based interfaces.

A base/aspect design rule should result in easy-to-use join points that give base designers considerable implementation freedom. For ADT design, the key to an easy-to-use interface that provides freedom in implementation is that the interface should place reasonably minimal assumptions (constraints) on both the implementation and the clients. For example, an associative memory abstraction should not have an accessor method that reports the average collision chain length. Nor should the interface make assumptions about clients' intents, say by calling it a Dictionary and limiting use to short string inputs and outputs. For base/aspect design rules, then, we provide the dictum that the quantification of join points and the states at those points should be characterized in terms of the application's own concepts and abstractions, rather in terms of implementation-dependent aspect or base code details that are subject to change.

In HyperCast, for example, the design rules are best stated not in terms of the logging or notification aspects, but rather in terms of interesting abstract states and behaviors of the system, e.g., the abstract states of the finite state machine that tracks and manages the configuration and use of the overlay network. This reflective, application-centric view allows the base code designer to be oblivious to logging and other aspects, per se, and creates options for several possible aspect-oriented extensions to HyperCast such as mirroring and caching. Because design rules are formulated in terms

of the abstract system model, the resulting syntactic, naming and other constraints are consistent with base code's natural OO ontology. Although there is scattered work that has to be carried out to satisfy the rule, and the result is a crosscutting interface, we hypothesize that the result is practically indistinguishable from pure OO design. The following is a description and comparative evaluation of our method through our case study application.

5.1 Experimental application to HyperCast

To create interfaces for the two crosscutting concerns of logging and notification, we formulated eight major design rules. They constitute five interfaces (with rules 1–4 comprising one interface).

1. **State Update.** DR1: HyperCast's functionality is driven by the abstract state transitions of the protocol's finite state machine (FSM). This rule ensures that these transitions are visible to clients of HyperCast and it prohibits clients from interfering with HyperCast's core FSM behavior.
2. **Update Logical Address.** DR2: The changing of a Node's logical address is essential to the transition function. This rules ensures that the changes of logical address are visible, and it prohibits interference with base behavior.
3. **Update Neighborhood.** DR3: The changing of overlay topology is essential to the transition function. This rule ensures that changes of topology are visible to HyperCast clients, and it prohibits them from interfering with base protocol behavior.
4. **Join and Leave Overlay.** DR4: Expose the leaving and joining of the overlay
5. **Update Message Store State.** DR5: HyperCast's services are driven by abstract state transitions in a component called the MessageStore FSM. This rule ensures that these transitions are visible to clients of HyperCast.
6. **Throw Exception.** DR6: Expose exceptions with context information. This rule ensures that context information for exceptions is available to clients, and it prohibits clients from interfering with the base code behavior.
7. **Error Handler.** DR7: This design rule provides a unified error handling approach in HyperCast.
8. **Non-Exception Error Handling.** DR8: This rule ensures that error handling is done by the approach defined in DR7. Thus all error states are exposed to HyperCast clients.

Figure 3 presents the resulting structure, with base code on the left, design rules in the middle, aspects on the right, and dependences of base and aspect codes on DRs indicated by arcs.

We refactored HyperCast to implement these rules. We modified 65 places in the Protocol module to follow DRs 1–4, mostly in six classes, especially in MessageArrivedFromAdapter. We modified 21 places in the service classes to follow DR 5. All the modifications are simple, as the finite state machines were already implemented and we only used methods with naming conventions to expose them. We also removed the old event notification method, which occurred in 41 places across 10 classes. For DRs 7–8, we first identified 14 error types according to the specification. We then create a template to implement DR 7 and DR 8. The refactoring involved 55 error occurrences across 18 classes.

Such design rules would be used extensively by the aspect designers. Unlike APIs, our interfaces do not have any explicit representation in the source code. Rather, they simply constrain the manner in which the code is written. We document these interfaces in a style reminiscent of design patterns [8], as shown in Figure 5.

