Generic Associative Array Implementation

Overview

This associative array implementation is an object container with the following properties:

1. Objects are opaque pointers. The implementation does not care where they point (if anywhere) or what they point to (if anything).

   > **Note**
   >
   > Pointers to objects _must_ be zero in the least significant bit.

2. Objects do not need to contain linkage blocks for use by the array. This permits an object to be located in multiple arrays simultaneously. Rather, the array is made up of metadata blocks that point to objects.
3. Objects require index keys to locate them within the array.
4. Index keys must be unique. Inserting an object with the same key as one already in the array will replace the old object.
5. Index keys can be of any length and can be of different lengths.
6. Index keys should encode the length early on, before any variation due to length is seen.
7. Index keys can include a hash to scatter objects throughout the array.
8. The array can iterated over. The objects will not necessarily come out in key order.
9. The array can be iterated over while it is being modified, provided the RCU readlock is being held by the iterator. Note, however, under these circumstances, some objects may be seen more than once. If this is a problem, the iterator should lock against modification. Objects will not be missed, however, unless deleted.
10. Objects in the array can be looked up by means of their index key.
11. Objects can be looked up while the array is being modified, provided the RCU readlock is being held by the thread doing the look up.

The implementation uses a tree of 16-pointer nodes internally that are indexed on each level by nibbles from the index key in the same manner as in a radix tree. To improve memory efficiency, shortcuts can be emplaced to skip over what would otherwise be a series of single-occupancy nodes. Further, nodes pack leaf object pointers into spare space in the node rather than making an extra branch until as such time an object needs to be added to a full node.

The Public API

The public API can be found in `<linux/assoc_array.h>`. The associative array is rooted on the following structure:

```
struct assoc_array {
        ...
};
```

The code is selected by enabling `CONFIG_ASSOCIATIVE_ARRAY` with:

```
./script/config -e ASSOCIATIVE_ARRAY
```

Edit Script

The insertion and deletion functions produce an 'edit script' that can later be applied to effect the changes without risking `ENOMEM`. This retains the preallocated metadata blocks that will be installed in the internal tree and keeps track of the metadata blocks that will be removed from the tree when the script is applied.

This is also used to keep track of dead blocks and dead objects after the script has been applied so that they can be freed later. The freeing is done after an RCU grace period has passed - thus allowing access functions to proceed under the RCU read lock.

The script appears as outside of the API as a pointer of the type:

```
struct assoc_array_edit;
```

There are two functions for dealing with the script:

1. Apply an edit script:

   ```
   void assoc_array_apply_edit(struct assoc_array_edit *edit);
   ```

This will perform the edit functions, interpolating various write barriers to permit accesses under the RCU read lock to continue. The edit script will then be passed to `call_rcu()` to free it and any dead stuff it points to.

2. Cancel an edit script:

```
void assoc_array_cancel_edit(struct assoc_array_edit *edit);
```

This frees the edit script and all preallocated memory immediately. If this was for insertion, the new object is _not_ released by this function, but must rather be released by the caller.

These functions are guaranteed not to fail.

Operations Table

Various functions take a table of operations:

```
struct assoc_array_ops {
        ...
};
```

This points to a number of methods, all of which need to be provided:

1. Get a chunk of index key from caller data:

```
unsigned long (*get_key_chunk)(const void *index_key, int level);
```

This should return a chunk of caller-supplied index key starting at the *bit* position given by the level argument. The level argument will be a multiple of `ASSOC_ARRAY_KEY_CHUNK_SIZE` and the function should return `ASSOC_ARRAY_KEY_CHUNK_SIZE bits`. No error is possible.

2. Get a chunk of an object's index key:

```
unsigned long (*get_object_key_chunk)(const void *object, int level);
```

As the previous function, but gets its data from an object in the array rather than from a caller-supplied index key.

3. See if this is the object we're looking for:

```
bool (*compare_object)(const void *object, const void *index_key);
```

Compare the object against an index key and return `true` if it matches and `false` if it doesn't.

4. Diff the index keys of two objects:

```
int (*diff_objects)(const void *object, const void *index_key);
```

Return the bit position at which the index key of the specified object differs from the given index key or -1 if they are the same.

