Fast Firewall Implementations for Software and Hardware-based Routers

Lili Qiu
liliq@microsoft.com
Microsoft Research

George Varghese
varghese@cs.ucsd.edu
University of California, San Diego

Subhash Suri [*]
suri@cs.ucsb.edu
University of California, Santa Barbara

Abstract

Routers must perform packet classification at high speeds to efficiently implement functions such as firewalls and diffserv. Classification can be based on an arbitrary number of fields in the packet header. Performing classification quickly on an arbitrary number of fields is known to be difficult, and has poor worst-case complexity.

In this paper, we re-examine two basic mechanisms that have been dismissed in the literature as being too inefficient: backtracking search and set pruning tries. We find using real databases that the time for backtracking search is much better than the worst-case bound; instead of $\Omega((logN)^{k-1})$, the search time is only roughly twice the optimal search time[1]*. Similarly, we find that set pruning tries (using a DAG optimization) have much better storage costs than the worst-case bound. We also propose several new techniques to further improve the two basic mechanisms. Our major ideas are (i) backtracking search on a small memory budget, (ii) a novel compression algorithm, (iii) pipelining the search, (iv) the ability to trade-off smoothly between backtracking and set pruning. We quantify the performance gain of each technique using real databases. We show that on real firewall databases our schemes, with the accompanying optimizations, are close to optimal in time and storage.*

Keywords: Firewall, packet classification, compression, pipeline, selective push, trade-off.

1. Introduction

Packet classification is a key mechanism that enables differentiation in a connectionless network. In packet classification routers, the route and resources allocated to a packet can be determined by the destination address as well as other header fields of the packet such as the source address and TCP/UDP port numbers. More importantly (at least today), many router products allow a firewall capability, such as Cisco Access Control Lists (ACLs), which allow packets to be blocked based on the same fields. Edge routers need ACLs to implement firewalls. However, even large backbone routers today implement ACLs to trace denial-of-service and flood attacks. Thus our paper concentrates on techniques for speeding up packet classification for firewalls using properties we have observed in real firewall databases. [2]

The current state of the art in most routers is to either use linear search of the filter database or to use hardware, such as ternary CAMs (content addressable memory) or other ASICs that perform parallel linear search (e.g., [4]). Such hardware solutions do not scale to large filter databases. Other solutions reported in literature that can be implemented in software (e.g., [9, 5]) are either slow or take too much storage. With the advent of software based routers (e.g., [7]), which are typically aimed at the edge router space where classification is particularly important, it is necessary to find software techniques for fast firewall implementations.

There is evidence that the general filter problem is a hard problem, and requires either $O(N^k)$ memory or $\Omega((\log N)^{K-1})$ search time, where N is the number of filters and K is the number of classified fields [4, 9]. However recent research [5, 6, 10] indicates that such worst-case behavior does not arise in real databases. Based on this observation, these papers introduce clever *new* techniques like pruned tuple search [10] and Recursive Flow Classification [5] that exploit the structure of existing databases. However, if real databases have regularities that can be exploited, perhaps even the simplest packet classification algorithms will do quite well.

This question motivates us to re-examine two of the simplest packet classification mechanisms: *backtracking search* and *set pruning tries*. There is an interesting duality between these two simple schemes: backtracking search re-

[*] The work of Subhash Suri was partially supported by National Science Foundation grants ANI 9813723 and CCR-9901958.

[1] The height of the multiplane trie is regarded as the optimal search time throughout the paper, unless otherwise specified.

[2] While we believe our techniques generalize to other filter databases such as diffserv, it is difficult to test this assertion because there are no models of diffserv databases that are generally agreed upon.

quires the least storage but can have poor worst-case search times; set pruning tries have minimal search times but have poor worst-case storage. Earlier researchers have dismissed backtracking search as being too slow [10], and dismissed set pruning tries as being suitable only for very small packet classifiers [2, 10].

However, we find that using real databases the time for backtracking search is much better than the worst-case bound. Instead of $(\log N)^{k-1}$, the search time is only a constant factor (often only a factor of two) worse than the optimal. Similarly, we find that set pruning tries (using a DAG optimization) have much better storage costs than the worst-case bound of N^k indicates.

We also propose several novel techniques to further improve the performance of backtracking and set pruning tries. First, most designers assume that backtracking is infeasible to implement in hardware because of its high computational state (large number of registers) and its inability to be pipelined (without having memory replicated for each pipeline stage). We show, perhaps surprisingly, that both these assumptions are false. We introduce a new backtracking search technique, which reduces the hardware register cost for doing backtracking in hardware to $k + 1$, where k is the number of dimensions. We also show that while some memory must be replicated at different stages, one can pipeline real firewall databases with a very small amount of replicated memory.

Second, we design a novel compression algorithm that applies to any multiplane trie. Our results indicate that compression reduces the lookup time by a factor of 2 - 5, and reduces the storage by a factor of 4 - 12.

Finally, given that backtracking search and set pruning tries are at two ends of a spectrum between the optimal storage and the optimal time, it makes sense to study the tradeoff between these two extremes. As the two schemes are structurally similar and use multiplane tries as their underlying basis, we show that it is possible to smoothly tradeoff storage for time using a new mechanism called selective pushing. Our results show that the tradeoff scheme offers more choices, and can improve the time of backtracking search with only modest increase in storage.

The paper is organized as follows. We give the problem definition in Section 2, and review related work in Section 3. In Section 4 we describe backtracking search and introduce some simple new optimizations to improve search time. We then evaluate its performance, and quantify the effects of each of the optimizations using the real firewall databases. In Section 5 we describe set pruning search, introduce some optimizations to improve storage, and present our experimental results. In Section 6 we propose a novel compression scheme, and evaluate its performance gain both in theory and with experiments. In Section 7, we examine pipelining the search, and evaluate its storage cost. In Section 8 we explore a tradeoff between time and space by starting with backtracking search and using selective pushing. We conclude in Section 9.

