"Look Ma, No Hashing, And No Arrays Neither"

Jiazhen Cai [1] *and Robert Paige* [2]

New York University/ Courant Institute
New York, NY 10012
and
University of Wisconsin
Madison, WI 53706

ABSTRACT

It is generally assumed that hashing is essential to many algorithms related to efficient compilation; e.g., symbol table formation and maintenance, grammar manipulation, basic block optimization, and global optimization. This paper questions this assumption, and initiates development of an efficient alternative compiler methodology without hashing or sorting. Underlying this methodology are several generic algorithmic tools, among which special importance is given to *Multiset Discrimination*, which partitions a multiset into blocks of duplicate elements. We show how multiset discrimination, together with other tools, can be tailored to rid compilation of hashing without loss in asymptotic performance. Because of the simplicity of these tools, our results may be of practical as well as theoretical interest. The various applications presented culminate with a new algorithm to solve iterated strength reduction folded with useless code elimination that runs in worst case asymptotic time and auxiliary space linear in the maximum text length of the initial and optimized programs.

1. Introduction.

An important practical and theoretical question in Computer Science is whether there are algorithms whose worst case performance can match the expected performance of solutions that utilize hashing. In the context of this broader question, we initiate an investigation of efficient compilation without hashing and, consequently, raise some doubts about the prevailing view that hashing (e.g., universal hashing [5]) is essential to the various aspects of compilation from symbol table management [2] to reduction in strength by hashed temporaries [6].

Aho, Sethi, and Ullman [2] present only two data structures for storing symbol tables - a linear linked list with linear search time and a hash table. They also propose these two data structures for methods to turn an expression tree into a dag and the more general basic block optimization of value numbering. Hashing is involved in preprocessing for global optimizations to perform constant propagation [22], global redundant code elimination [4], and code motion [8]. The best methods of strength reduction [3,6] rely on hashed temporaries to obtain efficient implementations.

There are several reasons why hashing is used in these applications. Hashing has O(1) expected time performance and linear auxiliary space. The method of Universal Hashing, due to Carter and Wegman [5], is especially desirable, since the expected O(1) time is independent of the input distribution. Universal hashing is well suited to applications such as compilation, where the hash tables do not persist beyond a single compilation run. In the applications mentioned above hashing leads to

1. The research of this author was partially supported by National Science Foundation grant CCR-9002428.
2. The research of this author was partially supported by Office of Naval Research Grant No. N00014-90-J-1890.

simple on-line algorithms supporting immediate storage/access. Consequently, various phases of compilation can be carried out incrementally with few passes and with good space utilization.

However, liberal use of hashing incurs certain costs. Even discounting the costs of collisions and rehashing, the calculation of a single hash operation, say $((ax + b) \ mod \ N) \ mod \ m$ for input x and constants a, b belonging to $\{0, ..., N\}$, is much greater that the cost of an array or pointer access. Mairson proved that for any 'minimal' class of universal hash functions there exists a bad input set on which every hash function will not perform much better than binary search [13]. The slow speed of SETL, observed in the SETL implemented ADA-ED compiler, has been attributed to an overuse of hashing. And a hash table implementation involving an array twice the size of the data set is another cost. Arrays lack the benefits offered by linked lists - namely, easy dynamic allocation, dynamic maintenance, and easy integration with other data structures. Finally, although on-line algorithms are vital to incremental compilation, batch processing may otherwise suffice.

In this paper we show that all of the hashed implementations of applications mentioned above can be replaced by algorithms with matching or superior worst case performance. This is achieved by using several simple algorithmic tools (that exclude sorting), the most important of which is multiset discrimination; i.e., finding all duplicate values in a multiset. Multiset discrimination is discussed for various types of elements including pointers, strings, numeric constants, subtrees, and dags. In this paper it is adapted to solve the following problems.

(i) Array or list-based Symbol tables can be formed during lexical scanning with unit-time curser or pointer storage/access.

(ii) Many grammar transformations can be implemented efficiently using multiset string discrimination. In this paper we exhibit a new linear time left factoring transformation based on a generalization of Horner's rule to multisets of strings. Simpler forms of this 'heuristic' transformation were previously studied by Stearns [19] and others (see also [2, 12]) to turn non-LL context free grammars into LL grammars.

(iii) An expression tree-to-dag transformation is implemented without hashing in a simpler way than before and in linear time and space. The numerous applications include one in which any linear pattern matching algorithm (e.g., [10]) can be turned into an efficient nonlinear matching algorithm, where each equality check takes unit-time.

(iv) A new hash-free basic block optimization by value numbering [2, 7] is given, which leads to a faster solution to the program equivalence problem used in integration by Yang, Horwitz, and Reps [23, 24].

(v) Although the main parts of algorithms for global constant propagation [22], global common subexpression detection [4], and code motion [8] do not use hashing, the preprocessing portions for each of these algorithms do. Such hashing can be eliminated without penalty by efficient construction and maintenance of the symbol table.

(vi) We regard the strength reduction transformation presented by Cocke and Kennedy [6] to be the most practical reduction in strength algorithm published in the literature. Although the transformation due to Allen, Cocke, and Kennedy [3] is more powerful (since it can reduce multivariate products) and analyzes control flow more deeply, this algorithm can degrade performance by introducing far too many sums in order to remove nests of products (as was shown in [14]).

We solve three progressively more complex versions of the Cocke and Kennedy transformation without hashing and with superior worst case time and space than the expected performance in previous hash-based solutions[6]. In particular, an algorithm is presented to solve iterated strength reduction folded with useless code elimination in worst case asymptotic time and auxiliary space linear in the maximum text length of the initial and optimized programs.

2. Partial Tool Kit for Algorithm Design Without Hashing

There are many simple combinatorial problems for which hashing seems like the natural, perhaps only, way to obtain an efficient solution. These include such basic computations as:

(i) set union, difference, and intersection;

(ii) multiset string discrimination; i.e., finding all duplicates in a multiset of strings;

(iii) computing $\{\{x,y\}: [x,y] \in S \times T\}$

Although hashing may seem like a panacea, it does incur costs, and we should not overlook the many contexts in which the preceding computations can be solved by an

efficient hash-free approach.

