LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-PROC-231453

Analyzing and Visualizing Whole Program Architectures

T. Panas, D. Quinlan, R. Vuduc

June 4, 2007

ICSE Workshop on Aerospace Software Engineering (AeroSE)
Minneapolis, MN, United States
May 21, 2007 through May 22, 2007

Analyzing and Visualizing Whole Program Architectures

Thomas Panas Dan Quinlan Richard Vuduc
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
{panas2, dquinlan, richie}@llnl.gov

1. Introduction

This paper describes our work to develop new tool support for analyzing and visualizing the architecture of complete large-scale (millions or more lines of code) programs. Our approach consists of (i) creating a compact, accurate representation of a whole C or C++ program, (ii) analyzing the program in this representation, and (iii) visualizing the analysis results with respect to the program's architecture. We have implemented our approach by extending and combining a compiler infrastructure and a program visualization tool, and we believe our work will be of broad interest to those engaged in a variety of program understanding and transformation tasks.

We have added new whole-program analysis support to ROSE [15, 14], a source-to-source C/C++ compiler infrastructure for creating customized analysis and transformation tools. Our whole-program work does not rely on procedure summaries; rather, we preserve all of the information present in the source while keeping our representation compact. In our representation, a million-line application fits in well less than 1 GB of memory. Because whole-program analyses can generate large amounts of data, we believe that abstracting and visualizing analysis results at the architecture level is critical to reducing the cognitive burden on the consumer of the analysis results. Therefore, we have extended Vizz3D [19], an interactive program visualization tool, with an appropriate metaphor and layout algorithm for representing a program's architecture. Our implementation provides developers with an intuitive, interactive way to view analysis results, such as those produced by ROSE, in the context of the program's architecture.

The remainder of this paper summarizes our approach to whole-program analysis (Section 2) and provides an example of how we visualize the analysis results (Section 3).

2. Whole-Program Analysis

Our source code analysis and transformation work is part of the ROSE project, sponsored by the U.S. Department of Energy (DOE). Although research in the ROSE project emphasizes performance optimization, ROSE contains many of the components and analyses common to any compiler infrastructure, and thus facilitates the development of a broad range of source-based analysis tools. ROSE routinely compiles million-line applications. This section describes our whole-program analysis implementation in ROSE.

2.1. Representing whole programs

Whole-program analysis is typically implemented either using procedure summaries or by embedding information into the object files to use whole-program context at link-time. In ROSE, we approach the problem differently, namely, by using a space-efficient representation of the complete source.

We use the Edison Design Group C++ front-end (EDG) [7] to parse C and C++ programs. EDG generates an abstract syntax tree (AST) and fully evaluates all types. We translate the EDG AST into our own object-oriented AST, SAGEIII. The SAGEIII intermediate representation (IR) has 240 types of IR nodes and can fully represent original structure of the application, including the preservation of comments and preprocessor control structure. From the IR, the original source may be reproduced completely.

The IR is space-efficient by design. In particular, we share parts of the AST (subtrees) that are determined to be identical. This technique is critical for C and C++. For example, a typical million-line application compiled by ROSE has about 1000 files containing approximately 1,000 lines of source code (LOC) per source file, where each source file includes in addition about 75,000 LOC from header files. In this scenario, the effective 76,000 LOC per source file generate 1,000 ASTs with each containing about 500,000 IR nodes. To support whole-program analysis, ROSE merges multiple ASTs from the compilation of different source files into a single AST. Merging 75,000 re-occuring LOC over each of the 1000 files saves 75 million LOC from being represented redundantly in the AST.

