Checked C: Making C Safe by Extension

Archibald Samuel Elliott, University of Washington
Andrew Ruef and Michael Hicks, University of Maryland
David Tarditi, Microsoft Research

Abstract—This paper presents Checked C, an extension to C designed to support spatial safety, implemented in Clang and LLVM. Checked C’s design is distinguished by its focus on backward-compatibility, incremental conversion, developer control, and enabling highly performant code. Like past approaches to a safer C, Checked C employs a form of checked pointer whose accesses can be statically or dynamically verified. Performance evaluation on a set of standard benchmark programs shows overheads to be relatively low. More interestingly, Checked C introduces the notions of a checked region and bounds-safe interfaces.

1 INTRODUCTION

Vulnerabilities that compromise memory safety are at the heart of many attacks. Memory safety has two aspects. Temporal safety is ensured when memory is never used after it is freed. Spatial safety is ensured when any pointer dereference is always within the memory allocated to that pointer. Buffer overruns—a spatial safety violation—still constitute a frequent and pernicious source of vulnerability, despite their long history. During the period 2012–2018, buffer overruns were the source of 9.7% to 18.4% of CVEs reported in the NIST vulnerability database [1], with the highest numbers occurring in 2017. During that time, buffer overruns were the leading single cause of CVEs.

As discussed in depth in Section 2, several efforts have attempted to make C programs safe. Static analysis tools [2], [3], [4] aim to find vulnerabilities pre-deployment, but may miss bugs, have trouble scaling, or emit too many alarms. Security mitigations, such as CFI [5] and DEP [6], can mute the impact of vulnerabilities by making them harder to exploit, but provide no guarantee; e.g., data leaks and mimicry attacks may still be possible. Several efforts have aimed to provide spatial safety by adding run-time checks; these include CCured [7], Softbound [8], and ASAN [9]. The added checks can add substantial overhead and can complicate interoperability with legacy code when pointer representations are changed. Lower overhead can be achieved by reducing safety, e.g., by checking only writes, or ignoring overruns within a memory region (e.g., from one stack variable to another, or one struct field to another). In the end, no existing approach is completely satisfying.

This paper presents a new effort towards achieving a spatially-safe C that we call Checked C. Checked C borrows many ideas from prior safe-C efforts but ultimately differs in that its design focuses on allowing incremental conversion while balancing control, interoperability, and high performance. Technically speaking, Checked C’s design has three key features. First, all pointers in Checked C are represented as in normal C—no changes to pointer layout are imposed. This eases interoperability. Second, the legal boundaries of pointed-to memory are specified explicitly; the goal here is to enhance human readability and maintainability while supporting efficient compilation and running times. Checked C supports pointers to single objects, arrays, and NUL-terminated arrays. The types and bounds are used by the compiler to either prove that an access is safe, or else to insert a run-time bounds check when such a proof is too difficult. Programmers can use dynamic checks and annotations to convince the compiler to eschew unnecessary checks in performance-critical code.

Finally, Checked C is designed to support incremental porting from legacy C. Programs may consist of a mix of checked and legacy code, and fully ported code can be annotated as within a checked region, which can be held blameless for any spatial safety violation. This guarantee is made possible by restricting any use of unchecked pointers and casts within the region. To allow existing unchecked code to be accessed by checked regions and with checked pointers, Checked C allows unchecked code to be annotated with bounds-safe interfaces. These describe the expected behavior and requirements of the code and can be added to parameters and return values of function declarations/definitions, function and record types, and global variables. Such interfaces support modular porting and use of legacy libraries. In the Checked C universe, programmers can add safety with each use of a checked pointer, and then extend the safety guarantee by expanding the scope of checked regions. At every step, they enjoy a working software artifact. Ultimately, a fully-ported program is assuredly safe, and in the meantime scrutiny can be focused on any unchecked regions.

We are implementing Checked C as an extension to the LLVM compiler infrastructure. Preliminary results are promising. On a standard benchmark suite, Checked C adds an average run-time overhead of 8.6%, with about 90% of the code in checked regions. Code changes were modest: 17.5% of benchmark lines of code were changed, and most (> 80%, on average) were trivial to perform.

Checked C is under active and ongoing development. We are porting more programs to evaluate its utility. We are developing a tool to rewrite legacy programs semi-automatically. We are also improving Checked C’s analysis, and formalizing a proof of its safety guarantee. Checked C’s code and in-depth specification are available at https://github.com/Microsoft/checkedc.

2 PRIOR WORK

There has been extensive research addressing out-of-bounds memory accesses in C [10]. The research falls into four categories:
C(-like) dialects, compiler implementations, static analyses, and security mitigations.