In [15] a different more general discussion of principles underlying hash-free algorithms for simple set operations is presented. Below we discuss a few sharper techniques with a focus on multiset discrimination. Unless otherwise stated, throughout this paper we will assume that sets and multisets are implemented as linked lists.

2.1. Multiset Discrimination of Pointers

We use the following notation for pointer manipulation. If variable x contains a pointer to variable y, expression $x\uparrow$ retrieves the value stored in y, and $\uparrow x$ is a pointer to the value stored in x.

Consider a multiset M of pointers to elements in a set S. For each element (i.e., symbolic address) x in M, we want to compute the set $f(x\uparrow)$ of pointers to all elements in M with the same value as x. (Note here that $x\uparrow$ is the value of the element in S that x points to.) Assume that $f(x\uparrow)$ is initially nil. Multiset M can be partitioned into blocks of duplicates using the following simple procedure

```
F := {}                    -- F will be the set underlying M
(for x ∈ M) -- linear search through M
 if f(x↑) = nil then
  F := F ∪ {x}
 end if
 f(x↑) := f(x↑) ∪ {↑x}
end
```

2.2. Multiset Discrimination of Strings

Solving multiset discrimination of strings is slightly more complicated. Let M be a multiset of n variable-length strings over a k-symbol alphabet $\Sigma = \{1, ..., k\}$. Assume for convenience that each string ends with a sentinel symbol 0.

Starting with an initial partition P containing only one block M, we can solve this problem by repeatedly splitting blocks of P until all the duplicates are found. For each string s we implement a *current position* where the *current symbol* for s is stored. Initially the current symbol for each string is the symbol in its first position. A block is a set of pointers to strings in M. Once we know that a block contains all duplicate strings we say that the block is *finished*; otherwise, we say the block is *unfinished*. Partition P contains two parts - a set of unfinished blocks initially containing the single block M, and a set of finished blocks initially empty.

A partition refinement of P can be implemented using a primitive operation *split*(B) that replaces block B by new blocks, each containing pointers to strings with the same current symbol. The technique implements a variant of multiset discrimination of pointers that makes use of an array of $k + 1$ buckets, where the i-th bucket contains the new block with current symbol i. Any new block found in the 0 bucket or containing only one pointer to a string is finished; otherwise it is unfinished. During execution of *split*(B), the current position is incremented in each string belonging to a new unfinished block. It is easy to implement *split*(B) in time $O(|B|)$ and space $O(|B| + k)$.

We can also implement *split*(B) exclusively with lists and list processing. For each symbol $i = 0, 1, ..., k$ form a special list called the i-list with no elements. Strings are represented as lists of pointers to i-lists. Buckets can then be formed using these i-lists instead of arrays, and multiset discrimination can be solved for pointers as in the preceding subsection.

The following algorithm makes use of *split*(B) to solve multiset discrimination of strings:

1. Form the initial partition $P = \{M\}$.
2. Repeat Step 3 until all of the blocks in P are finished.
3. Scan the set of unfinished blocks, and replace each such block B by *split*(B).

The algorithm runs in $O(m')$ time and $O(n+k)$ space, where m' is the total length of the prefixes needed to distinguish the strings in M. Both the theoretical time bounds and the simplicity of the implementation make it superior to lexicographic sorting for solving multiset discrimination.

Previously, the lexicographic sorting algorithm found in Aho, Hopcroft, and Ullman's book [1] was used to solve congruence closure [9] and also tree isomorphism [1]. Both these problems can be solved more simply by our solutions to multiset discrimination of strings. Their sorting algorithm has the theoretical disadvantage of an $\Theta(m)$ complexity in time and auxiliary space, where m is the total length of all strings in M. Their algorithm also has the practical disadvantage of a complex multi-pass implementation.

The array-based version of multiset discrimination of strings was used earlier by Paige and Tarjan to obtain improved solutions to lexicographic sorting [16] and DFA minimization for one symbol alphabets [17]. The implementation that uses pure list processing without

arrays has the advantage of easier memory management. Both implementations are simple, and involve one pass through the prefixes of the strings. Consequently, our proposed applications may be practical.

2.3. Multiset Discrimination of Numeric Constants

Multiset discrimination of numeric constants can be solved by treating these constants as strings over a k-bit alphabet for arbitrary k. By treating character strings as bit strings, we can also vary k to obtain space/time tradeoffs in solving string discrimination.

3. Applications

3.1. Symbol Tables

Multiset discrimination of strings can be used directly to implement a two-pass lexical scanner. First the string is scanned to produce tokens and initial pointers to symbol table entries. The symbol table is a multiset of lexemes (implemented as a doubly linked list), and an additional pass is needed to remove redundant entries, and redirect pointers (the lexical values) to distinct entries in the modified table. The performance is linear time and space in the length of the input string. Consequently, parsing and semantic analysis can proceed without hashing, since these processes can store and access the symbol table using pointers.

3.2. Fast Left Factoring

Left factoring is a context free grammar transformation that replaces productions of the form $A \rightarrow \alpha\beta_i$ $i = 1, .., n$ by productions $A \rightarrow \alpha\ C$ and $C \rightarrow \beta_i$ $i = 1, ..., n$, where C is a new nonterminal added to the grammar. It was studied by Stearns [19] and others [2, 12] as a tool for turning non-LL grammars into LL grammars. They did not describe optimal forms of factoring or algorithmic details. By making an analogy with Horner's rule for polynomials, we can define a left factoring transformation so that the new grammar has a minimal number of new productions and cannot be further factored.