Figure 1 (top) shows the AST for three example source

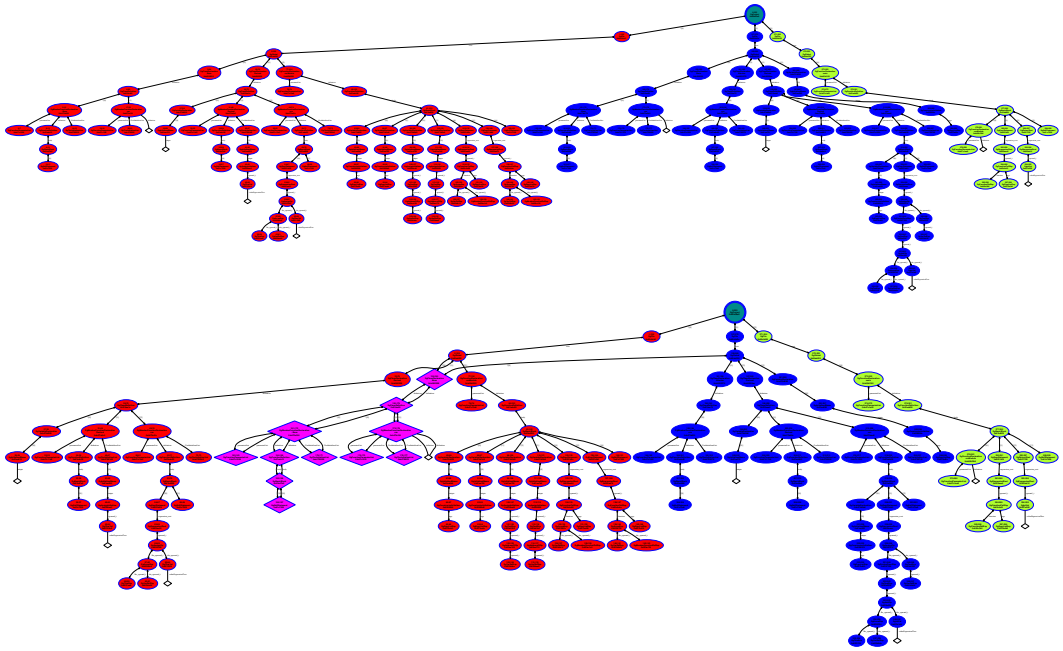

Figure 1. (*Top*) AST before merging. File 1 = green nodes, File 2 = blue nodes, File 3 = red nodes. (*Bottom*) AST after merging. The magenta subtree shows common (merged) structure.

files, with AST subtrees colored by file. The ASTs from the files are not shared. Figure 1 (bottom) shows the AST after the merge process, where the diamond shaped IR nodes of the AST indicate that those IR nodes are shared. To be shared, the declaration at the root of the subtrees had to generate the same internal name (in C++, this includes standard name mangling plus a number of other language specific details) *and* the subtrees had to pass the One-time Definition Rule (ODR) test of equivalence. For a more detailed description of our merge algorithm, see our recent paper [14].

In our initial experiments, we have merged the full source of SMG2000 [1], a semicoarsing multigrid code consisting of 53 files. The ROSE AST initially consists of 2,287,000 nodes, and is merged into just 779,607 nodes. This indicates a merge efficiency of 66%; the maximum possible merge efficiency is 100%, which would occur when merging two identical files.

We have estimated the theoretically best merge efficiency for SMG2000 to be 77%. For this, we compare the total amount of LOC in source and header files (including library headers) with the LOC of the preprocessed source files of SMG2000, embedding header files repeatedly into the source files. The numbers are 31,628 LOC and 138,456 LOC, respectively (with empty lines removed). Finally, for SMG2000, the time to merge all 2,287,000 nodes is about 30% of the time spent on the entire compilation process.

Using a 250 KLOC benchmark, we have estimated that a 1,000 KLOC application will fit into approximately 400 MB of memory after merging header files. The AST holding the million-line application can also be saved to and loaded from disk using a custom ROSE-specific binary file format; on current single-processor desktop machines, one of these binary files can be written in roughly 30 sec and read in under a minute. Simple traversals (for code analysis purposes) of the whole AST (in memory) are expected to take only a few seconds.

2.2. Analysis

ROSE internally implements a number of forms of procedural and interprocedural analysis, with much of this work in development. ROSE currently supports analyses such as:

Metrics in order to measure certain aspects of a software system to interactively or automatically evaluate design specifications, such as LOC per function, function complexity measures or arithmetic complexities. Arithmetic metrics help to detect computationally expensive functions and classes. This property is particularly important in scientific computing codes, since such functions should be the most robust and reliable pieces of the software.

Static Analyses aid the analysis of a software system without the system being executed. ROSE includes e.g. analyses that detect public declared variables (within

the scope of classes) and global variables (outside the scope of classes) and unsafe function calls that can lead to buffer overflows, page faults, and segmentation faults (such as `sprintf`, `scanf`, `strcpy`). ROSE supports also more sophisticated analyses for system dependence, call graph, control and data flow analysis. In collaboration with academic groups, we are extending the analysis infrastructure to interface with general analysis tools, including PAG [2], OpenAnalysis [17], as well as analysis tools specifically for automated debugging and security, MOPS for finite state machine-based temporal specification checking [4], and coverage analysis tools [6].

Dynamic Analyses are used to attribute the ROSE AST with runtime information. This allows us to detect and visualize program locations with high execution costs.

Binary Analysis is the analysis and inspection of program binaries. We can represent binary information in the ROSE IR in order to apply analysis mechanisms developed for static source analysis to binaries as well.

