xh

Spencer Tipping

March 22, 2014

Contents

I Language reference 2

1 Similarities to TCL 3

2 Similarities to Lisp 6

3 Dissimilarities from everything else I know of 7

4 Functions 9

II Bootstrap implementation 10

5 Self-replication 11

6 Perl-hosted evaluator 13

Part I

Language reference

Chapter 1

Similarities to TCL

Every xh value is a string. This includes lists, functions, closures, lazy expressions, scope chains, call stacks, and heaps. Asserting string equivalence makes it possible to serialize any value losslessly, including a running xh process.[1]

Although the string equivalence is available, most operations have higher-level structure. For example, the `$` operator, which performs string interpolation, interpolates values in such a way that two things are true:

1. No interpolated value will be further interpolated (idempotence).
2. The interpolated value will be read as a single list element.

For example:

```
(def bar bif)
(def foo "hi there \$bar!")
(def baz $foo)                      # no quoting necessary here by (2)
(identity $baz)
"hi there \$bar!"
(echo $baz)
hi there $bar!                      # $bar unevaluated by (1)
(identity $foo)
"hi there \$bar!"                   # $foo == $baz, of course
(echo $foo)
hi there $bar!
()
```

This interpolation structure can be overridden by using one of three alternative forms of `$`:

[1] Note that things like active socket connections and external processes will be proxied, however; xh can't migrate system-native things.

```
(def bar bif)
(def foo "hi there \$bar!")
(echo $!foo)                          # allow re-interpolation
hi there bif!
(count [$foo])                        # single element
1
(count [$@foo])                       # multiple elements
3
(nth [$@!foo] 2)                      # multiple and re-interpolation
bif!
()
```

All string values in xh programs are lifted into reader-safe quotations. This causes any "active" characters such as `$` to be prefixed with backslashes, a transformation you can mostly undo by using `$@!`. The only thing you can't undo is bracket balancing, which if undone would wreak havoc on your programs. You can see the effect of balancing by doing something like this:

```
(def foo "[[ [[")
(def bar [$@!foo])
(echo $bar)
["[[ [["]
()
```

1.1 Type hints

The string form of a value conveys its type. xh syntax supports the following structures:

```
[x y z ...]                           # array/vector
{x y z ...}                           # map
(x y z ...)                           # interpolated (active!) list
$x                                    # interpolated (active!) variable
bareword                              # string with interpolation
-42.0                                 # string with interpolation
"..."                                 # string with interpolation
'...'                                 # string with no interpolation
\x                                    # single-character string, no interp
```

When you ask xh about the type of a value, xh looks at the first byte and figures it out.[2] Because of this, not all strings are convertible to values despite all values being convertible to strings. You can easily convert between types by interpolating:

[2]Note that xh is in no way required to represent these values as strings internally. It just lies so convincingly that you would never know the difference.

```
(def list-form [1 2 3 4])
(def string-form "$@list-form")
(identity $list-form)
[1 2 3 4]
(identity $string-form)
"1 2 3 4"
(def map-form {$@list-form})
(identity $map-form)
{1 2 3 4}
()
```

Therefore the meaning of `$@x` could be interpreted as, "the untyped version of `x`," and `$!@x` could be, "eval the untyped version of `x`."

1.2 Laziness and localization

xh is a distributed runtime with serializable lazy values, which is a potential problem if you want to avoid proxying all over the place. Fortunately, a more elegant solution exists in most cases. Rather than using POSIX calls directly, xh programs access system resources like files through a slight indirection:

```
(def some-bytes (subs /etc/passwd 0 4096))
(echo $some-bytes)                # $some-bytes is lazy
```

This is clearly trivial if the `def` and `echo` execute on the same machine. But the `echo` can also be moved trivially by adding a hostname component to the file:

```
(def some-bytes (subs /etc/passwd 0 4096))
(identity $some-bytes)
(subs @host1/etc/passwd 0 4096)
```

This `@host1` namespace allows any remote xh runtime to negotiate with the original host, making lazy values fully mobile (albeit possibly slower).

Chapter 2

Similarities to Lisp

xh is strongly based on the Lisp family of languages, most visibly in its homoiconicity. Any string wrapped in `[]`, `{}`, or `()` is interpreted as a list of words, just as it is in Clojure. Also as in Lisp in general, `()` interpolates its result into the surrounding context:

```
(def foo 'hi there')
(echo $foo)
hi there
(echo (echo $foo))                  # similar to bash's $()
hi there
()
```

Any `()` list can be prefixed with `@` and/or `!` with effects analogous to `$`; e.g. `echo !@(echo hi there)`.