2. Problem Specification

Packet classification is performed using a packet classifier, which is a collection of filters (or rules in firewall terminology). Each filter specifies a class of packet headers based on some criterion on K fields of the packet header. Each filter has an associated directive, which specifies how to forward the packet matching this filter. We say that a packet P matches a filter F if each field of P matches the corresponding field of F. This can be either an exact match, a prefix match, or a range match. Since we can represent a range using multiple prefixes [10], we assume for the rest of the paper that each field in a rule is a prefix unless otherwise specified. [3]

Since a packet can match multiple filters in the database, we associate a cost for each filter to determine an unambiguous match. Thus each filter F in the database is associated with a non-negative number, $cost(F)$. Our goal is to find the least cost filter matching a packet's header. The key metric is classification speed. It is also important to reduce the size of the data structure to allow it to fit into high speed memory. The time to add or delete filters is often ignored in existing work, but can be important for dynamic filters.

3. Related Work

Many router vendors do a linear search of the filter database for each packet, which scales poorly with the number of filters. To improve the lookup time, some vendors cache the result of the search keyed against the whole header. Caching may work well and have high hit rates [11] but still requires a fast packet classification scheme to handle the $10 - 20\%$ cache misses. A hardware-only algorithm could employ a ternary CAM (content addressable memory). However ternary CAMs are still fairly small, inflexible and consume a lot of power.

[4] describes a scheme optimized for implementation in hardware. It works well for up to 8000 filters and should scale further with further hardware improvements. However, it requires specialized hardware. [10] proposes two solutions for multi-dimensional packet classification: grid-of-tries and crossproducting. The former scheme decomposes the multidimensional problem into several 2-dimensional planes, and uses a data structure, grid-of-tries, to solve the 2-dimensional problem. Crossproducting is more general

[3] There can be a large increase in the number of rules during range-to-prefix conversion. Our new compression algorithm will address this issue as shown in Section 6.1.

but either requires $O(N^k)$ memory or requires a caching scheme with non-deterministic performance.

[5] proposes a simple multi-stage classification algorithm, called recursive flow classification (RFC). It has much better (but still large) storage for real databases than cross-producting. [9] suggests searching through combinations of field lengths (tuples) and also suggest a heuristic of first doing prefix searches on the individual fields to prune the set of tuples to be searched. [6] suggests another heuristic based on geometrically partitioning the classification space, which produces fairly good search times and requires less memory than [5].

[3] considers a tradeoff between lookup time and storage cost. However, their experimental results are only for 2-dimensional databases, and it is not clear how the algorithm would perform on real higher-dimensional databases. We describe a completely different algorithm that trades storage for lookup time in Section 8 and evaluate its performance on 5-dimensional real databases.

4. Revisiting Backtracking Search

Given that real firewall filter databases contain considerable structure, in this section we revisit a simple algorithm: backtracking search. We start by reviewing the basic mechanism, and then show how it can be augmented by simple optimizations.

4.1. Basic Backtracking Search

A trie is a binary branching tree with each branch labeled 0 or 1. The prefix associated with a node u is the concatenation of all the bits from the root to the node u. It is straight-forward to extend a single dimensional trie to multiple dimensions [10]. We first build a trie on the prefixes of the first field, $\textit{Field}_1$. Each valid prefix in the $\textit{Field}_1$ trie points to a trie containing $\textit{Field}_2$ prefixes (that follow $Field_1$ prefixes in some filter). Similarly for the other dimensions.

Throughout the paper, the storage cost is assessed as the total number of nodes in the trie, and the lookup time is assessed as the total number of nodes visited during the search, which corresponds to the total number of memory references during the search. Since each filter is stored exactly once, the memory requirement for the structure is $O(NW)$, where N is the total number of filters, and W is the maximum number of bits specified in any of the three dimensions.

Backtracking search is essentially a depth first traversal of the tree which visits all the nodes satisfying the given constraints. For example, suppose we search for a header that matches $[D3, S2, P1]$ in the trie shown in Figure 1. We first traverse the D trie, and remember all the matches $D3$, $D2$, and $D1$. Then we traverse the S trie for $D3$ and remember $S1$ and $S2$. Next we traverse the P trie for $S2$. After reaching $P1$ (assume no match beyond $P1$), we do our first backtrack, and start exploring the P trie for $S1$. Once we reach $P1$, we do another backtrack, and start walking the S trie for $D2$. Finally, we explore the S trie for $D1$. In general, the amount we have to remember can be $W * K$, where W is the average number of matches per dimensions, and K is the number of dimensions.

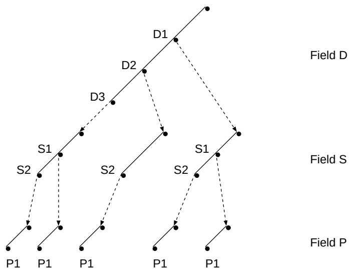


Figure 1. Backtracking search on a small memory budget.

The lookup time of backtracking search can be as large as $\theta(W^K)$, where K is the number of dimension. The application of switch pointers, introduced in [10], can help to avoid backtracking in the last two dimensions. This reduces the worst-case lookup time to $O(W^{K-1})$. Thus in the 2-dimensional case, the lookup time is $O(W)$.