**Safe C Dialects.** Cyclone [11] and Deputy [12, 13] are type-safe dialects of C. Cyclone’s key novelty is its support for GC-free temporal safety [14, 15]. Checked C differs from Cyclone by being backward compatible (Cyclone disallowed many legacy idioms) and avoiding pointer format changes (e.g., Cyclone used “fat” pointers to support arithmetic). Deputy keeps pointer layout unchanged by allowing a programmer to describe the bounds using other program expressions. Deputy incorporates the bounds information into the types of pointers by using dependent types. Deputy requires that values of all pointers stay in bounds so that they match their types. To enforce this invariant (and make type checking decidable), it inserts runtime checks before pointer arithmetic, not at memory accesses. Accessed C uses separate annotations that describe bounds invariants instead of incorporating bounds into pointer types and inserts runtime checks at memory accesses.

Non C-based programming languages like D [16] and Rust [17] also aim to support safe, low-level systems-oriented programming. Legacy programs would need to be ported wholesale to take advantage of these languages, which could be a costly affair (but could be partially automated [18]).

**Safe C implementations.** Rather than use a new language, several projects have looked at new ways to implement legacy C programs so as to make them spatially safe. The bcc source-to-source translator [19] and the rtcc compiler [20] changed the representations of pointers to include bounds. The rtcc-generated code was 3 times larger and about 10 times slower. Fail-Safe C [21] changed the representation of pointers and integers to be pairs. Benchmarks were 2 to 4 times slower. CCured [2] employed a whole-program analysis for transforming programs to be safe. Its transformation involved changes to data layout (e.g., fat and “wild” pointers), which could cause interoperation headaches. Compilation was all-or-nothing: unhandled code idioms in one compilation unit could inhibit compilation of the entire program.

Safety can also be offered by the loader and run-time system. “Red zones”, used by Purify [22, 23] are inserted before and after dynamically-allocated objects and between statically-allocated objects, where bytes in the red zone are marked as inaccessible (at a cost of 2 bits per protected byte). Red-zone approaches cannot detect out-of-bounds accesses that occur entirely within valid memory for other objects or intra-object buffer overruns (a write to an array in a struct that overwrites another member of the struct). Checked C detects accesses to unrelated objects and intra-object overruns.

Checking that accesses are to the proper objects can be done using richer side data structures that track object bounds and by checking that pointer arithmetic stays in bounds [24-29]. Baggy Bounds Checking [27] provides a fast implementation of object bounds by reserving 1/n of the virtual address space for a table, where n is the smallest allowed object size and requiring object sizes be powers of 2. It increases SPECINT 2000 execution time by 60% and memory usage by 20%. SoftBound [8] tracks bounds information by using a hash table or a shadow copy of memory. It increases execution time for a set of benchmarks by 67% and average memory footprint by 64%. SoftBound can check only writes, in which case execution time increases by 22%.

There is also work on adding temporal safety with different memory allocation implementations, e.g., via conservative garbage collection (GC) [30] or regions [14, 15]. Checked C focuses on spatial safety both due to its importance at stopping code injection style attacks as well as information disclosure attacks. Temporal safety can be added by linking Checked C programs with a conservative GC; overcoming the cost of doing so (e.g., as done in Cyclone or Rust) is something we plan to investigate in the future.

**Static analysis.** Static analysis tools take source or binary code and attempt to find possible bugs, such as out-of-bounds array accesses, by analyzing the code. Commercial tools include CodeSonar, Coverity Static Analysis, HP Fortify, IBM Security AppScan, Klocwork, Microsoft Visual Studio Code Analysis for C/C++, and Polyspace Static Analysis [2, 31, 32]. Static analysis tools have difficulty balancing precision and performance. To be precise, they may not scale to large programs. While imprecision can aid scalability, it can result in false positives, i.e., error reports that do not correspond to real bugs. False positives are a significant problem [4]. As a result, tools may make unsound assumptions (e.g., inspecting only a limited number of paths through function [31]) but the result is they may also miss genuine bugs (false negatives). Alternatively, they may focus on supporting coding styles that avoid problematic code constructs, e.g., pointer arithmetic and dynamic memory allocation [4, 33, 34, 35]. Or, they may require sophisticated side conditions on specifications, i.e., as pre- and post-conditions at function boundaries, so that the analysis can be modular, and thus more scalable [35].

Checked C occupies a different design point than static analysis tools. It avoids problems with false positives by deferring bounds checks to runtime—in essence, it trades run-time overhead for soundness and coding flexibility.