Chapter 3

Dissimilarities from everything else I know of

xh evaluates expressions outside-in:

1. Variable shadowing is not generally possible.
2. Expansion is idempotent for any set of bindings.
3. Unbound variables expand to active versions of themselves (a corollary of 2).
4. Laziness is implemented by referring to unbound quantities.
5. Bindings can be arbitrary list expressions, not just names (a partial corollary of 4).
6. No errors are ever thrown; all expressions that cannot be evaluated become `(error)` clauses that most functions consider to be opaque.
7. xh has no support for syntax macros.

Unbound names are treated as though they might at some point exist. For example:

```
(echo $x)
$x
(def x $y)
(echo $x)
$y
(def y 10)
(echo $x)
10
()
```

You can also bind expressions of things to express partial knowledge:

```
(echo (count $str))
(count $str)
(def (count $str) 10)
(echo $str)
$str
(echo (count $str))
10
()
```

This is the mechanism by which xh implements lazy evaluation, and it's also the reason you can serialize partially-computed lazy values.

Chapter 4

Functions

xh supports two equivalent ways to write function-like relations:

```
(def (foo $x) {echo hi there, $x!})
(foo spencer)
hi there, spencer!
()
```

This is named definition by destructuring, which works great for most cases. When you're writing an anonymous function, however, you'll need to describe the mappings individually:

```
(reduce {[$total +$x] (+ $total $x)
         [$total *$x] (* $total $x)} \
        0 \
        [+1 +2 *5 +1])
16
()
```

Part II

Bootstrap implementation

Chapter 5

Self-replication

Listing 5.1 boot/xh-header

```
#!/usr/bin/env perl
BEGIN {
print STDERR q{
NOTE: Development image

If you see this note after installing the shell, it's probably because
you're running a version that has not yet rebuilt itself (maybe you got the
wrong file from the Git repo?). You can do this, but it will be really
slow and may use a lot of memory. There are two ways to fix this:

1. Download the standard image from http://spencertipping.com/xh
2. Have this image recompile itself by running xh.recompile-in-place (this
   will take some time because it stress-tests your Perl runtime)

Note also that bootstrapping requires Perl 5.14 or later, whereas running a
compiled image just requires Perl 5.10.

};
}

BEGIN {eval(our $xh_bootstrap = q{
# xh: the X shell | https://github.com/spencertipping/xh
# Copyright (C) 2014, Spencer Tipping
# Licensed under the terms of the MIT source code license

# For the benefit of HTML viewers (long story):
# <body style='display:none'>
# <script src='http://spencertipping.com/xh/page.js'></script>
use 5.014;
```

```
30 package xh;
31 our %modules;
32 our @module_ordering;
33
34 our %compilers = (pl => sub {
35   my $package = $_[0] =~ s/\./::/gr;
36   eval "{package ::$package;\n$_[1]\n}";
37   die "error compiling module $_[0]: $@" if $@;
38 });
39
40 sub defmodule {
41   my ($name, $code, @args) = @_;
42   chomp($modules{$name} = $code);
43   push @module_ordering, $name;
44   my ($base, $extension) = split /\.(\w+$)/, $name;
45   die "undefined module extension '$extension' for $name"
46     unless exists $compilers{$extension};
47   $compilers{$extension}->($base, $code, @args);
48 }
49
50 chomp($modules{bootstrap} = $::xh_bootstrap);
51 undef $::xh_bootstrap;
```

At this point we need a way to reproduce the image. Since the bootstrap code is already stored, we can just wrap it and each defined module into an appropriate `BEGIN` block.

Listing 5.2 `boot/xh-header (continued)`

```
1  sub image {
2    my @pieces = "#!/usr/bin/env perl";
3    push @pieces, "BEGIN {eval(our \$xh_bootstrap = <<'_')}",
4                  $modules{bootstrap},
5                  '_';
6    push @pieces, "BEGIN {xh::defmodule('$_', <<'_')}",
7                  $modules{$_},
8                  '_' for @module_ordering;
9    push @pieces, "xh::main::main;\n__DATA__";
10   join "\n", @pieces;
11 }
12 })}
```

Chapter 6

Perl-hosted evaluator

xh is self-hosting, but to get there we need to implement an interpreter in Perl. This interpreter is mostly semantically correct but slow and shouldn't be used for anything besides bootstrapping the real compiler.