We expect our approach towards whole program analysis to be most valuable in terms of program analysis precision and scalability; which, in turn, allows us also to experiment with large scale application visualizations.

3. Architecture Visualization

The aim of architecture-level visualization is to rapidly summarize and communicate the architecture and design decisions of the overall software system. Architectural visualization is more abstract than low-level visualizations (from low-level analyses) [13], and therefore better suited to visualizations in the large. An architectural visualizations combined with metrics can help software developers to answer many questions about a software system.

Common examples of architectural visualizations are function call graphs, hierarchy graphs, and directory structures. There are many ways to present these graphs, such as UML diagrams and graph browsers. For example, Figure 2 shows various analysis results within one image, in order to reduce cognitive burden a viewer would otherwise experience when looking at multiple views [16]. However, this image shows a huge amount of information, making techniques for information reduction necessary.

To help a viewer better navigate and understand analysis results of large-scale applications, we have implemented a 3D city metaphor, a predictable layout, and abstraction and navigation mechanisms within our visualization tool, Vizz3D. We show an example of visualizing SMG2000 in Figure 3. This image shows exactly the same information as Figure 2, with the difference that a 3D city metaphor and our predictible layout algorithm is used.

More specifically, Figure 3 is a snapshot taken of Vizz3D while running SMG2000. The directories of this application

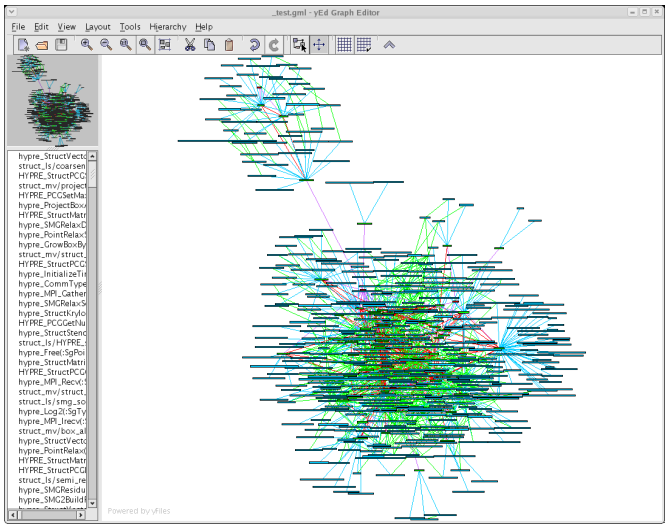

Figure 2. Architecture Visualization Example.

appear as "islands," individual files as "cities" within the island, and individual function definitions as "buildings." In addition, aggregate shaded edges between cities indicates that some function in one file (red end) calls some function in another file (dark end). Other user-selected metrics and analyses (whether static or dynamic) may be rendered as textures, colors, and icons in this view. We believe that the right choice of metaphor and layout are crucial, and we will investigate this claim in future studies.

4. Related Work

Whole-program analysis has traditionally been applied in performance optimization contexts [3, 18], but has recently also been used to find bugs and detect security flaws using global dataflow analyses [10, 11, 8]. Our techniques complement earlier work by providing the basic infrastructure for accurately representing the source of an entire program but for purposes of program understanding. Among other open C or C++ infrastructures [9, 12, 5] and C++ static analysis infrastructures [20], our basic mechanisms for building whole-program representations are unique.

Acknowledgements. This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livremore National Laboratory under Contract W-7405-Eng-48.

References

[1] The SMG2000 Benchmark, 2001. llnl.gov/asci/platforms/purple/rfp/benchmarks/limited/smg.