**Security mitigations.** Security mitigations employ runtime-only mechanisms that detect whether memory has been corrupted or prevent an attacker from taking control of a system after such corruption. They include data execution prevention (DEP) [6], software fault isolation (SFI) [36], address-space layout randomization (ASLR) [37, 38], stack canaries [39], shadow stacks [40, 41], and control-flow integrity (CFI) [5]. DEP, ASLR, and CFI focus on preventing execution of arbitrary code and control-flow modification. Stack protection mechanisms focus on protecting data or return addresses on the stack.

Checked C provides protection against data modification and data disclosure attacks, which the other approaches do not. For example, ASLR does not protect against data modification or data disclosure attacks. Data may be located on the stack adjacent to a variable that is subject to a buffer overrun; the buffer overrun can be used reliably to overwrite or read the data. Shadow stacks do not protect stack-allocated buffers or arrays, heap data, and statically-allocated data. Chen et al. [42] show that data modification attacks that do not alter control-flow (Heartbleed is an example) pose a serious long-term threat.

### 3 Checked C

This section presents an overview of Checked C’s main features.

#### 3.1 Basics

The Checked C extension extends the C language with two additional *checked pointer types* `Ptr<T>` and `Array_ptr<
void read_next (int *b, int idx, _Ptr<int>out) {
    int tmp = *(b+idx);
    *out = tmp;
}

Figure 1. Example use of _Ptr<T>

_T< T> The _Ptr<T> type indicates a pointer that is used for dereference only and has no arithmetic performed on it, while _Array_ptr<T> supports arithmetic with bounds declarations provided in the type. The compiler dynamically confirms that checked pointers are valid when they are dereferenced. In blocks or functions specifically designated as checked regions, it imposes stronger restrictions on uses of unchecked pointers that could corrupt checked pointers, e.g., via aliases. We also extend C with checked array types (including NUL-terminated ones), by prefixing array dimensions with the _Checked keyword, as in int buf_Checked[10]. We expect a Checked C program to involve a mix of both checked and unchecked regions, and a mix of checked and unchecked pointer and array types.

3.2 Simple pointers
Using _Ptr<T> is straightforward: any pointer to an object that is only referenced indirectly, without any arithmetic or array subscript operations, can be replaced with a _Ptr<T>. For example, one frequent idiom in C programs is an out parameter, used to indicate an object found or initialized during parsing. Figure 1 shows using a _Ptr<int> for the out parameter. When this function is called, the compiler will confirm that it is a given valid pointer, or null. Within the function, the compiler will insert a null check before writing to out. Such null checks are elided when the compiler can prove they are unnecessary.

3.3 Arrays
The _Array_ptr<T> type identifies a pointer to an array of values. Prior safe-C efforts sometimes involve the use of fat pointers, which consist both of the actual pointer and information about the bounds of pointed-to memory. Rather than changing the run-time representation of a pointer in order to support bounds checking, in Checked C the programmer associates a bounds expression with each _Array_ptr<T>-typed variable and member to indicate where the bounds are stored. The compiler inserts a run-time check that ensures that deferencing an _Array_ptr<T> is safe (the compiler may optimize away the runtime check if it can prove it always passes). Bounds expressions consist of non-modifying C expressions and can involve variables, parameters, and struct field members. For bounds on members, the bounds can refer only to other members declared in the same structure. Bounds declarations on members are type-level program invariants that can be suspended temporarily when updating a specific struct object.

Figure 2 shows using _Array_ptr<T> with declared bounds as parameters to a function. In particular, the types of the dst and src arrays have bound expressions that refer to the function’s other two respective parameters. In the body of the function, both src and dst are accessed as expected. The

1. We use the C++ style syntax for programmer familiarity, and precede the names with an underscore to avoid parsing conflicts in legacy libraries.

void append(
    _Array_ptr<char> dst : count(dst_count),
    _Array_ptr<char> src : count(src_count),
    size_t dst_count, size_t src_count)
{
    _Dynamic_check(src_count <= dst_count);
    for (size_t i = 0; i < src_count; i++) {
        if (src[i] == '\0') {
            break;
        }
        dst[i] = src[i];
        ++i;
    }
    dst[i] = '\0';
    return i;
}

Figure 2. Example use of _Array_ptr<T>

3.4 NUL-terminated Arrays
The _Nt_array_ptr<T> type identifies a pointer to an array of values (often chars) that ends with a NUL (\0). The bounds
expression identifies the known-to-be-valid range of the pointer. This range can be expanded by reading the character just past the bounds to see if it is NULL\textsuperscript{2}. If not, then the bounds can be expanded by one. Otherwise, the current bounds cannot be expanded, and only a ‘\texttt{\textbackslash \textbackslash 0}’ may be written to this location.\_\texttt{Nt\_array\_ptr}\texttt{<}\texttt{T}> types without explicit bounds default to bounds of \texttt{count(0)}, meaning that index 0 can be read safely. A \_\texttt{Nt\_array\_ptr}\texttt{<}\texttt{T}> can be cast to a \texttt{\_Array\_ptr}\texttt{<}\texttt{T}> safely—for an \texttt{\_Array\_ptr}\texttt{<}\texttt{T}>, the character just past the bounds can not be read or written, which ensures the zero-termination invariant is maintained for any aliases.

