



Kent Academic Repository

Rigger, Manuel, Marr, Stefan, Adams, Bram and Mössenböck, Hanspeter (2019) *Understanding GCC Builtins to Develop Better Tools*. In: ESEC/FSE 2019 Proceedings of the 2019 27th ACM. Proceedings of the 27th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. ESEC/FSE . pp. 74-85. ACM ISBN 978-1-4503-5572-8.

Downloaded from

<https://kar.kent.ac.uk/75258/> The University of Kent's Academic Repository KAR

The version of record is available from

<https://doi.org/10.1145/3338906.3338907>

This document version

Author's Accepted Manuscript

DOI for this version

Licence for this version

UNSPECIFIED

Additional information

Versions of research works

Versions of Record

If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

Author Accepted Manuscripts

If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in *Title of Journal*, Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries

If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our [Take Down policy](https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies) (available from <https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies>).

Understanding GCC Builtins to Develop Better Tools

Manuel Rigger
Johannes Kepler University Linz
Austria
manuel.rigger@jku.at

Bram Adams
Polytechnique Montréal
Canada
bram.adams@polymtl.ca

Stefan Marr
University of Kent
United Kingdom
s.marr@kent.ac.uk

Hanspeter Mössenböck
Johannes Kepler University Linz
Austria
hanspeter.moessenboeck@jku.at

ABSTRACT

C programs can use compiler builtins to provide functionality that the C language lacks. On Linux, GCC provides several thousands of builtins that are also supported by other mature compilers, such as Clang and ICC. Maintainers of other tools lack guidance on whether and which builtins should be implemented to support popular projects. To assist tool developers who want to support GCC builtins, we analyzed builtin use in 4,913 C projects from GitHub. We found that 37% of these projects relied on at least one builtin. Supporting an increasing proportion of projects requires support of an exponentially increasing number of builtins; however, implementing only 10 builtins already covers over 30% of the projects. Since we found that many builtins in our corpus remained unused, the effort needed to support 90% of the projects is moderate, requiring about 110 builtins to be implemented. For each project, we analyzed the evolution of builtin use over time and found that the majority of projects mostly added builtins. This suggests that builtins are not a legacy feature and must be supported in future tools. Systematic testing of builtin support in existing tools revealed that many lacked support for builtins either partially or completely; we also discovered incorrect implementations in various tools, including the formally verified CompCert compiler.

CCS CONCEPTS

• **Software and its engineering** → **Language features; Compilers.**

KEYWORDS

GCC builtins, compiler intrinsics, C GitHub projects

ACM Reference Format:

Manuel Rigger, Stefan Marr, Bram Adams, and Hanspeter Mössenböck. 2019. Understanding GCC Builtins to Develop Better Tools. In *Proceedings of the 27th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '19)*, August 26–30, 2019, Tallinn, Estonia. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3338906.3338907>

ESEC/FSE '19, August 26–30, 2019, Tallinn, Estonia

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *Proceedings of the 27th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '19)*, August 26–30, 2019, Tallinn, Estonia, <https://doi.org/10.1145/3338906.3338907>.

1 INTRODUCTION

Most C programs consist not only of C code, but also of other elements, such as preprocessor directives, freestanding assembly code files, inline assembly, compiler pragmas, and compiler builtins. While recent studies have highlighted the role of linker scripts [20] and inline assembly [53], compiler builtins have so far attracted little attention. Builtins resemble functions or macros; however, they are not provided by libc, but are directly implemented in the compiler. The following code fragment shows the usage of a GCC builtin that returns the number of leading zeros in an integer's binary representation:

```
int leading_zeros = __builtin_clz(INT_MAX); // returns 1
```

On Linux, we observed that GCC builtins are widely used and are also supported by other mature compilers, such as Clang [6] and ICC [42]. For developers working on tools that process C code, such as compilers as well as static and dynamic analysis tools, implementation and maintenance of GCC builtins is a large effort, as we identified a total number of 12,339 GCC builtins, all of which are potentially used by projects and thus need to be supported. Hence, to assist developers of tools that process C code, the goal of this study was to investigate the use of builtins and how current tools support them. To this end, we analyzed the builtin use of 4,913 projects from GitHub and implemented a builtin test suite, which we used to test popular tools employed by C developers.

By combining quantitative and qualitative analyses, we answer the following research questions (RQs):

RQ1: How frequently do projects use builtins? Knowing the prevalence of builtins helps tool writers to judge the importance of implementing support for them. We hypothesized that builtins are used by many projects, and that any program that processes C code will therefore encounter them, yet—similar to inline assembly [53]—we expected that they are used in only a few source-code locations.

RQ2: For what purposes are builtins used? Knowing the primary use cases for builtins helps tool developers to judge whether their tools can support them. For example, static analysis tools might lack support for multithreading and hence be unable to deal with atomic builtins used for synchronization.

RQ3: How many builtins must be implemented to support most projects? Tool authors who have decided to support GCC builtins would find it helpful to know the implementation order that would maximize the number of projects supported.

RQ4: How does builtin usage develop over time? Understanding the usage of builtins over time could tell us whether projects continue

to add builtins or remove them. If builtins were a legacy feature of compilers that projects sought to remove, the incentive of tool developers to implement them would be low.

RQ5: How well do tools support builtins? To determine the room for improvement in tools, we examined how well existing tools support builtins. Our assumption was that state-of-the-art compilers such as GCC, Clang, and ICC provide full support, while other tools provide partial or no support.

We found the following:

- 12,339 GCC builtins exist, but only 3,083 were used in our corpus of projects;
- 37% of the projects used builtins.
- Projects primarily used architecture-independent builtins, for example, to interact with the compiler, for bit-level operations, and for atomic operations. However, when a project would use an architecture-specific builtin, it would often be used many times in the same project.
- While mature compilers seem to provide full support for builtins, most other tools lack some builtins or have some implemented incorrectly. Notably, we found two incorrectly implemented GCC builtins in an unverified part of the formally verified CompCert compiler.
- The effort of supporting a specific number of projects rises exponentially; for example, to support half of the projects only 32 builtins are needed. Supporting 99% of the projects, however, requires about 1,600 builtins.
- Over time, most of the projects increasingly used builtins; nevertheless, a number of projects removed builtin uses to reduce maintenance effort.

Our results are expected to help tool developers in prioritizing implementation effort, maintenance, and optimization of builtins. Thus, this study facilitates the development of compilers such as GCC, Clang [24], ICC, and the formally verified CompCert compiler [25, 26]; of static-analysis tools such as the Clang Static Analyzer [66], splint [14, 15], Frama-C [65], and uno [17]; of semantic models for C [23, 31]; and of alternative execution environments and bug-finding tools such as KLEE [3], Sulong [52, 56], the LLVM sanitizers [58, 59], and SoftBound [36, 37]. For reproducibility and verifiability, we provide the database with GCC builtin usage, test suite, tools used for the analysis, and a record of the manual decisions on <https://github.com/jku-ssw/gcc-builtin-study>.

2 METHODOLOGY

To answer our research questions, we analyzed builtin use in a large number of C projects and populated a SQLite3 database with the extracted data. As detailed below, we downloaded and filtered projects from GitHub on which we performed a textual search for the names of the GCC builtins. To identify the builtin names, we extracted them from the official documentation and from the GCC source code. To exclude false-positive identifications of builtin use, we applied heuristics such as excluding builtin names inside string literals and comments.

Selecting the projects. We analyzed projects from the popular GitHub code-hosting service. Similar to other large empirical studies [5, 44, 64], we selected projects based on popularity, specifically the number of GitHub stars [2]. To obtain about 5,000 projects,

we downloaded projects down to 80 stars. This cutoff point was sufficiently large to prevent the inclusion of personal projects, homework assignments, and forks [19]. In total, we downloaded 4,998 GitHub projects that contained in total 1,124 million lines of C code. This strategy allowed us to obtain a diverse set of projects (see Table 1). To provide further evidence for the diversity of the projects, we computed Nagappan et al.’s coverage score [35]. For this, we used the manually validated GitHub project metadata of Munaiah et al.’s RepoReapers data set [33], which contained 2,329 of our studied projects. This subset alone obtained a coverage score of 0.966 with respect to the universe of 145,355 C projects in the RepoReapers data set. This indicates that our project sample is both representative and diverse.

Table 1: Overview of the projects obtained (after filtering); the first commit in 1984 stems from a project that was converted from another version-control system.

Metric	Minimum	Maximum	Average	Median
C LOC	100	37M	228k	10k
# commits	1	668k	4872	1147
# committers	1	17k	120	54
first commit	1984-02-21	2017-11-06	-	2011-04-12
last commit	2003-12-08	2017-11-24	-	2017-11-07

Filtering the projects. From the downloaded projects, we selected 4,913 by filtering out those that did not meet our needs. First, we filtered out all projects that had fewer than 100 LOC, as we considered them too small to constitute C projects. GCC, forks of GCC¹, and other C/C++ compilers (such as ROSE [45]) implement the GCC builtins themselves, use them internally, and exercise them in their test suites. Hence, to avoid a high number of false positives, we excluded these projects; they were easy to identify, as they contained the largest numbers of unique builtins.

