Compiling and Optimizing Scripting Languages

Paul Biggar and David Gregg

Department of Computer Science and Statistics
Trinity College Dublin

LLNL, 17th March, 2009

Motivation

- User needs web page in 0.5 seconds
  - Execution time
  - DB access
  - Network latency
  - Browser rendering
- Easier maintainance
- What if execution was:
  - 2x as fast?
  - 10x as fast?

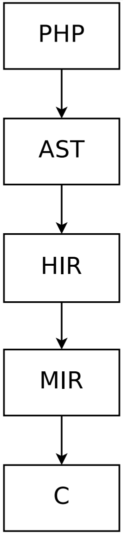
Outline

Outline

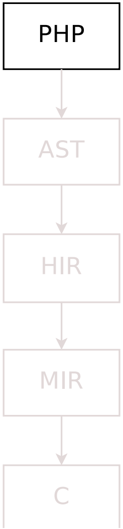
phc

- `http://phpcompiler.org`
- Ahead-of-time compiler for PHP
- Edsko de Vries, John Gilbert, Paul Biggar
- BSD license
- Latest release: 0.2.0.3 - compiles non-OO
- svn trunk: compiles most OO

Structure of **phc**

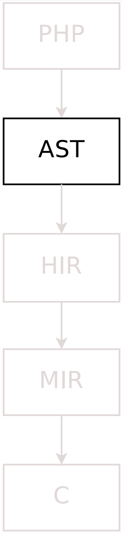


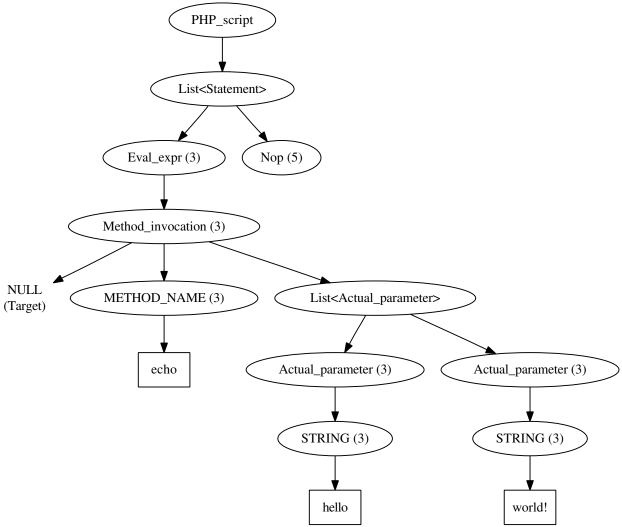
PHP



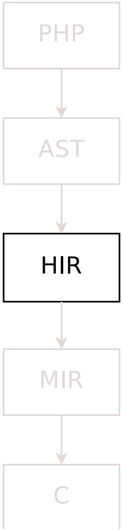
```
<?php

  echo "hello", "world!";

?>
```

AST


PHP
AST
HIR
MIR
C


PHP_script
List<Statement>
Eval_expr (3)
Nop (5)
Method_invocation (3)
NULL (Target)
METHOD_NAME (3)
List<Actual_parameter>
echo
Actual_parameter (3)
Actual_parameter (3)
STRING (3)
STRING (3)
hello
world!

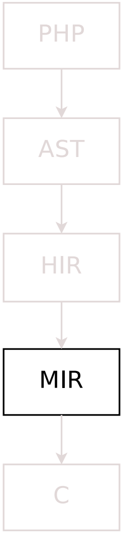
HIR



```
<?php
  $x = $a + $b + $c + $d;
?>
```

```
<?php
  $TLE0 = ($a + $b);
  $TLE1 = ($TLE0 + $c);
  $x = ($TLE1 + $d);
?>
```