4.2. Backtracking Search Optimizations

Before we introduce our major optimization ideas, we start by describing some simple (but new) optimizations for backtracking search: considering backtracking on a small memory budget, optimal field ordering, pruning based on cost, and generalizing switch pointers [10].

Backtracking using limited computation state: Backtracking is considered expensive (especially in hardware where the backtracking state has to be kept in limited register memory) because of the potential need to keep state for all potential backtrack points. We show that for a tree that does backtracking in multiple dimensions, we only need state proportional to the number of dimensions (e.g., 6 registers for the common case of IP 5-tuples).

As described in Section 4.1, using standard backtracking search we need to keep $W * K$ pieces of state. We now describe a solution that takes $K + 1$ pieces of state. The key modification is that instead of doing depth-first traversal while remembering all past backtracking points, we go to the first trail we see, and only then return for more exploration. In other words, we visit more general filters first, and more specific filters later.

In the previous example, we will explore the D trie until we reach $D1$. Since we see a pointer to a S trie here, we immediately follow it. When we reach $S1$, we see another trail to follow to P trie, we start walking P trie until we reach $P1$. When we come to the end of that, we have to do our first backtrack. We start walking the S trie for $D1$ further till we hit $S2$. Finally, after walking the P trie for $S2$, we once again try to walk the S trie for $D1$, and fail (no matches beyond $S2$ in the S trie). Then we back up and try to walk the D trie further. Thus the order of filters visited is $(D1, S1, P1)$, $(D1, S2, P1)$, $(D2, S2, P1)$, $(D3, S1, P1)$, and $(D3, S2, P1)$. The standard method visits them exactly in reverse order.

While the standard scheme does not keep a lot of state in this simple example, it can in the worst-case. In contrast, the modified backtracking search only takes $K + 1$ pieces of state, since we only need to keep track of the current dimension being explored, the pointers to the roots of the tries (in all dimensions) on the current search path, and the bit positions in the header to start the trie search. Since bit positions only take 5 bits and pointers are less than 27 bits, we can easily keep the state information in a set of $K + 1$ 32-bit registers.

Optimal Field Ordering: We have observed that the lookup time in backtracking search is sensitive to the ordering of fields in the trie. We evaluated the effect of different orderings on the query lookup time using real firewall databases, and find that using the best ordering improves the lookup time by 10% to 65%. Moreover, there are about half the orderings whose lookup times are within 10% different from that of the best ordering. This suggests one way to find a good ordering is to randomly try a few orderings, and pick the best one. A further enhancement is to use some heuristic (e.g. choose the orderings that place the field with the most prefix containment [4] as the last field) to pre-select some good orderings, and then randomly try some of the selected orderings. Other heuristics were suggested in [6, 3].

Pruning Based on Cost: The basic idea behind pruned backtracking is as follows. During the backtracking search, if we encounter a trie node such that the tree beneath the node does not contain any lower cost filter than the current match, then we do not need to search through the trie below that node. Refer to [8] for more details. (Pruning based on cost is used in other filter algorithms such as [5, 1].).

Switch Pointers: We further optimize backtracking search with switch pointers. Switch pointers were introduced in [10] but the technique is limited to 2-dimensional packet classification. We extend switch pointers to higher-dimensional packet classification by using it over the last 2 fields, or by avoiding the first backtracking, whichever is more beneficial. We omit the details for lack of space.

Database	# Rules	# Rules in the prefix format
Database 1	67	127
Database 2	158	418
Database 3	183	531
Database 4	279	949
Database 5	266	1640

Table 1. Firewall databases.

4.3. Performance Evaluation

In this section, we experimentally evaluate backtracking search to quantify the effects of the various optimizations described above. We use the total storage and the worst-case lookup time as our performance metrics. The total storage is computed as the total number of nodes in the multiplane trie. The worst-case lookup time is the total number of memory accesses in the worst-case assuming a 1 bit at a time traversal of each trie. Finding worst-case backtracking search times is non-trivial. Refer to [8] for details.

We use a set of 5 industrial firewall databases that we obtained from various sites for performance evaluation throughout the paper. For privacy reasons, we are not allowed to disclose the names of the companies, or the actual databases. Table 1 shows the number of rules in the databases. The rules in the firewall are specified either as exact match, or as prefix, or as ranges. In order to use a multiplane trie for filter classification, we need to convert all the rules to a prefix format. Rules specified as ranges are converted using the technique of [10]. Column 3 in Table 1 shows the number of rules after converting to the prefix format. The conversion leads to a factor of 2 - 6 increase in the number of rules. Our new compression algorithm, described in Section 6.1, will address this issue.

The databases have the following characteristics:

- *Prefix containments:* In our databases, no prefix contains more than 4 matching prefixes for each dimension. Most prefixes contain 1, 2, or 3 matching prefixes only. We believe our performance results will be applicable for other filter databases that have a similar number of prefix containments.
- *Prefix lengths:* The most popular source/destination prefix lengths are 0 (wildcard) and 32. There are also a number of prefixes with lengths 21, 23, 24, 28, and 30. This is very important for the performance of our compression algorithm, described in Section 6.1.
- *Port ranges:* 5% - 10% of the filters have port fields specified as ≥ 1024 [5]. Such a range is converted into 6 prefixes using [10], which contributes heavily to the increase in the prefix rules. This will be addressed by our compression algorithm, described in Section 6.1.
- *IP addresses:* The destination and source prefix fields have roughly half the rules that are wildcarded.

[4] Prefix containment is the number of prefixes that are prefixes of a given prefix. For example, suppose we have the following prefixes 0*, 00*, 001*, and 000*. Then the prefix containment of 000* is 2, since 0* and 00* are both its prefixes.