Identifying the builtins. Next, we identified the names of the available GCC builtins, to then perform a textual search on the GitHub projects. Identifying the list of names was difficult, since GCC builtins are not described or specified in a coherent manner, as they were added over a period of more than 30 years. Thus, we investigated both (I) builtins listed in the GCC documentation as well as (II) builtins internal to GCC, which we automatically extracted from GCC’s source code (including test cases for builtins).

(I) Builtins from the documentation. Initially, we considered only builtins described by the GCC documentation. The GCC documentation stated that some builtins are internal, which we initially did not want to include as we expected that other projects would not use them. While extracting the names of architecture-independent builtins worked well, GCC also provides builtins that are specific to an architecture. For example, `__builtin_ia32_paddq` allows the use of x86’s `paddq` instruction. In some cases, architecture-specific builtins were not described by the documentation, but referred to vendor documentation, for example, the ARM C Language Extensions. For these builtins, the documentation of GCC version 4.8 contained a list of builtins, which we used instead. However, for certain special-purpose architectures, obtaining such a list was impractical,

¹The projects filtered out included the GCC fork for the Xtensa processor (<https://github.com/jcmvbkbc/gcc-xtensa>), and a fork that is based on GCC to dump an XML description of C++ code (<https://github.com/gccxml/gccxml>).

for example, for the TILE-Gx and TILEPro processor builtins. As we expected little influence on the results—overall, architecture-specific builtins were used infrequently (see Section 3.2)—we omitted analyzing these special-purpose builtins. In total, this process yielded 6,040 builtins, of which 560 were architecture-independent and 5,480 were architecture-specific.

(II) Builtins from the GCC source code To verify that we did not omit any commonly used builtins, we searched the projects for strings starting with `__builtin_`. Since we found that many analyzed projects relied on a small number of GCC’s internal (i.e., undocumented) builtins (see below), we assumed that tool developers would also need to support these builtins. Hence, we added them to our search terms by including all additional `__builtin_` functions that we found in the GCC source code and test suite (6,299 additional builtins). In a number of cases, GCC implemented public builtins using undocumented internal builtins; this was a potential problem in our study, as public and internal builtins would be counted as separate even if they implemented the same semantics. However, since the number of internal builtins actually used was relatively small, we did not attempt to match public builtins with internal ones in our quantitative analysis.

Searching within the Projects. For each analyzed project, we searched all its C files for the names of the 12,339 builtins described by the GCC documentation or used in the GCC source code. Note that we considered only occurrences where the builtin name was not a substring of another identifier. For each builtin that we found, we created a record in our database, thus obtaining 659k builtin entries.

Excluding builtin use records. We used several strategies to eliminate false positives in the builtin use records. While investigating the projects with the highest numbers of unique builtins—mostly operating systems—we found that many of them included parts of the source code of Clang or GCC, even though the projects themselves were not compiler projects. Such projects were missed by our prior filtering. For these projects, we excluded directories whose name started with `gcc`, `clang` or `llvm` (excluding 45% of our records).

We also excluded builtin occurrences that were enclosed in double quotes, as this indicates that they are part of a string literal instead of part of the code (excluding 2% of the records). To exclude builtins in comments, we did not consider builtins found in lines that started with `/*`, `*`, or `//` (which excluded 1% of the records).

Finally, we manually inspected a number of randomly-selected uses for each distinct builtin, which we used to create a list of 1,272 one-line code fragments that indicated false positives (excluding 1% of the records). We consider this filtering step optional, since it did not significantly reduce the number of builtin uses. As part of this process, we detected that builtins not starting with the `__builtin_` prefix (i.e., machine-specific builtins) were likely to cause false positives, which is why we examined such builtin uses in detail. For example, the TI C6X architecture provided builtins like `_abs`, which often occurred in code that did not use builtins. As another example, inline assembly with an instruction mnemonic that corresponded to a builtin name often resulted in false positives. In total, these measures reduced the number of records to 319k (48% of the original number).

3 RESULTS

3.1 RQ1: How frequently are builtins used?

To answer RQ1, we considered both duplicate and unique builtin uses per project. Counting uses—even if they were duplicated within a project—allowed us to measure the overall prevalence of builtin use. Counting project-unique uses better reflected the implementation effort needed to support a project, because duplicates do not increase the implementation effort.

Overall use. In total, 1,842 of the projects (37% of all projects) used a common subset of 3,083 builtins. The frequency of compiler builtins varied strongly, depending on the project, and ranged from one builtin every 7 LOC to one every 1,680,582 LOC. The median frequency of builtins was one every 5,741 LOC (on average one builtin every 20417 LOC). Figure 1 shows boxplots to illustrate the builtin use by the projects, and breaks their use up into architecture-specific and architecture-independent uses, considering both unique and non-unique builtin occurrences within a project.

Non-unique occurrences. The median number of builtin calls in a project that used builtins was 9, the average was 173, indicating that there were outlier projects that used a large number of builtins. In projects that used builtins, architecture-specific builtins were employed in greater numbers (median = 69); in contrast, when architecture-independent builtins were used, their numbers were far lower (median = 7). However, since use of architecture-specific builtins is limited to fewer projects (see Section 3.2), the overall result is dominated by the architecture-independent builtins. We investigated the 15 projects with the highest numbers of builtins and found that audio/video players and codecs lead the ranking (9/15), followed by operating systems (3/15), a game engine, a software library specialized for ARM processors, and a libc implementation.

Unique occurrences. Of the 319k builtin calls, 30k were project-unique; that is, the others were duplicated within a project. The median number of unique builtins used by projects with builtins was low, with a median of 4 and an average of 17. As with non-unique builtins, projects that used architecture-specific builtins had more such builtins (median = 17) than projects that used architecture-independent builtins (median = 3). The projects that used the largest number of unique builtins were, again, in most cases audio/video players and codecs (6/15). However, operating systems (2/15), game engines (2/15), language implementations and compilers (2/15), a messenger, and an image codec also ranked among the top 15.

Reoccurring files. We observed that files with particular names, primarily header files, were more likely to contain calls to builtins. One reason for this was that, consistent with findings by Lopes et al. [28], files were copied from other projects. The majority of these files originated either from the GNU C library `glibc` or from Linux-based operating systems. While they were used primarily in operating system implementations, they were also copied to projects with application code. As another example, the frequently used `sqlite3.c` and `SDL_stdinc.h` files even contained the projects’ names as part of the file name: SQLite is a popular database, and SDL a commonly used media library. In other cases, duplicate file names indicated the use case for the builtin use. For example, builtin-based atomicity support was often implemented in files named `atomic.h`, and math builtins were used in files named `math.h`.

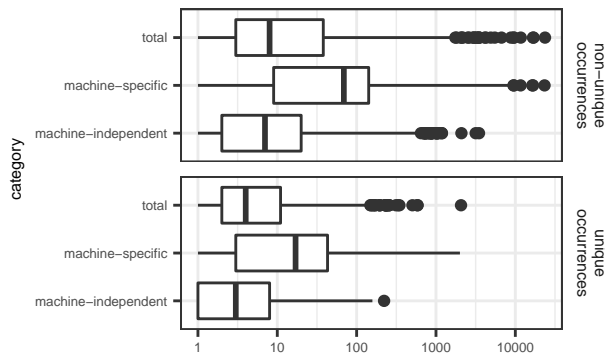


Figure 1: Number of architecture-specific and architecture-independent builtins per project that used builtins of the respective category (logarithmic scale).

Discussion. We found that 12,339 GCC builtins exist that tool developers potentially need to consider, but that only about 3,000 of these are used. Although a builtin is typically found only once every 5,741 lines of C code, 37% of all popular projects rely on compiler builtins, thus strongly incentivizing their implementation in analysis and other tools.

3.2 RQ2: For what purposes are builtins used?

To identify the purpose for which builtins are used, we explored their usage at different levels of granularity: First, we examined in detail the usage of architecture-independent builtins and of architecture-specific builtins, as summarized in Figure 2. Then, we analyzed builtins that remained unused in our corpus.

Architecture-specific and -independent builtins. The GCC documentation categorizes builtins into architecture-specific and architecture-independent ones, which we used as a basis for discussion. While 1,776 projects used at least one architecture-independent builtin, we found architecture-specific builtins in only 422 projects. That architecture-independent builtins are more common across project was unexpected, since we found only 85k architecture-independent builtin uses, but 213k architecture-specific ones. However, as discussed in Section 3.1, a project using architecture-specific builtins is likely to use more such builtins than projects that use architecture-independent builtins.

“Other” builtins. The builtin category “other”, which contained miscellaneous builtins, was the most common category of GCC builtins, even though it comprised only 68 builtins—21 of which were among the 50 most frequently used. Since these builtins were the most common, we further analyzed their use, and classified them into the following subcategories: (I) direct compiler interaction, (II) bit and byte operations, (III) special floating-point values, and (IV) dynamic stack allocation.