5. Free an object:

```
void (*free_object)(void *object);
```

Free the specified object. Note that this may be called an RCU grace period after `assoc_array_apply_edit()` was called, so `synchronize_rcu()` may be necessary on module unloading.

Manipulation Functions

There are a number of functions for manipulating an associative array:

1. Initialise an associative array:

```
void assoc_array_init(struct assoc_array *array);
```

This initialises the base structure for an associative array. It can't fail.

2. Insert/replace an object in an associative array:

```
struct assoc_array_edit *
assoc_array_insert(struct assoc_array *array,
                   const struct assoc_array_ops *ops,
                   const void *index_key,
                   void *object);
```

This inserts the given object into the array. Note that the least significant bit of the pointer must be zero as it's used to type-mark pointers internally.

If an object already exists for that key then it will be replaced with the new object and the old one will be freed automatically.

The `index_key` argument should hold index key information and is passed to the methods in the ops table when they are called.

This function makes no alteration to the array itself, but rather returns an edit script that must be applied. `-ENOMEM` is returned in the case of an out-of-memory error.

The caller should lock exclusively against other modifiers of the array.

3. Delete an object from an associative array:

```
struct assoc_array_edit *
assoc_array_delete(struct assoc_array *array,
                   const struct assoc_array_ops *ops,
                   const void *index_key);
```

This deletes an object that matches the specified data from the array.

The `index_key` argument should hold index key information and is passed to the methods in the ops table when they are called.

This function makes no alteration to the array itself, but rather returns an edit script that must be applied. `-ENOMEM` is returned in the case of an out-of-memory error. `NULL` will be returned if the specified object is not found within the array.

The caller should lock exclusively against other modifiers of the array.

4. Delete all objects from an associative array:

```
struct assoc_array_edit *
assoc_array_clear(struct assoc_array *array,
                  const struct assoc_array_ops *ops);
```

This deletes all the objects from an associative array and leaves it completely empty.

This function makes no alteration to the array itself, but rather returns an edit script that must be applied. `-ENOMEM` is returned in the case of an out-of-memory error.

The caller should lock exclusively against other modifiers of the array.

5. Destroy an associative array, deleting all objects:

```
void assoc_array_destroy(struct assoc_array *array,
                         const struct assoc_array_ops *ops);
```

This destroys the contents of the associative array and leaves it completely empty. It is not permitted for another thread to be traversing the array under the RCU read lock at the same time as this function is destroying it as no RCU deferral is performed on memory release - something that would require memory to be allocated.

The caller should lock exclusively against other modifiers and accessors of the array.

6. Garbage collect an associative array:

```
int assoc_array_gc(struct assoc_array *array,
                   const struct assoc_array_ops *ops,
                   bool (*iterator)(void *object, void *iterator_data),
                   void *iterator_data);
```

This iterates over the objects in an associative array and passes each one to `iterator()`. If `iterator()` returns `true`, the object is kept. If it returns `false`, the object will be freed. If the `iterator()` function returns `true`, it must perform any appropriate refcount incrementing on the object before returning.

The internal tree will be packed down if possible as part of the iteration to reduce the number of nodes in it.

The `iterator_data` is passed directly to `iterator()` and is otherwise ignored by the function.

The function will return `0` if successful and `-ENOMEM` if there wasn't enough memory.

It is possible for other threads to iterate over or search the array under the RCU read lock while this function is in progress. The caller should lock exclusively against other modifiers of the array.

Access Functions

There are two functions for accessing an associative array:

1. Iterate over all the objects in an associative array:

```
int assoc_array_iterate(const struct assoc_array *array,
                        int (*iterator)(const void *object,
                                        void *iterator_data),
                        void *iterator_data);
```

This passes each object in the array to the iterator callback function. `iterator_data` is private data for that function.

This may be used on an array at the same time as the array is being modified, provided the RCU read lock is held. Under such circumstances, it is possible for the iteration function to see some objects twice. If this is a problem, then modification should be locked against. The iteration algorithm should not, however, miss any objects.

