C provenance semantics: detailed semantics (for PNVI-plain, PNVI address-exposed, PNVI address-exposed user-disambiguation, and PVI models)

PETER SEWELL, University of Cambridge
KAYVAN MEMARIAN, University of Cambridge
VICTOR B. F. GOMES, University of Cambridge

1 INTRODUCTION

This note gives detailed mathematical semantics for four variants of C provenance semantics:

- **PNVI-plain**: a semantics that does not track provenance via integers, but instead, at integer-to-pointer cast points, checks whether the given address points within a live object and, if so, recreates the corresponding provenance.

- **PNVI-ae (PNVI exposed-address)**: a variant of PNVI that allows integer-to-pointer casts to recreate provenance only for storage instances that have previously been exposed. A storage instance is deemed exposed by a cast of a pointer to it to an integer type, by a read (at non-pointer type) of the representation of the pointer, or by an output of the pointer using `%p`.

- **PNVI-ae-udi (PNVI exposed-address user-disambiguation)**: a further refinement of PNVI-ae that supports roundtrip casts, from pointer to integer and back, of pointers that are one-past a storage instance. This is the currently preferred option in the C memory object model study group.

- **PVI**: a semantics that tracks provenance via integer computation, associating a provenance with all integer values (not just pointer values), preserving provenance through integer/pointer casts, and making some particular choices for the provenance results of integer and pointer +/- integer operations; or

We write PNVI-* for PNVI-plain, PNVI-ae, and PNVI-ae-udi. The PNVI-plain and PVI semantics were described in the POPL 2019/N2311 paper [Memarian et al. 2019]. PNVI-ae and PNVI-ae-udi have emerged from discussions in the C memory object model study group.

Changes for PNVI-ae from PNVI-plain are highlighted. Additional changes for PNVI-ae-udi are highlighted. This should be read together with the two companion notes, one giving a series of examples (N2363), and another giving detailed diffs to the C standard text (N2362).

The PNVI-ae and PNVI-ae-udi variants of PNVI permit bytewise copy of a pointer to an initially unexposed object, but leaves it marked as exposed. Additional machinery may well be desirable for PNVI-ae and PNVI-ae-udi to give programmers more control of the provenance of the results of byte manipulations, and of what is left marked as exposed. The design of that machinery should ideally be based on the treatment of representation-byte-accessed pointer values by existing compiler alias analyses and optimisations.

2 THE PNVI-AE-UDI, PNVI-AE, PNVI-PLAIN, AND PVI SEMANTICS

These semantic definitions are manually typeset mathematics simplified from the executable-as-test-oracle Cerberus source (expressed in the pure-functional Lem [Mulligan et al. 2014] definition language). We have removed most subobject details, function pointers, and some options. Neither the typeset models or the Lem source consider linking, or pointers constructed via I/O (e.g. via `%p` or representation-byte I/O).

The memory object semantics can be combined with a semantics for the thread-local semantics of the rest of C (expressed in Cerberus as a translation from C source to the Core intermediate language, together with an operational semantics for Core) to give a complete semantics for a large fragment of sequential C.

For simplicity, we assume that pointer representations are the two’s complement representation of their addresses (and identical to the two’s complement representations of their conversions to sufficiently wide integer types), assume NULL pointers have address (and representation) 0, and allow NULL pointers to be constructed from any empty-provenance integer zero, not just integer constant expressions.

At present the model does not include the ISO semantics that makes all pointers to an object or region indeterminate at the end of its lifetime, and it permits equality comparison between pointers irrespective of whether the objects of their provenances are live, but it does permit pointer subtraction, relational comparison, array offset, member offset, and casts to integer only for pointers to live objects for which the address is within or one past the object footprint. These are all debatable choices. One could instead check only that the addresses are within or one past the original object footprint (and not check the object is live), or go further towards a concrete-address view of pointer values and not check that either. Sketching out some of the options:
The interface involves types `pointer_value(p)`, `integer_value(x)`, `floating_value(x)`, and `mem_value(u)`, which are abstract as far as Core is concerned. Distinguishing pointer and integer values gives more precise internal types.