(I) **Direct compiler interaction.** These builtins allow direct interaction with the compiler, for example, to improve performance; the most frequently used builtin was `__builtin_expect`, which communicates expected branch probabilities to the compiler, which can exploit this information for optimization. The `__builtin_unreachable` builtin can be used to silence warnings by informing the compiler that code is unreachable, which

is useful when the compiler cannot deduce this. Some of the builtins in this subcategory can also be used for metaprogramming; the `__builtin_constant_p` builtin is resolved at compile time and allows programmers to query whether a pointer is known by the compiler to be constant. As another example, `__builtin_types_compatible_p` queries whether two input types passed to the builtin are the same. Plain C does not offer similar functionality.

(II) **Bit and byte operations.** These builtins process integers at the level of bits and bytes. The second-most frequently used builtin was `__builtin_clz`, which counts the leading zeroes in an unsigned int; its variants for other data types also ranked among the most commonly used builtins overall. Similarly frequent were builtins for computing the position of the least significant one-bit, for counting the number of one-bits in an integer, and for reversing the bytes of an integer. We believe that these builtins were used for convenience and performance optimizations, as the same functionality could be implemented in plain C.

(III) **Special floating-point values.** These builtins generate special values for various floating-point types. For example, the `__builtin_inf` builtin generates a positive infinity double value. As another example, `__builtin_nan` returns a not-a-number value. Recent C standards specify macros and functions for obtaining such values.

(IV) **Dynamic stack allocation.** The `__builtin_alloca` builtin allocates the specified number of bytes of stack memory. Since C99, variable length arrays have offered a similar functionality, as the size of an allocated array can depend on a run-time value.

Synchronization and atomics. After “other”, the next common builtin category was synchronization (“sync”) with 11 of the 50 most common builtins. In this category, the most frequently used builtin was `__sync_synchronize`, which issues a full memory barrier to restrict the order of execution in out-of-order CPUs. Builtins for atomically executing operations were also common (e.g., `__sync_fetch_and_add`). These builtins were designed for the Intel Itanium ABI and were deprecated in favor of the builtins contained in the “atomic” category. The builtins in the “atomic” category additionally allow specifying the memory order of the operation, but were not that frequently used; nevertheless 7 builtins of this category ranked among the 100 most common builtins. Note that C11 introduced synchronization primitives, which are alternatives to these builtins.

Libc functions. GCC provides builtins for many functions of the standard C library—4 such builtins were amongst the 100 most common builtins. An example is `__builtin_memcpy`, which implements the semantics of `memcpy`. The builtin version of the `libc` function is useful when compiling a program assuming a C dialect in which a function is not yet available; for example, when compiling under the C90 standard (`-std=c90`), the newer C99 function `log2` cannot be used; however, the prefixed version `__builtin_log2` can still be used. Furthermore, they enable bare-metal programs, which are compiled *freestanding* and therefore do not have access to `libc` functions, unless they use compiler builtins.

GCC internal functions. Several builtins were used by projects although they were not documented—4 ranked among the top 100 frequently used builtins. These most frequently used builtins, namely `__builtin_va_start`, `__builtin_va_end`,

`__builtin_va_arg`, and `__builtin_va_copy`, were used exclusively to implement the `vararg` macros of the C standard.

Function return address and offsetof. The “introspection” category—with 3 of the top 100 builtins—enables programmers to query (I) the address to which a function returns and (II) the address of the current frame (i.e., the area where local variables are stored). To this end, GCC provides `__builtin_return_address`, `__builtin_frame_address` and other builtins. Another, similar category is “`offsetof`” with a single builtin `__builtin_offsetof`, which was one of the top 100 builtins. It determines the offset of a struct or array member from the start address of the struct or array.

Object size and safe integer arithmetics. The builtin `__builtin_object_size` in the “object-size” category enables programmers to query the size of an object, which is useful when implementing bounds checks. To implement this builtin, GCC relies on static analysis to determine the size of an object where possible. The “overflow” category—of which no builtin ranked among the top 100—provides wraparound semantics for overflow in signed-integer operations (e.g., `__builtin_add_overflow` for addition), which would otherwise induce undefined behavior in C [9].

Usage of architecture-specific builtins. Of the 100 most-frequent builtins, 44 were specific to an architecture. Most frequent were the builtins for the PowerPC family—17 of which were among the top 100 builtins. The most frequent PowerPC builtins were those implementing vector operations such as `vec_perm`, which implements a vector permutation. The second category were ARM C NEON extensions—25 of which were among the top 100 builtins—that also implement vector operations. On x86, which ranked next, the most common builtin was `__builtin_cpu_supports` followed by `__builtin_cpu_init`, which allow programmers to query the availability of CPU features such as SIMD support. In x86-64 inline assembly, the equivalent `cquad` instruction ranked among the most commonly used instructions [53]. Other x86 builtins were quite diverse and less frequent. For brevity, the less frequently used architecture-specific builtin categories are omitted. However, they are included in the full list of commonly used builtins in the online appendix.

Unused builtins. To identify unused builtins, we considered only those described in the GCC documentation (i.e., the public ones). Surprisingly, we found that half of them, namely 3,033 (50%), were not used in our corpus. The distribution differed between architecture-specific and architecture-independent builtins. From the architecture-independent builtins, 379 of 560 were used, which corresponds to 32% unused builtins. We characterize these unused builtins below. From the architecture-specific builtins, only 2,627 of 5,480 builtins were used, which means that more than half of them (52%) were not used in any project; this is why we do not characterize them in detail.

We contacted the GCC developers to report our findings [51]; they responded that builtins could not be removed from the documentation due to vendor guarantees (for architecture-specific builtins) and because they might still be used in closed-source software or by projects not hosted on GitHub. While the possibility cannot be excluded that these builtins are used by some projects (or code yet to be written), the possibility that all of them are used

is rather low. Thus, our study provides a first step towards deprecating unused builtins, and removing those builtins from the public documentation that could be considered internal.

Unused architecture-independent builtins. None of the projects used any of the 11 bounds-checking builtins for controlling the Intel MPX-based pointer-bounds-checker instrumentation, which is based on a hardware extension in Intel processors. One reason for this is that they are used by a pass within GCC and have received only little further attention [54], as Intel MPX-based approaches perform only about as fast as pure software approaches [41]. Four of the object-size-checking builtins were not used, namely a subset of those for printing format strings (e.g., `__builtin___vfprintf_chk`). The builtins of this category were derived from library functions (e.g., `memcpy`), but require an additional size argument (e.g., `__builtin___memcpy_chk`). The intended use of these builtins is to prevent buffer overflow attacks, since object accesses that exceed the size of the object can be prevented. We speculate that these builtins were not frequently used because neither the C language nor builtins provide the functionality to reliably query the size of an object, which would require run-time support [55].

None of the 13 builtins of the Cilk Plus C/C++ language extensions [18], which offer a mechanism for multithreading, were used. In 2017, Cilk Plus was deprecated, and in November 2017 GCC removed its implementation [22]. Of the prefixed libc functions, 37% were unused. Most programs are probably compiled in hosted mode, where compilers can substitute calls to the libc functions with these builtins. Another reason could be that some of them are used only internally. Nevertheless, they were documented in the public API.

Of the unused builtins in the “other” category, the majority were narrowly specialized builtins such as `__builtin_inffn`, which generates an infinity value for the data type `_Floatn`. Further, `__builtin___clear_cache` for flushing the processor’s instruction cache remained unused. The unused `__builtin_call_with_static_chain` enables calls to languages that expect static chain pointers, such as Go.

Discussion. The use cases for builtins were diverse. The use of GCC builtins was dominated by architecture-independent builtins for direct interaction with the compiler, for bit-and-byte operations, atomic operations, and libc equivalents. Depending on the tool, different builtin categories could be supported to different degrees; for example, static analysis tools that do not analyze the semantics of multithreaded atomic operations might eschew implementing those. Architecture-specific builtins were used by fewer projects, but, within these projects, in greater number than architecture-independent builtins. They were used for SIMD instructions, to determine CPU features, and to access platform-specific registers.

