Towards a Software Product Line of Trie-Based Collections

Michael J. Steindorfer
Centrum Wiskunde & Informatica, The Netherlands
Michael.Steindorfer@cwi.nl

Jurgen J. Vinju
Centrum Wiskunde & Informatica, The Netherlands
Jurgen.Vinju@cwi.nl

Abstract

Collection data structures in standard libraries of programming languages are designed to excel for the average case by carefully balancing memory footprint and runtime performance. These implicit design decisions and hard-coded trade-offs do constrain users from using an optimal variant for a given problem. Although a wide range of specialized collections is available for the Java Virtual Machine (JVM), they introduce yet another dependency and complicate user adoption by requiring specific Application Program Interfaces (APIs) incompatible with the standard library.

A product line for collection data structures would relieve library designers from optimizing for the general case. Furthermore, a product line allows evolving the potentially large code base of a collection family efficiently. The challenge is to find a small core framework for collection data structures which covers all variations without exhaustively listing them, while supporting good performance at the same time.

We claim that the concept of Array Mapped Tries (AMTs) embodies a high degree of commonality in the sub-domain of immutable collection data structures. AMTs are flexible enough to cover most of the variability, while minimizing code bloat in the generator and the generated code. We implemented a Data Structure Code Generator (DSCG) that emits immutable collections based on an AMT skeleton foundation. The generated data structures outperform competitive hand-optimized implementations, and the generator still allows for customization towards specific workloads.

1. Introduction

Collection data structures that are contained in standard libraries of programming languages are popular amongst programmers. Almost all programs make use of collections. Therefore optimizing collections implies automatically increasing the performance of many programs. Optimizations within collection libraries are orthogonal to compiler and runtime improvements, because they usually focus on improving data structure encodings and algorithms.

Immutable collections represent key data structures in hybrid functional and object-oriented programming languages, such as Scala[1] and Clojure[2]. Immutability allows optimizations that exploit the fact that data does not change [16, 28], allows safe sharing of data in concurrent environments, and makes equational reasoning possible in object-oriented programming environments.

Collection data structures that are contained in standard libraries are mostly one-off solutions, aiming for reasonable performance for the general use case. Design decisions and trade-offs are preselected by the library engineer and turn collection data structures into hard-coded assets. This is problematic, since statically encoding data structure design decisions and trade-offs brings disadvantages for the library users and the library engineers. While the former do not have easy access to optimized problem-specific data structures, the latter cannot extend and evolve potentially large code bases of collection libraries efficiently.

A Large Domain with Variability. The various dimensions of collection libraries become apparent when looking at the module structures of languages such as Java or Scala. They provide many data structure variations that duplicate code and are split by several of the following dimensions:

Split by data type semantics: Interfaces and implementations for lists, sets, bags, maps, multi-maps, etcetera.

Split by ordering: Data structures can be ordered either due to data type semantics or temporal properties such as insertion order. Otherwise, data structures can be unordered by nature (e.g., sets), or due to hashing of the keys.

Split by update semantics: Data structures can allow mutation of their content over time, or remain immutable after initialization. Transient data structures represent the middle ground by allowing efficient initialization and batch updates on otherwise immutable data structures.

[1] `https://scala-lang.org`

[2] `https://clojure.org`

Split by processing semantics: Data structures are often divided into categories by their supported processing semantics. They can either support basic sequential processing, parallel processing (e.g., by splitting and merging data), or concurrent processing.

Split by encoding: Different encodings yield different performance characteristics. For example, a list data type allows implementations as an array, or as entries that are linked through references.

Split by content: Most collection data structures are designed to be type-safe by restricting elements to a single homogeneous generic type. Storing mixed content of various types is often only possible untyped.

Given the above (incomplete) indication of variability, collection libraries seem like an ideal case for generative programming in the traditional sense [5, 8, 19]. We expect to factor out commonalities for ease-of-maintenance, improve efficiency, and make variants available as context-specific solutions. Because of the large amount of variability, the challenge is to find a minimal core that is expressive enough to cover the domain while at the same time offer good performance. We claim that by fixing the dimension of update semantics to immutable (and transient), we can provide a minimal core, on basis of an Array Mapped Trie (AMT) skeleton, which is able to satisfy our performance requirements.

