uthash User Guide

REVISION HISTORY			
NUMBER	DATE	DESCRIPTION	NAME
1.8	September 2009		TDH

Contents

1 A hash in C

This document is written for C programmers. Since you're reading this, chances are that you know a hash is used for looking up items using a key. In scripting languages like Perl, hashes are used all the time. In C, hashes don't exist in the language itself. This software provides a hash table for C structures.

1.1 What can it do?

This software supports these operations on items in a hash table:

1. add
2. find
3. delete
4. count
5. iterate
6. sort
7. select (explained later)

1.2 Is it fast?

Add, find and delete are normally constant-time operations. This is influenced by your key domain and the hash function.

This hash aims to be minimalistic and efficient. It's around 900 lines of C. It inlines automatically because it's implemented as macros. It's fast as long as the hash function is suited to your keys. You can use the default hash function, or easily compare performance and choose from among several other built-in hash functions.

1.3 Is it a library?

No, it's just a single header file: `uthash.h`. All you need to do is copy the header file into your project, and:

```
#include "uthash.h"
```

Since uthash is a header file only, there is no library code to link against.

1.4 C/C++ and platforms

This software can be used in C and C++ programs. It has been tested on:

- Linux
- Mac OS X
- Solaris
- OpenBSD
- Cygwin
- MinGW

1.4.1 Non-GNU C++ compilers

When compiling C++ code it may be necessary to use the GNU compiler, g++, which supports the `typeof` extension. Alternatively some users of non-GNU C++ compilers simply hardcode the `TYPEOF` macro in `uthash.h` to the data type they want to use.

1.4.2 Test suite

You can run the test suite on these platforms, or any prospective Unix-like platform, in this way:

```
cd tests/
make
```

1.5 BSD licensed

This software is made available under the revised BSD license. It is free and open source.

1.6 Obtaining uthash

Please follow the link to download on the uthash website at http://uthash.sourceforge.net.

1.7 Getting help

Feel free to email the author at thanson@users.sourceforge.net.

1.8 Resources

News
: The author has a news feed for software updates (RSS).

Linked list macros
: uthash includes a separate, standalone header called utlist.h which has *linked list macros* for C structures, similar in style to the hash macros

1.9 Who's using it?

Since releasing uthash in 2006, it has been downloaded thousands of times, incorporated into commerical software, academic research, and into other open-source software.

2 Your structure

In uthash, a hash table is comprised of structures. Each structure represents a key-value association. One or more of the structure fields constitute the key. The structure pointer itself is the value.

Example 2.1 Defining a structure that can be hashed

```
#include "uthash.h"

struct my_struct {
    int id;                    /* key */
    char name[10];
    UT_hash_handle hh;         /* makes this structure hashable */
};
```

Note that, in uthash, your structure will never be moved or copied into another location when you add it into a hash table. This means that you can keep other data structures that safely point to your structure-- regardless of whether you add or delete it from a hash table during your program's lifetime.

2.1 The key

There are no restrictions on the data type or name of the key field. The key can also comprise multiple contiguous fields, having any names and data types.

> Any data type… really? Yes, your key and structure can have any data type. Unlike function calls with fixed prototypes, uthash consists of macros-- whose arguments are untyped-- and thus able to work with any type of structure or key.

2.1.1 Unique keys

As with any hash, every item must have a unique key. Your application must enforce key uniqueness. Before you add an item to the hash table, you must first know (if in doubt, check!) that the key is not already in use. You can check whether a key already exists in the hash table using `HASH_FIND`.

2.2 The hash handle

The `UT_hash_handle` field must be present in your structure. It is used for the internal bookkeeping that makes the hash work. It does not require initialization. It can be named anything, but you can simplify matters by naming it `hh`. This allows you to use the easier "convenience" macros to add, find and delete items.

3 Hash operations

This section introduces the uthash macros by example. For a more succinct listing, see Macro Reference.