3.3 RQ3: How many builtins must be implemented to support most projects?

In order to provide tool developers with a recommended implementation order for builtins, we considered two implementation scenarios. The first scenario considered all builtins as implementation candidates. The second considered only architecture-independent

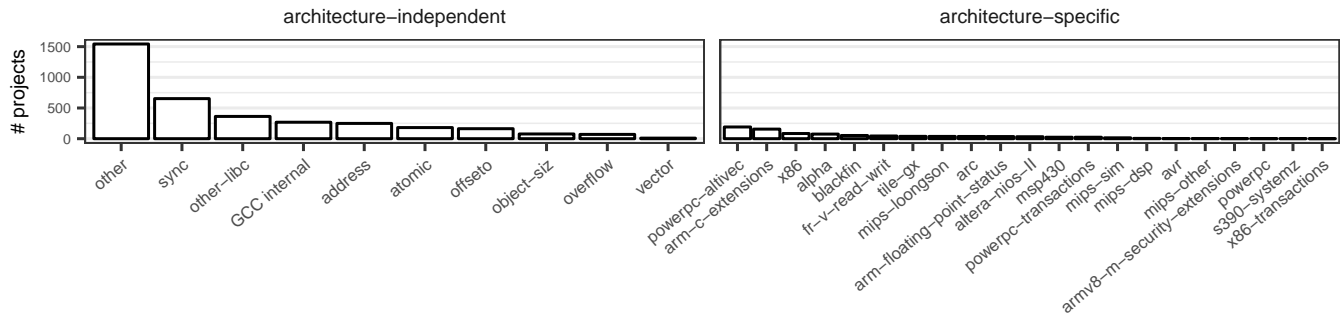


Figure 2: The number of projects that rely on architecture-independent and architecture-specific builtins.

builtins, which can be relevant when only a subset of architectures is to be supported. Additionally, we assumed two pragmatic strategies for the order of implementation: an order based on the frequency of builtins, and one based on a greedy algorithm. Note that this paper assumes equal weights across projects, since weights would have biased the results based on assumptions that might not hold for all tools.

Frequency order. Using this strategy, we assumed that the builtins used by the highest number of projects are to be implemented first. Thus, this strategy follows the order given by Table 2. This order is not generally optimal, because it does not take into account that, in order for a project to be supported, all builtins used must be implemented.

Greedy order. For rapid experimentation, it can be beneficial to quickly support as many projects as possible. To this end, we implemented a greedy order where the next builtin to be implemented is selected such that it enables support of the largest number of additional projects. If no such builtin exists, the next builtin is selected using the frequency order.

Results. Implementing builtins takes an exponential implementation effort in terms of number of builtins that must be implemented to support a specific number of projects (see Figure 3). The greedy order for implementing builtins performs better than the frequency order, a trend that is more clear-cut when considering all builtins rather than just architecture-independent ones. To support half of the projects, in both scenarios and using both strategies, no more than 32 builtins need to be implemented. Note that these builtins are all architecture-independent ones; this is expected, because, as described in Section 3.1, projects rely less frequently on architecture-specific builtins, but if they do, they use a larger number of such unique builtins.

Supporting 90% of the projects requires 106 builtins to be implemented for the greedy approach and 112 builtins for the frequency strategy when considering only architecture-independent builtins. When considering all builtins, more than 850 builtins must be implemented for the frequency strategy, and more than 600 for the greedy strategy. To support 99% of the projects, the greedy algorithm is better: when considering only architecture-independent builtins, around 250 instead of 300 builtins must be implemented, compared to 1,600 instead of 3,000 builtins when considering all builtins. Thus, we suggest that tool developers use a greedy approach when implementing builtins.

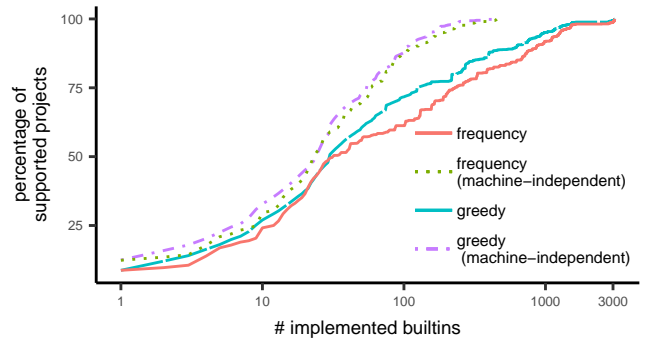


Figure 3: The numbers of builtins needed to support an increasing number of projects increases exponentially; note the exponential x-axis.

Table 2: The 10 most frequent builtins.

builtin	category	projects
__builtin_expect	other (compiler interaction)	890 / 48.3%
__builtin_clz	other (bitwise operation)	536 / 29.1%
__builtin_bswap32	other (bitwise operation)	483 / 26.2%
__builtin_constant_p	other (compiler interaction)	430 / 23.3%
__builtin_alloca	other (stack allocation)	373 / 20.2%
__sync_synchronize	sync	356 / 19.3%
__builtin_bswap64	other (bitwise operation)	347 / 18.8%
__sync_fetch_and_add	sync	332 / 18.0%
__builtin_ctz	other (bitwise operation)	324 / 17.6%
__builtin_bswap16	other (bitwise operation)	304 / 16.5%

3.4 RQ4: How does builtin usage develop over time?

To understand whether builtin usage is an ongoing concern of software projects or just a form of technical debt (introduced temporarily before being removed), we studied the development of builtin usage over time in the projects that used builtins. For this, we analyzed all commits by iterating from the latest commit to the oldest commit—including merge commits (represented by the union of all commits that are merged)—always by following the first parent (i.e., staying on the master branch). We considered only those projects for further inspection that had at least five commits that introduced or removed calls to builtins, since projects with fewer commits made it difficult to judge a project’s development trend. This left us with 677 projects, 37% of the projects for which we processed the builtin history.

Table 3: Builtin trends in projects.

trend	classification	#/% projects	median commits	
Increasing	mostly increasing	250	37%	18
	stable, then increasing	17	3%	14
Stagnant	increasing, then stable	140	21%	12
	spike, then stable	24	4%	8
	mostly stable	6	1%	10
Decreasing		93	14%	16
Inconclusive		147	22%	14

Manual inspection methodology. We manually classified the 677 projects based on their trends of adding and removing builtins. Since manual classification of trends is partly subjective, we performed manual classification based on “negotiated agreement” [4, 32]. Basically, the three authors jointly open-coded the qualitative data sources, arriving at a classification through consensus. Given the lack of pre-existing classifications, such an approach seems justified. In particular, the three authors first independently classified a fixed set of 15% randomly-selected projects with respect to their builtin trends. In 36% of the cases, all three authors agreed on the classification. In 46% of the cases, two authors agreed. In 18% of the cases, all authors disagreed. Subsequently, the three authors discussed diverging classifications and came to a consensus for each of them. As with other studies [10, 27], this initial classification served as a “calibration phase” for a single author to classify the remaining trends.

Classification Results. The final classification consisted of four main categories of trends (see Table 3). Most prevalent was the *Increasing* trend, which we assigned to projects that mostly added builtins (39%). The majority of those showed a clear increasing trend (37%), while few had an initial stable period that was followed by an increasing trend (3%). The second most common trend was the *Stagnant* trend (25%) for those projects that initially had builtin-related commits, but then did not show any or few further changes to the usage of builtins. Most *Stagnant* projects initially added builtins, then became stagnant (21%). Others initially added builtin uses, but then removed all or many of them shortly afterwards—a development to which we refer as a spike—and subsequently showed none or few further changes (4%). A low number of *Stagnant* projects exhibited a mostly stable trend overall (1%). We assigned the *Decreasing* trend to projects that initially had an increasing trend followed by a decreasing trend (i.e., the removal of builtin uses, 14%). Finally, we assigned the *Inconclusive* trend to projects for which we could not clearly assign a trend (e.g., because they exhibited a combination of trends, 22%).

Reasons for builtin additions or removal. We attempted to find reasons for changes in the numbers of builtins, for which we analyzed commit messages and commit changes, then identified common cases.

Builtin additions. The majority of sharp increases in the number of builtins was caused by the inclusion of third-party libraries that call builtins internally, as indicated by commits such as “*update packaged sqlite to 3.8.11.1*” or “*Added latest stb_image*.” In some cases, only single existing header files were included, as indicated

by commit messages such as “*add atomic.h that wraps GCC atomic operations*” or “*Copy over stdatomic.h from FreeBSD*.”

Builtins, both architecture-specific and -independent ones, were often used for performance optimizations. Example architecture-independent optimizations are “*popcount() optimization for speed*” (using `__builtin_popcount`), “*Use __builtin_expect in scanline drawers to help gcc predict branching*”, and “*A prefetch of status->last_alloc_tslot saved 5%*” (using `__builtin_prefetch`). Examples of architecture-specific builtin commits were “*VP9 common for ARMv8 by using NEON intrinsics*” and “*30% encoding speedup: use NEON for QuantizeBlock()*”.

Builtins were also used when they conveniently provided required functionality in commits such as “*bitmap – Add few helpers for [bit] manipulations*”. They were often used for atomics, as in “*GCC 4.1 builtin atomic operations*” and “*Adding atomic bitwise operations api and rwlocks support*”. They enabled metaprogramming techniques, for example, by enabling macros to handle various data types: “*util: Ensure align_power2() works with things other than uint. This uses a [cascading] set of if (__builtin_types_compatible_p()) statements to pick the correct alignment function tailored to a specific type [...]*”.

Finally, builtins were employed to reduce the usage of assembly and inline assembly in commits such as “*avoid inline assembly in favor of gcc builtin functions*” and “*Padlock engine: make it independent of inline assembler*”, or as an alternative to architecture-specific system libraries, such as “*alloca fallback for gcc*”, which added a use of `__builtin_alloca` when the platform did not provide a header file that implements `alloca`.