Without loss of generality, AMTs do allow the generation of mutable collections. However, early experiments showed that these generally exhibit weaker performance characteristics than competing array-based data structures. We limit our motivation and claims in this paper to immutable data.

Contributions. We contribute a domain analysis that covers variability in collection data structures, and the application of AMT skeletons in our domain specific code generator, factoring out commonalities while enabling performance.

2. Related Work

Product Lines and Dynamic Adaptation. We take a (static) Software Product Line (SPL) [7] perspective on collections to enable software reuse. Features of collections and variability are typically known at design time. Dynamic Software Product Lines [13] in contrast concentrate on variability at program runtime and share commonalities with research goal of the Run-Time Adaptation [1] community. AMTs are amenable to run-time variability as well; which we consider future work.

Data Structure Selection at Run-Time. SETL pioneered automatic data structure selection [22]. On the Java Virtual Machine (JVM), Shacham et al. [23] introduced Chameleon, a dynamic analysis tool that lets programmers choose the most efficient implementation within a collection library for a given collection Application Program Interface (API). While SETL and Chameleon support selection of appropriate data types within a product family, both are not concerned with our goal of encoding commonalities of data types.

Generating complex collection data structures. Declaratively synthesizing complex collection data structures by component composition goes back to DiSTiL [24].

Hawkins et al. worked on declarative and provable specifications and synthesis of data structures with complex sharing, both for the sequential [14] and concurrent [15] case.

Loncaric et al. [18] extend the work of Hawkins et al. by adding support for order among elements and complex retrieval operations. They generate *intrusive* data structures that avoid a layer of indirection by storing auxiliary pointers in domain elements directly, trading flexibility of generic collections for a potential increase in performance. In contrast, our approach natively supports sharing of sub-structures and focuses on non-intrusive collections, however we do not integrate formal methods for making correctness claims.

All previously discussed papers have one approach in common: they synthesize complex data structures by composing basic collection data structures (e.g., array-list, linked-list, hash-map, etcetera). None of these results tackle the generation of basic collection API like the current paper does.

Specializing for Primitive Data Types. Ureche et al. [31] added automatic specializations for primitive JVM data types to the Scala compiler. Combinatorial code-bloat is tackled by specializing for the largest primitive type `long` and by automatically coercing smaller-sized primitives.

State of the Art of Trie Data Structures. Trie data structures were invented 1959 by Briandais [9] and named a year later by Fredkin [12]. An AMT [2, 6] is a trie variant where lookup time is independent from the number of keys stored in the trie. AMTs eliminate empty array slots of nodes by using one bit in a bitmap for each valid outgoing trie branch.

Functional Unordered Collections based on AMTs. A Hash-Array Mapped Trie (HAMT) [3] is a space-efficient trie that encodes the hash code prefixes of elements. HAMTs constitute the basis for purely functional collections that are incrementally constructed and may refer to the unaltered parts of previous states [11, 20]. In previous work we introduced Compressed Hash-Array Mapped Prefix-tree (CHAMP) [27], a cache-aware and canonical HAMT variant that improves runtime efficiency over its predecessor of iteration (1.3–6.7 x) and equality checking (3–25.4 x) at microbenchmarks and real-word benchmarks, while reducing memory footprints.

Functional Lists and Vectors Inspired by HAMTs. Immutable vector are primarily based on principles of AMTs, because they resulting prefix trees cover densely filled lists. Bagwell and Rompf [4] published a technical report about efficient immutable vectors that improved runtimes of split and merge operations to a logarithmic bound. Stucki et al. [30] improved upon the latter and added a broad scale evaluation.

Concurrent HAMTs. Prokopec et al. [21] worked on mutable concurrent HAMTs that feature iterators with snapshot semantics, which preserve enumeration of all elements that were present when the iterator was created.