The function will return `0` if no objects were in the array or else it will return the result of the last iterator function called. Iteration stops immediately if any call to the iteration function results in a non-zero return.

2. Find an object in an associative array:

```
void *assoc_array_find(const struct assoc_array *array,
                       const struct assoc_array_ops *ops,
                       const void *index_key);
```

This walks through the array's internal tree directly to the object specified by the index key..

This may be used on an array at the same time as the array is being modified, provided the RCU read lock is held.

The function will return the object if found (and set `*_type` to the object type) or will return `NULL` if the object was not found.

Index Key Form

The index key can be of any form, but since the algorithms aren't told how long the key is, it is strongly recommended that the index key includes its length very early on before any variation due to the length would have an effect on comparisons.

This will cause leaves with different length keys to scatter away from each other - and those with the same length keys to cluster together.

It is also recommended that the index key begin with a hash of the rest of the key to maximise scattering throughout keyspace.

The better the scattering, the wider and lower the internal tree will be.

Poor scattering isn't too much of a problem as there are shortcuts and nodes can contain mixtures of leaves and metadata pointers.

The index key is read in chunks of machine word. Each chunk is subdivided into one nibble (4 bits) per level, so on a 32-bit CPU this is good for 8 levels and on a 64-bit CPU, 16 levels. Unless the scattering is really poor, it is unlikely that more than one word of any particular index key will have to be used.

Internal Workings

The associative array data structure has an internal tree. This tree is constructed of two types of metadata blocks: nodes and shortcuts.

A node is an array of slots. Each slot can contain one of four things:

- A NULL pointer, indicating that the slot is empty.
- A pointer to an object (a leaf).
- A pointer to a node at the next level.
- A pointer to a shortcut.

Basic Internal Tree Layout

Ignoring shortcuts for the moment, the nodes form a multilevel tree. The index key space is strictly subdivided by the nodes in the tree and nodes occur on fixed levels. For example:

```
Level: 0               1               2               3
       =============== =============== =============== ===============
                                                       NODE D
                       NODE B          NODE C  +------>+---+
               +------>+---+   +------>+---+   |       | 0 |
       NODE A  |       | 0 |   |       | 0 |   |       +---+
       +---+   |       +---+   |       +---+   |       :   :
       | 0 |   |       :   :   |       :   :   |       +---+
       +---+   |       +---+   |       +---+   |       | f |
       | 1 |---+       | 3 |---+       | 7 |---+       +---+
       +---+           +---+           +---+
       :   :           :   :           | 8 |---+
       +---+           +---+           +---+   |       NODE E
       | e |---+       | f |           :   :   +------>+---+
       +---+   |       +---+           +---+           | 0 |
       | f |   |                       | f |           +---+
       +---+   |                       +---+           :   :
               |       NODE F                          +---+
               +------>+---+                           | f |
                       | 0 |           NODE G          +---+
                       +---+   +------>+---+
                       :   :   |       | 0 |
                       +---+   |       +---+
                       | 6 |---+       :   :
                       +---+           +---+
                       :   :           | f |
                       +---+           +---+
                       | f |
                       +---+
```

In the above example, there are 7 nodes (A-G), each with 16 slots (0-f). Assuming no other meta data nodes in the tree, the key space is divided thusly:

```
KEY PREFIX      NODE
==========      ====
137*            D
138*            E
13[0-69-f]*     C
```

```
1[0-24-f]*      B
e6*             G
e[0-57-f]*      F
[02-df]*        A
```

So, for instance, keys with the following example index keys will be found in the appropriate nodes:

```
INDEX KEY       PREFIX  NODE
=============== ======= ====
13694892892489  13      C
13795289025897  137     D
13889dde88793   138     E
138bbb89003093  138     E
1394879524789   12      C
1458952489      1       B
9431809de993ba  -       A
b4542910809cd   -       A
e5284310def98   e       F
e68428974237    e6      G
e7fffcbd443     e       F
f3842239082     -       A
```

To save memory, if a node can hold all the leaves in its portion of keyspace, then the node will have all those leaves in it and will not have any metadata pointers - even if some of those leaves would like to be in the same slot.