An example use of \texttt{\_Nt\_array\_ptr}\texttt{<}\texttt{T}> is given in Figure\texttt{5}. It implements the \texttt{strncpy} libC routine, which copies at most \texttt{dst\_sz} characters from \texttt{src} to \texttt{dst}. We must alias \texttt{src} into the local variable \texttt{s} so that its count, \texttt{i}, can grow dynamically as the loop executes.

\texttt{Nt\_terminated\_arrays} can be declared by using the keyword \texttt{\_Nt\_checked} instead of \texttt{\_Checked}. An implicit conversion of an \texttt{\_Nt\_checked} array to a pointer produces an \texttt{\_Nt\_array\_ptr}\texttt{<}\texttt{T}>.

### 3.5 Checked and unchecked regions

The safety provided by checked pointers can be thwarted by unsafe operations, such as writes to traditional pointers. For example, consider this variation of the code in Figure\texttt{1}.

```c
void more(int *b, int idx, \_Ptr<int >*out) {
    int oldidx = idx, c;
    do {
        c = readvalue();
        b[idx++] = c;
    } while (c != 0);
    *out = b+idx-oldidx;
}
```

This function repeatedly reads an input value into \texttt{b} until a 0 is read, at which point it returns an updated \texttt{b} pointer via the checked \texttt{out} parameter. While we might expect that writing to \texttt{out} is safe, since it is a checked pointer, it will not be safe if the loop overflows \texttt{b} and in the process modifies \texttt{out} to point to invalid memory.

In a program with a mix of checked and unchecked pointers we cannot and should not expect complete safety. However, we would like to provide some assurance about which code is possibly dangerous, i.e., whether it could be the source of a safety violation. Code review and other efforts can then focus on that code. For this purpose Checked C provides \texttt{checked regions} of code. Such code is designated specifically at the level of a file (using a pragma), a function (by annotating its prototype), or a single block (by labeling that block, similar to an \texttt{asm} block).

An example checked block is shown in Figure\texttt{4}. Outside of the \texttt{\_Checked}-annotated region, an unchecked pointer is cast to a checked one. This cast is a potential source of problems (if \texttt{out} is bogus) and so would not be permitted in checked regions. Within the checked block, checked pointers declared inside and outside the block can be freely manipulated and the compiler performs the expected checks. The compiler also treats uses of the address-of operator \texttt{s} in a checked block as producing a checked pointer, not an unchecked one. When doing this to a \texttt{struct} field, the bounds are defined as the extent of that field.

\texttt{Figure 4. Example _\Checked\_ block, and _\Checked\_ array}

```c
size_t fwrite(void *p : byte_count(s*n),
                size_t s,
                size_t n,
                FILE *st : itype(_Ptr<FILE>));
```

### 3.6 Bounds safe interfaces

Checked C permits ascribing checked types and bounds expressions to unchecked functions, members, and global variables via \texttt{bounds safe interfaces}. This allows checked regions to use legacy libraries and for the interactions to be checked. As an example, the type we give to the \texttt{fwrite} standard library function is shown in Figure\texttt{5}. The first argument to the function is the target buffer whose size (in bytes) is given by the second and third arguments. The presence of a bounds expression indicates that an argument with type \texttt{T*} should have a checked type \texttt{\_Array\_ptr}\texttt{<}\texttt{T}> ascribed to it. For \texttt{p} this is \texttt{\_Array\_ptr<}

---

\textsuperscript{2}This means that bounds of \texttt{count(n)} requires allocating \texttt{n+1} bytes.
We plan to support flow-sensitive declarations of bounds, so that
will insert any needed checks to ensure the access is legal.
The problem is that the bounds declared for
variables can have different bounds at different program points.
variable is declared. It is possible, however, that a variable may
acquire different bounds at different points in the program. For
example, we might like to replace the loop in Figure 2 with:

```
for ( ; i < src_count; i++) {
  if (*src == '0') break;
  *dst = *src;
  src++; dst++;
}
```

The problem is that the bounds declared for `src` are tantamount
to the range `(src, src+src_count)`. This means that updating
`src` to `src+1` would invalidate them, as the upper bound
would be off by one. We would like to declare `src` to have
new bounds before entering the loop, such as

```
(src - i, src + src_count - i)
```

We plan to support flow-sensitive declarations of bounds, so that
variables can have different bounds at different program points.