3. A Stable Data Type Independent Encoding

Efficient collection data structures on the JVM are typically coded as array-based hashtables. The array core complicates separating commonality from variability to construct a product family. In particular, arrays imply that either all elements are primitives or they are all references. For primitive collections, the absence of a value requires additional encoding (sentinels or bitmaps) to represent **`null`**. AMT-based collections on the other hand do allow fine-grained memory layout choices (per internal node) and are therefore more amenable for encoding a product family of collection data structures. While the API operations and details may differ between variants, we explain how to use the AMT as a fundamental skeleton to support many kinds of efficient immutable collections.

The remainder of this section describes the core concepts of trie-based collections in Feature Description Language (FDL) notation [10]. The full model has been archived [25]. It describes the variability in the domain of collections, making commonalities and differences of configurations explicit, as well as constraints among them.

```
features trie
  EncodingType      : one-of(data, hashOfData)
  EncodingLength    : one-of(bounded, unbounded)
  EncodingDirection : one-of(prefix, postfix)
  ChunkUnit   : one-of(bit, char)
  ChunkLength : int
  DataDensity : one-of(sparse, dense)
  Content     : one-of(mixedNodes, dataAsLeafs)
```

A `trie` is an ordered tree data structure. It is like a Deterministic Finite Automaton (DFA) without any loops, where the transitions are steps of a search path, the internal nodes encode prefix sharing, and the accept nodes hold the stored values. Like with a DFA, a single path represents a single `data` value by concatenating the labels of the edges. An example would be a vector data structure where the index is stored in the path. When we store `hashOfData` however, like in unordered map collections, usually we store a copy at the accept nodes to cater for possible hash collisions. The features `ChunkUnit`, `ChunkLength` and `EncodingDirection` determine the granularity of information encoded by the edges. Encoding direction `prefix` starts at the least-significant bit, whereas `postfix` starts at the most significant bit.

The `trie` model describes the common core characteristics of trie-based collections: each flavor encodes prefixes of either bounded (e.g. integers) or unbounded length (e.g. strings) with a particular stepping size. Based on any particular `trie` configuration, a code generator can derive the storage and lookup implementation using different (bitlevel) operations to split values across the respective paths.

The above describes how the *keys* of a collection are stored in an ordered or unordered collection, but we also cater for more general collections such as maps and relations. To do this we store `Payload` tuples (specification elided) at the accept nodes with variable arity and content. To achieve the required top-level API, a code generator will wrap the internal trie nodes using different visitors to collect the stored data in the required form (e.g., `java.util.Map.Entry`).

The following partial configuration characterises AMT. An AMT-based vector maps from a prefix-encoded `index` $\mapsto$ `element`. The `prefix` code direction ensures space efficiency for `dense` vectors, because vector indices usually occupy the least-significant bits:

```
config amt-vector requires EncodingType::data,
  EncodingDirection::prefix, DataDensity::dense
```

A HAMT based unordered collection on the other hand looks slightly different:

```
config hamt-unordered requires
  EncodingType::hashOfData,
  EncodingLength::bounded, DataDensity::sparse
```

Efficient immutable hash data structures are typically implemented as HAMTs, mapping from `hash(key)` $\mapsto$ `key/value`, in case of a hash-map. In Java, default hash codes are bound in size (32 bit) and assumed to have an almost uniform distribution, so the `EncodingDirection` is not constrained. The size of collections is usually `sparse`, compared to the 2^{32} space of possible hash codes. The previous two listings describe viable default configurations for vectors and hash-maps of collection libraries. Yet, a feature model allows for customization towards specific workloads (e.g., sparse vectors). For certain efficiency trade-offs it is important to distinguish between HAMT encodings which store `dataAsLeafs` and encodings which allow for `mixedNodes` internally [27].