> Convenience vs. general macros: The uthash macros fall into two categories. The *convenience* macros can be used with integer or string keys (and require that you chose the conventional name `hh` for the `UT_hash_handle` field). The convenience macros take fewer arguments than the general macros, making their usage a bit simpler for these common types of keys.
>
> The *general* macros can be used for any types of keys, or for multi-field keys, or when the `UT_hash_handle` has been named something other than `hh`. These macros take more arguments and offer greater flexibility in return. But if the convenience macros suit your needs, use them-- your code will be more readable.

3.1 Declare the hash

Your hash must be declared as a `NULL`-initialized pointer to your structure.

```
struct my_struct *users = NULL;    /* important! initialize to NULL */
```

3.2 Add item

Allocate and initialize your structure as you see fit. The only aspect of this that matters to uthash is that your key must be initialized to a unique value. Then call `HASH_ADD`. (Here we use the convenience macro `HASH_ADD_INT`, which offers simplified usage for keys of type `int`).

Example 3.1 Add an item to a hash

```
int add_user(int user_id, char *name) {
    struct my_struct *s;

    s = malloc(sizeof(struct my_struct));
    s->id = user_id;
    strcpy(s->name, name);
    HASH_ADD_INT( users, id, s );  /* id: name of key field */
}
```

The first parameter to HASH_ADD_INT is the hash table, and the second parameter is the *name* of the key field. Here, this is `id`. The last parameter is a pointer to the structure being added.

> Wait.. the field name is a parameter? If you find it strange that `id`, which is the *name of a field* in the structure, can be passed as a parameter, welcome to the world of macros. Don't worry- the C preprocessor expands this to valid C code.

3.2.1 Key must not be modified while in-use

Once a structure has been added to the hash, do not change the value of its key. Instead, delete the item from the hash, change the key, and then re-add it.

3.3 Find item

To look up a structure in a hash, you need its key. Then call HASH_FIND. (Here we use the convenience macro HASH_FIND_INT for keys of type `int`).

Example 3.2 Find a structure using its key

```
struct my_struct *find_user(int user_id) {
    struct my_struct *s;

    HASH_FIND_INT( users, &user_id, s );  /* s: output pointer */
    return s;
}
```

Here, the hash table is `users`, and `&user_id` points to the key (an integer in this case). Last, `s` is the *output* variable of HASH_FIND_INT. The final result is that `s` points to the structure with the given key, or is `NULL` if the key wasn't found in the hash.

> **Note**
> The middle argument is a *pointer* to the key. You can't pass a literal key value to HASH_FIND. Instead assign the literal value to a variable, and pass a pointer to the variable.

3.4 Delete item

To delete a structure from a hash, you must have a pointer to it. (If you only have the key, first do a HASH_FIND to get the structure pointer).

Example 3.3 Delete an item from a hash

```
void delete_user(struct my_struct *user) {
    HASH_DEL( users, user);  /* user: pointer to deletee */
    free(user);              /* optional; it's up to you! */
}
```

Here again, `users` is the hash table, and `user` is a pointer to the structure we want to remove from the hash.

3.4.1 uthash never frees your structure

Deleting a structure just removes it from the hash table-- it doesn't `free` it. The choice of when to free your structure is entirely up to you; uthash will never free your structure.

3.4.2 Delete can change the pointer

The hash table pointer (which initially points to the first item added to the hash) can change in response to `HASH_DEL` (i.e. if you delete the first item in the hash table).

3.5 Delete all items

To delete all the structures in a hash table, you can either iteratively delete the items (if you plan to free each one or need to other per-element work), or you can delete all the items in a single operation.

3.5.1 Iterative deletion

It's easy: just keep deleting the first item. If you plan to free it, copy the pointer beforehand since the delete will advance the "first item" to the next.

Example 3.4 Delete all items from a hash

```
void delete_all() {
  user_struct *current_user;

  while(users) {
    current_user = users;          /* copy pointer to first item     */
    HASH_DEL(users,current_user);  /* delete; users advances to next */
    free(current_user);            /* optional- if you want to free  */
  }
}
```

3.5.2 All-at-once deletion

If you only want to delete all the items, but not free them or do any per-element clean up, you can do this more efficiently in a single operation:

```
HASH_CLEAR(hh,users);
```

Afterward, the list head (here, `users`) will be set to `NULL`.