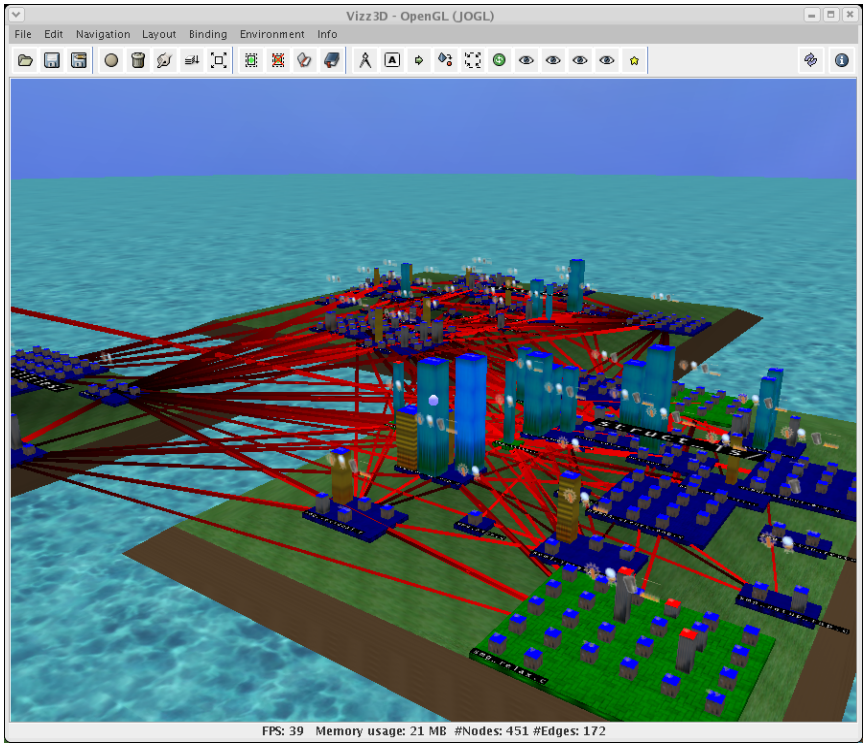

Figure 3. Our Architecture Visualization.

[2] AbsInt, Inc. PAG: The Program Analysis Generator, 2006. absint.com/pag.

[3] D. C. Atkinson and W. G. Griswold. The design of whole-program analysis tools. In *Proc. International Conference on Software Engineering*, Berlin, Germany, March 1996.

[4] H. Chen, D. Dean, and D. Wagner. Model checking one million lines of C code. In *Proc. Network and Distributed System Security Symposium*, San Diego, CA, USA, 2004.

[5] S. Chiba. Macro processing in object-oriented languages. In *TOOLS Pacific '98, Technology of Object-Oriented Languages and Systems*, 1998.

[6] O. Edelstein, E. Farchi, E. Goldin, Y. Nir, G. Ratsaby, and S. Ur. Testing multithreaded Java programs. *IBM Systems Journal: Special Issue on Software Testing*, February 2002.

[7] Edison Design Group. EDG front-end. `edg.com`.

[8] D. Engler and M. Musuvathi. Static analysis versus software model checking for bug finding. In *Proc.International Conference on Verification, Model Checking, and Abstract Interpretation*, Venice, Italy, 2004.

[9] F. S. Foundation. GNU Compiler Collection, 2005. gcc.gnu.org.

[10] S. Z. Guyer, E. D. Berger, and C. Lin. Detecting errors with configurable whole-program dataflow analysis. In *Proc. Conference on Programming Language Design and Implementation*, Berlin, Germany, 2002.

[11] D. L. Heine and M. S. Lam. A practical flow-sensitive and context-sensitive C and C++ memory leak detector. In *Proc. Conference on Programming Language Design and Implementation*, pages 168–181, June 2003.

[12] G. Keating. Inter-module analysis in GCC. In *Proc. GCC Developers' Summit*, Ottowa, Canada, June 2005.

[13] T. Panas. *Towards a Generic Framework for Reverse Engineering*. Licentiate thesis, Växjö University, Sweden, November 2003.

[14] D. Quinlan, R. Vuduc, T. Panas, J. Härdtlein, and A. Sæbjørnsen. Support for whole-program analysis and verification of the One-Definition Rule in C++. In *Proc. Static Analysis Summit*, Gaithersburg, MD, USA, June 2006. National Institute of Standards and Technology Special Publication.

[15] M. Schordan and D. Quinlan. A source-to-source architecture for user-defined optimizations. In *Proc. Joint Modular Languages Conference*, 2003.

[16] M.-A. D. Storey, F. D. Fracchia, and H. A. Mueller. Cognitive design elements to support the construction of a mental model during software visualization. In *Proc. of the 5th Int. Workshop on Program Comprehension (WPC '97)*, Washington, DC, USA, 1997. IEEE Computer Society.

[17] M. M. Strout, J. Mellor-Crummey, and P. D. Hovland. Representation-independent program analysis. In *ACM SIGPLAN-SIGSOFT Workshop on Program Analysis for Software Tools and Engineering*, September 2005.

[18] S. Triantafyllis, M. J. Bridges, E. Raman, G. Ottoni, and D. I. August. A framework for unrestricted whole-program optimization. In *Proc. on Programming Language Design and Implementation*, Ottowa, Canada, June 2006.

[19] Vizz3D. Available at: `http://vizz3d.sourceforge.net`, July 2006.

[20] D. Wilkerson. OINK: A collection of composable C++ static analysis tools, 2005. freshmeat.net/projects/oink.