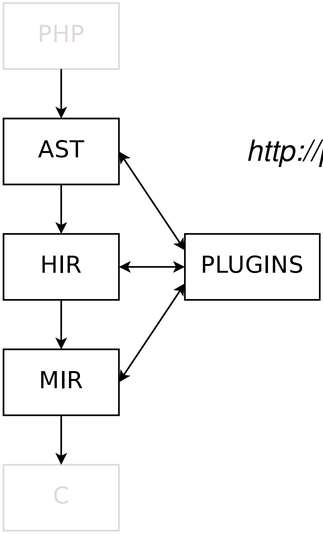
MIR



```
<?php
  while ($cond)
    echo "hello", "world!";
?>
```

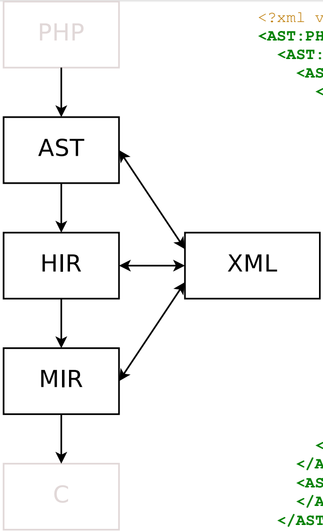
```
<?php
L7:
  $TLE0 = !$cond;
  if ($TLE0) goto L3 else goto L6;
L6:
  print('hello');
  print('world!');
  goto L7;
L3:
?>
```

Plugins



http://phpcompiler.org/doc/latest/devmanual.html

XML



```
<?xml version="1.0"?>
<AST:PHP_script xmlns:AST="http://www.phpcompiler.org/phc-1.1">
  <AST:Statement_list>
    <AST:Eval_expr>
      <AST:Method_invocation>
        <AST:Target xsi:nil="true" />
        <AST:METHOD_NAME>
          <value>echo</value>
        </AST:METHOD_NAME>
        <AST:Actual_parameter_list>
          <AST:Actual_parameter>
            <bool><!-- is_ref -->false</bool>
            <AST:STRING>
              <value>hello</value>
            </AST:STRING>
          </AST:Actual_parameter>
          <AST:Actual_parameter>
            <bool><!-- is_ref -->false</bool>
            <AST:STRING>
              <value>world!</value>
            </AST:STRING>
          </AST:Actual_parameter>
        </AST:Actual_parameter_list>
      </AST:Method_invocation>
    </AST:Eval_expr>
    <AST:Nop>
    </AST:Nop>
  </AST:Statement_list>
</AST:PHP_script>
```

Outline

SAC 2009

A Practical Solution for Scripting Language Compilers

Paul Biggar, Edsko de Vries and David Gregg

Department of Computer Science and Statistics
Trinity College Dublin

ACM Symposium on Applied Computing - PL track
12th March, 2009

Sneak peak

- Problem: Scripting languages present “unique” problems (in practice)
- Solution: Re-use as much of the ~~*Canonical*~~ *Reference Implementation* as possible.

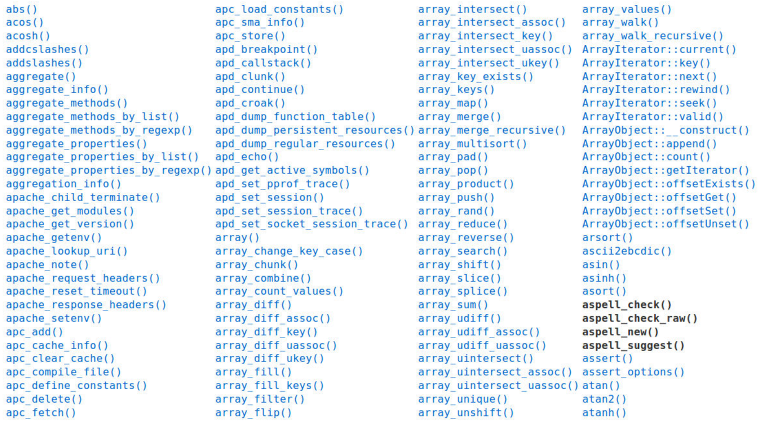
Outline

Undefined

The PHP group claim that they have the final say in the specification of PHP. This group's specification is an implementation, and there is no prose specification or agreed validation suite. There are alternate implementations [...] that claim to be compatible (they don't say what this means) with some version of PHP.