3.6 Count items

The number of items in the hash table can be obtained using HASH_COUNT:

Example 3.5 Count of items in the hash table

```
unsigned int num_users;
num_users = HASH_COUNT(users);
printf("there are %u users\n", num_users);
```

Incidentally, this works even the list (users, here) is NULL, in which case the count is 0.

3.7 Iterating and sorting

You can loop over the items in the hash by starting from the beginning and following the hh.next pointer.

Example 3.6 Iterating over all the items in a hash

```
void print_users() {
    struct my_struct *s;

    for(s=users; s != NULL; s=s->hh.next) {
        printf("user id %d: name %s\n", s->id, s->name);
    }
}
```

There is also an hh.prev pointer you could use to iterate backwards through the hash, starting from any known item.

> A hash is also a doubly-linked list. Iterating backward and forward through the items in the hash is possible because of the hh.prev and hh.next fields. All the items in the hash can be reached by repeatedly following these pointers, thus the hash is also a doubly-linked list.

If you're using uthash in a C++ program, you need an extra cast on the for iterator, e.g., s=(struct my_struct*)s→hh.next.

3.7.1 Sorted iteration

The items in the hash are, by default, traversed in the order they were added ("insertion order") when you follow the hh.next pointer. But you can sort the items into a new order using HASH_SORT. E.g.,

```
HASH_SORT( users, name_sort );
```

The second argument is a pointer to a comparison function. It must accept two arguments which are pointers to two items to compare. Its return value should be less than zero, zero, or greater than zero, if the first item sorts before, equal to, or after the second item, respectively. (Just like strcmp).

Example 3.7 Sorting the items in the hash

```
int name_sort(struct my_struct *a, struct my_struct *b) {
    return strcmp(a->name,b->name);
}

int id_sort(struct my_struct *a, struct my_struct *b) {
    return (a->id - b->id);
}

void sort_by_name() {
    HASH_SORT(users, name_sort);
}

void sort_by_id() {
    HASH_SORT(users, id_sort);
}
```

When the items in the hash are sorted, the first item may change position. In the example above, `users` may point to a different structure after calling `HASH_SORT`.

3.8 A complete example

We'll repeat all the code and embellish it with a `main()` function to form a working example.

If this code was placed in a file called `example.c` in the same directory as `uthash.h`, it could be compiled and run like this:

```
cc -o example example.c
./example
```

Follow the prompts to try the program, and type `Ctrl-C` when done.

Example 3.8 A complete program (part 1 of 2)

```
#include <stdio.h>   /* gets */
#include <stdlib.h>  /* atoi, malloc */
#include <string.h>  /* strcpy */
#include "uthash.h"

struct my_struct {
    int id;                    /* key */
    char name[10];
    UT_hash_handle hh;         /* makes this structure hashable */
};

struct my_struct *users = NULL;

int add_user(int user_id, char *name) {
    struct my_struct *s;

    s = malloc(sizeof(struct my_struct));
    s->id = user_id;
    strcpy(s->name, name);
    HASH_ADD_INT( users, id, s );  /* id: name of key field */
}

struct my_struct *find_user(int user_id) {
    struct my_struct *s;

    HASH_FIND_INT( users, &user_id, s );  /* s: output pointer */
    return s;
}

void delete_user(struct my_struct *user) {
    HASH_DEL( users, user);  /* user: pointer to deletee */
    free(user);
}

void delete_all() {
  struct my_struct *current_user;

  while(users) {
    current_user = users;          /* grab pointer to first item */
    HASH_DEL(users,current_user);  /* delete it (users advances to next) */
    free(current_user);            /* free it */
  }
}

void print_users() {
    struct my_struct *s;

    for(s=users; s != NULL; s=s->hh.next) {
        printf("user id %d: name %s\n", s->id, s->name);
    }
}

int name_sort(struct my_struct *a, struct my_struct *b) {
    return strcmp(a->name,b->name);
}

int id_sort(struct my_struct *a, struct my_struct *b) {
    return (a->id - b->id);
}
```