A node can contain a heterogeneous mix of leaves and metadata pointers. Metadata pointers must be in the slots that match their subdivisions of key space. The leaves can be in any slot not occupied by a metadata pointer. It is guaranteed that none of the leaves in a node will match a slot occupied by a metadata pointer. If the metadata pointer is there, any leaf whose key matches the metadata key prefix must be in the subtree that the metadata pointer points to.

In the above example list of index keys, node A will contain:

```
SLOT    CONTENT         INDEX KEY (PREFIX)
====    =============== ==================
1       PTR TO NODE B   1*
any     LEAF            9431809de993ba
any     LEAF            b4542910809cd
e       PTR TO NODE F   e*
any     LEAF            f3842239082
```

and node B:

```
3   PTR TO NODE C   13*
any LEAF            1458952489
```

Shortcuts

Shortcuts are metadata records that jump over a piece of keyspace. A shortcut is a replacement for a series of single-occupancy nodes ascending through the levels. Shortcuts exist to save memory and to speed up traversal.

It is possible for the root of the tree to be a shortcut - say, for example, the tree contains at least 17 nodes all with key prefix `1111`. The insertion algorithm will insert a shortcut to skip over the `1111` keyspace in a single bound and get to the fourth level where these actually become different.

Splitting And Collapsing Nodes

Each node has a maximum capacity of 16 leaves and metadata pointers. If the insertion algorithm finds that it is trying to insert a 17th object into a node, that node will be split such that at least two leaves that have a common key segment at that level end up in a separate node rooted on that slot for that common key segment.

If the leaves in a full node and the leaf that is being inserted are sufficiently similar, then a shortcut will be inserted into the tree.

When the number of objects in the subtree rooted at a node falls to 16 or fewer, then the subtree will be collapsed down to a single node - and this will ripple towards the root if possible.

Non-Recursive Iteration

Each node and shortcut contains a back pointer to its parent and the number of slot in that parent that points to it. None-recursive iteration uses these to proceed rootwards through the tree, going to the parent node, slot N + 1 to make sure progress is made without the need for a stack.

The backpointers, however, make simultaneous alteration and iteration tricky.

Simultaneous Alteration And Iteration

There are a number of cases to consider:

1. Simple insert/replace. This involves simply replacing a NULL or old matching leaf pointer with the pointer to the new leaf

after a barrier. The metadata blocks don't change otherwise. An old leaf won't be freed until after the RCU grace period.

2. Simple delete. This involves just clearing an old matching leaf. The metadata blocks don't change otherwise. The old leaf won't be freed until after the RCU grace period.

3. Insertion replacing part of a subtree that we haven't yet entered. This may involve replacement of part of that subtree - but that won't affect the iteration as we won't have reached the pointer to it yet and the ancestry blocks are not replaced (the layout of those does not change).

4. Insertion replacing nodes that we're actively processing. This isn't a problem as we've passed the anchoring pointer and won't switch onto the new layout until we follow the back pointers - at which point we've already examined the leaves in the replaced node (we iterate over all the leaves in a node before following any of its metadata pointers).

   We might, however, re-see some leaves that have been split out into a new branch that's in a slot further along than we were at.

5. Insertion replacing nodes that we're processing a dependent branch of. This won't affect us until we follow the back pointers. Similar to (4).

6. Deletion collapsing a branch under us. This doesn't affect us because the back pointers will get us back to the parent of the new node before we could see the new node. The entire collapsed subtree is thrown away unchanged - and will still be rooted on the same slot, so we shouldn't process it a second time as we'll go back to slot + 1.

Note

Under some circumstances, we need to simultaneously change the parent pointer and the parent slot pointer on a node (say, for example, we inserted another node before it and moved it up a level). We cannot do this without locking against a read - so we have to replace that node too.

However, when we're changing a shortcut into a node this isn't a problem as shortcuts only have one slot and so the parent slot number isn't used when traversing backwards over one. This means that it's okay to change the slot number first - provided suitable barriers are used to make sure the parent slot number is read after the back pointer.

Obsolete blocks and leaves are freed up after an RCU grace period has passed, so as long as anyone doing walking or iteration holds the RCU read lock, the old superstructure should not go away on them.