D. M. Jones. Forms of language specification: Examples from commonly used computer languages. ISO/IEC JTC1/SC22/OWG/N0121, February 2008.

Batteries included



abs()	apc_load_constants()	array_intersect()	array_values()
acos()	apc_sma_info()	array_intersect_assoc()	array_walk()
acosh()	apc_store()	array_intersect_key()	array_walk_recursive()
addcslashes()	apd_breakpoint()	array_intersect_uassoc()	ArrayIterator::current()
addslashes()	apd_callstack()	array_intersect_ukey()	ArrayIterator::key()
aggregate()	apd_clunk()	array_key_exists()	ArrayIterator::next()
aggregate_info()	apd_continue()	array_keys()	ArrayIterator::rewind()
aggregate_methods()	apd_croak()	array_map()	ArrayIterator::seek()
aggregate_methods_by_list()	apd_dump_function_table()	array_merge()	ArrayIterator::valid()
aggregate_methods_by_regexp()	apd_dump_persistent_resources()	array_merge_recursive()	ArrayObject::__construct()
aggregate_properties()	apd_dump_regular_resources()	array_multisort()	ArrayObject::append()
aggregate_properties_by_list()	apd_echo()	array_pad()	ArrayObject::count()
aggregate_properties_by_regexp()	apd_get_active_symbols()	array_pop()	ArrayObject::getIterator()
aggregation_info()	apd_set_pprof_trace()	array_product()	ArrayObject::offsetExists()
apache_child_terminate()	apd_set_session()	array_push()	ArrayObject::offsetGet()
apache_get_modules()	apd_set_session_trace()	array_rand()	ArrayObject::offsetSet()
apache_get_version()	apd_set_socket_session_trace()	array_reduce()	ArrayObject::offsetUnset()
apache_getenv()	array()	array_reverse()	arsort()
apache_lookup_uri()	array_change_key_case()	array_search()	ascii2ebcdic()
apache_note()	array_chunk()	array_shift()	asin()
apache_request_headers()	array_combine()	array_slice()	asinh()
apache_reset_timeout()	array_count_values()	array_splice()	asort()
apache_response_headers()	array_diff()	array_sum()	**aspell_check()**
apache_setenv()	array_diff_assoc()	array_udiff()	**aspell_check_raw()**
apc_add()	array_diff_key()	array_udiff_assoc()	**aspell_new()**
apc_cache_info()	array_diff_uassoc()	array_udiff_uassoc()	**aspell_suggest()**
apc_clear_cache()	array_diff_ukey()	array_uintersect()	assert()
apc_compile_file()	array_fill()	array_uintersect_assoc()	assert_options()
apc_define_constants()	array_fill_keys()	array_uintersect_uassoc()	atan()
apc_delete()	array_filter()	array_unique()	atan2()
apc_fetch()	array_flip()	array_unshift()	atanh()

Jeff Atwood, Coding Horror, May 20th, 2008
http://www.codinghorror.com/blog/archives/001119.html