Example 3.9 A complete program (part 2 of 2)

```
void sort_by_name() {
    HASH_SORT(users, name_sort);
}

void sort_by_id() {
    HASH_SORT(users, id_sort);
}

int main(int argc, char *argv[]) {
    char in[10];
    int id=1;
    struct my_struct *s;
    unsigned num_users;

    while (1) {
        printf("1. add user\n");
        printf("2. find user\n");
        printf("3. delete user\n");
        printf("4. delete all users\n");
        printf("5. sort items by name\n");
        printf("6. sort items by id\n");
        printf("7. print users\n");
        printf("8. count users\n");
        gets(in);
        switch(atoi(in)) {
            case 1:
                printf("name?\n");
                add_user(id++, gets(in));
                break;
            case 2:
                printf("id?\n");
                s = find_user(atoi(gets(in)));
                printf("user: %s\n", s ? s->name : "unknown");
                break;
            case 3:
                printf("id?\n");
                s = find_user(atoi(gets(in)));
                if (s) delete_user(s);
                else printf("id unknown\n");
                break;
            case 4:
                delete_all();
                break;
            case 5:
                sort_by_name();
                break;
            case 6:
                sort_by_id();
                break;
            case 7:
                print_users();
                break;
            case 8:
                num_users=HASH_COUNT(users);
                printf("there are %u users\n", num_users);
                break;
        }
    }
}
```

This program is included in the distribution in `tests/example.c`. You can run `make example` in that directory to compile it easily.

4 Standard key types

This section goes into specifics of how to work with different kinds of keys. You can use any type of key-- integers, strings, pointers, etc.

4.1 Integer keys

The preceding examples demonstrated use of integer keys. To recap, use the convenience macros `HASH_ADD_INT` and `HASH_FIND_IN` for structures with integer keys. (The other operations such as `HASH_DELETE` and `HASH_SORT` are the same for all types of keys).

4.2 String keys

If your structure has a string key, the operations to use depend on whether your structure *points to* the key (`char *`) or the string resides `within` the structure (`char a[10]`).

---

char[ ] vs. char*
The string is *within* the structure in the first example below-- `name` is a `char[10]` field. If instead the structure *points* to the key (i.e., `name` is declared `char *`), as in the second example below, use `HASH_ADD_KEYPTR` described in the Macro reference.

4.2.1 String *within* structure

Example 4.1 A string-keyed hash (string within structure)

```
#include <string.h>  /* strcpy */
#include <stdlib.h>  /* malloc */
#include <stdio.h>   /* printf */
#include "uthash.h"

struct my_struct {
    char name[10];             /* key (string is WITHIN the structure) */
    int id;
    UT_hash_handle hh;         /* makes this structure hashable */
};


int main(int argc, char *argv[]) {
    char **n, *names[] = { "joe", "bob", "betty", NULL };
    struct my_struct *s, *users = NULL;
    int i=0;

    for (n = names; *n != NULL; n++) {
        s = malloc(sizeof(struct my_struct));
        strcpy(s->name, *n);
        s->id = i++;
        HASH_ADD_STR( users, name, s );
    }

    HASH_FIND_STR( users, "betty", s);
    if (s) printf("betty's id is %d\n", s->id);
}
```

This example is included in the distribution in `tests/test15.c`. It prints:

```
betty's id is 2
```

4.2.2 String *pointer* in structure

Now, here is the same example but using a `char *` key instead of `char [ ]`:

Example 4.2 A string-keyed hash (structure points to string)

```
#include <string.h>  /* strcpy */
#include <stdlib.h>  /* malloc */
#include <stdio.h>   /* printf */
#include "uthash.h"

struct my_struct {
    char *name;               /* key (structure POINTS TO string */
    int id;
    UT_hash_handle hh;        /* makes this structure hashable */
};


int main(int argc, char *argv[]) {
    char **n, *names[] = { "joe", "bob", "betty", NULL };
    struct my_struct *s, *users = NULL;
    int i=0;

    for (n = names; *n != NULL; n++) {
        s = (struct my_struct*)malloc(sizeof(struct my_struct));
        s->name = *n;
        s->id = i++;
        HASH_ADD_KEYPTR( hh, users, s->name, strlen(s->name), s );
    }

    HASH_FIND_STR( users, "betty", s);
    if (s) printf("betty's id is %d\n", s->id);
   return 0;
}
```

4.3 Binary keys

We're using the term "binary" here to simply mean an arbitrary byte sequence. Your key field can have any data type. To uthash, it is just a sequence of bytes. We'll use the general macros `HASH_ADD` and `HASH_FIND` to demonstrate usage of a floating point key of type `double`.