The algorithm proceeds by repeatedly partitioning the set of grammar productions starting with an initial partition P in which every set of productions with the same left-hand-side nonterminal forms a block. The data structure for P makes use of a set M of pointers to all the right-hand-sides of productions in the grammar. This set can be obtained by multiset discrimination of all the right-hand-side strings. Each block is represented by a subset $B \subseteq M$, a nonterminal symbol A, and an interval $[i, j]$, where the substring from the i-th to j-th symbol of every string in B must be the same. Initially, for each nonterminal A, P has a block containing the set of pointers to all right-hand-sides *rhs* such that $A \rightarrow rhs$ is a production. A is the grammar symbol for this block, and [1,0] is the interval, which represents the empty string λ.

In presenting the algorithm below we use the notation $s_{i..j}$ to denote the substring of string s from the i-th to the j-th symbol; $s_{i..}$ denotes the substring of s from the i-th symbol to the last symbol of s. After initialization the algorithm computes the new grammar G as described below:

```
G := {}
( while P ≠ {})
 remove block [B, [i,j], A] from P
--
-- find a left factor for strings in block B starting
-- from the ith position
--
 j := j +1
 split (B)
 (case 1: B was not refined into more than one block
          by split (B))
--
-- part of a nonempty left factor is found for strings
-- in B
--
  if sentinel 0 is reached then
--
-- every string in B is identically the left factor
-- rhs_{i..j-1}
--
        G := G ∪ {A → rhs_{i..j-1}}
  else
--
-- rhs_{i..j} is part of the left factor yet to be found
--
        add [B, [i,j], A] to P
  end if
 end case 1
 (case 2: B was refined into more than one block by
          split (B))
--
-- complete left factor is found for strings in B
--
  if i = j then
--
-- left factor is the empty string, and B must be an
-- initial block
--
    C := A   -- C represents nonterminal A
  else
--
-- nonempty left factor is rhs_{i..j-1}
--
```

```
    create new nonterminal C
    G := G ∪ {A → rhs_{i..j-1} C}
   end if
   (for each new block D that results from split(B))
--
-- nothing more to factor
--
    if it is a 0 block then G := G ∪ {C → λ }
--
-- D contains just a single string, which is the trivial
-- factor rhs_{j..}
--
    elseif |D | = 1 then G := G ∪ {C → rhs_{j...} }
--
-- try to find the left factor for strings in D
--
    elseif |D | > 1, then add [D,[j,j],C] to P
   end if
  end for
  end case 2
end while
```

THEOREM 1. *The preceding algorithm is correct and runs in linear time in the sum of the grammar symbols contained in all the input productions.*

3.3. Multiset Discrimination of Trees and Applications

Suppose we have a forest of syntax trees produced by syntactic analysis. Suppose also that the nodes of the syntax tree contain pointers to symbol table entries for function symbols, constants, and variables. There are various applications in which we want to find duplicate subtrees. Multiset subtree discrimination can be solved in a new way without hashing by combining multiset string discrimination with multiset pointer discrimination.

Let T be a forest of n nodes. We identify each subtree rooted in node j by a string of length 1 + the number of children of j and with symbols ranging over the alphabet $\{1, ..., n\}$. First we solve multiset pointer discrimination on the symbol table pointers in all the nodes of T. Next, we assign successive integers, called *local* numbers, starting with 1 to the distinct pointers of T. The local number at each node j will be the initial symbol of the string that identifies j.

To obtain the remaining symbols of the subtree identifier, we exploit the idea that subtrees at different heights must be distinct. This allows us to solve multiset subtree discrimination separately for all nodes of the same height bottom-up starting from the leaves to the tree height d. That is,

(1) Solve multiset string discrimination for the leaves, and identify each distinct local number with new numbers, called *value numbers*, with successive values starting with 1.

(2) For height $i = 2, 3, ..., d$, repeat steps 3 and 4:

(3) Identify each node j at height i with a string formed from the local number of j followed by the value numbers of the children of j.

(4) Solve multiset string discrimination on the strings described in step 3. This solves the multiset subtree discrimination problem at height i. Then identify each distinct subtree at height i with new successive value numbers, starting from the last value number assigned to a subtree at height i-1.

The preceding algorithm requires $O(n)$ time and space and is a great deal simpler than the previous best algorithm based on lexicographic sorting. It can be used to obtain new hash-free solutions to many applications including tree-to-dag compression, turning an arbitrary linear tree pattern matching algorithm [10] into a non-linear matching algorithm [18], deciding structural equivalence of type denotations [2], and many more.

3.4. Multiset Dag Discrimination and Acyclic Coarsest Partitioning

The solution to multiset tree discrimination extends without modification to solve multiset discrimination for dags with m edges and n nodes in time $O(m)$ and space $O(n)$. Recall that this space bound improves the $O(m)$ space bound that could be obtained to solve this problem using Aho, Hopcroft, and Ullman's lexicographic sorting algorithm [1]. We show how multiset dag discrimination can be used to obtain an improved solution to acyclic instances of the many-function coarsest partition problem.

The many-function coarsest partition problem, used by Hopcroft to model the problem of DFA minimization [11], has applications in program optimization and program integration. It can be formulated as follows. Given a directed multi-graph $(V, E_1, ..., E_k)$ (where V is the set of vertices, and $E_1, ..., E_k$ are sets of edges), and an initial partition $P = \{V_1, ..., V_s\}$ of V, find a coarsest refinement P' of P such that for each block C in P' and each $i = 1, ..., k$, there exists a block C_0 in P' such that the image set $E_i[C] \subseteq C_0$, where $E_i[C] = \{y: [x, y] \in E_i$ and $x \in C\}$. Here we assume that for each $i = 1, ..., k$, the outdegree of each vertex $v \in V$ in (V, E_i) is at most 1.

An algorithm was given in[11] that solves this problem in time $\Theta(k|V|\log|V|)$ and space $\Theta(k|V|)$ in the worst case, which is true even when the graph $(V, E_1 \cup ... \cup E_k)$ is acyclic. However, when the graph $(V, E_1 \cup ... \cup E_k)$ is acyclic, we can solve the problem in time and space $O(k|V|)$ using a solution to multiset discrimination for dags.