Builtin removals. Removals of third-party libraries accounted for the most significant number of removals of builtins, as indicated by commits such as “*Remove thirdparties*” or “*Removed outdated headers and libraries*.” Individual files or functions that used builtins were removed as side effects of refactoring or cleanup in commits with messages such as “*General cleanup of the codebase, remove redundant files*.” or “*tools: Remove unused code*.” Auto-generated files were removed, for instance, in the commit “*Removed getdate.c as it is regenerated from getdate.y*”.

A number of removals were related to technical debt [7]. Projects removed builtins for old architectures for which they dropped support, for instance, in “*avr32: Retire AVR32 for good. AVR32 is gone. [...]*” or “*Blackfin: Remove. The architecture is currently unmaintained, remove*”. In other cases, builtins for certain architectures were removed due to their maintenance effort: “*Remove support for altivec using gcc builtins, since these keep changing across gcc versions. [...]*”. Uses of builtins were hidden behind a macro, to concentrate their use to a single location in the source code: “*Convert remaining __builtin_expect to likely/unlikely [...]*” (for `__builtin_expect`) and “*Use the new sol-atomic.h API instead of directly GCC intrinsics*” (for atomic operations).

In other cases, a use of `__builtin_expect` was removed because it did not improve performance: “[...] *It had no reliably measurable performance improv[e]ment, at least on an i7 960 and within a microbenchmark*.”

Case study. Finally, we examined the builtin development in four projects whose trends we considered both representative and insightful for our case study (see Figure 4). First, we selected `libucl`,

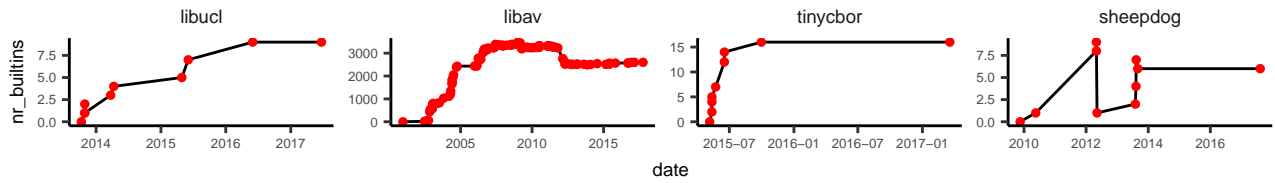


Figure 4: Builtin development in tinybcor, hashcat, libav, and libucl.

a configuration library parser, which is representative of the *Increasing* trend. Like the majority of projects that we examined, it added a small number of builtin calls for various tasks. We selected libav, a collection of cross-platform tools to process multimedia formats and protocols, to represent the *Decreasing* category. As is typical of a media library, it contained a number of builtin-related commits that improved performance by adding calls to architecture-specific builtins, but also systematically removed them to reduce maintenance effort. We selected tinybcor, a library for encoding and decoding the CBOR format, to represent the *Stagnant* trend. Specifically, we classified it as *increasing, then stable*. Finally, we selected sheepdog, a distributed storage system for the QEMU virtual machine, which we classified as *Inconclusive*, due to its “pit shape” in the center of the plot.

libucl (Increasing). The builtin additions in libucl were in most cases related to hashing. The first two builtin-related commits of libucl imported a hash algorithm from third-party libraries that used `__builtin_clz` and `__builtin_swap32` in their hashing computations. Subsequently, a third-party library hashing implementation was replaced with a custom implementation, removing a builtin use. Subsequent commits were also related to finding better hashing algorithms, resulting in additions of calls to byteswap builtins and checks for SIMD support using `__builtin_cpu_init` and `__builtin_cpu_supports`. Additionally, the library added a reference-counting scheme to free memory when an allocation is no longer referenced, whose implementation depended on atomics.

libav (Decreasing). In the first half of libav’s development, its use of builtins mainly increased, mostly due to AltiVec-specific builtins used to optimize computation-intensive operations, but also due to architecture-specific builtins of other architectures such as PowerPC or ARM. In a few cases, calls to architecture-independent builtins were added, for example for atomics. In the second half of the project, refactorings reduced the number of builtin calls. In 2009, calls to 236 AltiVec-specific builtins were removed to reduce technical debt and improve the maintainability of the Snow codec (which was removed in 2012): “Remove AltiVec optimizations for Snow. They are hindering the development of Snow, which is still in flux.” In 2012, calls to 233 builtins were removed as part of a cleanup that dropped an unused function; in the same year, a library was removed that used 469 builtins. In 2013, another smaller, but interesting, commit removed calls to 23 Alpha-specific builtins, as the platform was no longer considered important: “Remove all Alpha architecture optimizations. Alpha has been end-of-lived and no more test machines are available.”

tinybcor (Stagnant) The initial commit introduced macros for performance optimizations that used `__builtin_expect` to communicate branch probabilities to the compiler; one of the macros was used to annotate an error handling case as unlikely. Similarly,

the `__builtin_unreachable` builtin was used to annotate the case that should not happen as undefined, allowing the compiler to generate more efficient code. To support byteswap operations on non-Linux systems, where the `endian.h` header file is typically not present, a use of `__builtin_bswap64` was added. A subsequent commit also introduced byteswap uses for Linux systems, with the commit message stating that it was more efficient and made cross-building the project easier. The `__builtin_add_overflow` was added to implement an addition that does not cause undefined behavior on overflow [9]. Three commits that did not change the number of used builtins adjusted the conditions when builtins were used due to portability reasons. For example, according to the commit messages, `builtin_bswap16` was added with GCC 4.8 and ICC did not support `__builtin_add_overflow`, making it necessary to check for these cases using macros. While the last builtin-related commit was in 2015, the project continued to be active until 2017.

sheepdog (Inconclusive) In sheepdog, the prominent increase before the pit was caused by a commit that replaced mutex locks by equivalent synchronization builtins, as it was stated to make the execution faster. The uses were then replaced with calls to an external library that offered equivalent functionality, resulting in the sharp decrease. Other commits replaced an assembly fragment that obtained the address of the frame pointer with `__builtin_frame_address` for logging. The builtin in turn was replaced by invoking `gdb` to perform this action. The performance of logging was improved by `__builtin_expect`, which was used to annotate code to assume the standard logging level. Besides, bit operations were simplified using `__builtin_clzl` and `__builtin_ffsl`.

Discussion. We analyzed the development history of builtins in projects and found that many projects mostly added calls to builtins. They were added for performance optimizations, atomic implementations, to enable metaprogramming techniques, and others; they were removed, for example, due to their maintenance cost and through refactorings. Overall, it seems that compiler builtins are not a legacy feature from times when compilers applied less sophisticated optimizations; tool developers must expect that contemporary and future code will use them.

The four representative projects gave insights into how projects added and removed builtin uses. Like the majority of projects we examined, libucl, tinybcor, and sheepdog had few commits related to architecture-independent builtins. These builtins were used in various use cases, for instance, to improve the performance of code, to test for CPU features, to implement hash computations, and as a fallback when architecture-specific builtins were missing. Libav was one of the relatively few projects that had a large number of commits related to architecture-specific builtins, and it reduced their number during code refactorings. For sheepdog and libav, builtins were also removed to reduce technical debt; in sheepdog,

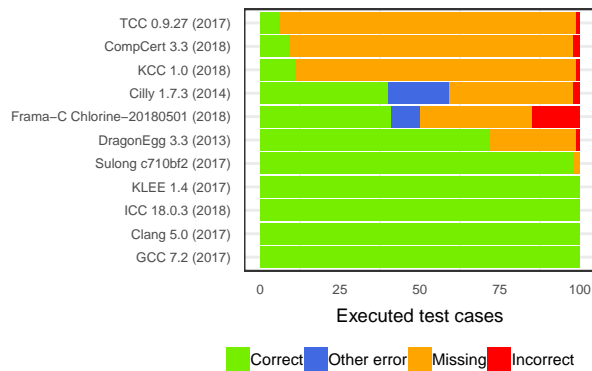


Figure 5: The test-case results before bugs were fixed or builtins implemented. The order of the legend items corresponds to the order of the stacked plots. “Other errors” refers to failed test cases unrelated to builtin support.

builtins were replaced by using an external library instead, and in libav they were removed since an outdated architecture was no longer supported.

3.5 RQ5: How well do tools support builtins?

To determine how well current tools support GCC builtins, we manually implemented a builtin test suite for the 100 most commonly used architecture-independent builtins (cf. RQ2), which would support the architecture-independent portion of almost 90% of the builtin-using projects (see Section 3.3). For each builtin, we used its documentation to determine both typical inputs and corner cases, then wrote test cases for them. As tools to be tested, we selected popular and widely used mature compilers, special-purpose compilers, source-to-source translators, alternative execution environments, and static analysis tools. Figure 5 shows the results.

Mature compilers. We tested the most widely used open-source compilers on Linux, GCC and Clang [24], as well as the commercial ICC. They all executed the test cases successfully.