Example 4.3 A key of type double

```
#include <stdlib.h>
#include <stdio.h>
#include "uthash.h"

typedef struct {
    double veloc;
    /* ... other data ... */
    UT_hash_handle hh;
} veloc_t;

int main(int argc, char *argv[]) {
    veloc_t *v, *v2, *veloc_table = NULL;
    double x = 1/3.0;

    v = malloc( sizeof(*v) );
    v->veloc = x;
    HASH_ADD(hh, veloc_table, veloc, sizeof(double), v);
    HASH_FIND(hh, veloc_table, &x, sizeof(double), v2 );

    if (v2) printf("found (%.2f)\n", v2->veloc);
}
```

Note that the general macros require the name of the `UT_hash_handle` to be passed as the first argument (here, this is `hh`). The general macros are documented in Macro Reference.

5 Advanced Topics

5.1 Compound keys

Your key can even comprise multiple contiguous fields.

Example 5.1 A multi-field key

```
#include <stdlib.h>    /* malloc       */
#include <stddef.h>    /* offsetof     */
#include <stdio.h>     /* printf       */
#include <string.h>    /* memset       */
#include "uthash.h"

#define UTF32 1

typedef struct {
  UT_hash_handle hh;
  int len;
  char encoding;      /* these two fields */
  int text[];         /* comprise the key */
} msg_t;

int main(int argc, char *argv[]) {
    int keylen;
    msg_t *msg, *msgs = NULL;
    struct { char encoding; int text[]; } *lookup_key;

    int beijing[] = {0x5317, 0x4eac};   /* UTF-32LE for &#x5317;&#x4eac; */

    /* allocate and initialize our structure */
    msg = malloc( sizeof(msg_t) + sizeof(beijing) );
    memset(msg, 0, sizeof(msg_t)+sizeof(beijing)); /* zero fill */
    msg->len = sizeof(beijing);
    msg->encoding = UTF32;
    memcpy(msg->text, beijing, sizeof(beijing));

    /* calculate the key length including padding, using formula */
    keylen =   offsetof(msg_t, text)       /* offset of last key field */
             + sizeof(beijing)             /* size of last key field */
             - offsetof(msg_t, encoding);  /* offset of first key field */

    /* add our structure to the hash table */
    HASH_ADD( hh, msgs, encoding, keylen, msg);

    /* look it up to prove that it worked :-) */
    msg=NULL;

    lookup_key = malloc(sizeof(*lookup_key) + sizeof(beijing));
    memset(lookup_key, 0, sizeof(*lookup_key) + sizeof(beijing));
    lookup_key->encoding = UTF32;
    memcpy(lookup_key->text, beijing, sizeof(beijing));
    HASH_FIND( hh, msgs, &lookup_key->encoding, keylen, msg );
    if (msg) printf("found \n");
    free(lookup_key);
}
```

This example is included in the distribution in `tests/test22.c`.

If you use multi-field keys, recognize that the compiler pads adjacent fields (by inserting unused space between them) in order to fulfill the alignment requirement of each field. For example a structure containing a `char` followed by an `int` will normally have 3 "wasted" bytes of padding after the char, in order to make the `int` field start on a multiple-of-4 address (4 is the length of the int).

Calculating the length of a multi-field key: To determine the key length when using a multi-field key, you must include any intervening structure padding the compiler adds for alignment purposes.

An easy way to calculate the key length is to use the `offsetof` macro from `<stddef.h>`. The formula is:

```
key length =   offsetof(last_key_field)
             + sizeof(last_key_field)
             - offsetof(first_key_field)
```

In the example above, the `keylen` variable is set using this formula.