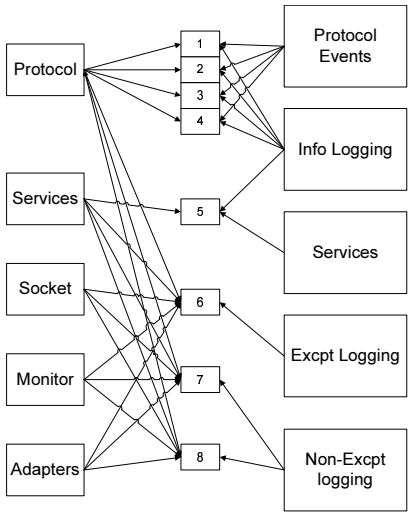

Figure 3: The crosscutting interface in HyperCast. The dependences of base and aspects on DRs is indicated by the arcs. Base functionality is on the left, design rules in the middle, and aspects on the right.

Figure 4 shows how the design rules decouple the base and aspect elements. Compared to the oblivious design's DSM in Figure 2, we observe that all dependences between the basic modules and aspect modules are removed. Instead, both the aspect modules and the basic modules now depend on the design rules.

We now compare the aspects implemented over the two different versions of the base code. In the oblivious AO design, the aspect modules are an average 240 lines each in length, whereas for the design-rules AO design the aspects average only 30 lines each for the same functionality. For example, Figure 6 shows the aspects for handling logical address events, using the two approaches. We observe that the aspect in design rules approach is much simpler because specific naming conventions were followed and interesting abstract states were implicitly exposed in all protocol modules. Without design rules, the aspect had to compute complex pointcuts by going through lots of details of base code. Moreover, the design-rule pointcut will capture newly-coded address changes, because it is quantified and base code designers are constrained to write new code following the design rule.

6. QUANTITATIVE ANALYSIS WITH NET OPTION VALUE

Sullivan et al. [24] and Lopes [17], have previously used *net option value* analysis [3] to quantitatively compare software designs modeled by DSMs. In this section, we evaluate the HyperCast system similarly based on the DSMs introduced in previous sections.

Baldwin and Clark's theory is based on the idea that modularity provides a portfolio of options. They define a model for reasoning about the value added to a system by its modularity. Splitting a design into N modules increases its base value S_0 by a fraction obtained by summing the net option values (NOV_i) of the resulting options. NOV is the expected payoff of exercising a search and substitute option optimally, accounting for both the benefits and cost of exercising options:

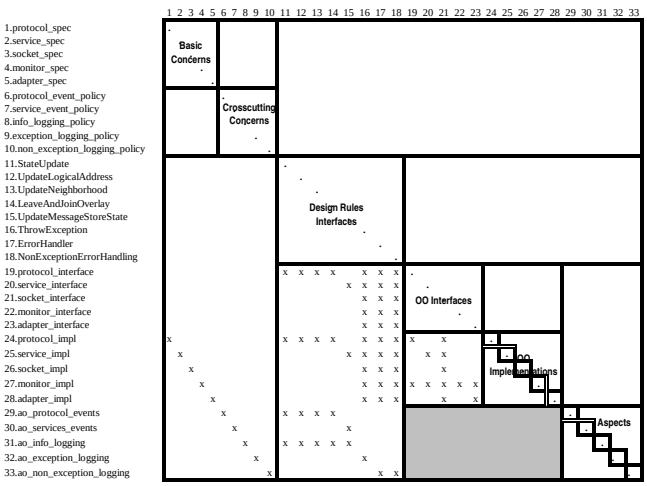

Figure 4: Design Rules Aspect-Oriented Design of HyperCast.

Name: State Update

Rationale: HyperCast's functionality is driven by the abstract state transitions of the protocol FSM. This design rule ensures that these transitions are visible to clients of HyperCast and alert clients that they may not interfere with HyperCast's code behavior.