Change between releases

```
<?php
  var_dump (0x9fa0ff0b);
?>
```

PHP 5.2.1 (32-bit)

```
int(2147483647)
```

PHP 5.2.3 (32-bit)

```
float(2678128395)
```

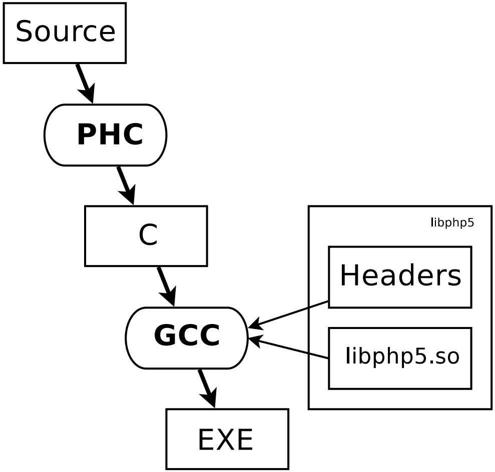
Run-time code generation

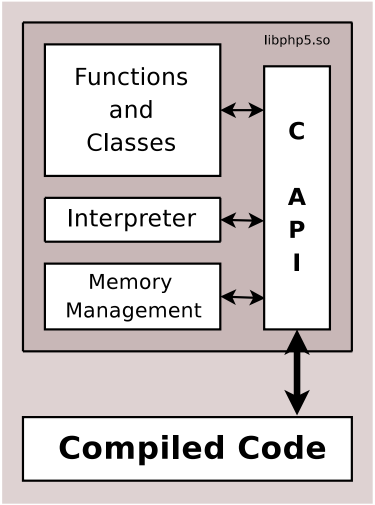
```
<?php
  eval ($argv[1]);
?>
```

```
<?php
  include ("mylib.php");
  ...
  include ("plugin.php");
  ...
?>
```

Outline

Use C API


Source
PHC
C
GCC
EXE
libphp5
Headers
libphp5.so


libphp5.so
Functions and Classes
Interpreter
Memory Management
C API
Compiled Code

More detail

PHP	zval
Python	PyObject
Ruby	VALUE
Lua	TValue

H. Muhammad and R. Ierusalimschy. C APIs in extension and extensible languages. Journal of Universal Computer Science, 13(6):839–853, 2007.

Simple listings: $i = 0

```
// $i = 0;
{
  zval* p_i;
  php_hash_find (LOCAL_ST, "i", 5863374, p_i);
  php_destruct (p_i);
  php_allocate (p_i);
  ZVAL_LONG (*p_i, 0);
}
```

Example: $i = 0

```
// $i = 0;
{
  if (local_i == NULL)
  {
    local_i = EG (uninitialized_zval_ptr);
    local_i->refcount++;
  }
  zval **p_lhs = &local_i;

  zval *value;
  if ((*p_lhs)->is_ref)
  {
    // Always overwrite the current value
    value = *p_lhs;
    zval_dtor (value);
  }
  else
  {
    ALLOC_INIT_ZVAL (value);
    zval_ptr_dtor (p_lhs);
    *p_lhs = value;
  }

  ZVAL_LONG (value, 0);
}
```

Example: $i = $j

```
// $i = $j;
{
  if (local_i == NULL)
  {
    local_i = EG (uninitialized_zval_ptr);
    local_i->refcount++;
  }
  zval **p_lhs = &local_i;

  zval *rhs;
  if (local_j == NULL)
    rhs = EG (uninitialized_zval_ptr);
  else
    rhs = local_j;

  if (*p_lhs != rhs)
  {
    if ((*p_lhs)->is_ref)
    {
      // First, call the destructor to remove any data structures
      // associated with lhs that will now be overwritten
      zval_dtor (*p_lhs);
      // Overwrite LHS
      (*p_lhs)->value = rhs->value;
      (*p_lhs)->type = rhs->type;
      zval_copy_ctor (*p_lhs);
    }
    else
    {
      zval_ptr_dtor (p_lhs);
      if (rhs->is_ref)
      {
        // Take a copy of RHS for LHS
        *p_lhs = zvp_clone_ex (rhs);
      }
      else
      {
        // Share a copy
        rhs->refcount++;
        *p_lhs = rhs;
      }
    }
```