We currently generate unordered set, map, and multi-map data structures based on the state-of-the-art HAMT variants: HAMT [3], CHAMP [27], and HHAMT [29]. The latter is a generalization of the former two and supports multiple heterogeneous payload categories simultaneously. A subset of the generated collections is distributed with the *capsule* library.[3] In future work we plan to support vectors and concurrency.

4. Intermediate Generator Abstractions

We use a form of these feature models to configure a domain specific Data Structure Code Generator (DSCG) that actually implements each variant. The DSCG is implemented in Rascal, a Domain-Specific Language (DSL) designed for analyzing, processing, transforming and generating source code [17]. We represent variants in trie implementation details using abstract tree grammars with Rascals **`data`** declarations. In the following section we detail the core intermediate abstractions, necessary to efficiently implement each configuration.

[3] `https://github.com/usethesource/capsule/`

```
list[Partition] champ_partition_configuration(int bound) = [
  slice("payload", sequence([ generic("K"), generic("V") ]), range(0, bound), forward()),
  slice("node", specific("Node"), range(0, bound), backward()) ];
```

Listing 1. ADT term for the partitioning of a set of family members called CHAMP, parametrized by a size bound (i.e. 32).

```
list[PartitionCopy] applyManipulation(Partition p, Manipulation m:copyAndInsert()) {
  list[PartitionCopy] operations = [ rangeCopy (p, m.beginExpr, m.indexExpr, indexIdentity, indexIdentity),
                                     injection (p, m.indexExpr, valueList = m.valueList),
                                     rangeCopy (p, m.indexExpr, p.lengthExpr, indexIdentity, indexPlus1) ];
  return p.direction == forward() ? operations : reverse(operations);
}
```

Listing 2. Linearization and transformation from domain specific `copyAndInsert` primitive to intermediate abstraction.

Modeling Trie Node Data Layouts and Transformations. The skeleton design is that the out edges of the trie nodes are stored in a array, at least conceptually. Depending on the feature configuration, order, sequence, and types of the elements in the array may differ. For example, these arrays can mix payload and sub-nodes in arbitrary order, or group elements per content category together [27]. We model this variability in array content as follows:

```
data Partition
 = slice (Id,Type,Range,Direction)
 | stripe(Id,Type,Range,Direction,list[Partition]);
```

A partition describes a typed sequence of elements that is limited to a size `Range` (lower and upper bounds). A `slice` is the atomic unit, whereas a `stripe` joins two or more adjacent slices together. The two `Direction` values, `forward` or `backward`, allow advanced slice configurations that —similar to heap and stack— grow from separate fixed bases, to omit the necessity of dynamic partition boundary calculations [27].

Listing 1 shows the partition configuration of a hash-map encoded in CHAMP [27]. CHAMP splits a node's content into two homogeneously typed groups —payload and sub-nodes— that are indexed from different directions. Each partition is delimited in growth (`bound`). Furthermore, a domain specific invariant guarantees space sharing: the sum of sizes of all partitions together must not exceed the `bound`.

DSCG reduces the partition layout to a minimal set of physical arrays, e.g., by grouping adjacent slices of reference types together into a single untyped `stripe`. To reduce memory footprints further, DSCG supports specialization approaches that are specific to AMTs [26, 29].

Synthesizing linearized update operations. DSCG supports twelve primitives for manipulating logical partitions of AMT-based data structures. These primitives cover (lazy) expansion of prefix structures, insert/update/deletion on partitions, migration of data between partitions and canonicalization on insert and delete. However, the cost of manipulating data on top of logical partitions increases with added data categories, and furthermore different encoding directions break linearity of copying operations as shown for `copyAndInsert` (in Java):

```
for (int i = 0; i < index; i++)
  dst.setPayload(i, src.getPayload(i));

dst.setPayload(index, new Tuple(key, val));

for (int i = index; i < src.payloadLength(); i++)
  dst.setPayload(i + 1, src.getPayload(i));

for (int i = src.nodeLength(); i >= 0; i--)
  dst.setNode(i, src.getNode(i));
```