When dealing with a multi-field key, you must zero-fill your structure before `HASH_ADD`'ing it to a hash table, or using its fields in a `HASH_FIND` key.

In the previous example, `memset` is used to initialize the structure by zero-filling it. This zeroes out any padding between the key fields. If we didn't zero-fill the structure, this padding would contain random values. The random values would lead to `HASH_FIND` failures; as two "identical" keys will appear to mismatch if there are any differences within their padding.

5.2 Items in several hash tables

A structure can be added to more than one hash table. A few reasons you might do this include:

- each hash table may use an alternative key;
- each hash table may have its own sort order;
- or you might simply use multiple hash tables for grouping purposes. E.g., you could have users in an `admin_users` and a `users` hash table.

Your structure needs to have a `UT_hash_handle` field for each hash table to which it might be added. You can name them anything. E.g.,

```
UT_hash_handle hh1, hh2;
```

5.3 Items with alternative keys

You might create a hash table keyed on an ID field, and another hash table keyed on username (if usernames are unique). You can add the same user structure to both hash tables (without duplication of the structure), allowing lookup of a user structure by their name or ID. The way to achieve this is to have a separate `UT_hash_handle` for each hash to which the structure may be added.

Example 5.2 A structure with two alternative keys

```
struct my_struct {
    int id;                    /* usual key */
    char username[10];         /* alternative key */
    UT_hash_handle hh1;        /* handle for first hash table */
    UT_hash_handle hh2;        /* handle for second hash table */
};
```

In the example above, the structure can now be added to two separate hash tables. In one hash, `id` is its key, while in the other hash, `username` is its key. (There is no requirement that the two hashes have different key fields. They could both use the same key, such as `id`).

Notice the structure has two hash handles (`hh1` and `hh2`). In the code below, notice that each hash handle is used exclusively with a particular hash table. (`hh1` is always used with the `users_by_id` hash, while `hh2` is always used with the `users_by_name` hash table).

Example 5.3 Two keys on a structure

```
    struct my_struct *users_by_id = NULL, *users_by_name = NULL, *s;
    int i;
    char *name;

    s = malloc(sizeof(struct my_struct));
    s->id = 1;
    strcpy(s->username, "thanson");

    /* add the structure to both hash tables */
    HASH_ADD(hh1, users_by_id, id, sizeof(int), s);
    HASH_ADD(hh2, users_by_name, username, strlen(s->username), s);

    /* lookup user by ID in the "users_by_id" hash table */
    i=1;
    HASH_FIND(hh1, users_by_id, &i, sizeof(int), s);
    if (s) printf("found id %d: %s\n", i, s->username);

    /* lookup user by username in the "users_by_name" hash table */
    name = "thanson";
    HASH_FIND(hh2, users_by_name, name, strlen(name), s);
    if (s) printf("found user %s: %d\n", name, s->id);
```

5.4 Several sort orders

It comes as no suprise that two hash tables can have different sort orders, but this fact can also be used advantageously to sort the *same items* in several ways. This is based on the ability to store a structure in several hash tables.

Extending the previous example, suppose we have many users. We have added each user structure to the `users_by_id` hash table and the `users_by_name` hash table. (To reiterate, this is done without the need to have two copies of each structure). Now we can define two sort functions, then use `HASH_SRT`.

```
int sort_by_id(struct my_struct *a, struct my_struct *b) {
  if (a->id == b->id) return 0;
  return (a->id < b->id) ? -1 : 1;
}

int sort_by_name(struct my_struct *a, struct my_struct *b) {
  return strcmp(a->username,b->username);
}

HASH_SRT(hh1, users_by_id, sort_by_id);
HASH_SRT(hh2, users_by_name, sort_by_name);
```

Now iterating over the items in `users_by_id` will traverse them in id-order while, naturally, iterating over `users_by_name` will traverse them in name-order. The items are fully forward-and-backward linked in each order. So even for one set of users, we might store them in two hash tables to provide easy iteration in two different sort orders.