Finally, the design for checking statically that declared
bounds are valid is incomplete. We could fall back on a combina-
tion of static and dynamic checking if necessary. We describe
the state of our implementation in the next section.

### 3.7 Restrictions and limitations

Checked C is in progress and currently has several
limitations. First, to ensure that checked pointers are valid by
construction, we require that checked pointer variables and
arrays containing checked pointers are initialized when
they are declared. In addition, heap-allocated memory that con-
tains checked pointers (like a struct or array of checked pointers)
or is pointed to by a `_Ptr<T>` must use `calloc`
to ensure safe initialization. We plan to employ something akin
to Java’s `definite initialization` analysis to relax this requirement.

Second, we disallow taking the address of variables/`struct`
members with bounds, variables used in bounds expressions,
and members used in member bounds expressions. Such point-
ers could be used to subvert the validity of bounds checks.

Third, `_Array_ptr<T>` values can be dereferenced following
essentially arbitrary arithmetic, e.g., if `x` is an `_Array_ptr<int>` we could dereference it via `*(x+y-n+1)` and the compiler
will insert any needed checks to ensure the access is legal. However, updates to `_Array_ptr<T>` variables are currently
more limited. The bounds for a variable are declared when the
variable is declared. It is possible, however, that a variable may
need different bounds at different points in the program. For
example, we might like to replace the loop in Figure 2 with:

```
size_t i = 0;
for ( ; i < src_count; i++) {
  if (*src == '0') break;
  *dst = *src;
  src++; dst++;
}
```

The type used during type checking depends on whether
the function is called from a checked or unchecked region.
For a checked region, it is the ascribed the checked type. For
an unchecked region, an argument can have the unchecked or
unchecked type. The type checker tries both types when
checking the call. If the checked type is used, the argument must
meet the requirements of a parameter’s bounds expression also.

Bounds-safe interfaces may also be used to give checked
types to as-yet unconverted code. As a project is converted, some
files can be left alone, but have their externally visible signatures
converted so that invoking their functions or using the types
in checked blocks is allowed. For example, callers in a checked
region could use the interface in Figure 3 to provide checked
pointer types as parameters, while callers in an unchecked region
can provide parameters with either unchecked or checked types.

### 4 IMPLEMENTATION

This section briefly describes our current implementation.

**Compiler Implementation.** We have been implementing
Checked C as an extension to the Clang/LLVM compiler. The
extension is enabled as a flag passed to Clang. We extended
the C grammar [43] to support bounds declarations, the new
`_Ptr<T>,` `_Array_ptr<T>,` and `T _Checked[N]` types, and
adding `_Checked` or `_Unchecked` annotations to blocks and
functions. We chose reserved identifiers so that they will not
conflict with identifiers in existing code.

We include a set of checked headers which ascribe bounds-safe
interfaces to C standard library functions. These are used in lieu
of the standard headers when the Checked C flag is present.

In order to support Checked C’s new types, we extended
Clang’s type representation and type checker. We added a
pointer kind discriminator to Clang’s compile-time pointer type
representation, and a “checked” flag to Clang’s compile-time
array type representation.

Checked C’s bounds expressions provide a static
description of the bounds on a pointer. We check statically that the sub-
expressions of a bounds expression are non-modifying expressions:
they do not contain any assignment, increment or decrement op-
erators, or function calls. This ensures that using the expressions
at bounds checks does not cause unexpected side-effects.

Checked C performs inference to compute a bounds ex-
pression that conservatively describes the bounds on a pointer-
typed expression. Inference uses bounds expressions normalized into
`bounds(l, u)` form. The inferred bounds are used to check
memory accesses using the value of the pointer expression.

For pointer variables, the inferred bounds are the declared
bounds. For pointer arithmetic expressions, the inferred bounds
are those of the pointer-typed subexpression. When taking the
address of a `struct`’s member (`&p->f`), the bounds are those of
the particular field. On the other hand, the address of an array
element retains the bounds of the whole array. For example, the bounds of `int x[5]` are

```
bounds(x, x+5*sizeof(int))
```

as are the bounds of `&x[3]`, rather than (say)

```
bounds(x+3*sizeof(int), x+4*sizeof(int))
```

The compiler must statically ensure that bounds declarations
are valid after assignments and initialization. This requires two
steps. First, a subsumption check confirms that assigning to a
variable (an `lvalue`, more generally) meets the bounds required
of pointers stored in the variable (`lvalue`). The required pointer
bounds must be within (subsumed) by the inferred bounds
of the right-hand expression. (Subsumption also applies to
initialization and function parameter passing.) This check allows
assignment to narrow, but not to widen, the bounds of the right-
hand side value. Determining the required bounds is generally
straightforward. In the simplest case, the bounds for pointers
stored in a variable (`lvalue`) are directly declared, e.g., for a
local variable or function parameter. For assignments to struct
members, uses of struct members within the bounds expression
for the member are replaced with an appropriate struct access,
For example, given

```
struct S {
  int len;
  _Array_ptr<int> buf : count(len);
};
```
the required bounds for an assignment to `a.buf` are `a.len`.