In PNVI-ae, PNVI, and PVI, a provenance \( \pi \) is either \( @i \) where \( i \) is a storage-instance ID, or the empty provenance `@empty`. In PNVI-ae-udi a provenance can also be a symbolic storage instance ID \( i \) (iota), initially associated to two storage instance IDs and later resolved to one or the other.

A pointer value can either be `null` or a pair \( (\pi, a) \) of a provenance \( \pi \) and address \( a \). In PNVI*, an integer value is simply a mathematical integer (within the appropriate range for the relevant C type), while in PVI, an integer value is a pair \( (\pi, n) \) of a provenance \( \pi \) and a mathematical integer \( n \).

Memory values are the storable entities, either a pointer, integer, floating-point, array, struct, or union value, or `unspec` for unspecified values, each together with their C type.

### 3 THE MEMORY OBJECT MODEL STATE

In both PVI and PNVI*, a memory state is a pair \( (A, M) \). The \( A \) is a partial map from storage-instance IDs to either `killed` or storage-instance metadata \( (n, r_{opt}, a, f, k, t) \):

- size \( n \)
- optional C type \( t \) (or `none` for allocated regions)
- base address \( a \)
- permission flag \( f \in \{ \text{readWrite}, \text{readOnly} \} \)
- kind \( k \in \{ \text{object, region} \} \)
- for PNVI-ae and PNVI-ae-udi, a taint flag \( f \in \{ \text{unexposed, exposed} \} \)

In PNVI-ae-udi, A also maps all symbolic storage instance IDs \( i \), to sets of either one or two (non-symbolic) storage instance IDs. One might also need to record a partial equivalence relation over symbolic storage instance IDs, to cope with the pointer subtraction and relational comparison cases where one learns that two provenances are equal but both remain ambiguous, but that is debatable and not spelt out in this document.

The \( M \) is a partial map from addresses to abstract bytes, which are triples of a provenance \( \pi \), either a byte \( b \) or `unspec`, and an optional integer pointer-byte index \( j \) (or `none`). The last is used in PNVI* to distinguish between loads of pointer values that were written as whole pointer writes vs those that were written byte-wise or in some other way.

### 3.1 Mappings between abstract values and representation abstract-byte sequences

The \( M \) models the memory state in terms of low-level abstract bytes, but `store` and `load` take and return the higher-level memory values. We relate the two with functions `repr(v)`, mapping a memory value to a list of abstract bytes, and `abst(r, bs)`, mapping a list of abstract bytes `bs` to its interpretation as a memory value with C type \( r \).
3.2 Memory operations

The successful semantics of memory operations is expressed as a transition relation between memory states, with transitions labelled by the operation (including its arguments) and return value:

\[(A, M) \xrightarrow{\text{LABEL}} (A', M')\]

For example, the transitions

\[(A, M) \xrightarrow{\text{load}(\tau, p) = v} (A', M')\]

describe the semantics of a load(\(\tau, p\)) in memory state \((A, M)\), returning value \(v\) and with resulting memory state \((A', M')\). The semantics also defines when each operation flags an out-of-memory (OOM) or undefined behaviour (UB) in a memory state \((A, M)\).

**Storage instance creation** When a new storage instance is created, either with allocate_region (for the results of malloc, calloc, and realloc, i.e. heap-allocated regions), or with allocate_object (for objects with automatic or static storage duration, i.e. global and local variables), in non-const and const variants: a fresh storage-instance ID \(i\) is chosen; an address \(a\) is chosen from newAlloc(\(A, al, n\)), defined to be the set of addresses of blocks of size \(n\) aligned by \(al\) that do not overlap with 0 or any other allocation in \(A\); and the pointer value \(p = (@i, a)\) is returned. In all three cases the storage-instance metadata \(A\) is updated with a new record for \(i\), and this is initially marked as unexposed. In the allocate_object case the size \(n\) of the allocation is the representation size of the C type \(\tau\). In the allocate_region(\(al, \tau, readOnly(v)\)) case, the last of the three rules, the memory \(M\) is updated to contain the
representation of \( v \) at the addresses \( a..a + \text{sizeof}(\tau) - 1 \).