Example: printf ($f)

```c
static zend_fcall_info printf_fci;
static zend_fcall_info_cache printf_fcic = { 0, NULL, NULL, NULL };

// printf($f);
{
  if (!printf_fcic->initialized)
  {
    zval fn;
    INIT_PZVAL (&fn);
    ZVAL_STRING (&fn, "printf", 0);
    int result = zend_fcall_info_init (&fn, &printf_fci, &printf_fcic TSRMLS_CC);
    if (result != SUCCESS)
    {
      phc_setup_error (1, "listings_source.php", 8, NULL TSRMLS_CC);
      php_error_docref (NULL TSRMLS_CC, E_ERROR,
          "Call to undefined function %s()", function_name);
    }
  }

  zend_function *signature = printf_fcic.function_handler;
  zend_arg_info *arg_info = signature->common.arg_info; // optional

  int by_ref[1];
  int abs_index = 0;
  // TODO: find names to replace index
  if (arg_info)
  {
    by_ref[abs_index] = arg_info->pass_by_reference;
    arg_info++;
  }
  else
    by_ref[abs_index] = signature->common.pass_rest_by_reference;

  abs_index++;


  // Setup array of arguments
  // TODO: i think arrays of size 0 is an error
  int destruct[1];
  zval *args[1];
  zval **args_ind[1];

  int af_index = 0;
  destruct[af_index] = 0;
  if (by_ref[af_index])
  {
    if (local_f == NULL)
    {
      local_f = EG (uninitialized_zval_ptr);
      local_f->refcount++;
    }
    zval **p_arg = &local_f;

    // We don't need to restore ->is_ref afterwards,
    // because the called function will reduce the
    // refcount of arg on return, and will reset is_ref to
    // 0 when refcount drops to 1.  If the refcount does
    // not drop to 1 when the function returns, but we did
    // set is_ref to 1 here, that means that is_ref must
    // already have been 1 to start with (since if it had
    // not, that means that the variable would have been
    // in a copy-on-write set, and would have been
    // seperated above).
    (*p_arg)->is_ref = 1;

    args_ind[af_index] = p_arg;

    assert (!in_copy_on_write (*args_ind[af_index]));
    args[af_index] = *args_ind[af_index];
  }
  else
  {
    zval *arg;
    if (local_f == NULL)
      arg = EG (uninitialized_zval_ptr);
    else
      arg = local_f;

    args[af_index] = fetch_var_arg (arg, &destruct[af_index]);
    if (arg->is_ref)
    {
      // We dont separate since we don't own one of ARG's references.
      arg = zvp_clone_ex (arg);
      destruct[af_index] = 1;

      // It seems we get incorrect refcounts without this.
      // TODO This decreases the refcount to zero, which seems wrong,
      // but gives the right answer. We should look at how zend does
      // this.

      arg->refcount--;
    }
    args[af_index] = arg;

    args_ind[af_index] = &args[af_index];
  }
  af_index++;


  phc_setup_error (1, "listings_source.php", 8, NULL TSRMLS_CC);

  // save existing parameters, in case of recursion
  int param_count_save = printf_fci.param_count;
  zval ***params_save = printf_fci.params;
  zval **retval_save = printf_fci.retval_ptr_ptr;

  zval *rhs = NULL;

  // set up params
  printf_fci.params = args_ind;
  printf_fci.param_count = 1;
  printf_fci.retval_ptr_ptr = &rhs;

  // call the function
  int success = zend_call_function (&printf_fci, &printf_fcic TSRMLS_CC);
  assert (success == SUCCESS);

  // restore params
  printf_fci.params = params_save;
  printf_fci.param_count = param_count_save;
  printf_fci.retval_ptr_ptr = retval_save;

  // unset the errors
  phc_setup_error (0, NULL, 0, NULL TSRMLS_CC);

  int i;
  for (i = 0; i < 1; i++)
  {
    if (destruct[i])
    {
      assert (destruct[i]);
      zval_ptr_dtor (args_ind[i]);
    }
  }


  // When the Zend engine returns by reference, it allocates a zval into
  // retval_ptr_ptr. To return by reference, the callee writes into the
  // retval_ptr_ptr, freeing the allocated value as it does.  (Note, it may
  // not actually return anything). So the zval returned - whether we return
  // it, or it is the allocated zval - has a refcount of 1.

  // The caller is responsible for cleaning that up (note, this is unaffected
  // by whether it is added to some COW set).

  // For reasons unknown, the Zend API resets the refcount and is_ref fields
  // of the return value after the function returns (unless the callee is
  // interpreted). If the function is supposed to return by reference, this
  // loses the refcount. This only happens when non-interpreted code is
  // called. We work around it, when compiled code is called, by saving the
  // refcount into SAVED_REFCOUNT, in the return statement. The downside is
  // that we may create an error if our code is called by a callback, and
  // returns by reference, and the callback returns by reference. At least
  // this is an obscure case.
  if (signature->common.return_reference
      && signature->type != ZEND_USER_FUNCTION)
  {
    assert (rhs != EG (uninitialized_zval_ptr));
    rhs->is_ref = 1;
    if (saved_refcount != 0)
    {
      rhs->refcount = saved_refcount;
    }
    rhs->refcount++;
  }
```