Depends upon: none

Base code scope: implements `edu.virginia.cs.mng.hypercast.I_Node+`

Design Rule: *Provides:* Call to `void setState(byte)` at the conclusion of performing a state transition.
Requires: No changes to the trace of `edu.virginia.cs.mng.hypercast.I_Node+`

Example: A pointcut for advising all state transitions might be:

```
pointcut NodeStateChanged () :
  call (void I_Node+.setState(*));
```

Figure 5: Specification of DR 1, Update State.

$$V = S_0 + NOV_1 + NOV_i + ... + NOV_m$$
$$NOV_i = max_{k_i}\{\sigma_i n_i^{1/2} Q(k_i) - C_i(n_i)k_i - Z_i\}$$

For module i, $\sigma_i n_i^{1/2} Q(k_i)$ is the expected benefit to be gained by accepting the best positive-valued candidate generated by k_i independent experiments. $C_i(n_i)k_i$ is the cost to run k_i experiments as a function C_i of the module complexity n_i. $Z_i = \Sigma_{j \text{sees} i} c n_j$ is the cost of ripple effects of changes due to module dependences. The max picks the experiment that maximizes the gain for module i. Details of the NOV model can be found in the literature [3, 24, 17]. The two most important parameters for NOV analysis are *technical potential*, σ, and *complexity*, n.

Technical potential is the expected variance on the rate of return on an investment in producing a variant of a module implementation—that is, risk with commensurate rewards for success. On the assumption that the prevailing implementation of a module is adequate, the expected rate of return on investments is proportional to changes in requirements that drive the evolution of the module's specification. Consequently, a module whose specification does not change has low technical potential. For the evaluation of the benefits of the aspect-oriented designs, we operationalize technical potential as a stream of incoming change requests, with a given percentage of the changes impinging on the base modules, and the rest on the aspect modules. (We do not consider changes that crosscut between base functionality and the aspects.) We do not expect to see large advantages in the aspect-oriented designs, because only two crosscutting concerns have been modeled compared to 20 base concerns. Numerous other crosscutting concerns would need to be modeled for a considerable incremental value to be observed.

The standard method for estimating the complexity parameter for each module is to use the size of the artifact as a proportion of the overall system. Recognizing that lines of code (LOC) can include disproportionate inessential complexity for any given module, we decided to minimize this effect by using the LOC of the smallest version of each module amongst our three systems. (Recall that the oblivious aspects were on average eight times larger than the design rule aspects.) The design-rules version of the modules in this system invariably produced this effect, so we used its sizes in computing the proportions.

With these two estimations in hand, we computed the NOV for each design for several scenarios, from a low change rate of the base functionality relative to the aspects to a high change rate relative to the aspects. The comparison result is presented in Table 1. The precise values in the table cannot be taken as absolute truth. It remains an open challenge to justify precise estimates for real op-

```
privileged aspect EventTest {
 // have to compute pointcuts for each protocols
 pointcut DTLogicalAddrChanged():
  (execution(* DT_Neighborhood.
      DT_randomShiftMyCoordinates())||
  execution(* DT_Neighborhood.
      updateNodeAddress(DT_LogicalAddress)));

 pointcut HCLogicalAddrChanged():
  set(I_LogicalAddress HC_Node.MyLogicalAddress)&&
  (withincode (void HC_Node.
      messageArrivedFromAdapter(I_Message))||
   withincode(void HC_Node.timerExpired(Object))||
   withincode(void HC_Node.resetNeighborhood()));

 pointcut  SPTLogicalAddrChanged():
   execution (* SPT_Node.setLogicalAddress
              (I_LogicalAddress));

 // advice
 after() : DTLogicalAddrChanged() || ... {
    // handle address changes here
  }
// for other events...
}
```

(a)

```
privileged aspect DREventTest {
 pointcut logicalAddrChanged(I_LogicalAddress):
  execution(* I_Node+.
       setMyLogicalAddress(I_LogicalAddress));

 after() : logicalAddrChanged() {
  // handle address changes here
 }
// for other events...
}
```

(b)

Figure 6: Aspect for LogicalAddressChanged Events, (a) without design rules, and (b) with design rules.

tions in software design. As discussed in our earlier work, estimating technical potential is difficult [24]. However, as a back-of-the envelope model, it provides ballpark figures and useful insights.