If we transform update operations such that they operate on a linearized view of the underlying physical array instead on logical partitions, we can further reduce the number of back-end generator primitives to two —`rangeCopy` that supports index shifts, and `injection` of payload— as shown in Listing 2. A linearized view effectively turns copy operations into stream processing operations, where the source and destination arrays are traversed with monotonous growing indices front to back. Adjacent `rangeCopy` operations can be fused together to increase efficiency as shown below (in Java):

```
offset += rangeCopy (src, dst, offset, index);
delta  += injection (dst, offset, key, val);
offset += rangeCopy (src, offset, dst, offset +
  delta, length - index);
```

5. Conclusion

The Array Mapped Tries skeleton is a common framework for generating fast immutable collection data structures. Our feature model covers both variants that occur in the wild, and supports novel heterogeneous variants [29]. The generated code is efficient, overall outperforming competitive state-of-the-art collections [27, 29], and —when specialized for primitive data types— they match the memory footprints of best-of-breed primitive collections [29].

Based on this evidence of the efficacy of the feature model and the intermediate abstractions for DSCG, we will extend it further to generate a complete Software Product Line of trie-based immutable collections.

References

[1] V. Alves, D. Schneider, M. Becker, N. Bencomo, and P. Grace. Comparitive study of variability management in software product lines and runtime adaptable systems. In *Third International Workshop on Variability Modelling of Software-Intensive Systems, Seville, Spain, January 28-30, 2009. Proceedings*, pages 9–17, 2009. URL `http://www.vamos-workshop.net/proceedings/VaMoS_2009_Proceedings.pdf`.

[2] P. Bagwell. Fast And Space Efficient Trie Searches. Technical Report LAMP-REPORT-2000-001, Ecole polytechnique fédérale de Lausanne, 2000.

[3] P. Bagwell. Ideal Hash Trees. Technical Report LAMP-REPORT-2001-001, Ecole polytechnique fédérale de Lausanne, 2001.

[4] P. Bagwell and T. Rompf. RRB-Trees: Efficient Immutable Vectors. Technical Report EPFL-REPORT-169879, Ecole polytechnique fédérale de Lausanne, 2011.

[5] T. J. Biggerstaff. A Perspective of Generative Reuse. *Annals of Software Engineering*, 5(1):169–226, Jan. 1998.

[6] R. Bird. Two dimensional pattern matching. *Information Processing Letters*, 6(5):168 – 170, 1977. ISSN 0020-0190.

[7] P. Clements and L. Northrop. *Software Product Lines: Practices and Patterns*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2001. ISBN 0-201-70332-7.

[8] K. Czarnecki and U. W. Eisenecker. *Generative Programming: Methods, Tools, and Applications*. ACM Press, 2000.

[9] R. De La Briandais. File Searching Using Variable Length Keys. In *IRE-AIEE-ACM '59 (Western): Papers Presented at the March 3-5, 1959, Western Joint Computer Conference*. ACM, Mar. 1959.

[10] A. v. Deursen and P. Klint. Domain-specific language design requires feature descriptions. *Journal of Computing and Information Technology*, 10(1):1–17, 2002.

[11] J. R. Driscoll, N. Sarnak, D. D. Sleator, and R. E. Tarjan. Making Data Structures Persistent. In *STOC '86: Proceedings of the Eighteenth Annual ACM Symposium on Theory of Computing*. ACM, Nov. 1986.

[12] E. Fredkin. Trie Memory. *Communications of the ACM*, 3(9): 490–499, Sept. 1960.

[13] S. Hallsteinsen, M. Hinchey, S. Park, and K. Schmid. Dynamic software product lines. *Computer*, 41(4):93–95, April 2008. ISSN 0018-9162.

[14] P. Hawkins, A. Aiken, K. Fisher, M. Rinard, and M. Sagiv. Data Representation Synthesis. In *PLDI '11: Proceedings of the 32nd ACM SIGPLAN Conference on Programming Language Design and Implementation*. ACM, 2011.