THEOREM 2. *If $(V, E_1 \cup ... \cup E_k)$ is acyclic, then the many function coarset partition problem can be solved by Hopcroft's algorithm in time $\Theta(k|V|\log|V|)$ and space $\Theta(k|V|)$ in the worst case, and by multiset dag discrimination in time $O(k|V|)$ and space $O(|V|)$.*

3.5. The Sequence Congruence Problem

The sequence congruence problem [23,24] arises in the context of program integration. It asks how to partition program components into classes whose members have equivalent execution behaviors. The algorithm presented in [23,24] solves this problem in two phases: the program components are first partitioned *w.r.t* the flow dependence graph, and then refined *w.r.t.* the control graph. Hopcroft's coarsest partition algorithm is used in both phases, giving the $O(m_1 \log m_1 + m_2 \log m_2)$ time complexity, where m_1 and m_2 are the sizes of the flow dependence graph and control graph respectively. Since their control graph is essentially acyclic, the linear time multiset dag discrimination method can be used for the second phase to improve their time bound to $O(m_1 \log m_1 + m_2)$.

3.6. Value Numbering Without Hashing

Value numbering is a standard optimization technique of determining equalities within basic blocks to avoid redundant computations[2,7]. Although the technique is mostly implemented with hashing, multiset discrimination can be used to obtain an efficient implementation without hashing.

Consider a basic block B consisting of a sequence of assignment statements $s_1, ..., s_k$, each of the form *lhs* := *rhs*, where *lhs* is a variable, and *rhs* is either a constant, a variable, or an expression of the form x *op* y. Here, *op* is some binary operator, and x, y can be constants or variables. Assume that B is lexically scanned, and that variables and constants are represented by pointers to a symbol table as described previously. We want to assign an integer (i.e. a value number) to each occurrence of an expression in the statements of B so that if two occurrences of expressions have the same value number, then they must have the same run-time value. We compute value numbers in three steps as follows:

1. Construct an initial dag representation $D = (V, E_1, E_2)$ of B, where the vertices of D represent the values of subexpressions. Each vertex has a label representing an approximate initial value. Leaves are associated with variables and constants, and are labeled by pointers to symbol table entries. Each internal vertex is associated with a right-hand-side *rhs* of the form x *op* y, and is labeled *op*.

We construct D by scanning the statements in B in order from s_1 to s_k. During this scan, the vertex *node*(v), representing the current dag vertex for variable or constant v, is accessed by pointer from the symbol table entry for v. Initially *node*(v) is nil.

When scanning statement *lhs* := *rhs*, for each right-hand-side argument x in which *node*(x) is nil, assign a new vertex to *node*(x) labeled by a pointer to the symbol table entry for x. If *rhs* is a variable or constant y, then *rhs* is represented by vertex v=*node*(y). Otherwise, if *rhs* is of the form x *op* y, then create a new vertex v labeled *op* and having two children *node*(x) and *node*(y), representing the most recent value of x and y respectively. The edge $[v,$ *node*$(x)]$ belongs to E_1, and the edge $[v,$ *node*$(y)]$ belongs to E_2. Finally, assign v to *node*(*lhs*), so that *lhs* and *rhs* are represented by the same vertex. Two different vertices of D may have the same label and children during this step, but they will be merged in the final step.

2. Propagate constants in D bottom up. If a vertex w in D is labeled *op* and the labels for its two children point to constants c_1 and c_2, then label w with a pointer to the computed value of c_1 *op* c_2. Then delete the edges leaving w from E_1 and E_2.

3. Compute the value numbers. We first partition V based on the labels of its elements. Call the resulting partition P_0. We solve the coarsest partition problem with the input (V, E_1, E_2) and P_0 by multiset dag discrimination to get a final partition $P = \{B_1, ..., B_t\}$ of V. Then for $i = 1, ..., t$, we assign integer i to each element of B_i as its value number.

Each of the above steps can be done in linear time without hashing. Note that greater accuracy can be achieved by using explicit constant labels (instead of pointers to constants), folding step 3 with step 2, and performing multiset discrimination of all the constant labels before performing step 3.

4. Reduction in Strength

The final three examples use the preceding techniques to obtain new solutions to strength reduction with worst case performance asymptotically better than the expected performance of the previous best algorithms. Ironically, the efficiency obtained seems to stem from using batch techniques to implement strength reduction, which itself uses incremental techniques to improve program performance.

4.1. Basic Strength Reduction

First we consider a new hash-free algorithm that implements Cocke and Kennedy's strength reduction transformation [6]. The algorithm runs in worst case time/space linear in the length of the final program text, which, as we shall show, can be as much as two orders of magnitude better than their hash-based algorithm. Like their algorithm we are careful not to compute the potentially costly data flow relation.

Cocke and Kennedy's transformation is concerned with replacing hidden costs of linear polynomials involved in the array access formula used in programming languages like Fortran or Algol. As was suggested by Allen, Cocke, and Kennedy [3], the earlier transformation [6] can be improved by sharper analysis of control flow and taking safety of code motion into account. However, such improvement is orthogonal to the solution presented here.

The strength reduction transformation of [6] may be defined as follows. Let L be a strongly connected region of code. We assume that this code consists of assignments to simple variables of the form $z{:=}op(x,y)$ or $z{:=}op(x)$ and conditional branches with boolean valued variables as predicates. We assume implicit assignment to certain designated input variables, and implicit output variables that are printed whenever they are assigned a new value. All concern for control flow is simplified by taking a most conservative position that L forms a clique; i.e., that every two statements in L can be executed one after the other.

If c is either a region constant variable of L or a constant, and if i is a variable that is defined in L, then product $i{\times}c$ is *reducible* if all definitions to i occurring in L are among the following forms: $i := j$, $i := -j$, $i := j + k$, $i := j - k$, $i := -j + k$, or $i := -j - k$, where in each such form $j{\times}c$ and $k{\times}c$ must also be reducible. For each reducible product $i{\times}c$ occurring in L, strength reduction transforms L as follows:

(i) Replace each occurrence of $i{\times}c$ in L by a new variable t_{ic} uniquely associated with text expression $i{\times}c$.