We first observe that the more likely it is for the crosscutting concerns to change (the left side of the table), the more relative value is added by either AO design. The value of the modularity in the OO design goes to zero because its modularity is unused. The value of the two AO designs similarly converges to the same value, because they only differ in how they relate to the base code.

Towards the right of the table, as the relative change rate of the base functionality goes above 90%, the oblivious design becomes even worse than the OO design. This is because the oblivious design's aspects are dependent on the base modules, which are changing at a high rate, thus producing a high Z.

In the middle to upper middle part of the table, presumably closer to where the relative base change rate is likely to lie, we see that the design-rule design outperforms the oblivious design, entirely because its Z is zero when the base functionality evolves, due to the join point design rules. Both aspect-oriented designs outperform the OO design.

7. CONCLUSION

In his seminal paper on information hiding [19], Parnas showed the deleterious effects of scattered dependences on design decisions that are complex or likely to change. He showed that they make programs hard to understand, develop in parallel, and change at reasonable cost. These technical difficulties ultimately manifest themselves in ways that matter to people: poorer dependability, lower productivity, and fewer valuable choices. Parnas then showed designers how to do better using abstract interfaces to decouple decisions. In this paper, we have revisited Parnas's ideas in light of the addition of join point models and quantified (crosscutting) advising to the programmer's toolkit.

An AOP language like AspectJ implicitly publishes a vast array of join points for a given base program. Like the exposed data structures of Parnas's functional design, they create valuable opportunities for aspect designers. The problem is that, if not managed, they invite scattered dependences on unstable, complex design decisions. Oblivious aspect-oriented design, however, dictates that the base code designer should make no preparations for aspects. The difficulty, then, is that the aspect designer simply has to live with the arbitrary decisions of the base code designer, and cannot count on the availability, simplicity, or semantics of join points.

This paper provides a practical alternative criterion. Identify the crosscutting concerns that are likely to change; for each, define a crosscutting interface in the form of constraints on the exposure and semantics of certain join points. The evidence from our case study on the complex, modern HyperCast system supports the claim that this criterion provides qualitatively and quantitatively better modularity than obliviousness (or regular OO design). The principal difference is that design rule interfaces symmetrically decouple base code and aspect code, whereas oblivious design creates a sequential dependence of aspects on base code. We sacrifice design obliviousness, but our application-centric approach to imposing design rules preserves feature obliviousness. Designers are aware of crosscutting in their application, but not of what aspect behaviors advise the crosscuts.

Parnas's interfaces are procedural and hierarchical. Ours fall in the more general class of design rules: constraints that decouple otherwise coupled design processes. Our non-procedural and non-hierarchical (crosscutting, join point) interfaces modularize design decisions that would be scattered by OO design or subject to revision when the base code evolves in oblivious design. It's better for our interfaces to crosscut than the design decisions themselves. First, our interfaces are stable and simple; the design decisions are not. Second, our interfaces say only how to write code, not to write more code. No extra or tangled code is left in the base code.

Recently, in a paper on the nature of crosscutting interfaces and modularity in AO design, Kiczales and Mezini wrote [13]:

> *Aspects cut new interfaces through the primary decomposition of a system. This implies that in the presence of aspects, the complete interface of a module can only be determined once the complete configuration of modules in the system is known. While this may seem anti-modular, it is an inherent property of crosscutting concerns, and using aspect-oriented programming enables modular reasoning in the presence of such concerns.*

This definition supposes that aspects determine the interfaces of modules by the join points they actually use. What Kiczales and Mezini compute are actually *dependences* on join points in an overall system. In the absence of agreement on the assumptions that aspect designers *may* make about join points, revealing those that they *do* make is an important enabler of modular reasoning and change. With our approach, sets of permissible assumptions are specified explicitly as interfaces; there is no need for a concept of unstable interfaces inferred from whole system configurations.