\[
\text{[LABEL: allocate_region(al, n) = p]} \\
i \notin \text{dom}(A) \quad a \in \text{newAlloc}(A, al, n) \\
p = (@i, a) \\
A, M \rightarrow A[i \mapsto (n, none, a, \text{readWrite}, \text{region, unexposed})], M
\]

\[
\text{[LABEL: allocate_object(al, } \tau, \text{readWrite) = p]} \\
i \notin \text{dom}(A) \quad a \in \text{newAlloc}(A, al, n) \\
n = \text{sizeof}(\tau) \quad p = (@i, a) \\
A, M \rightarrow A(i \mapsto (n, \tau, a, \text{readWrite, object, unexposed})), M
\]

\[
\text{[LABEL: allocate_object(al, } \tau, \text{readOnly}(v) = p]} \\
i \notin \text{dom}(A) \quad a \in \text{newAlloc}(A, al, n) \\
n = \text{sizeof}(\tau) \quad p = (@i, a) \\
A, M \rightarrow A(i \mapsto (n, \tau, a, \text{readOnly, object, unexposed})), M([a..a + n - 1] \mapsto \text{repr}(v))
\]

**Storage instance lifetime end** When the storage instance of a pointer value \((@i, a)\) is killed, either by a `free()` for a heap-allocated region or at the end of lifetime of an object with automatic storage duration, the storage-instance metadata \(A\) of storage instance \(i\) is updated to record that \(i\) has been killed.

\[
\text{[LABEL: kill(p, k)]} \\
p = (@i, a) \quad k = k' \\
A(i) = (n,., a, f, k',..) \\
A, M \rightarrow A(i \mapsto \text{killed}), M
\]

**Load** To load a value \( v \) of type \( \tau \) from a pointer value \( p = (@i, a) \), there must be a live storage instance for \( i \) in \( A \), the footprint of \( \tau \) at \( a \) must be within the footprint of that allocation, and the value \( v \) must be the abstract value obtained from the appropriate memory bytes from \( M \).

\[
\text{[LABEL: load(r, } p, \text{ = v)]} \\
p = (@i, a) \quad A(i) = (n,., a', f, k,_.) \\
[a..a + \text{sizeof}(\tau) - 1] \subseteq [a'.a' + n - 1] \\
v = \text{abst}(r, M[a..a + \text{sizeof}(\tau) - 1]) \\
A, M \rightarrow A, M
\]

For PNVI-ae and PNVI-ae-udi, if the recursive-on-\( \tau \) computation of \( \text{abst}(r, M[a..a + \text{sizeof}(\tau) - 1]) \) involves a call of abst at any non-pointer scalar type for a region of \( M \) including an abstract byte with non-empty provenance, and the corresponding storage instance is live, it is marked as exposed. This applies e.g. for reads of pointer values via `char` pointers, and for union type punning reads at `uintptr_t` of pointer values.

**Store** To store a value \( v \) of type \( \tau \) to a pointer value \( p = (@i, a) \), there must be a live storage instance for \( i \) in \( A \), which must be writable, and the footprint of \( \tau \) at \( a \) must be within the footprint of that allocation. The memory \( M \) is updated with the representation bytes of the value \( v \).

\[
\text{[LABEL: store(r, } \tau, p, v)\text{]} \\
p = (@i, a) \quad A(i) = (n,., a', \text{readOnly}, k,_.) \\
[a..a + \text{sizeof}(\tau) - 1] \subseteq [a'.a' + n - 1] \\
A, M \rightarrow A, M([a..a + \text{sizeof}(\tau) - 1] \mapsto \text{repr}(v))
\]

For PNVI-ae-udi, the kill, load, and store rules above must be adapted. If \( p = (i, a) \) and \( A(i) = \{i\} \), the other premises and conclusion of the appropriate above rule apply. If \( A(i) = \{i_1, i_2\} \) and the premises are satisfied for one of the two, say \( i_j \), the rest of the rule applies except that in the final state \( A \) is additionally updated to map \( t \) to \( \{i_j\} \).