(ii) If variable i is live on entry to L, then introduce assignment $t_{ic} := i{\times}c$ in a unique entry block (a detail we add to their transformation for correctness), which must be entered before entering L.

(iii) Within L and just prior to each definition to i of the forms either $i := \pm j$ or $i := \pm j \pm k$, insert the code $t_{ic} := \pm j{\times}c$ or $t_{ic} := \pm j{\times}c \pm k{\times}c$ respectively.

(iv) If any of the products introduced in step (iii) has been previously eliminated by either code motion or strength reduction, replace it by its associated temporary variable. Remove all other products introduced in step (iii) by either code motion or recursive application of strength reduction as appropriate.

Like Cocke and Kennedy we assume that strength reduction is performed after redundant code elimination, constant propagation, and code motion. Given a strongly connected program region L as input, our solution shares the first four steps of the Cocke and Kennedy algorithm; i.e.,

(i) Compute the set RC of region constant variables of L and a set $Defs(v)$ of all definitions in L to each variable v defined in L.

(ii) Compute the set IV of *induction variables*; that is, the set of all variables x with definitions occurring in L such that any product $x{\times}c$ would be reducible. This procedure was also described by Cocke and Schwartz [7].

(iii) Find the set *Cands* of all reducible products $x{\times}c$ actually appearing in L, and the associated places where they occur.

(iv) For each induction variable x, compute the set $Afct(x) = \{x\} \cup \{y$: y is a variable or constant on the right-hand-side of any assignment to x in $L\}$.

The preceding steps can be performed in worst case time and space linear in the program text. If $Afct$ is regarded as a binary relation and $Afct^*$ represents its transitive closure, then the following fact immediately follows from Cocke and Kennedy's paper.

LEMMA 1. *The set of all expressions removed from L by strength reduction is defined by $Rm = \{j{\times}c$: $i{\times}c \in Cands$, $j \in Afct^*(i)\}$.*

Calculation of Rm is central to the implementation of strength reduction, and it is important to observe three

sources of redundancy in computing this set naively.

(i) when $i \times c$ and $j \times c$ belong to *Cands* and $Afct^*(i) \cap Afct^*(j)$ is nonempty;

(ii) when $i \times c_1$ and $j \times c_2$ belong to *Cands*, $c_1 \in Afct^*(j)$, and $c_2 \in Afct^*(i)$;

(iii) when two different products of constants evaluate to the same constant

Because only the first source of redundancy can lead to an asymptotic blowup in time and space, we avoid it during the calculation of *Rm*. Because the other two sources of redundancy only contribute constant factors in complexity, we avoid them during a postpass cleanup. Our approach combines multiset discrimination with data structuring techniques.

It is at this point that our solution differs from Cocke and Kennedy. They go on to compute the transitive closure $Afct^*$ in time $\Theta(n^3+m)$ using, say, Warshall's algorithm[21] (see also[1]), where n is the number of variables and constants contained in *Afct*, and m is the number of assignments to induction variables. They also use a greedy strategy committed to hashing each product removed by strength reduction. In contrast, we compute the strong component decomposition of *Afct* inverse (i.e., we consider decomposition of a graph with directed edge $i \rightarrow j$ iff $i \in Afct(j)$) in $\Theta(m)$ time and space using Tarjan's algorithm [20]. The dag structure *Scd* of this decomposition is used to efficiently compute *Rm* in time O(final text length). The algorithm rests on the following obvious fact:

LEMMA 2. *Let* $Cs = \{c: i \times c \in Cands\}$. *For each* $c \in Cs$ *let* $Cmps(c)$ *be the set of strong components containing some variable* i *for which* $i \times c \in Cands$. *If* c *is any region constant variable or constant, then the set of all expressions* $j \times c$ *removed by strength reduction is defined by* $Rm(c) = \{j \times c: j$ *belongs to a component of Scd from which there is a path in Scd to any component of* $Cmps(c)\}$.

The remaining steps of the algorithm are given just below:

(v) Compute the set *Cs* using multiset pointer discrimination. At the same time, for each constant $c \in Cs$, form a set of pointers to strong components $Cmps(c)$ as described in Lemma 2, and mark variables v within these components such that $v \times c$ belongs to *Cands*.

(vi) Initialize an empty multiset *Mrc* of subtrees and an empty multiset *Mc* of numeric constants. For each constant $c \in Cs$ repeat steps (vii) and (viii)

(vii) Compute the set $Scd_c = \{v: v \times c \in Rm(c)\}$ using a depth-first-search through dag *Scd* in the reverse direction of its edges and starting from components belonging to $Cmps(c)$. Observe that for each strong component of *Scd*, if it has no edges leading in, then its entries are constants or region constant variables; otherwise, its entries are induction variables. Link each induction variable $v \in Scd_c$ to a new symbol table entry containing unique identifier t_{vc}, and insert assignment $t_{vc} := v \times c$ on entry to L if v is live on entry to L. If v is marked, indicating that $v \times c \in Cands$, then replace each occurrence of $v \times c$ in L by a pointer to the symbol table entry for t_{vc}. Link each region constant variable $v \in Scd_c$ to a new entry in *Mrc* containing subtree $v \times c$. For each constant $c' \in Scd_c$, if c is a region constant variable, then link c' to a new entry in *Mrc* containing subtree $c \times c'$; otherwise, link c' to a new entry in *Mc* containing the computed value of $c \times c'$.

(viii) For each induction variable v in Scd_c and each assignment to v in $Defs(v)$, introduce update code to t_{vc} according to the definition of the strength reduction transformation described earlier. Replace products that are introduced within this update code by references to the symbol table, *Mrc*, or *Mc* as is indicated by the links in Scd_c.

(ix) Use multiset subtree discrimination and constant discrimination to find duplicate region constant expressions and constants in *Mrc* and *Mc*, and augment the symbol table with new variables for each distinct item in *Mrc* and *Mc*. At the same time readjust pointers inside L to the symbol table, and insert an assignment $t_{c_1c_2} := c_1 \times c_2$ on entry to L for each product $c_1 \times c_1 \in Mrc$.