Special-purpose compilers. We tested the special-purpose compilers CompCert [25, 26] and TCC. CompCert is a compiler used in safety-critical applications and has been formally verified to be correct, which, however, excludes its implementation of builtins. We found that CompCert correctly executed only 9 builtin test cases, supporting 5 out of the 10 most frequently used builtins. Both `__builtin_clz1` and `__builtin_ctz1` computed an incorrect result for large input values [49]. After reporting the bugs detected by our test suite, they were fixed within a day with the note that “we need more testing here”.

The TCC compiler is a small compiler developed to compile code quickly. It successfully ran only six builtin test cases. While most tests failed with a build error, the `__builtin_types_compatible_p` builtin produced an incorrect result when comparing enumerations [50].

C front end. The C Intermediate Language (CIL) [39] is a front end for the C language that facilitates program analysis and transformation. We tested its driver, called `cilly`, which can also be used as a drop-in replacement for GCC. It successfully executed 40 builtin test cases. The `__builtin_bswap16` and

`__builtin_types_compatible_p` builtins produced incorrect results [46]. `Cilly` also failed on 34 atomic test cases, on 15 test cases due to a failure to parse a system library, on 5 test cases due to unrecognized builtins, and on 4 test cases due to warnings for the long double type.

Source-to-source translators. We evaluated DragonEgg, which compiles source languages supported by GCC to LLVM IR. Although it has not been updated for several years, it successfully executed more than two thirds of the test cases. It failed to translate more recent builtins (e.g., from the “atomic” category) that were added to GCC after the last commit in DragonEgg.

Static analysis. We tested Frama-C [8, 21], a static-analysis framework. By default, it assumes code to be portable, and supports compiler extensions only with an option. For 41 test cases, Frama-C’s analysis did not trigger a warning or error [48]. 9 test cases failed because its standard library lacked macros for INFINITY and NAN, which were used in the test cases. 14 test cases for `__sync` builtins were generally supported, but incorrectly implemented for the long type. Furthermore, `__builtin_object_size` referred to an undefined variable in its macro, which resulted in an error.

Alternative execution environments. We tested Sulong [52, 56], an interpreter with dynamic compiler for LLVM-based languages, and KCC [12, 16], a commercial interpreter for C that was automatically derived from a formal semantics for C and detects Undefined Behavior. Sulong successfully executed all but two test cases, namely for `__builtin_fabs1` and `__builtin___clear_cache`, which were not implemented [43]. Note that we found these errors with a preliminary version of the test suite, and consequently contributed implementations for the two missing builtins.

KCC successfully executed test cases for 10 builtins, but, since it is based on CIL, it had the same error in the implementation of the `__builtin_types_compatible_p` builtin. The KCC developers also mentioned that they have “recently been trying to add more supports for gnuc builtins.” [47].

Symbolic execution engine. We tested KLEE [3], a symbolic execution for LLVM-based languages. KLEE executed all test cases successfully when executed with concrete inputs.

Discussion. Our findings indicate that mature compilers support builtins, which is expected, since many projects rely on them. However, many other tools lack builtin implementations or have errors in their implementations. Note that working builtin implementations can typically not be reused by other tools due to their differences in use cases and implementation languages. For example, while GCC translates builtin usages to efficient machine code in its C/C++ source code, Frama-C abstractly reasons about them using OCaml. Tools based on existing mature compiler infrastructure—such as KLEE and Sulong, which are based on LLVM—seem to have a better builtin support, partly because some builtins are handled by the compiler’s front end.

4 THREATS TO VALIDITY

Internal Validity. The main threat to internal validity (i.e., risk of confounding variables) is that we relied on a source-based heuristic approach to determine the usage of GCC builtins, namely by searching for identifiers of known builtins in the source files. We could have mistakenly recorded a builtin use when the builtin was

enclosed in a comment, or when an identifier with the same name as a builtin was used for another purpose. However, as described, we used several mitigation strategies to address such “deceiving” uses. Conversely, we could have missed builtin uses if their names consisted of strings that were concatenated by using preprocessor macros; however, we expect such uses to be uncommon. RQ4 required manual effort for classifying projects based on their use of builtins as they evolved, and selecting representative projects, which is both difficult to reproduce and subjective. To address this, we had a calibration phrase where three authors performed a classification of 15% of the projects.

Construct Validity. The main threat to construct validity is that the implementation order we suggest might not reflect the needs of developers. In all implementation scenarios, we assumed that each project has equal weights. In practice, tool developers might prioritize one project domain over another, and might thus implement builtins in a different order. For example, developers of security analyzers would likely first want to support projects with a high attack surface, while compilers for embedded systems would implement support only for embedded software. We consider equal weights as a neutral and simple metric; if not applicable, developers can use our artifact to determine a better-suited order.

External Validity. Several threats to external validity (i.e., whether our results are generalizable) are related to the scope of our analyses. First, besides C code, C++ code also can access GCC builtins, which we considered beyond our scope, so our results cannot be generalized to C++ projects. We analyzed open-source GitHub projects, hence our findings might not apply to proprietary projects. Furthermore, they do not necessarily apply to projects hosted on sites other than GitHub; this biases our results as, for example, GNU projects other than GCC are often hosted on Savannah and could potentially rely more strongly on GCC builtins. Additionally, our results cannot be generalized to the builtins of compilers other than GCC. Finally, we investigated the usage of builtins at the source level, which might be different from the usage in the compiled binary (e.g., because their usage could be influenced by macro metaprogramming) and the usage during execution of the program.

5 RELATED WORK

Studies of inline assembly and linkers. Besides compiler builtins, C projects also contain other elements not specified by the C standard. Rigger et al. found that around 30% of popular C projects use x86-64 inline assembly [53]. The current paper demonstrates that GCC builtins are used more frequently than inline assembly, which provides even stronger incentives to implement support by C tools. Other studies focused on the role of linkers [20] and the preprocessor [13]. C projects are often built using Makefiles, whose feature usage has also been investigated [29].

Studies of other language features. This paper fits into a recent stream of empirical studies of programming language feature usage, all of which share a methodology of mining software repositories to determine the popularity of features in large sets of open-source projects and/or evaluate the “harmfulness” of features in terms of potential for bugs. Most of this work has focused on general-purpose programming languages, and research has evolved

from more common to lesser known features. For example, for Java, the usage of general language features [11, 44], fields [62], inheritance [61], exception handling [1, 38, 57], lambda features [30] and async constructs on Android [40] have been studied. For C++ projects, the usage of templates [63], generic constructs [60], concurrency constructs [64] and asserts [5] have been studied. The latter also considered C projects, similar to Nagappan et al.’s study [34] of the usage and harmfulness of the goto construct. However, to the best of our knowledge, a study of the usage of compiler builtins has not yet been conducted, and as such fits into the line of research into C programming language features.

6 CONCLUSIONS

We have presented an empirical study of the usage of GCC builtins in a corpus of 4,912 open-source C projects retrieved from GitHub. To the best of our knowledge, this is the first study of compiler builtins despite them having existed in GCC for 30 years. We believe that they warrant investigation, since more than 12,000 builtins exist that tools could support and because even safety-critical tools such as the CompCert compiler have bugs in the implementations of common builtins.

Implications for tool builders. Since 37% of all popular projects relied on compiler builtins, any tool that processes C code needs to deal with them. However, since only about 3,000 of 12,339 builtins were used, it might not be necessary to implement all builtins. In fact, we found that architecture-independent builtins are most commonly used and by implementing only 32 of such core builtins, half of the projects can be supported. Since the majority of projects mostly added builtin usages, tools are likely still expected to support builtins for code yet to be written.

Implications for the GCC developers. We think that our study is also informative for compiler developers, especially those of GCC. This study demonstrated the large scope of GCC builtin usage and might encourage compiler developers to add, maintain, and document builtins in a consistent and structured way. In particular, we think that public and internal builtins should be strictly separated. Since our study highlighted the builtins that remained unused in our data set, such builtins could potentially be considered as deprecation candidates. As part of our future work, we want to engage with the GCC developers to discuss how the identified problems could be addressed.

Implications for application developers. Furthermore, our study informs application developers about downsides of using builtins. Although supported by mature compilers, GCC builtins are a language extension that are not supported by all tools (e.g., CompCert, TCC, KCC, and Frama-C). Thus, a reliance on builtins means that fewer code analysis and other tools can be used on such applications. Furthermore, projects that rely on internal builtins—such as those for variadic arguments handling—add an additional level of technical debt, as they should only be used within GCC and can, in theory, change without notice. Thus, we recommend application developers to use builtins with caution.

Implications for language designers. We believe that our results are also useful to language designers, as they show which functionality plain C lacks, and what potential implications adding compiler builtins has on the projects developed in a given language.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. We thank Ingrid Abfalter for proofreading an early draft of this paper. The authors from Johannes Kepler University Linz are funded in part by a research grant from Oracle.