[5] This is common because of the convention that the well known ports for standard services such as email etc, use port numbers < 1024.

Database	Trie Depth	# Memory Accesses			Storage (# nodes)		
		BB	PB	SB	BB	PB	SB
1	86	146	119	117	1848	1782	1782
2	102	153	145	143	4914	4914	4914
3	102	169	149	149	3949	3743	3743
4	102	202	170	170	6785	26006	6630
5	102	208	198	196	6555	13724	16493

Table 2. Performance of backtracking search using one 5-dimensional trie, where BB, PB, and SB stand for basic backtracking, pruned backtracking, and pruned backtracking with the extended switch pointer optimization (described in Section 4.2).

4.3.1. Performance Results

Our performance results for backtracking are summarized in Table 2. Note that our results are based on searching one bit at a time. A simple extension is to search multiple bits at a time. Clearly, if we search 4 bits at a a time, then the memory accesses are reduced to $\frac{1}{4}$ of the values reported here, but the storage could increase by a factor of up to 16.

We compare three algorithms: basic backtracking, pruned backtracking, and pruned backtracking with the switch pointer optimization [6]. We list the results for the best ordering in Table 2. (The best ordering for all forms of backtracking search is the ordering that minimizes memory accesses.)

As we can see, the three backtracking search algorithms have small memory requirements. The exact storage requirements are sometimes different because the best ordering for the three schemes is not necessarily the same. For all five databases, even the basic backtracking cost is around twice the height of the trie or less. This is somewhat surprising, since backtracking is usually regarded as too slow for packet classification. Thus for real databases with limited number of prefix containments, we believe backtracking can be affordable in practice. Using say an 8 bits at a time trie traversal, backtracking requires around 18 - 26 memory accesses. This is better or as good as any firewall implementations we know while using much less storage.

5. Revisiting Set Pruning Tries

Set pruning tries were initially proposed in [2] and briefly examined (and then discarded) in [10]. As with backtracking search, set pruning tries work using multiplane tries. Set pruning tries, however, differ from backtracking search tries by fully specifying all search paths so that no backtracking is necessary. However this is done at the cost of increasing storage. In the worst-case, set pruning tries may take up to $O(N^K)$ storage.

We first review some standard terminology, and then explain the process of converting a backtracking search trie to its corresponding set pruning trie.

We say that string S' is a *descendant* of string S if S is a prefix of S'. We say that filter A is a descendent of filter B if for all dimensions $j = \{1, 2, ..., k\}$, string $A(j)$ is a descendant of $B(j)$. (Note that $A(j)$ is allowed to be equal to $B(j)$ and still be a descendant of $B(j)$.)

Converting a backtracking search trie to a set pruning trie is essentially replacing a general filter with its descendent filters. In other words, for every filter F, we "push" F down to all its descendent filters, and then delete F. For instance, in Figure 2, filter [*, *, *] is pushed down to the places corresponding to [0*, 0*, *], [0*, 0*, 0*], [0*,1*, *], [1*, 0*, *], and [1*, 1*, *], and [1*, 1*, 1*], all of which are descendent filters of [*, *, *]. Note that after [*, *, *] is pushed down, the filter that is stored in the node A changes from F_1 to $min(F_1, F_2)$; similarly the filter that is stored in the node B changes from F_3 to $min(F_3, F_2)$. Since [*, *, *] is replaced with all its descendent filters during the push-down step, we can simply delete it. As we can see, the push-down step may potentially lead to memory blow up. In the worst-case, we may need $O(N^K)$ storage for K dimensional filters.

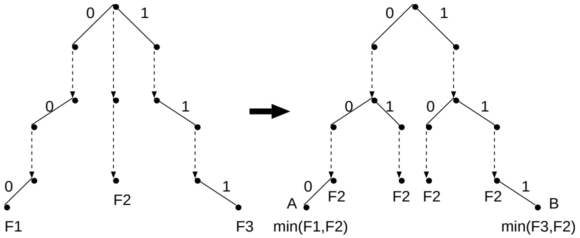


Figure 2. Backtracking search trie versus set pruning trie.

We now consider two optimizations to reduce storage in set pruning tries: the use of DAGs, and the use of optimal field orderings.

Optimizing storage using DAGs: A natural technique to reduce the memory of set pruning tries is to change the *tree* structure of a multiplane set pruning trie to a *Directed Acylic Graph (DAG)*. This was first suggested in [2]. We illustrate the idea using a 3-dimensional trie. As shown in Figure 3, the tries that node A and B (both in the first field) are pointing to are identical. So instead of keeping several copies of identical tries, we only need to keep one copy, and have both nodes point to the same 2-dimensional trie. This is done at all field boundaries.

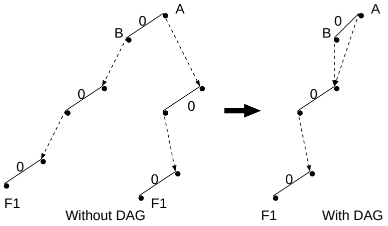


Figure 3. Set pruning trie with the DAG optimization.

Our results show the DAG optimization helps to reduce

[6] Our evaluation is based on the standard backtracking search. The results of modified backtracking search on a small memory budget are the same for basic backtracking, and similar for the other types of backtracking.

Database	# memory accesses	Storage (# nodes)
Database 1	86	5541
Database 2	102	51785
Database 3	102	59180
Database 4	102	123951
Database 5	102	165920

Table 3. Performance of set pruning tries (with the DAG optimization) using one 5-dimensional trie.

memory blow up by two orders of magnitude in the complete set pruning trie. We also experimented with using the DAG optimization at the same bit position *within* a field, but found that the additional saving was insignificant, usually around $5 - 10\%$.