THEOREM 3. *The preceding algorithm is correct and has worst case time and space O(length of the final program text).*

4.2. Strength Reduction With Cleanup

Cocke and Kennedy noted that after strength reduction is applied, it is necessary to apply global cleanup transformations such as useless code elimination (i.e., elimination of statements not contributing to the output) and variable subsumption (i.e., eliminating useless copy operations). In this section we show how to fold useless code elimination together with strength reduction. Our hash-free solution runs in worst case time and space linear in the sum of the lengths of the initial

and final program texts.

As before we assume that L is a strongly connected region of code, and $Defs(v)$ is a set of all definitions in L to each variable v defined in L. Instead of computing *Cands* directly, we compute the set *Prods* of all products appearing in L and the places where they occur. Also, the set *IV* of induction variables is not computed explicitly, but is detected implicitly in a simpler way.

By a *spoiler* we mean any variable v for which $Defs(v)$ contains a definition not amongst the forms $v := \pm j$ or $v := \pm j \pm k$. We compute the set *Spoilers* of all such variables. Finally, we generalize relation *Afct* so that $Afct(x)$ is defined for each variable x (and not just induction variables) that is assigned within L. Let *Afcti* denote *Afct* inverse. As before we compute the strong component decomposition dag *Scd* of *Afcti*.

Recall that those single node components of *Scd* with no edges leading in contain only constants and region constant variables. Also, any product $x \times c \in Prods$ is reducible (i.e., belongs to *Cands*) iff there is no path in *Scd* from a spoiler to x. If we mark all strong components containing spoilers, and mark all other components reachable from these marked components, then the unmarked portion of *Scd* corresponds precisely to the data structure at the heart of the strength reduction algorithm in the preceding subsection. Recall that the induction variables are all those variables contained in unmarked strong components with edges leading in.

Consequently, we can proceed to solve strength reduction starting with step (v) of the previous algorithm. We now have an alternative linear time strength reduction algorithm, where the first four steps of Cocke and Kennedy's solution are simplified. This new algorithm can also be extended to support efficient analysis for useless code.

Consider how the new strong component dag Scd_{new} of the program loop L after strength reduction differs from the initial dag Scd_{old}.

LEMMA 3. *i. The subdag of Scd induced by unmarked strong components and the subdag of Scd induced by marked strong components are both invariant with respect to strength reduction. ii. The only new components in* Scd_{new} *are ones containing only new temporaries; the only edges incident to these components are between them and from them to marked components. iii. Edges only go from unmarked to marked components, and these can only be deleted by strength reduction.*

Proof Strength Reduction alters loop L in the following ways:

i Assignments are introduced within L to modify compiler-generated temporaries t_{xc}. The right-hand-side of any such assignment must contain only compiler-generated temporaries. Hence, these assignments cannot create new edges from Scd_{old} to any strong components in Scd_{new} containing compiler-generated variables.

ii An assignment $z := x \times c$ appearing in L can be replaced by assignment $z := t_{xc}$. In this case, variable z must be a spoiler that belongs to a marked component $Scd_z \in Scd_{old}$, and x must appear in an unmarked component $Scd_x \in Scd_{old}$. Moreover, there must be an edge from Scd_x to Scd_z. After replacement, there would be an edge from the strong component in Scd_{new} containing t_{xc} to Scd_z. If the edge count between Scd_x and Scd_z after replacement becomes zero, indicating no assignments in L from a right-hand-side variable in Scd_x to a variable in Scd_z, then this edge is deleted in Scd_{new}. ■

Let *inputs* be the set of input variables, *outputs* be the set of output variables, and *controls* be the set of predicate variables of control statements. We will assume that these variables are all *useful*, and that the strong components of *Scd* containing them, which we call the critical set *crit*, are also useful. The useful components include *crit* and all strong components of *Scd* that can reach the components in *crit*.

If we assume that all statements in L are initially useful, then after strength reduction is applied to L once, only induction variables, region constants, and constants can become useless. Temporaries generated by strength reduction must all be useful. Consequently, only the replacement of products by temporaries can create useless code. And all statements that undergo such replacement will be useful in the end.

Hence, we can modify steps (vii)-(ix) of the algorithm in the previous subsection to facilitate useless code elimination as follows. In step (vii), for each assignment $z := v \times c$ replaced by assignment $z := t_{vc}$, decrement the edge count from Scd_v to Scd_z. If the edge count reaches zero, then delete the edge from Scd_v to Scd_z. This is implemented using a pointer linking a record for assignment $z := v \times c$ into the adjacency list for *Scd*. Also, add edges from t_{vc} to z in *Afcti*. In step (viii) introduce a new edge in *Afcti* for each assignment to a temporary introduced. In step (ix) multiset discrimination will determine the new vertices corresponding to new temporaries in *Afcti*. Add a final step (x) in which the useful components of *Scd* are computed. Within L all assign-

ments to variables not in useful components can be removed.

By the preceding discussion we have

THEOREM 3. *The preceding algorithm is correct and has worst case time and space O(length of the initial plus final program text).*

4.3. Iterated Strength Reduction

Cocke and Kennedy noted that after strength reduction is applied, the new compiler generated variables t_{vc} and other variables can become new induction variables, and new products defined in terms of these variables can be removed by further applications of strength reduction[6]. In this section we show how iterated strength reduction folded with useless code elimination can be solved in worst case time and space linear in the maximum length of the initial and final program texts.

Note, first of all, that iterated strength reduction terminates, because each iteration except the last must eliminate at least one product in the original strongly connected region L. In order to achieve the promised linear time complexity, we must be careful to generate only temporaries that are not useless. We will also exploit the observation that the portion of graph *Afcti* involving only temporaries generated by strength reduction is an exact copy of the core part of *Afcti* involving original program variables only. Consequently, a key idea underlying the new algorithm will be to avoid altering the original graph and its strong component decomposition *Scd*. By attaching labels to the components of *Scd*, we will be able to interpret these components as being formed from either original program variables or from temporaries generated in some *i*th application of strength reduction.