Listing 6.1 modules/interpreter.pl

```
BEGIN {xh::defmodule('xh::interpreter.pl', <<'_')}
use Memoize qw/memoize/;
use List::Util qw/max/;

sub active_regions {
  # Returns a series of numbers that describes, in pre-order, regions of
  # the given string that should be interpolated. The numeric list has the
  # following format:
  #
  # (offset << 32 | len), (offset << 32 | len) ...

  my @pieces = split /(\\.|\$@?!?\w+|\$@?!?\{[^}]+\}|@?!?\(|[')])/s, $_[0];
  my $offset = 0;
  my @result;
  my @quote_offsets;

  for (@pieces) {
    if (@quote_offsets && substr($_[0], $quote_offsets[-1], 1) eq "'") {
      # We're inside a hard-quote, so ignore everything except for the next
      # hard-quote.
      pop @quote_offsets if /^'/;
    } else {
      if (/^'/ || /^@?!?\(/) {
        push @quote_offsets, $offset;
      } elsif (/^\$/) {
        push @result, $offset << 32 | length;
```

```
      } elsif (/^\)/) {
        my $start = pop @quote_offsets;
        push @result, $start << 32 | $offset + 1 - $start;
      }
    }
    $offset += length;
  }

  sort {$a <=> $b} @result;
}

memoize 'active_regions';

our %closers = ('(' => ')', '[' => ']', '{' => '}');
sub element_regions {
  # Returns integer-encoded regions describing the positions of list
  # elements. The list passed into this function should be unwrapped; that
  # is, it should have no braces.
  my ($xs)   = @_;
  my $offset = 0;
  my @pieces = split / ( "(?:\\.|[^"])*"
                       | '(?:\\.|[^'])*'
                       | \\.
                       | [({\[\]})]
                       | \s+ ) /xs, $_[0];
  my @paren_offsets;
  my @parens;
  my @result;
  my $item_start = -1;

  for (@pieces) {
    unless (@paren_offsets) {
      if (/\s+/ || /^[)\]}]/) {
        # End any item if we have one.
        push @result, $item_start << 32 | $offset - $item_start
        if $item_start >= 0;
        $item_start = -1;
      } else {
        # Start an item unless we've already done so.
        $item_start = $offset if $item_start < 0;
      }
    }

    # Update bracket tracking.
    if ($_ eq $closers{$parens[-1]}) {
      if (@parens) {
```

```
        pop @paren_offsets;
        pop @parens;
      } else {
        die 'illegal closing brace: ... '
          . substr($xs, max(0, $offset - 10), 20)
          . ' ...'
          . "\n(whole string is $xs)";
      }
    } elsif (/^[(\[{]/) {
      push @paren_offsets, $offset;
      push @parens, $_;
    }

    $offset += length;
  }

  push @result, $item_start << 32 | $offset if $item_start >= 0;
  @result;
}

memoize 'element_regions';

sub xh_list_box {
  $_[0] !~ /^[({\[]/ && element_regions(0, $_[0]) > 1
    ? "[$_[0]]"
    : $_[0];
}

sub xh_list_unbox {
  return $1 if $_[0] =~ /^\[(.*)\]$/
            || $_[0] =~ /^\((.*)\)$/
            || $_[0] =~ /^\{(.*)\}$/;
  $_[0];
}

sub parse_list {
  my $unboxed = xh_list_unbox $_[0];
  map xh_list_box(substr $unboxed, $_ >> 32, $_ & 0xffffffff),
      element_regions 0, $unboxed;
}

sub into_list  {'(' . join(' ', map xh_list_box($_), @_) . ')'}
sub into_vec   {'[' . join(' ', map xh_list_box($_), @_) . ']'}
sub into_block {'{' . join(' ', map xh_list_box($_), @_) . '}'}

sub xh_vecp   {$_[0] =~ /^\[.*\]$/}
```

```
sub xh_listp  {$_[0] =~ /^\(.*\)$/}
sub xh_blockp {$_[0] =~ /^\{.*\}$/}
sub xh_varp   {$_[0] =~ /^\$/}

sub xh_count {
  scalar element_regions 0, xh_list_unbox $_[0];
}

sub xh_nth {(parse_list $_[0])[$_[1]]}

sub xh_nth_eq {
  # FIXME
  my ($copy, $i, $v) = @_;
  my @regions        = element_regions 0, $copy;
  my $r              = $regions[$i];
  substr($copy, $r >> 32, $r & 0xffffffff) = $v;
  $copy;
}

sub destructuring_bind;
sub destructuring_bind {
  # Both $pattern and $v should be quoted; that is, the string character [
  # should be encoded as \[.
  my ($pattern, $v) = @_;
  my @pattern_elements = element_regions 0, $pattern;
  my @v_elements       = element_regions 0, $v;
  my %bindings;

  # NOTE: no $@ matching
  return undef unless @v_elements == @pattern_elements;

  # NOTE: no foo$bar matching (partial constants)
  for (my $i = 0; $i < @pattern_elements; ++$i) {