Optimal field ordering: As in backtracking search, we also find that field ordering affects storage requirements significantly. We evaluated the effect of different orderings using the real databases, and found that the best ordering cuts down the storage cost by a factor of 2 - 7. On the other hand, there are a handful of orderings that give comparable performance (within 30% difference) to the best ordering. Therefore we can use similar heuristics as discussed in Section 4.2 to choose a good ordering.

5.1. Experimental Evaluation

In this section, we experimentally evaluate the performance of set pruning trees using the same 5 industrial firewall databases described in Section 4.3.

We list the results for the best ordering in Table 3. (The best ordering for a set pruning trie is the ordering that minimizes storage.) Again our results are based on searching one bit at a time. Compared to several forms of backtracking search as shown in Table 2, set pruning tries provide an optimal number of memory accesses at the cost of large storage requirement. In particular, the storage requirement increases to 3 - 25 times as large as what is minimally required by the corresponding backtracking search trie. Note that the databases 2, 3, 4, and 5 all have the same worst-case lookup time of 102, since this is the maximum number of nodes we can ever visit without backtracking [7].

6. Compression

We further improve backtracking search and set pruning tries using a novel form of compression. A standard compression scheme for tries (e.g., [2]) is to remove all single branching paths so that no node has more than one child). Figure 4 shows an example: the algorithm compresses the trie on the left into the one on the right by collapsing multiple nodes into a single node when multiple edges succeed each other without any branching. For the performance of this standard compression algorithm, we have the following theorem.

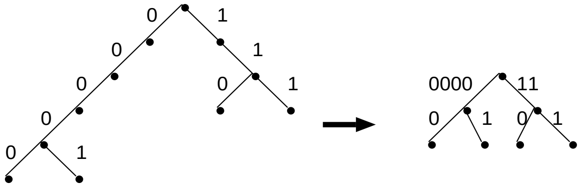


Figure 4. A standard compression algorithm: merge a single branch into one trie node.

Theorem 6.1 *Consider a 1-dimensional trie with N leaf nodes, and only leaf nodes are associated with filters. After compression, it has $2N - 1$ nodes.*

6.1. General Compression Algorithm

The standard compression scheme is efficient when there is no *redundancy* in the trie nodes. A trie has redundancy when many trie nodes have the same pointer value: either pointing to the same node in the next dimensional trie, or pointing to the same filter. Such redundancy especially arises when filters specified in ranges are converted into those specified in prefixes, or when a more general filter is pushed down to several more specific filters. The standard compression scheme cannot exploit such redundancy. For example, it fails to compress anything for the trie shown in Figure 5(a), since no node in the trie has only a single branch.

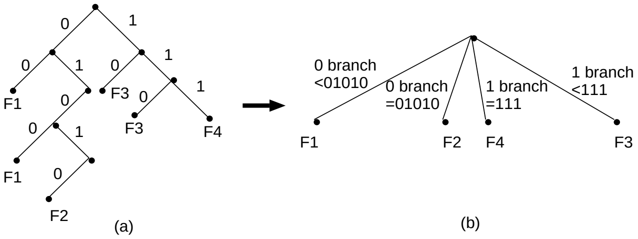


Figure 5. Our new compression algorithm.

However, a closer examination of the trie in Figure 5(a) reveals an interesting property: all nodes on the left of path 01010* (and also on the branches starting with 0) point to filter 1. If we can use range comparison as well as an equality test, we can compress all the branches starting with 0 by creating one *center branch* pointing to filter 2 with value 01010, and one *side branch* pointing to filter 1 with value < 01010. Similarly, we can compress all the branches starting with 1 by creating one *center branch* pointing to filter 4 with value 111, and one *side branch* pointing to filter 3 with value < 111. This leads to a more compact trie as shown on Figure 5(b).

To generalize the above example, if a path AB satisfies the following property, called the *Compressible Property*:

[7] We have 32 bits each for source/destination address, and 16 bits each for source/destination port. In each database there are less than 16 different protocols, so we use 4 bits to distinguish them. Except the protocol trie, which we search 4 bits at a time, all the other tries are 1-bit at a time. So the maximum number of nodes we can visit without backtracking is 102 (including 5 roots in each dimension).

(i) all nodes on its left point to the same place L (either the same filter or the same node in the trie), and (ii) all nodes on its right point to the same place R, then we can compress it as follows. Let $\delta(AB)$ denote the string labeling the path from node A to node B. We compress the entire branches by creating three edges: one *center branch* with value $\delta(AB)$ pointing to what B used to point to, and one *side branch* with value $< \delta(AB)$ pointing to L, and another *side branch* with value $> \delta(AB)$ pointing to R.

To simplify the following discussion, we use the data structure shown in Figure 6 to represent a compressed node. We add 3 fields to the original uncompressed trie node: $value$, len, and $rangePtr$ as shown in Figure 6. Since some of the elements may be empty, in practice we can have variable size trie nodes which are just large enough to hold the non-empty elements.

```
#define MAX_CHILD 2 // for binary trie
struct NODE {
    struct NODE * Child[MAX_CHILD]; // center branch
    long value[MAX_CHILD];
    uchar len[MAX_CHILD];
    struct NODE * rangePtr[2*MAX_CHILD]; // side branch
    ... //other data members used in uncompressed trie node
}
```

Figure 6. Data structure for compressed node.

The way we use the above data structure is as follows. If the current input is 0, we check to see if it matches $value[0]$ after taking the appropriate bit mask. If so, we follow the pointer to $Child[0]$. Otherwise, we follow its $rangePtr[0]$ if it is less than $value[0]$, or $rangePtr[1]$ if it is larger than $value[0]$. Similarly for input 1.