The memory operations flag out-of-memory (OOM) and undefined behaviour (UB) as follows:
allocate_region(al, n) / allocate_object(al, r, r, readwrite) / allocate_object(al, r, r, readonlyonly(v));

OOM out of memory  if newAlloc(A, al, n) = {} or newAlloc(A, al, sizeof(r)) = {}

load(r, p) / store(r, p, v) / kill(p):

UB null pointer  if p = null

UB empty provenance  if p = (@empty, a)

UB killed provenance  if p = (@i, a) and A(i) is killed

load(r, p) / store(r, p, v):

UB out of bounds  if p = (@i, a) and A(i) = (n, . , a’f,k,_) and [a..a’ + sizeof(r) - 1] ⊊ [a’..a’ + n - 1]

store(r, p, v):

UB read-only  if p = (@i, a) and A(i) = (n, . , a’, readonlyonly, k, _)

kill(p):

UB non-alloc-address  if p = (@i, a) and A(i) = (n, . , a’, f, k, _) and a ≠ a’

For PNVI-ae-udi, the rules above must be adapted. In the case where p = (i, a) and A(i) is {i}, the semantics is exactly as for p = (i, a), while if A(i) = {i1, i2}, one has UB only if the conditions above apply to both i1 and i2.

3.3 Pointer / Integer operations

Pointer subtraction  Points p = (@i, a) and p’ = (@i’, a’) can be subtracted if they have the same provenance (i = i’), there is a live storage instance for i in A, and both a and a’ are within or one-past the footprint of that allocation (in ISO C the last will always hold, otherwise UB would have been flagged in earlier pointer arithmetic).

Otherwise UB. The result is the numerical difference a - a’ divided by sizeof(dearray(r)), where dearray(r) returns r if it is not an array type, and otherwise returns its element type. Note that this disallows subtraction for which one or both arguments are null pointers, which is the ISO semantics but may be a debatable choice.

This rule is stated for PNVI and PNVI-ae, returning pure integer. For PVI, diff_ptrval constructs the same integer but with @empty provenance. For PNVI-ae, this is because subtraction of pointers with different provenance should be UB:

- if both the two pointers have either a provenance @i (resp. @i’) or a symbolic storage instance ID i (resp. i’) mapped by A to a singleton {i} (resp. {i’}), then i = i’, otherwise UB.
- if one of the two pointers has a symbolic storage instance ID i, mapped by A to {i1, i2}, while the other either has a provenance @i’ or an i’ mapped to a singleton {i’}, then i’ must be either i1 or i2, and i is resolved to that in the A of the final state. Otherwise UB.
- If both pointers are ambiguous, say mapped to {i1, i2} and {i1, i’2}, then if those two sets share exactly one element which satisfies the other rule preconditions, both symbolic storage instance IDs are resolved to that. Otherwise UB.
- If both pointers are ambiguous and those sets share two elements that satisfy the other conditions (which we believe can only happen if the addresses are equal), then subtraction is permitted but the symbolic storage instance IDs are left unresolved. Otherwise UB.

For example, suppose p and q have been produced by separate casts from an integer which is ambiguously one-past one allocation and at the start of another. Then after p - q or p<q we know they must have been the same provenance, but we still don’t know which. (Alternatively, we could change the semantics to record an identity relation over symbolic storage instance IDs, and additional modifications to the rules below beyond what is in this draft, but that seems to be unwarranted complexity).

\[
\text{LABEL: diff_ptrval(r, p, p') = x} \\
p = (@i, a) \quad p' = (@i', a') \quad i = i' \quad A(i) = (n, . , \hat{a}, f, k, _) \\
x = (a - a')' / \text{sizeof(dearray(r))} \quad a \in [\hat{a}, \hat{a} + n] \quad a' \in [\hat{a}, \hat{a} + n] \\
A, M \rightarrow A, M
\]

Pointer relational comparison  Points p = (@i, a) and p’ = (@i’, a’) can be compared with a relational operator (≤, <, etc.) if they have the same provenance (i = i’). The result is the boolean result of the mathematical comparison of a and a’. To make this analogous to pointer subtraction, we also require (though this is debatable) that there is a live storage instance for i in A, and both a and a’ are within or one-past the footprint of that allocation. Otherwise UB. Note that this disallows relational comparison against null pointers; a debatable choice.