Second, the compiler must ensure bounds expressions are still valid after a statement modifies a variable used in a bounds expression. For example, in Figure 3 the bounds of `s` is `count(i)`, but `i` is modified in a loop that iterates over `s` looking for a NUL terminator. For `_Array_ptr<T>` types, the modification is justified by subsumption: The updated bounds can be narrowed but not widened. For `_Ptr<_Array_ptr<T>` types, we can widen the bounds by 1 byte if we know that the rightmost byte is ‘\0’, e.g., due to a prior check, as is the case in Figure 5.

At the moment subsumption checking is rather primitive. Some subsumption checks for bounds declarations that could be statically proven are not. Currently, the static analysis can only reason about bounds expressions that are syntactically equivalent (modulo constant-folding and ignoring non-value changing operations) and bounds expressions that are constant-sized ranges (syntactically equivalent base expressions +/- constant offsets). The main issue is the need to perform a more sophisticated dataflow analysis (at the Clang AST level) to gather and consider relevant facts about relationships between variables (such as equalities and inequalities).

The compiler complains when it cannot (dis)prove a subsumption check in checked code. In our experimental evaluation, we manually review the warnings. We insert the code in an `_Unchecked` block (for checks that are trivially obvious) or perform a dynamic subsumption check with `_DynamicBoundsCast` (which eliminates the error).

We have designed but not yet implemented the analysis that checks assignments to variables used in bounds declarations. For our experimental evaluation, we verified by hand that such assignments do not happen.

The Checked C compiler inserts run-time checks into the evaluation of `lvalue` expressions whose results are derived from checked pointers and that will be used to access memory. The code for these checks is handed to LLVM, which we allow to remove checks if it can prove they will always pass. In general, such checks are the only source of Checked C run-time overhead (aside from programmer use of `_DynamicCheck`).

Before any `_Ptr<T>` accesses the compiler inserts a run-time check that the pointer is non-null. Before any `_Array_ptr<T>` accesses the compiler inserts a non-null check followed by the required bounds check computed from the inferred bounds. The compiler does not perform any range checks during pointer arithmetic (unless the arithmetic accesses memory).

5 Preliminary Experimental Evaluation

We (mostly) converted two existing C benchmarks as an initial evaluation of the consequences of porting code to Checked C. We quantify both the changes required for the code to become checked, and the overhead imposed on compilation, running time, and executable size.

We chose the Olden [44] and Ptdist [45] benchmark suites, described in Table 1, because they are specifically designed to test pointer-intensive applications, and they are the same benchmarks used to evaluate both Deputy [12] and Ccured [7]. We did not convert bc from the Ptdist suite and voronoi from the Olden suite for lack of time.

The evaluation results are presented in Table 2. These were produced using a 12-Core Intel Xeon X5650 2.66GHz, with 24GB of RAM, running Red Hat Enterprise Linux 6. All compilation and benchmarking was done without parallelism. We ran each benchmark 21 times with and without the Checked C changes using the test sizes from the LLVM versions of these benchmarks. We report the median; we observed little variance.

Code Changes. On average, we modified around 17.5% of benchmark lines of code. Most of these changes were in declarations, initializers, and type definitions rather than in the program logic. In the evaluation of Deputy [13], the reported figure of lines changed ranges between 0.5% and 11% for the same benchmarks, showing they have a lower annotation burden than Checked C.

We modified the benchmarks to use checked blocks and the top-level checked pragma. We placed code that could not be checked because it used unchecked pointers in unchecked blocks. On average, about 9.3% of the code remained unchecked after conversion, with a minimum and maximum of 3.9% and 20.4%. The cause was almost entirely variable-argument `printf` functions.

We manually inspected changes and divided them into easy changes and hard changes. Easy changes include: replacing included headers with their checked versions; converting a `T` to a `_Ptr<T>`; adding the `_Checked` keyword to an array declaration; introducing a `_Checked` or `_Unchecked` region; adding an initializer; and replacing a call to `malloc` with a call to `calloc`. The remaining “hard” changes include changing a `T` to a `_Array_ptr<T>` and adding a bounds declaration, adding structs, struct members, and local variables to represent run-time bounds information, and code modernization.