Applicability

- Everything
  - Perl
  - PHP
  - Ruby
  - Tcl – *I think*

Applicability

- Everything
  - Perl
  - PHP
  - Ruby
  - Tcl – *I think*
- Except specification
  - Lua
  - Python

Applicability

- Everything
  - Perl
  - PHP
  - Ruby
  - Tcl – *I think*
- Except specification
  - Lua
  - Python
- Not at all
  - Javascript

Outline

Original Speed-up

0.1x

(10 times slower than the PHP interpreter)

The problem with copies

```php
<?php
  for ($i = 0; $i < $n; $i++)
    $str = $str . "hello";
?>
```

```php
<?php
  for ($i = 0; $i < $n; $i++)
  {
    $T = $str . "hello";
    $str = $T;
  }
?>
```

Optimization

- Constant folding

```
<?php
  ...
  $T = "5" + true;
  ...
?>
```

```
<?php
  ...
  $T = 6;
  ...
?>
```

Optimization

- Constant folding
- Constant pooling

```
<?php
  $sum = 0;
  for ($i = 0; $i < 10; $i=$i+1)
  {
    $sum .= "hello";
  }
?>
```

Optimization

- Constant folding
- Constant pooling
- Function caching

```
// printf ($f);
static php_fcall_info printf_info;
{
  php_fcall_info_init ("printf", &printf_info);

  php_hash_find (
    LOCAL_ST, "f", 5863275, &printf_info.params);

  php_call_function (&printf_info);
}
```

Optimization

- Constant folding
- Constant pooling
- Function caching
- Pre-hashing

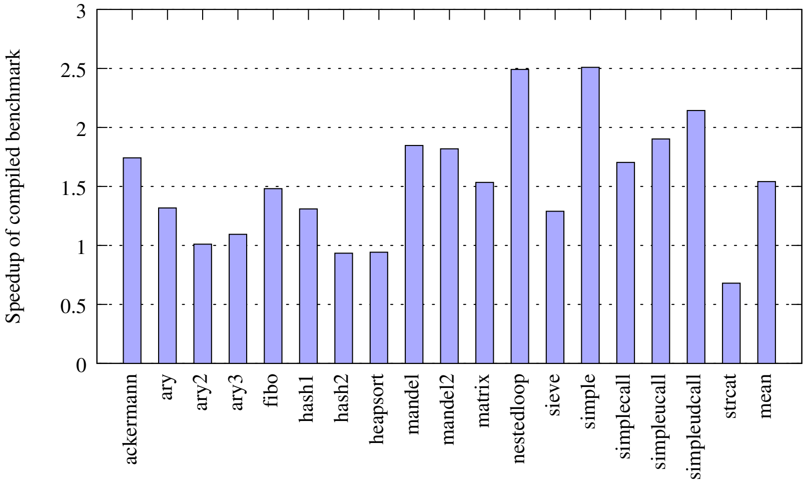
```
// $i = 0;
{
  zval* p_i;
  php_hash_find (LOCAL_ST, "i", 5863374, p_i);
  php_destruct (p_i);
  php_allocate (p_i);
  ZVAL_LONG (*p_i, 0);
}
```

Optimization

- Constant folding
- Constant pooling
- Function caching
- Pre-hashing
- Symbol-table removal

```
// $i = 0;
{
  php_destruct (local_i);
  php_allocate (local_i);
  ZVAL_LONG (*local_i, 0);
}
```