For PNVI-ae-udi, this has to be adapted in much the same way as the pointer subtraction rule above.

\[
\text{LABEL: rel_op_ptrval(p, p', op) = b} \\
p = (@i, a) \quad p' = (@i', a') \quad i = i' \quad A(i) = (n, . , \hat{a}, f, k, _) \\
b = op(a, a') \quad a \in [\hat{a}, \hat{a} + n] \quad a' \in [\hat{a}, \hat{a} + n] \quad op \in \{\leq, <, >\} \\
A, M \rightarrow A, M
\]
Relational comparison is used in practice between pointers to different objects. A variant which would allow that, which we call `ALLOW-INTER-OBJECT-RELATIONAL-OPERATORS TRUE`, removes the `i = i'` test above and (in the `ZOMBIE-POUNTERS-BECOME-INDETERMINATE` and `ZOMBIE-POUNTERS-ALLOW-EQUALITY-ONLY` variants) additionally checks that `i'` maps to a live object with in-range address.

**Pointer equality comparison**  Pointers `p` and `p'` can always be compared with an equality operator (`==`, `!=`). The result is true if they are either both null or both non-null and have the same provenance and address; nondeterministically either `a = a'` or `false` if they are both non-null and have different provenances; and false otherwise. For PNVI-ae-udi, because equality comparison is permitted (without UB) irrespective of the provenances of the pointers, if the two pointers both have determined single provenances after looking up any symbolic IDs in `A`, this should give `true`, otherwise the middle (nondeterministic) clause should apply. The final `A` should not resolve any symbolic IDs.

\[
\begin{align*}
\text{[LABEL: eq_op_ptrval(p, p') = b]} \\
\{ b = \text{true} \quad& \text{if } p = p' \\
\{ b \in ((a = a'), \text{false}) \quad& \text{if } p = (\pi, a), p' = (\pi', a'), \text{and } \pi \neq \pi' \\
\{ b = \text{false} \quad& \text{otherwise} \\
\} A, M \rightarrow A, M
\end{align*}
\]

Note that the above nondeterminism appears to be necessary to admit the observable behaviour of current compilers, but a simpler provenance-oblivious semantics is arguably desirable:

\[
\begin{align*}
\text{[LABEL: eq_op_ptrval(p, p') = b]} \\
\{ b = \text{true} \quad& \text{if } p = p' = \text{null} \\
\{ b = \text{true} \quad& \text{if } p = (\pi, a), p' = (\pi', a'), \text{and } a = a' \\
\{ b = \text{false} \quad& \text{otherwise} \\
\} A, M \rightarrow A, M
\end{align*}
\]

We call these two options `POINTER-EQUALITY-PROVENANCE-NONDET TRUE` and `false`.

**Pointer array offset**  Given a pointer `p` at C type `\(\tau\)`, the result of offsetting `p` by integer `x` (either by array indexing or explicit pointer/integer addition) is as follows, where `x = n` in PNVI\(^*\), or `x = (\pi, n)` in PVI. For the operation to succeed, `p` must be some non-null `(@i, a)`. Then there must be a live storage instance for `i`, and the numeric result of the addition of `a + n \times \text{sizeof}(\tau)` must be within or one-past the footprint of that storage instance. Otherwise the operation flags UB. For PNVI-ae-udi, if `p` is ambiguous (i.e., `p = (i, a)` and `A(i) = \{i_1, i_2\}` then if `x` is non-zero this should only be defined behaviour for (at most) one of the two, and then `i` should be resolved to that one in the final state. If `x = 0` it does not resolve the ambiguity.

\[
\begin{align*}
\text{iso_array_offset_ptrval}(A, p, \tau, x) &= \begin{cases} \\
\text{UB: out of bounds} & \text{if all except the last conjunct} \\
\text{above hold} & \\
\text{UB: empty prov} & \text{if } p = (@\text{empty}, a) \\
\text{UB: killed prov} & \text{if } p = (@i, a) \text{ and } A(i) = \text{killed} \\
\text{UB: null pointer} & \text{if } p = \text{null} \\
\text{if } p = (@i, a) \text{ and } \\
\quad a' = a + n \times \text{sizeof}(\tau) \text{ and } \\
\quad A(i) = (\pi', \ldots, a'', \ldots, \ldots) \text{ and } \\
\quad a'' \in \{a''', a''', \ldots, \ldots\} \end{cases}
\end{align*}
\]
analogous to pointer array offset.