We now examine the details of our general compression scheme. We need only consider the 1-dimensional case, since compressing a higher-dimensional trie can be achieved by compressing one dimension at a time.

A trie node has up to *MAX_CHILD* paths. Each of these paths can be compressed independently. (This is the same as the standard compression scheme.) So we only need to solve the problem of compressing one path. The basic idea is that at each node we try to decide whether the next immediate step to take can be compressed out. The next step is compressible if and only if it satisfies the invariant that all the nodes on the left of the center path (i.e. the path that will be converted to a center branch) have the same pointer value, and all the nodes on the right of the center path have the same pointer value. Below we refer to the invariant as the *compressible invariant*. To decide if the compressible invariant is maintained, we need to look at the characteristics of the node's $Child$. In particular, we classify a node into the following categories:

1. It has only one child, and none of its children are internal nodes (i.e. nodes whose $Child$ are not empty).
2. It has more than one child, and none of its children are internal nodes.
3. Exactly one of its children is an internal node;
4. It has more than one child which are internal nodes.

It is clear that we cannot compress the nodes in case 4, since the compressed path can retain the information of either of the paths, but not both. For the other cases, we need to check further to make sure the *compressible invariant* holds. This involves two steps: (i) finding the center path, and (ii) verifying the invariant. The main issue is the first step, finding the center path. The second step is straightforward once the center path is given. The center path is easy to identify in cases 1 and 3, since there is only one path we can possibly take. In case 1, the center path is the branch between the current node and its non-empty child node; in case 3, the center path is the branch between the current node and the child node that is an internal node. In case 2, there is more than one candidate, and we pick one that satisfies the compressible invariant. We prove the correctness of the algorithm, and analyze its performance in [8]. In particular, we have the following theorem:

Theorem 6.2 *For a trie representing N points in the range, we can compress it to $2N-1$ nodes using the general compression algorithm.*

6.2. Experimental results for Compression

In this section, we evaluate our new compression algorithm applied to both backtracking search and set pruning tries using the same five databases shown in Table 1.

We list the performance results for the best ordering (defined in Section 4.3.1 and Section 5.1). Compared with the performance results before compression, as shown in Table 2 and Table 4, it is evident that compression improves performance significantly. More specifically, compression reduces memory accesses and storage cost by a factor of 2 - 5 for backtracking search with and without cost based pruning. (We haven't implemented compression with switch pointers, but we expect similar performance gain as with the other types of backtracking search.) [8] For the set pruning trie, compression cuts down storage by a factor of 4 - 12, and cuts down lookup time by a factor of 1.6 - 4.

It is interesting to note that with compression, the lookup time of backtracking search is close to that of set pruning tries, and sometimes even better. This is because the high compression ratio of the paths in the backtracking search trie offsets the cost of a small number of backtracks. Refer to [8] for more details.

[8] Since the compressed trie nodes are bigger than standard trie nodes, each access to a trie node should strictly be charged twice the number of memory accesses shown (to access the value and then follow the pointer). However, since most processors prefetch a whole cache line, the second access should be essentially zero cost as long as the node fits in a cache line. We use the same assumption for linear search.

Database	# memory accesses				Storage (# nodes)			
	Set Pruning	BB	PB	Linear Search	Set Pruning	BB	PB	Linear Search
Database 1	22	30 (21)	28 (21)	67	1510	429	429	67
Database 2	57	51 (34)	39 (34)	158	5778	912	912	158
Database 3	59	49 (36)	43 (45)	183	6914	1261	1666	183
Database 4	64	98 (58)	85 (58)	279	30322	2951	2951	279
Database 5	55	59 (29)	58 (29)	266	13556	2815	2815	266

Table 4. Performance of compressing backtracking search and set pruning tries (with the DAG optimization) using one 5-dimensional trie, where BB and PB stand for basic backtracking and pruned backtracking (described in Section 4.2). The numbers in the parentheses are the height of the compressed trie. Sometimes the height is larger than the memory accesses of pruned backtracking, since some nodes can not be reached after pruning. For linear search, we assume each filter fits in a cache line, and accessing a filter rule takes 1 memory reference.

We also compare the performance of linear search with backtracking search and set pruning tries in Table 4. As we would expect, both backtracking search trie and linear search have low storage cost. On the other hand, the lookup time of linear search is 2 - 5 times as large as what is required by backtracking search or set pruning tries. If we use multi-bit tries, the performance of backtracking search and set pruning tries can be even better. [9]

Furthermore, the lookup time of linear search increases linearly with the size of databases. In contrast, the lookup time of set pruning tries is constant, at most the maximum number of bits in the header fields; the lookup time of backtracking search is a constant factor of what is required by the set pruning tries for databases with limited prefix containments. Therefore both backtracking search and set pruning tries have more scalable lookup time than linear search.

A few comments follow. First, although the storage cost of a compressed node is larger than an uncompressed node (40% larger in our implementation), the total storage is actually much smaller after the compression, since compression cuts down the number of nodes by a factor of 2 - 12. Second, search on our compressed tries involves both equality test and range comparison. Without special optimization, the CPU time spent per node almost doubles after compression [10]. However, the total CPU time is actually less because compression cuts down the number of nodes visited by more than 2 times. Moreover, the lookup time is dominated by the number of memory accesses, which is much less after the compression. Therefore the performance gain of compression is almost the amount of the reduction in memory accesses, and the additional CPU overhead with compression is negligible.

To summarize, we described a new compression algorithm that can reduce the storage cost by a factor of 4 - 12, and reduce lookup time by a factor of 2 - 5.