    my $pi = xh_nth $pattern, $i;
    my $vi = xh_nth $v,       $i;

    return undef if $pi !~ /^\$/ && $pi ne $vi;

    my @pattern_regions = element_regions 0, $pi;
    my @v_regions       = element_regions 0, $vi;
    return undef unless @pattern_regions == 1 && $pi =~ /^\$/
                     || @pattern_regions == @v_regions;

    if (xh_vecp $pi) {
      my $sub_bind = destructuring_bind $pi, $vi;
      return undef unless ref $sub_bind;
```

```
165       my %sub_bindings = %$sub_bind;
166       for (keys %sub_bindings) {
167         return undef if exists $bindings{$_}
168                      && $bindings{$_} ne $sub_bindings{$_};
169         $bindings{$_} = $sub_bindings{$_};
170       }
171     } elsif (xh_listp $pi) {
172       die "TODO: implement list binding for $pi";
173     } elsif ($pi =~ /^\$\{?(\w+)\}?$/) {
174       return undef if exists $bindings{$1} && $bindings{$1} ne $vi;
175       $bindings{$1} = $vi;
176     } elsif ($pi =~ /^\$/) {
177       die "illegal binding form: $pi";
178     } else {
179       return undef unless $pi eq $vi;
180     }
181   }
182
183   {%bindings};
184 }
185
186 sub invoke;
187 sub interpolate;
188 sub interpolate {
189   # Takes a string and a compiled binding hash and interpolates all
190   # applicable substrings outside-in. This process may involve full
191   # evaluation if () subexpressions are present, and is in general
192   # quadratic or worse in the length of the string.
193   my $bindings             = $_[0];
194   my @interpolation_regions = active_regions $_[1];
195   my @result_pieces;
196
197   for (@interpolation_regions) {
198     my $slice = substr $_[0], $_ >> 32, $_ & 0xffffffff;
199
200     # NOTE: no support for complex ${} expressions
201     if ($slice =~ /^\$(@?!?)\{?(\w+)\}?$/) {
202       # Expand a named variable that may or may not be defined yet.
203       push @result_pieces,
204            exists ${$bindings}{$2} ?
205                $1 eq ''  ? xh_listquote(xh_deactivate $bindings->{$2})
206              : $1 eq '@' ? xh_deactivate($bindings->{$2})
207              : $1 eq '!' ? xh_listquote($bindings->{$2})
208              :             $bindings->{$2}
209            : "\$$slice";
210     } elsif ($slice =~ /^\((.*)\)$/s) {
```

```
      push @result_pieces, invoke $bindings, parse_list interpolate $1;
    } else {
      push @result_pieces, $slice;
    }
  }

  join '', @result_pieces;
}

sub xh_function_cases {
  # FIXME
  my @result;
  my @so_far;
  for (parse_vlist $_[0]) {
    my ($command, @args) = parse_list $_;
    if (xh_vecp $command) {
      push @result, into_block @so_far if @so_far;
      @so_far = ($command, into_list @args);
    }
  }
  push @result, into_block @so_far if @so_far;
  @result;
}

sub evaluate;
sub invoke {
  # NOTE: no support for (foo bar $x)-style conditional destructuring;
  # these are all rewritten into lambda forms
  my ($bindings, $f, @args) = @_;
  my $args = into_vec @args;

  # Resolve f into a lambda form if it's still in word form.
  $f = $bindings->{$f} if exists $bindings->{$f};

  # Escape into perl
  return $f->($bindings, @args) if ref $f eq 'CODE';

  my %nested_bindings = %$bindings;
  for (xh_function_cases $f) {
    my ($formals, @body) = parse_block $_;
    if (my $maybe_bindings = destructuring_bind $formals, $args) {
      $nested_bindings{$_} = $$maybe_bindings{$_}
      for keys %$maybe_bindings;
      return evaluate {%nested_bindings}, into_block @body;
    }
  }
```

```
257
258   return into_list $f, @args;
259 }
260
261 sub evaluate {
262   my ($bindings, $block) = @_;
263   my @statements         = parse_block $block;
264   my $result;
265
266   # NOTE: this function updates $bindings in place.
267   for (@statements) {
268     # Each statement is an invocation, which for now we assume all to be
269     # functions.
270     #
271     # NOTE: this is semantically incomplete as we don't consider
272     # macro-bindings.
273     $result = invoke $bindings, parse_list interpolate $bindings, $_;
274   }
275   $result;
276 }
277 _
```