\[
\text{member\_offset\_ptrval}(p, r, m) = \begin{cases} 
(\pi, a'), & \text{if } p = (@i, a) \text{ and } \\
\text{A}(i) = (n'' \ldots a'' \ldots) \text{ and } \\
{a'} \in [a''..a'' + n''] \\
(\text{@empty}, \text{offsetof\_ival}(r, m)), & \text{if } p = \text{null}.
\end{cases}
\]

**Casts (PNVI-plain)** In PNVI-plain, a cast of a pointer value \( p \) to an integer value (at type \( r \)) just converts null pointers to zero and non-null pointer values to the address \( a \) of the pointer, if that is representable in \( r \), otherwise flagging UB. The provenance of the pointer is discarded. At present we require that the object is live and that its address is within bounds.

\[
\text{cast\_ptrval\_to\_ival}(r, p) = \begin{cases} 
0, & \text{if } p = \text{null}; \\
a, & \text{if } p = (@i, a) \text{ and } \\
\text{A}(i) = (n'' \ldots a'' \ldots) \text{ and } \\
a \in [a''..a'' + n''] \text{ and } a \in \text{value\_range}(r) \\
\text{UB}, & \text{otherwise}
\end{cases}
\]

In PNVI-plain, an integer-to-pointer cast of 0 returns the null pointer. For a non-0 integer \( x \), casting to a pointer to \( r \), if there is a storage instance \( i \) in the current memory model state \( (A, M) \) for which the address of the pointer would be properly within the footprint of the storage instance, it returns a pointer \( (@i, x) \) with the provenance of that storage instance. (The "properly within" prevents the one-past ambiguous case.) If there is no such storage instance, it returns a pointer with empty provenance.

\[
\text{cast\_ival\_to\_ptrval}(r, x) = \begin{cases} 
\text{null}, & \text{if } x = 0 \\
(@i, x), & \text{if } \text{A}(i) = (n\ldots a, f, k, \text{exposed}) \text{ and } x \in [a..a + n - 1] \\
(\text{@empty}, x), & \text{if there is no such } i
\end{cases}
\]

**Casts (PNVI-ae)** In PNVI-ae, the result of a cast of a pointer value \( p \) to an integer value is exactly as in PNVI-plain. In addition, for a cast of pointer value \( p = (@i, a) \) with provenance \( @i \), where \( \text{A}(i) = (n, \tau_{opt}, a, f, k, t) \) is the storage instance metadata for \( i \), the memory state \( (A, M) \) is updated to \( (A \mapsto (n, \tau_{opt}, a, f, k, \text{exposed}), M) \) to mark the that storage instance as exposed.

In PNVI-ae, an integer-to-pointer cast of 0 returns the null pointer. For a non-0 integer \( x \), casting to a pointer to \( r \), if there is a storage instance \( i \) in the current memory model state \( (A, M) \) for which the address of the pointer would be properly within the footprint of the storage instance, and storage instance \( i \) is exposed, it returns a pointer \( (@i, x) \) with the provenance of that storage instance. If there is no such storage instance, it returns a pointer with empty provenance.

\[
\text{cast\_ival\_to\_ptrval}(r, x) = \begin{cases} 
\text{null}, & \text{if } x = 0 \\
(@i, x), & \text{if } \text{A}(i) = (n\ldots a, f, k, \text{exposed}) \text{ and } x \in [a..a + n - 1] \\
(\text{@empty}, x), & \text{if there is no such } i
\end{cases}
\]

**Casts (PNVI-ae-udi)** In PNVI-ae-udi, a cast of a pointer value \( p \) to an integer is just like PNVI-ae. Unlike PNVI-ae, PNVI-ae-udi permits a cast of a one-past pointer to integer and back to recover the original provenance, replacing the integer-to-pointer check that \( x \) is properly within the footprint of the storage instance by a check that it is properly within or one-past:

\[
\text{cast\_ival\_to\_ptrval}(r, x) = \begin{cases} 
\text{null}, & \text{if } x = 0 \\
(@i, x), & \text{if } \text{A}(i) = (n\ldots a, f, k, \text{exposed}) \text{ and } x \in [a..a + n] \\
(\text{@empty}, x), & \text{if there is no such } i
\end{cases}
\]