REFERENCES

- [1] Muhammad Asaduzzaman, Muhammad Ahasanuzzaman, Chanchal K. Roy, and Kevin A. Schneider. 2016. How Developers Use Exception Handling in Java?. In *Proceedings of the 13th International Conference on Mining Software Repositories (MSR '16)*. ACM, New York, NY, USA, 516–519. <https://doi.org/10.1145/2901739.2903500>
- [2] Hudson Borges, André C. Hora, and Marco Tulio Valente. 2016. Understanding the Factors That Impact the Popularity of GitHub Repositories. In *2016 IEEE International Conference on Software Maintenance and Evolution, ICSME 2016, Raleigh, NC, USA, October 2-7, 2016*. 334–344. <https://doi.org/10.1109/ICSME.2016.31>
- [3] Cristian Cadar, Daniel Dunbar, and Dawson Engler. 2008. KLEE: unassisted and automatic generation of high-coverage tests for complex systems programs. In *Proceedings of the 8th USENIX conference on Operating systems design and implementation (OSDI'08)*. USENIX Association, Berkeley, CA, USA, 209–224.
- [4] John L. Campbell, Charles Quincy, Jordan Osserman, and Ove K. Pedersen. 2013. Coding In-depth Semistructured Interviews: Problems of Unitization and Inter-coder Reliability and Agreement. *Sociological Methods & Research* 42, 3 (2013), 294–320.
- [5] Casey Casalnuovo, Prem Devanbu, Abilio Oliveira, Vladimir Filkov, and Baishakhi Ray. 2015. Assert Use in GitHub Projects. In *Proceedings of the 37th International Conference on Software Engineering - Volume 1 (ICSE '15)*. IEEE Press, Piscataway, NJ, USA, 755–766.
- [6] Clang Team. 2018. Clang Language Extensions. Builtin Functions. <https://clang.llvm.org/docs/LanguageExtensions.html>
- [7] Ward Cunningham. 1993. The WyCash portfolio management system. 4 (04 1993), 29–30.
- [8] Pascal Cuoq, Florent Kirchner, Nikolai Kosmatov, Virgile Prevosto, Julien Signoles, and Boris Yakobowski. 2012. Frama-C. In *Software Engineering and Formal Methods, George Eleftherakis, Mike Hinchey, and Mike Holcombe (Eds.)*. Springer Berlin Heidelberg, Berlin, Heidelberg, 233–247.
- [9] Will Dietz, Peng Li, John Regehr, and Vikram Adve. 2012. Understanding Integer Overflow in C/C++. (2012), 760–770.
- [10] Marcia W. DiStaso and Denise Sevick Bortree. 2012. Multi-method analysis of transparency in social media practices: Survey, interviews and content analysis. *Public Relations Review* 38, 3 (2012), 511–514. <https://doi.org/10.1016/j.pubrev.2012.01.003> Public Relations History.
- [11] Robert Dyer, Hridesh Rajan, Hoan Anh Nguyen, and Tien N. Nguyen. 2014. Mining Billions of AST Nodes to Study Actual and Potential Usage of Java Language Features. In *36th International Conference on Software Engineering (ICSE'14)*. 779–790.
- [12] Chucky Ellison and Grigore Rosu. 2012. An Executable Formal Semantics of C with Applications. In *Proceedings of the 39th Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL '12)*. ACM, New York, NY, USA, 533–544. <https://doi.org/10.1145/2103656.2103719>
- [13] Michael D. Ernst, Greg J. Badros, and David Notkin. 2002. An Empirical Analysis of C Preprocessor Use. *IEEE Trans. Softw. Eng.* 28, 12 (Dec. 2002), 1146–1170. <https://doi.org/10.1109/TSE.2002.1158288>
- [14] David Evans, John Guttag, James Horning, and Yang Meng Tan. 1994. LCLint: A Tool for Using Specifications to Check Code. (1994), 87–96. <https://doi.org/10.1145/193173.195297>
- [15] David Evans and David Larochelle. 2002. Improving Security Using Extensible Lightweight Static Analysis. *IEEE Softw.* 19, 1 (Jan. 2002), 42–51. <https://doi.org/10.1109/52.976940>
- [16] Chris Hathhorn, Chucky Ellison, and Grigore Rosu. 2015. Defining the Undefinedness of C. In *Proceedings of the 36th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI '15)*. ACM, New York, NY, USA, 336–345. <https://doi.org/10.1145/2737924.2737979>
- [17] Gerard J Holzmann. 2002. UNO: Static source code checking for user-defined properties. In *Proc. IDPT*, Vol. 2.
- [18] Intel. [n.d.]. Cilk Plus. <https://www.cilkplus.org/>
- [19] Eirini Kalliamvakou, Georgios Gousios, Kelly Blincoe, Leif Singer, Daniel M. German, and Daniela Damian. 2014. The Promises and Perils of Mining GitHub. In *Proceedings of the 11th Working Conference on Mining Software Repositories (MSR 2014)*. ACM, New York, NY, USA, 92–101. <https://doi.org/10.1145/2597073.2597074>
- [20] Stephen Kell, Dominic P. Mulligan, and Peter Sewell. 2016. The Missing Link: Explaining ELF Static Linking, Semantically. In *Proceedings of the 2016 ACM SIGPLAN International Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA 2016)*. ACM, New York, NY, USA, 607–623. <https://doi.org/10.1145/2983990.2983996>
- [21] Florent Kirchner, Nikolai Kosmatov, Virgile Prevosto, Julien Signoles, and Boris Yakobowski. 2015. Frama-C: A software analysis perspective. *Formal Aspects of Computing* 27, 3 (May 2015), 573–609. <https://doi.org/10.1007/s00165-014-0326-7>
- [22] Julia Koval. 2017. remove cilk-plus. <https://gcc.gnu.org/ml/gcc-patches/2017-11/msg01345.html>.
- [23] Robbert Krebbers and Freek Wiedijk. 2015. A Typed C11 Semantics for Interactive Theorem Proving. In *Proceedings of the 2015 Conference on Certified Programs and Proofs (CPP '15)*. ACM, New York, NY, USA, 15–27. <https://doi.org/10.1145/2676724.2693571>
- [24] Chris Lattner and Vikram Adve. 2004. LLVM: A Compilation Framework for Lifelong Program Analysis & Transformation. In *Proceedings of the International Symposium on Code Generation and Optimization: Feedback-directed and Runtime Optimization (CGO '04)*. IEEE Computer Society, Washington, DC, USA, 75–.
- [25] Xavier Leroy. 2009. Formal Verification of a Realistic Compiler. *Commun. ACM* 52, 7 (July 2009), 107–115. <https://doi.org/10.1145/1538788.1538814>
- [26] Xavier Leroy. 2009. A formally verified compiler back-end. *Journal of Automated Reasoning* 43, 4 (2009), 363–446.
- [27] Matthew Lombard, Jennifer Snyder-Duch, and Cheryl Campanella Bracken. [n.d.]. Content Analysis in Mass Communication: Assessment and Reporting of Inter-coder Reliability. *Human Communication Research* 28, 4 ([n.d.]), 587–604. <https://doi.org/10.1111/j.1468-2958.2002.tb00826.x>
- [28] Cristina V. Lopes, Petr Maj, Pedro Martins, Vaibhav Saini, Di Yang, Jakub Zitny, Hitesh Sajani, and Jan Vitek. 2017. DéjàVu: A Map of Code Duplicates on GitHub. *Proc. ACM Program. Lang.* 1, OOPSLA (Oct. 2017), 84:1–84:28. <https://doi.org/10.1145/3133908>
- [29] Douglas H. Martin, James R. Cordy, Bram Adams, and Giulio Antonilo. 2015. Make It Simple: An Empirical Analysis of GNU Make Feature Use in Open Source Projects. In *Proceedings of the 2015 IEEE 23rd International Conference on Program Comprehension (ICPC '15)*. IEEE Press, Piscataway, NJ, USA, 207–217.
- [30] Davood Mazinanian, Ameya Ketkar, Nikolaos Tsantalis, and Danny Dig. 2017. Understanding the Use of Lambda Expressions in Java. *Proc. ACM Program. Lang.* 1, OOPSLA (Oct. 2017), 85:1–85:31. <https://doi.org/10.1145/3133909>
- [31] Kayvan Memarian, Justus Matthiesen, James Lingard, Kyndylan Nienhuis, David Chisnall, Robert N. M. Watson, and Peter Sewell. 2016. Into the Depths of C: Elaborating the De Facto Standards. In *Proceedings of the 37th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI '16)*. ACM, New York, NY, USA, 1–15. <https://doi.org/10.1145/2908080.2908081>
- [32] Samim Mirhosseini and Chris Parnin. 2017. Can Automated Pull Requests Encourage Software Developers to Upgrade Out-of-date Dependencies?. In *Proceedings of the 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE 2017)*. IEEE Press, Piscataway, NJ, USA, 84–94. <http://dl.acm.org/citation.cfm?id=3155562.3155577>
- [33] Nuthan Munaiah, Steven Kroh, Craig Cabrey, and Meiyappan Nagappan. 2017. Curating GitHub for Engineered Software Projects. *Empirical Softw. Engg.* 22, 6 (Dec. 2017), 3219–3253. <https://doi.org/10.1007/s10664-017-9512-6>
- [34] Meiyappan Nagappan, Romain Robbes, Yasutaka Kamei, Éric Tanter, Shane McIntosh, Audris Mockus, and Ahmed E. Hassan. 2015. An Empirical Study of Goto in C Code from GitHub Repositories. In *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering (ESEC/FSE 2015)*. ACM, New York, NY, USA, 404–414. <https://doi.org/10.1145/2786805.2786834>
- [35] Meiyappan Nagappan, Thomas Zimmermann, and Christian Bird. 2013. Diversity in Software Engineering Research. In *Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering (ESEC/FSE 2013)*. ACM, New York, NY, USA, 466–476. <https://doi.org/10.1145/2491411.2491413>
- [36] Santosh Nagarakatte, Jianzhou Zhao, Milo M.K. Martin, and Steve Zdancewic. 2009. SoftBound: Highly Compatible and Complete Spatial Memory Safety for C. In *Proceedings of the 30th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI '09)*. ACM, New York, NY, USA, 245–258. <https://doi.org/10.1145/1542476.1542504>
- [37] Santosh Nagarakatte, Jianzhou Zhao, Milo M.K. Martin, and Steve Zdancewic. 2010. CETS: Compiler Enforced Temporal Safety for C. *Proceedings of the International Symposium on Memory Management* 45, 8 (June 2010), 31–40.
- [38] Suman Nakshatri, Maithri Hegde, and Sahithi Thandra. 2016. Analysis of Exception Handling Patterns in Java Projects: An Empirical Study. In *Proceedings of the 13th International Conference on Mining Software Repositories (MSR '16)*. ACM, New York, NY, USA, 500–503. <https://doi.org/10.1145/2901739.2903499>
- [39] George C. Necula, Scott McPeak, Shree Prakash Rahul, and Westley Weimer. 2002. CIL: Intermediate Language and Tools for Analysis and Transformation of C Programs. In *Proceedings of the 11th International Conference on Compiler Construction (CC '02)*. Springer-Verlag, London, UK, UK, 213–228.
- [40] Semih Okur, Danny Dig, and Yu Lin. 2015. Study and Refactoring of Android Asynchronous Programming. (2015).
- [41] Aleksii Oleksenko, Dmitrii Kuvaitskii, Pramod Bhatotia, Pascal Felber, and Christof Fetzer. 2017. Intel MPX Explained: An Empirical Study of Intel MPX and Software-based Bounds Checking Approaches. *CoRR* abs/1702.00719 (2017).