In all of our benchmarks, we found the majority—more than 80% on average—of changes were easy. In six of the benchmarks, the only “hard” changes were adding bounds annotations relating to the parameters of `main`.

In three benchmarks—em3d, mst, and yacr2—we had to add intermediate structs so that we could represent the bounds on `_Array_ptr<T>`s nested inside arrays. In mst we also had to add a member to a struct to represent the bounds on an `_Array_ptr<T>`. In the first case, this is because we cannot represent the bounds on nested `_Array_ptr<T>`s, in the second case this is because we only allow bounds on members to reference other members in the same struct. In em3d and anagram we also added local temporary variables to represent bounds information. In yacr2 there are a lot of bounds declarations that are all exactly the same where global variables are passed as arguments, inflating the number of “hard” changes.

Running time overhead. The average run-time overhead introduced by added dynamic checks was 8.6%. In more than half of the benchmarks the overhead was less than 1%. We believe this to be an acceptably low overhead that better static analysis may reduce even further.

In all but two benchmarks—tread2 and ft—the added overhead matches (is within 2%) or betters that of Deputy. For yacr2 and em3d, Checked C does substantially better than Deputy, whose overheads are 98% and 56%, respectively. Checked C’s overhead betters or matches that reported by Ccured in every case but ft.

Compile-time overhead. On average, the overhead added to compilation time by using Checked C is 24.3%. The maximum overhead is 83.1%, and the minimum is 4.9% faster than compiling with C. Across the benchmarks, there is an average 7.4% code size overhead from the introduction of dynamic checks. Ten of the programs have a code size increase of less than 10%.
We have presented working to expand this formalism to include features such with type
Ptrdist suite. Descriptions are from [44], [45]. We did not convert
Such pointers are binary-compatible with legacy, unchecked
a program may contain, and benefit from, checked pointers. Such pointers are binary-compatible with legacy, unchecked
points but have explicitly annotated and enforced bounds. Code units annotated as checked regions provide guaranteed
safety: The code within may not use unchecked pointers or unsafe casts that could result in spatial safety violations. Checked
C’s bounds-safe interfaces provide checked types to unchecked
code, which is useful for retrofitting third party and standard
libraries. Together, these features permit incrementally adding
safety to a legacy program, rather than making it an all-or-
nothing proposition. Our implementation of Checked C as an
LLVM extension enjoys good performance, with relatively low
run-time and compilation overheads. It is freely available at
https://github.com/Microsoft/checke
ded and continues to be actively developed.

6 Future Work
Developing a practical, industrial-strength language is a substan-
tial effort. As such, we have several threads of ongoing work.

Automatically rewrite C programs to Checked C: We are in
the process of porting more and larger programs to Checked C.
To help, we have been developing a tool that uses a simple
analysis (inspired by that of CCured [7]) to convert pointer
types to use _Ptr<T>. We are working on extending this tool to
support both the placement of casts and bounds safe interfaces,
as well as inferring bounds for array accesses that will allow
conversion to _Array_ptr<T> types as well.

Improve the checking of bounds declarations: A key de-
gign goal of Checked C is to keep static checking fast and pre-
dictable. To nevertheless support sophisticated static reasoning,
we are working on an analysis that infers relational invariants
and adds them as source-level assertions. These assertions can
be easily proved by the static checker and then used to justify
eliminating costly dynamic checks.

Supporting polymorphism with safe casts: Polymorphic
data structures such as lists and trees often declare their contents
with type void*, casting to/from that type when inserting/remov-
data. Casts are also used to implement “poor man’s subtyping” for struct types. We are extending Checked C to
support these and other forms of safe cast, for greater flexibility.

Formalization: We have mechanized a formalization of a
core subset of Checked C and used it to prove that any violation
of spatial safety is due, directly or indirectly, to code occurring
in an unchecked region; checked code cannot be blamed. We are
working to expand this formalism to include features such
dynamically sized arrays and flow-sensitive updates to array
lengths due to pointer assignments or NUL checks.

7 Conclusion
We have presented Checked C, an extension to C to help ensure
spatial safety. Checked C’s design is focused on interoperability
with legacy C, usability, and high performance. Any part of
a program may contain, and benefit from, checked pointers.