Following Cocke and Schwartz[7], we say that a temporary $t_{vc_1 \cdots c_j}$ is *available* in program region L if, whenever it is referenced during execution of L, it stores the value of $v \times c_1 \times \ldots \times c_j$. By default, $t_v = v$, and t_v is said to be available in L if v is not useless. The main task of the algorithm is to determine all the temporaries that need to be kept available in L in order to eliminate all candidate products. It is then straightforward to generate the code to keep them available.

First consider the preprocessing. For iterated strength reduction we relax our definition of spoiler to be any variable z in which there is an assignment to z in L that is not amongst the forms $v := \pm j$, $v := \pm j \pm k$, or $v := j \times c$, where $j \times c$ is a candidate product.

Compute *Scd* as in the preceding subsection. We say that a component of *Scd* is *clean* if it has no spoilers; it is *reducible* if all its ancestors in *Scd* are clean. The set *Cands* consists of all those products $x \times c$ occurring in L, such that x belongs to a reducible component. It is straightforward to compute the spoilers, the reducible components, and *Cands* in a single topological search through *Scd* in the direction of its edges. For each component C, if it has no spoilers and if its predecessors are all reducible, then it is reducible; otherwise it is not reducible. The search continues until no more reducible components can be detected.

Next consider analysis for all temporaries needed to remove the products from *Cands*. We have the following observations:

LEMMA 4. *Let C be a component in Scd, $v \in C$, and $s = c_1 \cdots c_j$ for $j \geq 0$. (When $j=0$, then s is the empty string, denoted by λ.)*

1. If $v_1 \in C$, then temporary t_{vs} is kept available iff $t_{v_1 s}$ is kept available.

2. If $s \neq \lambda$, then t_{vs} is kept available iff either of the following conditions hold: i. there exists a successor $C_1 \in$ Scd of C and an assignment $u := d_1 \times c_1$ in L such that $u \in C_1$, $d_1 \in C$, and $t_{uc_2 \cdots c_j}$ is kept available; or ii. there exists a successor $C_1 \in$ Scd of C and an assignment $u := d_1$ op d_2 in L such that op is not multiplication, $u \in C_1$, $d_1 \in C$, and t_{us} is kept available.

Define *labelc* (C) to be the set of strings s such that for all $v \in C$, t_{vs} is kept available. To determine all temporaries to be kept available, we only need to compute the set *labelc* (C) for all $C \in$ *Scd*. For this purpose, we also associate a set of strings *labele* with each edge $e = [C_1, C_2]$ in *Scd* as follows: for each assignment $v := d_1 \times d_2$ of a candidate product in L such that $v \in C_2$ and $d_1 \in C_1$, then d_2 is in *labele* (e); if e is associated with any other kind of assignment, then λ is in *labele* (e).

According to Lemma 4, we can compute the mapping *labelc* as following:

1. For each nonreducible component C, set *labelc* $(C) = \{ \lambda \}$.

2. Let *ready* be the set of minimal reducible components;

3. While *ready* is not empty, select an arbitrary component C and delete it from *ready*; then do steps 4, 5, 6;

4. Let *succ* (C) be the set of successors of C in *Scd*. Set *labelc* $(C) = \bigcup_{C_i \in succ(C)} \{ x\|s: x \in labele([C,C_i]), s \in labelc(C_i) \}$.

5. If C contains output variables or control variables, add λ to *labelc* (C).

6. Mark C *processed*. Put the predecessors of C whose successors are all processed into *ready*.

In order to keep temporaries available, following actions are added to step 4:

4'. Just before each assignment $u := d_1\ op\ d_2$ in L such that $u \in C$, insert the code $t_{us} := t_{d_1 s}\ op\ t_{d_2 s}$.

Although multiset string discrimination could be used in computing the union in step 4, the $\Omega(|s|)$ worst case cost contributed by each string s in *lablec* (C) is too slow. More efficient is to modify the preceding algorithm to generate all strings of a given length before applying multiset discrimination. We will compute *labelc* by generating these strings in order of ascending length.

Initially, *labelc* (C) is empty for all reducible components C. Then for length $i = 0, 1, ..., k$, we generate the strings of length i in *labelc* (C) for all reducible components C. When $i=0$ useful program variables are detected. We assume no new strings are generated in round k. After each round i, $i = 1, ..., k-1$, we assign a unique identifier for each distinct string of length i. Thus, in round $i+1$, each newly generated string $c_1 c_2 \cdots c_{i+1}$ can be represented by a pair $[c_1, n_1]$, where n_1 is the name for $c_2 \cdots c_{i+1}$. Consequently, in order to determine distinct strings generated in each ith round $i>1$, multiset string discrimination is only needed for strings of length 2.

It is easy to compute strings of length $i+1$ from strings of length i using the following two rules:

1. If *lablec* (C) contains a string s of length i, and if for some predecesssor C_i of C, *labele*$([C_i, C])$ contains c, then $c\|s$ is a string of length $i+1$ that must be contained in *labelc* (C_i);

2. If *lablec* (C) contains a string s of length $i+1$, and if for some predecesssor C_i of C, *labele*$([C_i, C])$ contains λ, then s must also be contained in *labelc* (C_i).

The above representation of strings can also be used to initialize constants and temporaries. If $s = c_1 \cdots c_k$ is a new string in *labelc* (C) for some $k>1$ and $C \in Scd$, and if n_1 is the name of $c_2 \cdots c_k$, then we use t_s to store the value of $c_1 \times ... \times c_k$, and insert an assignment $t_s := c_1 \times t_{n_1}$ at the end of the initialization block. Once all the constants are initialized, we insert an assignment $t_{vs} := v \times t_s$ at the end of the initialization block for each temporary t_{vs}.