- [42] John O'Neill. 2006. Intel Compilers for Linux: Compatibility with GNU Compilers.
- [43] Thomas Pointhuber. 2017. Add some testcases for builtins, little bugfixes (Sulong GitHub Pull Request). <https://github.com/graalvm/sulong/pull/807>.
- [44] Dong Qiu, Bixin Li, Earl T. Barr, and Zhendong Su. 2017. Understanding the syntactic rule usage in java. *Journal of Systems and Software* 123 (2017), 160–172. <https://doi.org/10.1016/j.jss.2016.10.017>
- [45] Dan Quinlan. 2000. ROSE: Compiler Support for Object-Oriented Frameworks. In *Proceedings of Conference on Parallel Compilers (CPC2000)*, Aussois, France (Parallel Processing Letters), Vol. 10. Springer Verlag.
- [46] Manuel Rigger. 2018. GCC Builtin Support (CIL Github Bug Report). <https://github.com/cil-project/cil/issues/44>.
- [47] Manuel Rigger. 2018. GCC Builtins (C-Semantics GitHub Bug Report). <https://github.com/kframework/c-semantics/issues/318>.
- [48] Manuel Rigger. 2018. GCC Builtins Support (Frama-C Mailing List). <https://lists.gforge.inria.fr/pipermail/frama-c-discuss/2018-July/005483.html>.
- [49] Manuel Rigger. 2018. Support of GCC Builtins (CompCert GitHub Bug Report). <https://github.com/AbsInt/CompCert/issues/243>.
- [50] Manuel Rigger. 2018. Support of GCC Builtins (Tiny C Compiler Mailing List). <http://lists.nongnu.org/archive/html/tinycc-devel/2018-07/msg00007.html>.
- [51] Manuel Rigger. 2018. Unused GCC builtins (GCC Mailing List). <https://gcc.gnu.org/ml/gcc/2018-01/msg00166.html>.
- [52] Manuel Rigger, Matthias Grimmer, Christian Wimmer, Thomas Würthinger, and Hanspeter Mössenböck. 2016. Bringing Low-level Languages to the JVM: Efficient Execution of LLVM IR on Truffle. In *Proceedings of the 8th International Workshop on Virtual Machines and Intermediate Languages (VMIL 2016)*. ACM, New York, NY, USA, 6–15. <https://doi.org/10.1145/2998415.2998416>
- [53] Manuel Rigger, Stefan Marr, Stephen Kell, David Leopoldseder, and Hanspeter Mössenböck. 2018. An Analysis of x86-64 Inline Assembly in C Programs. In *Proceedings of the 14th ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments (VEE '18)*. ACM, New York, NY, USA, 84–99. <https://doi.org/10.1145/3186411.3186418>
- [54] Manuel Rigger, Daniel Pekarek, and Hanspeter Mössenböck. 2018. Context-Aware Failure-Oblivious Computing as a Means of Preventing Buffer Overflows. (2018), 376–390. https://doi.org/10.1007/978-3-030-02744-5_28
- [55] Manuel Rigger, Roland Schatz, Rene Mayrhofer, Matthias Grimmer, and Hanspeter Mössenböck. 2018. Introspection for C and its Applications to Library Robustness. *The Art, Science, and Engineering of Programming* 2 (2018). <https://doi.org/10.22152/programming-journal.org/2018/2/4>
- [56] Manuel Rigger, Roland Schatz, René Mayrhofer, Matthias Grimmer, and Hanspeter Mössenböck. 2018. Sulong, and Thanks for All the Bugs: Finding Errors in C Programs by Abstracting from the Native Execution Model. In *Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '18)*. ACM, New York, NY, USA, 377–391. <https://doi.org/10.1145/3173162.3173174>
- [57] Demóstenes Sena, Roberta Coelho, Uirá Kulesza, and Rodrigo Bonifácio. 2016. Understanding the Exception Handling Strategies of Java Libraries: An Empirical Study. In *Proceedings of the 13th International Conference on Mining Software Repositories (MSR '16)*. ACM, New York, NY, USA, 212–222. <https://doi.org/10.1145/2901739.2901757>
- [58] Konstantin Serebryany, Derek Bruening, Alexander Potapenko, and Dmitriy Vyukov. 2012. AddressSanitizer: A Fast Address Sanity Checker. In *USENIX Annual Technical Conference*. 309–318.
- [59] Evgeniy Stepanov and Konstantin Serebryany. 2015. MemorySanitizer: fast detector of uninitialized memory use in C++. In *Code Generation and Optimization (CGO), 2015 IEEE/ACM International Symposium on*. IEEE, 46–55.
- [60] Andrew M. Sutton, Ryan Holeman, and Jonathan I. Maletic. 2010. Identification of Idiom Usage in C++ Generic Libraries. In *The 18th IEEE International Conference on Program Comprehension, ICPC 2010, Braga, Minho, Portugal, June 30-July 2, 2010*. 160–169. <https://doi.org/10.1109/ICPC.2010.37>
- [61] Ewan Tempero, James Noble, and Hayden Melton. 2008. How Do Java Programs Use Inheritance? An Empirical Study of Inheritance in Java Software. In *Proceedings of the 22Nd European Conference on Object-Oriented Programming (ECOOP '08)*. Springer-Verlag, Berlin, Heidelberg, 667–691. https://doi.org/10.1007/978-3-540-70592-5_28
- [62] Ewan D. Tempero. 2009. How Fields are Used in Java: An Empirical Study. In *20th Australian Software Engineering Conference (ASWEC 2009), 14-17 April 2009, Gold Coast, Australia*. 91–100. <https://doi.org/10.1109/ASWEC.2009.19>
- [63] Di Wu, Lin Chen, Yuming Zhou, and Baowen Xu. 2014. An empirical study on the adoption of C++ templates: Library templates versus user defined templates. In *The 26th International Conference on Software Engineering and Knowledge Engineering, Hyatt Regency, Vancouver, BC, Canada, July 1-3, 2013*. 144–149.
- [64] Di Wu, Lin Chen, Yuming Zhou, and Baowen Xu. 2016. An extensive empirical study on C++ concurrency constructs. *Information & Software Technology* 76 (2016), 1–18. <https://doi.org/10.1016/j.infsof.2016.04.004>
- [65] Deng Xu. 2011. [Frama-c-discuss] inline assembly code. <https://lists.gforge.inria.fr/pipermail/frama-c-discuss/2011-March/002589.html> (Accessed October 2017).
- [66] Zhongxing Xu, Ted Kremenek, and Jian Zhang. 2010. A Memory Model for Static Analysis of C Programs. In *Proceedings of the 4th International Conference on Leveraging Applications of Formal Methods, Verification, and Validation - Volume Part I (ISoLA'10)*. Springer-Verlag, Berlin, Heidelberg, 535–548.