7. Pipelining Backtracking

So far all the algorithms in this paper could be implemented in either hardware or software. However, hardware implementors often wish to use extra hardware gates to improve speed by using parallelism. A natural way to speed up an algorithm that has sequential dependencies (such as backtracking) is to use pipelining.

First, we assume the use of our new backtracking scheme that uses limited computation state; this is crucial because it reduces the amount of *register* memory passed between pipeline stages. However, we also have to worry about the amount of main memory (especially if each pipeline stage requires its own memory, and the memory is fast SRAM). SRAM, which corresponds to cache memory in CPUs, is expensive and must be minimized for a feasible solution.

A simple way to pipeline the backtracking search is to store the entire backtracking search trie at every pipelining stage. During the backtracking search, we pass the node to be visited after the previous pipelining stage to the next pipelining stage, so that the search in the next pipelining stage will start from that node. In this case, the main memory storage requirement increases linearly with the number of pipelining stages. However since not all nodes in the backtracking trie will be visited in every pipelining stage, a natural enhancement to the previous approach is to have the pipelining stage i store only the trie nodes that will be visited in the stage i. Refer to [8] for the details of how to partition the backtracking trie into different pipelining stages.

Figure 7 shows the main memory storage requirement as a function of the total number of pipelining stages for both uncompressed and compressed tries [11]. As we can see, the storage increases only moderately with the number of pipelining stages in both cases.

Similarly, we can also apply pipelining to set pruning tries for improvement in search time. The total storage requirement in this case is equal to the size of the original set pruning tries, since every node is stored at exactly one pipelining stage.

[9] We are working on using multi-bit lookup on the compressed trie. We expect using k bits at a time, the speedup in the lookup time of backtracking search will be close to, but smaller than, a factor of k.

[10] The performance result is based on looking up one header filter 1000000 times on an otherwise idle UNIX machine. We conducted the experiments for different header filters and using different databases, and the performance is similar.

[11] Our evaluation is based on the modified backtracking search on a small register memory budget. Similar results are observed for the standard backtracking search.

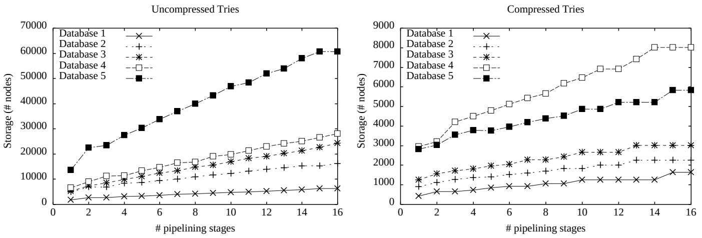


Figure 7. Storage requirement for pipeline.

8. Trading Storage for Time

In the previous section, we showed that real databases contain significant structure, and that simple mechanisms like backtracking search and set pruning tries can perform much better than the worst case bounds, especially using several optimization techniques (in particular, generalized compression) we proposed. These two algorithms are at two extremes: one has small storage requirement with suboptimal lookup times, and the other offers good lookup time at the expense of a suboptimal memory requirement. Ideally we would like to have a smooth tradeoff between the two extremes. That is, if we can afford larger memory, we would like to have correspondingly better lookup times. Similarly if the lookup time requirement is relatively large, we would like to be able to use cheap machines (with less memory) to do filter search. Such a *tunable algorithm* would give designers more choices and flexibility.

Note that, with compression, backtracking search may sometimes have better lookup time than set pruning tries. In this case, we can still use the same technique, described below, to tradeoff memory for lookup time. The only difference is that compressed set pruning tries are no longer in the tradeoff region, for they neither yield the best storage requirement nor the best lookup time.

8.1. Selective Pushing

As we have seen earlier in Section 5, set pruning tries eliminate all backtracking by “pushing” down *all* filters to its descendents. So searching for a filter in a set pruning trie is simply searching for the longest matching prefix in every dimension. There is no need to do any backtracking. An important observation is that eliminating all backtracking can be potentially storage intensive. If we are willing to afford a small amount of backtracking, however, we can “push” down fewer filters, which can reduce storage requirements. Below we describe a heuristic called *selective pushing* that decides which filters to push down.

The basic idea is that we only “push down” the filters with high worst-case backtracking times, and leave the other filters intact. The code in Figure 8 shows the skeleton of the selective push algorithm. Basically it computes the search cost for each header class, where a header class is defined as a set of headers that follow the same path in backtracking search. If the search time for the header class (itself represented as a filter) exceeds our required time bound, then we insert the filter into the trie. Note that the header class filters may be different from the original filters in the database.

```
foreach header class (represented as Filter(i))
    cost = BacktrackSearch(trie,header);
    if ($cost > bound$)
        insertFilter(trie,Filter(i));
        annotate the leaf so that search can stop at this leaf
end
```

Figure 8. Find worst-case search time.

After filter F is inserted, then our search for F will be exactly the same as in a set pruning trie: we simply search for the longest prefix match in each dimension, and the leaf will be the matching filter; no backtracking is necessary. Therefore we must annotate the leaf of the pushed down filter to indicate that there is no need for backtracking search after encountering this leaf.