But then a PNVI-ae-udi cast of an integer value to a pointer can create a pointer with ambiguous provenance (as in the definition of repr): if it could be within or one-past two live storage instances, with IDs \( i_1 \) and \( i_2 \), and both storage instances have been marked as exposed, the resulting pointer value is given a fresh symbolic storage instance ID \( i \), and \( A \) is updated to map \( i \) to \([i_1, i_2]\). This can only happen if the two storage instances are adjacent and the address is one-past the first and at the start of the second.
Casts (PVI)

\[
\text{cast}_\text{ival}_\text{to}_\text{ptrval}(\tau, x) = \begin{cases} 
\text{null}, & \text{if } x = (@\text{empty}, 0) \\
(\tau, n), & \text{otherwise, where } x = (\pi, n) 
\end{cases}
\]

\[
\text{cast}_\text{ptrval}_\text{to}_\text{ival}(\tau, p) = \begin{cases} 
(\pi, a), & \text{if } p = (\pi, a) \text{ and } a \in \text{value_range}(\tau) \\
\text{UB}, & \text{otherwise}
\end{cases}
\]

Integer operations (PVI) In PVI one also has to define the provenance results of all the other operations returning integer values. Below we do so for the basic operations, though this would also be needed for all the integer-returning library functions. Most would give integers with empty provenance. One might or might not also want to require that the objects of those provenances are live.

\[
\pi \odot \pi' = \begin{cases} 
\pi, & \text{if } \pi = \pi' \text{ or } \pi' = @\text{empty}; \\
\pi', & \text{if } \pi = @\text{empty}; \\
@\text{empty}, & \text{otherwise.}
\end{cases}
\]

\[
op\text{ival}(op, (\pi, n), (\pi', m)) = (\pi \odot \pi', op(n, m)), \text{ where } op \in \{+, *, /, \%, &, |, \land\}
\]

\[
op\text{ival}(-, (\pi, n), (\pi', m)) = \begin{cases} 
(\pi', n - m), & \text{if } \pi = @i \text{ and } \pi' = @i', \text{ whether } i = i' \text{ or not}; \\
@\text{empty}, & \text{if } \pi = @\text{empty};
\end{cases}
\]

\[
eq\text{ival}((\pi, n), (\pi', m)) = (n = m)
\]

\[
\lt\text{ival}((\pi, n), (\pi', m)) = (n < m)
\]

\[
\leq\text{ival}((\pi, n), (\pi', m)) = (n \leq m)
\]

3.4 No-expose annotation

For PNVI-ae and PNVI-ae-udi, to permit implementations, e.g. of memcpy-like functions, to operate on representation bytes but without needlessly leaving all the storage instances that were pointed to in those bytes exposed, we envisaged some "no-expose" annotation that users could apply to such code. But now it's not so clear how that could work. We can turn off exposure during execution of annotated code easily enough (though Jens points out that this might not be the right thing for code which is passed a function pointer). But if the user-memcpy code copies bytes via a char * pointer, then the resulting abstract types in memory still have empty provenance (because we’re not tracking provenance via the intervening integer values), so when a pointer value is read (after the user-memcpy) from the copy, it will still get empty provenance.

3.5 Provenance of other operations

In addition to the operations defined above, some operations are desugared/elaborated to simpler expressions by the Cerberus pipeline. Their PVI results have provenance as follows; their PNVI’ results are the same except that there integers have no provenance:

- the result of address-of ($) has the provenance of the object associated with the lvalue, for non-function-pointers, or empty for function pointers.
- prefix increment and decrement operators follow the corresponding pointer or integer arithmetic rules.
- the conditional operator has the provenance of the second or third operand as appropriate; simple assignment has the provenance of the expression; compound assignment follows the pointer or integer arithmetic rules; the comma operator has the provenance of the second operand.
- integer unary +, unary -, and ~ operators preserve the original provenance; logical negation ! has a value with empty provenance.
- sizeof and Alignof operators give values with empty provenance.
- bitwise shifts has the provenance of their first operand.
- Jens Gustedt highlights that atomic operations have their own specific provenance properties, not yet discussed here, as do some library functions.

REFERENCES


Draft of April 1, 2019


Draft of April 1, 2019