After round 1, when all temporaries for strings of length 1 are determined, replace all occurrences of candidate products in L by their associated temporaries. Because the algorithm processes components *Scd* in the reverse direction of edges, only useful variables and temporaries are introduced to the final program. Consequently, we have

THEOREM 4. *The iterated strength reduction problem with useless code elimination can be solved in time and auxiliary space linear in the maximum length of the initial and final program texts.*

4.4. Extensions

Two possible approaches that exploit commutative and associative laws of products may reduce the number of strings, and therefore temporaries, generated in the preceding strength reduction algorithms. One approach is to use a weak form of the Paige/Tarjan lexicographic sorting algorithm[16] to generate strings of constants in some arbitrarily chosen order. Another more effective, but less efficient, approach, would be to actually compute the product of constants identifying each temporary, and to use multiset constant discrimination.

We are currently investigating these ideas as well as extensions that implement a more powerful transformation integrating strength reduction of sums, products, quotients, exponentiations, and multivariate expressions. Such extensions would allow different kinds of spoilers for different arguments of candidate expressions. Development of simpler hash-based algorithms is another promising direction.

5. Conclusion

We have suggested hash-free methods for solving various aspects of optimizing compilation. These methods have been based in large part on efficient algorithms for solving multiset discrimination for different datatypes. Multiset discrimination of ordered flow graphs, unordered trees and dags, and unordered graphs with respect to given depth-first-search spanning trees are straightforward. An empirical investigation comparing our hash-free alternatives with their conventional hash-based counterparts would be worthwhile future work.

6. Acknowledgements

We are grateful to Alan Siegel, whose independent investigation of lexicographic sorting and great interest in its applications to algorithm design provided motivation for our work. We thank Bob Tarjan for describing a list based data structure known to Knuth for implementing fast string matching, which is related to our list based implementation of multiset descrimination. We also thank Ralph Wachter, whose workshop on randomized algorithms brought to our attention questions about randomized versus deterministic algorithms, which, we felt, raised similar questions about hash-based versus hash-free algorithms, and led to the current paper.

References

1. Aho, A., Hopcroft, J., and Ullman, J., *Design and Analysis of Computer Algorithms,* Addison-Wesley, 1974.
2. Aho, A., Sethi, R. and Ullman, J., *Compilers,* Addison-Wesley, 1986.
3. Allen, F. E., Cocke, J., and Kennedy, K., "Reduction of Operator Strength," in *Program Flow Analysis,* ed. Muchnick, S. and Jones, N., pp. 79-101, Prentice Hall, 1981.
4. Alpern, B., Wegman, N., and Zadeck, K., "Detecting Equality of Variables in Programs," in *Proc. 15th ACM POPL*, Jan, 1988.
5. Carter, J. and Wegman, M., "Universal Classes of Hash Functions," *JCSS*, vol. 18, no. 2, pp. 143-154, 1979.
6. Cocke, J. and Kennedy, K., "An Algorithm for Reduction of Operator Strength," *CACM*, vol. 20, no. 11, pp. 850-856, Nov., 1977.
7. Cocke, J. and Schwartz, J. T., *Programming Languages and Their Compilers,* Lecture Notes, CIMS, New York University, 1969.
8. Cytron, R., Lowry, A., and Zadeck, K., "Code Motion of Control Structures in High-level Languages," IBM Research Center/Yorktown Heights, 1985.
9. Downey, P., Sethi, R., and Tarjan, R., "Variations on the Common Subexpression Problem," *JACM*, vol. 27, no. 4, pp. 758-771, Oct., 1980.
10. Hoffmann, C. and O'Donnell, J., "Pattern Matching in Trees," *JACM*, vol. 29, no. 1, pp. 68-95, Jan, 1982.
11. Hopcroft, J., "An n log n Algorithm for Minimizing States in a Finite Automaton," in *Theory of Machines and Computations*, ed. Kohavi and Paz, pp. 189-196, Academic Press, New York, 1971.
12. Lewis, F., Rosencrantz, D., and Stearns, R., *Compiler Design Theory,* Addison-Wesley , 1976.
13. Mairson, H., "The Program Complexity of Searching a Table," in *24th IEEE FOCS*, pp. 40-47, Nov., 1983.
14. Paige, R., "Symbolic Finite Differencing - Part I," in *Proc. ESOP 90*, ed. N. Jones, Lecture Notes in Computer Science, vol. 432, Springer-Verlag, 1990.
15. Paige, R., "Real-time Simulation of a Set Machine on a RAM," in *ICCI '89*, ed. W. Koczkodaj, Computing and Information, Vol II, pp. 69-73, 1989.
16. Paige, R and Tarjan, R., "Three Efficient Algorithms Based on Partition Refinement," *SIAM Journal on Computing*, vol. 16, no. 6, Dec., 1987.
17. Paige, R., Tarjan, R., and Bonic, R., "A Linear Time Solution to the Single Function Coarsest Partition Problem," *TCS*, vol. 40, no. 1, pp. 67-84, Sep, 1985.
18. Pelegri-Llopart, E., *Rewrite Systems, Pattern Matching, and Code Generation,* U. of CA - Berkeley, 1987. Ph.D. Dissertation
19. Stearns, R., "Deterministic top-down parsing," in *Proc. 5th Princeton Conf. on Information Sciences and Systems*, pp. 182-188, 1971.
20. Tarjan, R., "Depth first search and linear graph algorithms," *SIAM J. Comput*, vol. 1, no. 2, pp. 146-160, 1972.
21. Warshall, S., "A Theorem on Boolean matrices," *JACM*, vol. 9, no. 1, pp. 11-12, 1962.
22. Wegman, M. N. and Zadeck, F. K., "Constant Propagation with Conditional Branches," in *Proc. 12th ACM POPL*, Jan, 1985.
23. Yang, W., Horwitz, S., and Reps, T., "Detecting program components with equivalent behaviors," TR-840, Computer Sciences Dept., Univ. of Wisconsin, Madison, WI, April 1989.
24. Yang, W., "A new algorithm for semantics-based program integration," Ph.D. Dissertation, TR 962, Computer Sciences Dept., Univ. of Wisconsin, Madison, WI, August 1990.