5.5 Bloom filter (faster misses)

Programs that generate a fair miss rate (`HASH_FIND` that result in `NULL`) may benefit from the built-in Bloom filter support. This is disabled by default, because programs that generate only hits would incur a slight penalty from it. Also, programs that do deletes should not use the Bloom filter. While the program would operate correctly, deletes diminish the benefit of the filter. To enable the Bloom filter, simply compile with `-DHASH_BLOOM=n` like:

```
-DHASH_BLOOM=27
```

where the number can be any value up to 32 which determines the amount of memory used by the filter, as shown below. Using more memory makes the filter more accurate and has the potential to speed up your program by making misses bail out faster.

Table 1: Bloom filter sizes for selected values of n

n	Bloom filter size (per hash table)
16	8 kilobytes
20	128 kilobytes
24	2 megabytes
28	32 megabytes
32	512 megabytes

Bloom filters are only a performance feature; they do not change the results of hash operations in any way. The only way to gauge whether or not a Bloom filter is right for your program is to test it. Reasonable values for the size of the Bloom filter are 16-32 bits.

5.6 Select

An experimental *select* operation is provided that inserts those items from a source hash that satisfy a given condition into a destination hash. This insertion is done with somewhat more efficiency than if this were using HASH_ADD, namely because the hash function is not recalculated for keys of the selected items. This operation does not remove any items from the source hash. Rather the selected items obtain dual presence in both hashes. The destination hash may already have items in it; the selected items are added to it. In order for a structure to be usable with HASH_SELECT, it must have two or more hash handles. (As described here, a structure can exist in many hash tables at the same time; it must have a separate hash handle for each one).

```
user_t *users=NULL, *admins=NULL; /* two hash tables */

typedef struct {
    int id;
    UT_hash_handle hh;  /* handle for users hash */
    UT_hash_handle ah;  /* handle for admins hash */
} user_t;
```

Now suppose we have added some users, and want to select just the administrator users who have id's less than 1024.

```
#define is_admin(x) (((user_t*)x)->id < 1024)
HASH_SELECT(ah,admins,hh,users,is_admin);
```

The first two parameters are the *destination* hash handle and hash table, the second two parameters are the *source* hash handle and hash table, and the last parameter is the *select condition*. Here we used a macro `is_admin()` but we could just as well have used a function.

```
int is_admin(void *userv) {
  user_t *user = (user_t*)userv;
  return (user->id < 1024) ? 1 : 0;
}
```

If the select condition always evaluates to true, this operation is essentially a *merge* of the source hash into the destination hash. Of course, the source hash remains unchanged under any use of HASH_SELECT. It only adds items to the destination hash selectively.

The two hash handles must differ. An example of using HASH_SELECT is included in `tests/test36.c`.

5.7 Built-in hash functions

Internally, a hash function transforms a key into a bucket number. You don't have to take any action to use the default hash function, currently Jenkin's.

Some programs may benefit from using another of the built-in hash functions. There is a simple analysis utility included with uthash to help you determine if another hash function will give you better performance.

You can use a different hash function by compiling your program with `-DHASH_FUNCTION=HASH_xyz` where `xyz` is one of the symbolic names listed below. E.g.,

```
cc -DHASH_FUNCTION=HASH_BER -o program program.c
```

Table 2: Built-in hash functions

Symbol	Name
JEN	Jenkins (default)
BER	Bernstein
SAX	Shift-Add-Xor
OAT	One-at-a-time
FNV	Fowler/Noll/Vo
SFH	Paul Hsieh
MUR	MurmurHash (see note)

MurmurHash
A special symbol must be defined if you intend to use MurmurHash. To use it, add `-DHASH_USING_NO_STRICT_ALIASING` to your `CFLAGS`. And, if you are using the gcc compiler with optimization, add `-fno-strict-aliasing` to your `CFLAGS`.

5.7.1 Which hash function is best?

You can easily determine the best hash function for your key domain. To do so, you'll need to run your program once in a data-collection pass, and then run the collected data through an included analysis utility.