[15] P. Hawkins, A. Aiken, K. Fisher, M. Rinard, and M. Sagiv. Concurrent Data Representation Synthesis. In *PLDI '12: Proceedings of the 33rd ACM SIGPLAN Conference on Programming Language Design and Implementation*. ACM, 2012.

[16] P. Helland. Immutability changes everything. *Commun. ACM*, 59(1):64–70, Dec. 2015. ISSN 0001-0782.

[17] P. Klint, T. van der Storm, and J. Vinju. Rascal: A Domain Specific Language for Source Code Analysis and Manipulation. In *Proceedings of Ninth IEEE International Working Conference on Source Code Analysis and Manipulation*. IEEE, 2009.

[18] C. Loncaric, E. Torlak, and M. D. Ernst. Fast Synthesis of Fast Collections. In *PLDI '16: Proceedings of the 37th ACM SIGPLAN Conference on Programming Language Design and Implementation*. ACM, 2016.

[19] D. McIlroy. Mass-Produced Software Components. In P. Naur and B. Randell, editors, *Proceedings of NATO Software Engineering Conference*, pages 138–155, Oct. 1968.

[20] C. Okasaki. *Purely Functional Data Structures*. Cambridge University Press, June 1999.

[21] A. Prokopec, N. G. Bronson, P. Bagwell, and M. Odersky. Concurrent tries with efficient non-blocking snapshots. In *PPoPP '12: Proceedings of the 17th ACM SIGPLAN symposium on Principles and Practice of Parallel Programming*. ACM, 2012.

[22] E. Schonberg, J. T. Schwartz, and M. Sharir. Automatic Data Structure Selection in SETL. In *POPL '79: Proceedings of the 6th ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages*. ACM, 1979.

[23] O. Shacham, M. Vechev, and E. Yahav. Chameleon: Adaptive selection of collections. In *PLDI '09: Proceedings of the 30th ACM SIGPLAN Conference on Programming Language Design and Implementation*. ACM, 2009.

[24] Y. Smaragdakis and D. Batory. Distil: A transformation library for data structures. In *DSL'97: Proceedings of the Conference on Domain-Specific Languages*. USENIX Association, 1997.

[25] M. J. Steindorfer. Towards a feature model of trie-based collections. To appear on arXiv, July 2016. URL `http://michael.steindorfer.name/drafts/collections-feature-model`.

[26] M. J. Steindorfer and J. J. Vinju. Code Specialization for Memory Efficient Hash Tries (Short Paper). In *GPCE '14: Proceedings of the International Conference on Generative Programming: Concepts and Experiences*. ACM, 2014.

[27] M. J. Steindorfer and J. J. Vinju. Optimizing Hash-array Mapped Tries for Fast and Lean Immutable JVM Collections. In *OOPSLA '15: Proceedings of the ACM SIGPLAN International Conference on Object-Oriented Programming, Systems, Languages, and Applications*. ACM, 2015.

[28] M. J. Steindorfer and J. J. Vinju. Performance Modeling of Maximal Sharing. In *ICPE '16: Proceedings of the 7th ACM/SPEC on International Conference on Performance Engineering*. ACM, 2016.

[29] M. J. Steindorfer and J. J. Vinju. Fast and Lean Immutable Multi-Maps on the JVM based on Heterogeneous Hash-Array Mapped Tries. To appear on arXiv, July 2016. URL `http://michael.steindorfer.name/drafts/hamt-heterogeneous`.

[30] N. Stucki, T. Rompf, V. Ureche, and P. Bagwell. RRB Vector: A Practical General Purpose Immutable Sequence. In *ICFP '15: Proceedings of the 20th ACM SIGPLAN International Conference on Functional Programming*. ACM, 2015.

[31] V. Ureche, C. Talau, and M. Odersky. Miniboxing: Improving the speed to code size tradeoff in parametric polymorphism translations. In *OOPSLA '13: Proceedings of the ACM SIGPLAN International Conference on Object Oriented Programming Systems Languages and Applications*. ACM, 2013.