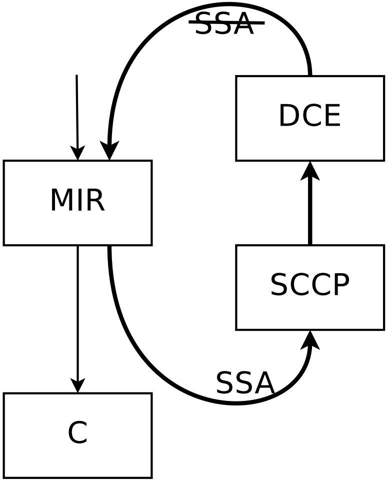
Current speed-up



Outline

Outline

Intra-procedural optimizations



- Dead-code elimination
- Sparse-conditional **constant propagation**

Type-inference

```
<?php

  function a ($x, $y)
  {
    $str = $x . $y;
    ...
    return $str;
  }

?>
```

User-space handlers

- __toString
- __get
- __set
- __isset
- __unset
- __sleep
- __wake
- __call
- __callStatic
- ...

C API handlers

- read_property
- read_dimension
- get
- set
- cast_object
- has_property
- unset_property
- ...

Unknown types propagate

- local symbol table
- global symbol table
- return values
- reference parameters
- callee parameters

Outline

Analysis design

- Must model types precisely
    - (Possibly unnamed) fields, arrays, variables and method calls

Analysis design

- Must model types precisely
  - (Possibly unnamed) fields, arrays, variables and method calls
- Uses and definitions incomplete
  - Can't use *def-use chains*
  - Can't use *SSA*

Analysis design

- Must model types precisely
  - (Possibly unnamed) fields, arrays, variables and method calls
- Uses and definitions incomplete
  - Can't use *def-use chains*
  - Can't use *SSA*
- Imprecise callgraph

Algorithm

- Abstract Execution / Interpretation

Algorithm

- Abstract Execution / Interpretation
- Points-to analysis
  - *-sensitive

Algorithm

- Abstract Execution / Interpretation
- Points-to analysis
  - *-sensitive
- Constant-propagation
  - Precision
  - Array-indices/field names
  - Implicit conversions

A. Pioli. Conditional pointer aliasing and constant propagation. Master's thesis, SUNY at New Paltz, 1999.

Algorithm

- Abstract Execution / Interpretation
- Points-to analysis
    - *-sensitive
- Constant-propagation
    - Precision
    - Array-indices/field names
    - Implicit conversions
- Type-inference
    - Virtual calls
    - Function annotations

Complex cases

- Hashtables
- Implicit conversions
- Variable-variables
- $GLOBALS
- Static includes
- $SESSION
- Compiler temporaries

Interesting thoughts

- Strip off first loop iteration

Interesting thoughts

- Strip off first loop iteration
- JITs or Gal/Franz Tracing?

Interesting thoughts

- Strip off first loop iteration
- JITs or Gal/Franz Tracing?
- Use string transducer analysis

Outline

Security

- Davis - if we include it, we'll do better

Sound and Precise Analysis of Web Applications
for Injection Vulnerabilities
Gary Wassermann, Zhendong Su, PLDI'07.

Static approximation of dynamically generated Web pages
Yasuhiko Minamide, WWW 2005

Security

- Davis - if we include it, we'll do better
- Tuwien/Pixy - taint analysis (literal analysis + points to)

Security

- Davis - if we include it, we’ll do better
- Tuwien/Pixy - taint analysis (literal analysis + points to)
- Utrecht/Stanford - dont remember

Summary

- Re-use existing run-time for language
- Better yet: standardize libraries (and language?), including FFI

- Analysis needs to be precise, and whole-program
- Pessimistic assumptions spread

- Language, implementation and community need to be fixed
  - All related?

Thanks

phc needs contributors

- contribute: `http://phpcompiler.org/contribute.html`
- mailing list: `phc-general@phpcompiler.org`

- slides: `http://www.cs.tcd.ie/~pbiggar/`
- contact: `paul.biggar@gmail.com`

Complex cases

- Hashtables
- Implicit conversions
- Variable-variables
- $GLOBALS
- Static includes
- $SESSION
- Compiler temporaries