Aldrich has recently proposed scoping constructs and a formal semantics that relate to our design rules. The Open Modules system focuses on the exposure of join points such that module state that is intended to be hidden cannot be advised [1]. Simply, a module has to declare a pointcut in order to export join points on private state. Thus, it permits the evolution of a module implementation without requiring rework of aspects. The semantics also supports automatically checking these properties. However, the resulting interfaces are not crosscutting, so Open Modules cannot regularize the exposure of join points across non-hierarchical parts of a system. Thus, aspect developers cannot be guaranteed that they can write the aspects required, or that they will not have to change when new base functionality is added.

Today, prominent industrial development teams are reporting successes with AOP. How can this be if the oblivious approach has such problems? Based on our informal discussions with insiders, it seems that some teams are formulating and following what amounts to rules for base code designers that aspect designers can depend upon. This work recognizes such good practice and raises it to the level of a general software engineering principle.

Crosscutting interfaces are not wholly new. Designers of distributed and real-time systems have long provided programming rules to ensure that protocols, although not modularized, are written in a mutually consistent manner so as to achieve the desired result. The naming and coding conventions enforced on a project also function as crosscutting interfaces to an extent, aiding both comprehensibility and the use of tools to manipulate crosscutting code [9]. Based on our experience with HyperCast, we feel that the consistency enforced by our design rules have similar positive effects on comprehension.

The need for aspects as a distinct abstraction mechanism from classes is a matter of on-going debate in the AOP community. Our comparative study shows that dichotmizing base functionality and aspects in design can be counterproductive. Rajan and Sullivan similarly showed that the base/aspect dichotomy in programming languages with advising has technical drawbacks [21], although it reportedly helped promote early adoption. Not all AOP approaches manifest a dichotomy, for example Eos [21] and the HyperSlice

Basic σ	1%	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	99%
OO NOV	0	0	0	0.00164	0.077	0.17	0.31	0.50	0.78	1.09	1.26	1.402
Ob. AO NOV	0.537	0.483	0.423	0.375	0.375	0.41	0.49	0.62	0.83	1.10	1.255	1.391
% impr.	N/A	N/A	N/A	2179.55	386.89	141.18	58.06	24.00	6.41	0.92	-0.40	-0.78
DR AO NOV	0.537	0.483	0.423	0.385	0.389	0.42	0.51	0.66	0.87	1.14	1.29	1.422
% impr.	N/A	N/A	N/A	2244.62	404.78	147.06	64.52	32.00	11.54	4.59	2.38	1.43

Table 1: Net option values for the OO, oblivious AO, and design rules AO designs. Each basic column represents the assumption that the percentage of requested software improvements is confined to the base functionality (versus the aspect functionality) is σ.

compile-time weaving approach [25]. Without this a priori distinction, the question remains: how best to decouple concerns in the design process when given the unique mechanism of quantified advising over a join point model? Our design rules approach provides the basis for an answer.

Open questions remain. By what notation and semantics should our interfaces be specified? Section 5 suggests one possibility, but it has not been fleshed out or widely tested. To what extent can our interfaces be checked, and how? AspectJ's `declare error` construct could check some rules, especially prohibitions. Can integrated programming environments like Eclipse's AJDT make these crosscutting rules more visible and ease their maintenance? AJDT has an effective model for exposing advising; what about rules governing advising, if they are written as pointcuts?

8. REFERENCES

[1] J. Aldrich. Open modules: Modular reasoning about advice. In *2005 European Conference on Object-Oriented Programming (ECOOP'05)*, page To Appear, July 2005.

[2] AspectJ project. http://www.eclipse.org/aspectj/.

[3] C. Y. Baldwin and K. B. Clark. *Design Rules: The Power of Modularity*. MIT Press, Cambridge, MA, 2000.

[4] C. Constantinides and T. Skotiniotis. Reasoning about a classification of cross-cutting concerns in object-oriented systems. In P. Costanza, G. Kniesel, K. Mehner, E. Pulvermüller, and A. Speck, editors, *Second Workshop on Aspect-Oriented Software Development of the German Information Society*. Institut für Informatik III, Universität Bonn, Feb. 2002. Technical report IAI-TR-2002-1.