<table>
<thead>
<tr>
<th>Name</th>
<th>LoC</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>bh</td>
<td>1,162</td>
<td>Barnes &amp; Hut N-body force computation</td>
</tr>
<tr>
<td>bisort</td>
<td>262</td>
<td>Sorts using two disjoint bitonic sequences</td>
</tr>
<tr>
<td>em3d</td>
<td>476</td>
<td>Simulates electromagnetic waves in 3D</td>
</tr>
<tr>
<td>health</td>
<td>338</td>
<td>Simulates Columbian health-care system</td>
</tr>
<tr>
<td>mst</td>
<td>325</td>
<td>Minimum spanning tree using linked lists</td>
</tr>
<tr>
<td>perimeter</td>
<td>399</td>
<td>Perimeter of quad-tree encoded images</td>
</tr>
<tr>
<td>power</td>
<td>452</td>
<td>The Power System Optimization problem</td>
</tr>
<tr>
<td>treadd</td>
<td>180</td>
<td>Sums values in a tree</td>
</tr>
<tr>
<td>tsp</td>
<td>415</td>
<td>Estimates Traveling-salesman problem</td>
</tr>
<tr>
<td>voronoi</td>
<td>814</td>
<td>Voronoi diagram of a set of points</td>
</tr>
<tr>
<td>anagram</td>
<td>346</td>
<td>Generates anagrams from a list of words</td>
</tr>
<tr>
<td>bc</td>
<td>5,194</td>
<td>An arbitrary precision calculator</td>
</tr>
<tr>
<td>ft</td>
<td>893</td>
<td>Fibonacci heap Minimum spanning tree</td>
</tr>
<tr>
<td>ks</td>
<td>549</td>
<td>Schweikert-Kernighan partitioning</td>
</tr>
<tr>
<td>yacr2</td>
<td>2,529</td>
<td>VLSI channel router</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>LM %</th>
<th>EM %</th>
<th>LU %</th>
<th>RT ±</th>
<th>CT ±</th>
<th>ES ±</th>
</tr>
</thead>
<tbody>
<tr>
<td>bh</td>
<td>10.0</td>
<td>76.7</td>
<td>5.2</td>
<td>+0.2</td>
<td>+23.8</td>
<td>+6.2</td>
</tr>
<tr>
<td>bisort</td>
<td>21.8</td>
<td>84.3</td>
<td>7.0</td>
<td>0.0</td>
<td>+7.3</td>
<td>+3.8</td>
</tr>
<tr>
<td>em3d</td>
<td>35.3</td>
<td>66.4</td>
<td>16.9</td>
<td>+0.8</td>
<td>+18.0</td>
<td>-0.4</td>
</tr>
<tr>
<td>health</td>
<td>24.0</td>
<td>97.8</td>
<td>9.3</td>
<td>+2.1</td>
<td>+18.5</td>
<td>+6.7</td>
</tr>
<tr>
<td>mst</td>
<td>30.1</td>
<td>75.0</td>
<td>19.3</td>
<td>0.0</td>
<td>+6.3</td>
<td>-5.0</td>
</tr>
<tr>
<td>power</td>
<td>9.8</td>
<td>92.3</td>
<td>5.2</td>
<td>0.0</td>
<td>+4.9</td>
<td>+0.8</td>
</tr>
<tr>
<td>treadd</td>
<td>15.0</td>
<td>69.2</td>
<td>3.9</td>
<td>0.0</td>
<td>+21.6</td>
<td>+8.5</td>
</tr>
<tr>
<td>tsp</td>
<td>17.2</td>
<td>92.3</td>
<td>20.4</td>
<td>+8.3</td>
<td>+83.1</td>
<td>+7.0</td>
</tr>
<tr>
<td>anagram</td>
<td>9.9</td>
<td>94.5</td>
<td>10.3</td>
<td>0.0</td>
<td>+47.6</td>
<td>+4.6</td>
</tr>
</tbody>
</table>

Table 1
Compiler Benchmarks. Top group is the Olden suite, bottom group is the Ptridge suite. Descriptions are from [44], [45]. We did not convert voronoi from the Olden suite and bc from the Ptridge suite.

References
Gros, A. Kamsky, S. McPeak, and D. Engler, “A few billion lines of code
later: Using static analysis to find bugs in the real world,” Commun.
[3] Mathworks, “Polyspace code prover: prove the absence of run-
time errors in software,” http://www.mathworks.com/products/
integrity,” in ACM Conference on Computer and Communications Security,
2005.
functionality in microsoft windows xp service pack 2, part 3: Memory
“Ccured: type-safe retrofitting of legacy software,” ACM Transactions
on Programming Languages and Systems (TOPLAS), vol. 27, no. 3, pp.
Highly compatible and complete spatial memory safety for C,” in
Proceedings of the 30th ACM SIGPLAN Conference on Programming
Sanitizer: A fast address sanitizer checker,” in Proceedings of the 2012