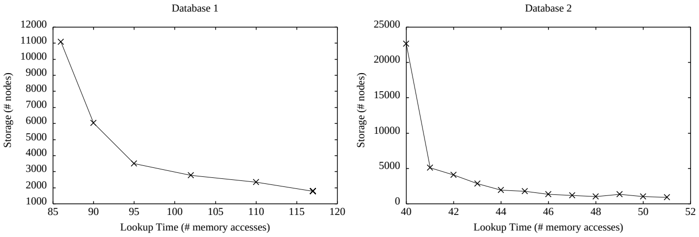
Pushing down a filter makes search time for that particular filter $O(KW)$, where K is the number of fields, and W is the maximum length of any field. However, as a side effect, adding some more paths to the trie (during the push-down) may make searching for some other filters longer. Therefore we need to *iteratively* push down: we first get rid of the longest path; if this push-down produces new long paths, then we need to get rid of these as well. The algorithm stops either when the worst-case lookup time is below the required time bound (specified as an input), or the memory grows to the size of a set pruning trie, corresponding to the state where all filters get pushed down.

Selective pushing can be applied to both uncompressed and compressed tries. The side effects of applying selective pushing for the compressed trie are two fold: (i) adding more branches may increase the search time for other filters, and (ii) adding more branches may reduce the compression ratio, which may in turn increase search time for a large number of filters. Therefore we need to be more conservative when applying selective pushing to a compressed trie. Our experiments suggest that a better heuristic in this case is to push the filters with largest search time in the current iteration, and do it iteratively.

8.2. Performance Results of Selective Pushing

We evaluate the performance of selective pushing applied to both uncompressed and compressed tries using the same 5 databases. For the following evaluation, we use the best field orderings as defined in Section 4.3 and Section 5.1.

For the uncompressed trie, we apply selective pushing on basic backtracking augmented with cost-based pruning and extended switch pointers, described in Section 4.2. The results are shown in Figure 9(a). As we can see, we can reduce the query lookup time with little increase in storage when the lookup time is large. For example, for database 1, the lookup time is reduced from 117 to 95 with moderate increase in storage. Further decrease in the lookup time is achieved at higher cost in storage after we reach the "knee" of the curve. Similar behavior is observed for the other databases. Also for all databases, when the lookup time comes close to the optimal, the storage costs saturate at those of the corresponding set pruning trie, as we would expect.



(a) an uncompressed trie (b) a compressed trie

Figure 9. Selective pushing applied to both uncompressed and compressed tries.

For compressed tries, we evaluate the performance of selective pushing on basic backtracking search. (We haven't implemented it for other types of backtracking, but we believe the performance should be similar if not better.) The results are shown in Figure 9(b). As before, when the lookup time is large, it drops rapidly at very small cost in storage. For instance, in database 2, we reduce the lookup time from 51 to 44 with little extra storage. Further decrease in lookup time incurs higher cost in storage after we reach the "knee" of the curve. Similar behavior is observed for the other databases.

To conclude, in this section we use selective pushing to smoothly tradeoff storage for lookup time and find it useful to improve backtracking search times by around 10-20% with only a small increase in storage.

9. Conclusion

This paper has four contributions. First, we showed experimentally that the performance of simple trie based filter schemes is much better than worst-case figures predict. We also improved the basic trie based filter schemes with a number of optimizations, including a new backtracking search technique, which requires only $D + 1$ pieces of state, where D is the number of dimensions. Second, we proposed a novel compression algorithm that further reduces the lookup time and storage cost. Third, we investigated pipelining the search, and found the storage cost for pipelining increases only moderately with the number of pipelining stages. Finally, we introduced a simple mechanism for trading memory to improve the search time of backtracking search.

Despite the fact that the storage numbers for optimized set pruning tries are reasonable, our final message is that backtracking (with compression, selective pushing, and possible hardware support) offers a more reasonable time-space tradeoff than set pruning tries, and should be seriously considered by router implementors.

10. Acknowledgments

We would like to thank Geoff Voelker, Pankaj Gupta, and anonymous reviewers for their helpful comments.

References

[1] M. M. Buddhikot, S. Suri, and M. Waldvogel. Space Decomposition Techniques for Fast Layer-4 Switching. IFIP Sixth International Workshop on Protocols For High-Speed Networks, Aug. 1999.

[2] D. Decasper, Z. Dittia, G. Parulkar, B. Plattner. Router Plugins: A Software Architecture for Next Generation Routers. In *Proc. of SIGCOMM'98*, 1998.

[3] A. Feldmann and S. Muthukrishnan. Tradeoffs for Packet Clasification. In *Proc. INFOCOM'2000*, March 2000.

[4] T. V. Lakshman and D. Stidialis. High Speed Policy-based Packet Forwarding Using Efficient Multi-dimensional Range Matching. In *Proc. of SIGCOMM'98*, Sept. 1998.

[5] P. Gupta and N. McKeown. Packet Classification on Multiple Fields. In *Proc. of SIGCOMM'99*, Sept. 1999.

[6] P. Gupta and N. McKeown. Packet Classification using Hierarchical Intelligent Cuttings, Proceedings Hot Interconnects VII, Aug. 1999.

[7] R. Morris, E. Kohler, J. Jannotti, M. F. Kaashoek. The Click Modular Router. In *17th ACM Symposium on Operating Systems Principles*, Dec. 1999.

[8] L. Qiu, G. Varghese, and S. Suri. Fast Firewall Implementations for Software and Hardware-based Routers. Microsoft Research Technical Report MSR-2001-61. June 2001.

[9] V. Srinivasan, S. Suri, and G. Varghese. Packet Classification using Tuple Space Search. In *Proc. of SIGCOMM'99*, Sept. 1999.

[10] V. Srinivasan, G. Varghese, S. Suri, and M. Waldvogel. Fast Scalable Level Four Switching. In *Proc. of SIGCOMM'98*, Sept. 1998.

[11] J. Xu, M. Singhal, and J. Degroat. A Novel Cache Architecture to Support Layer-Four Packet Classification at Memory Access Speeds. In Proc of INFOCOM'2000, March 2000.