[5] T. Elrad, R. E. Filman, and A. Bader. Aspect-oriented programming: Introduction. *Commun. ACM*, 44(10):29–32, 2001.

[6] R. E. Filman and D. P. Friedman. Aspect-oriented programming is quantification and obliviousness. In *Workshop on Advanced Separation of Concerns (OOPSLA 2000)*, Oct. 2000.

[7] R. E. Filman and D. P. Friedman. Aspect-oriented programming is quantification and obliviousness. In *Aspect-Oriented Software Development*, pages 21–35. Addison-Wesley, 2005.

[8] E. Gamma, R. Helm, J. Vlissides, and R. E. Johnson. *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley, Reading, MA, 1995.

[9] W. G. Griswold. Coping with crosscutting software changes using information transparency. In *Reflection 2001: The Third International Conference on Metalevel Architectures and Separation of Crosscutting Concerns*, pages 250–265, Kyoto, Sept. 2001.

[10] Hypercast project. http://www.cs.virginia.edu/ mngroup/hypercast/.

[11] S. Katz. Diagnosis of harmful aspects using regression verification. In C. Clifton, R. Lämmel, and G. T. Leavens, editors, *FOAL: Foundations Of Aspect-Oriented Languages*, pages 1–6, Mar. 2004.

[12] G. Kiczales, E. Hilsdale, J. Hugunin, M. Kersten, J. Palm, and W. G. Griswold. An overview of AspectJ. In *15th European Conference on Object-Oriented Programming (ECOOP 2001)*, pages 327–353, June 2001.

[13] G. Kiczales and M. Mezini. Aspect-oriented programming and modular reasoning. In *ICSE '05: Proceedings of the 27th international conference on software engineering*, 2005.

[14] R. Laddad. *AspectJ in Action: Practical Aspect-Oriented Programming*. Manning Publications, 2003.

[15] J. Liebeherr and T. K. Beam. Hypercast: A protocol for maintaining multicast group members in a logical hypercube topology. In *Networked Group Communication*, pages 72–89, 1999.

[16] J. Liebeherr, M. Nahas, and W. Si. Application-layer multicasting with delaunay triangulation overlays. *EEE Journal on Selected Areas in Communications*, 20(8), oct 2002.

[17] C. V. Lopes and S. K. Bajracharya. An analysis of modularity in aspect oriented design. In *AOSD '05: Proceedings of the 4th international conference on Aspect-oriented software development*, pages 15–26. ACM Press, 2005.

[18] J. Marshall, D. Orleans, and K. J. Lieberherr. DJ: Dynamic structure-shy traversal in pure Java. Technical report, Northeastern University, May 1999.

[19] D. L. Parnas. On the criteria to be used in decomposing systems into modules. *Commun. ACM*, 15(12):1053–1058, Dec. 1972.

[20] H. Rajan and K. Sullivan. Aspect language features for concern coverage profiling. In *the Fourth International Conference on Aspect-Oriented Software Development (AOSD 2005)*, March 2005.

[21] H. Rajan and K. Sullivan. Classpects: Unifying aspect- and object-oriented language design. In *Proceedings of the 27th International Conference on Software Engineering ICSE 2005)*, page To appear, May 2005.

[22] G. Steele. *Common LISP: The Language*. Digital Press, 2nd edition, 1990.

[23] D. V. Steward. The design structure system: A method for managing the design of complex systems. 28(3):71–84, 1981.

[24] K. J. Sullivan, W. G. Griswold, Y. Cai, and B. Hallen. The structure and value of modularity in software design. *SIGSOFT Softw. Eng. Notes*, 26(5):99–108, 2001.

[25] P. Tarr, H. Ossher, W. Harrison, and S. M. Sutton. N degrees of separation: Multi-dimensional separation of concerns. In *Proceedings of the 1999 International Conference on Software Engineering*, pages 107–119, May 1999.