First you must build the analysis utility. From the top-level directory,

```
cd tests/
make
```

We'll use `test14.c` to demonstrate the data-collection and analysis steps (here using `sh` syntax to redirect file descriptor 3 to a file):

Example 5.4 Using keystats

```
% cc -DHASH_EMIT_KEYS=3 -I../src -o test14 test14.c
% ./test14 3>test14.keys
% ./keystats test14.keys
fcn  ideal%     #items   #buckets  dup%  fl   add_usec  find_usec  del-all usec
---  ------ ---------- ---------- -----  -- ---------- ----------  ------------
SFH   91.6%       1219        256    0%  ok         92        131            25
FNV   90.3%       1219        512    0%  ok        107         97            31
SAX   88.7%       1219        512    0%  ok        111        109            32
OAT   87.2%       1219        256    0%  ok         99        138            26
JEN   86.7%       1219        256    0%  ok         87        130            27
BER   86.2%       1219        256    0%  ok        121        129            27
```

Note
The number 3 in `-DHASH_EMIT_KEYS=3` is a file descriptor. Any file descriptor that your program doesn't use for its own purposes can be used instead of 3. The data-collection mode enabled by `-DHASH_EMIT_KEYS=x` should not be used in production code.

Usually, you should just pick the first hash function that is listed. Here, this is `SFH`. This is the function that provides the most even distribution for your keys. If several have the same `ideal%`, then choose the fastest one according to the `find_usec` column.

5.7.2 keystats column reference

fcn
symbolic name of hash function

ideal%
The percentage of items in the hash table which can be looked up within an ideal number of steps. (Further explained below).

#items
the number of keys that were read in from the emitted key file

#buckets
the number of buckets in the hash after all the keys were added

dup%
the percent of duplicate keys encountered in the emitted key file. Duplicates keys are filtered out to maintain key uniqueness. (Duplicates are normal. For example, if the application adds an item to a hash, deletes it, then re-adds it, the key is written twice to the emitted file.)

flags
this is either `ok`, or `nx` (noexpand) if the expansion inhibited flag is set, described in Expansion internals. It is not recommended to use a hash function that has the `noexpand` flag set.

add_usec
the clock time in microseconds required to add all the keys to a hash

find_usec
the clock time in microseconds required to look up every key in the hash

del-all usec
the clock time in microseconds required to delete every item in the hash

5.7.3 ideal%

What is ideal%? The *n* items in a hash are distributed into *k* buckets. Ideally each bucket would contain an equal share *(n/k)* of the items. In other words, the maximum linear position of any item in a bucket chain would be *n/k* if every bucket is equally used. If some buckets are overused and others are underused, the overused buckets will contain items whose linear position surpasses *n/k*. Such items are considered non-ideal.

As you might guess, `ideal%` is the percentage of ideal items in the hash. These items have favorable linear positions in their bucket chains. As `ideal%` approaches 100%, the hash table approaches constant-time lookup performance.

5.8 hashscan

A **Linux-only** utility called `hashscan` is included in the `tests/` directory. It is built automatically when you run `make` in that directory. This tool examines a running process and reports on the uthash tables that it finds in that program's memory. It can also save the keys from each table in a format that can be fed into `keystats`.

Here is an example of using `hashscan`. First ensure that it is built:

```
cd tests/
make
```

Since `hashscan` needs a running program to inspect, we'll start up a simple program that makes a hash table and then sleeps as our test subject:

```
./test_sleep &
pid: 9711
```

Now that we have a test program, let's run `hashscan` on it:

```
./hashscan 9711
Address            ideal    items  buckets mc fl bloom/sat fcn keys saved to
------------------ ----- -------- -------- -- -- --------- --- -------------
0x862e038            81%    10000     4096 11 ok 16    14% JEN
```

If we wanted to copy out all its keys for external analysis using `keystats`, add the `-k` flag:

```
./hashscan -k 9711
Address            ideal    items  buckets mc fl bloom/sat fcn keys saved to
------------------ ----- -------- -------- -- -- --------- --- -------------
0x862e038            81%    10000     4096 11 ok 16    14% JEN /tmp/9711-0.key
```

Now we could run `./keystats /tmp/9711-0.key` to analyze which hash function has the best characteristics on this set of keys.

5.8.1 hashscan column reference

Address
: virtual address of the hash table

ideal
: The percentage of items in the table which can be looked up within an ideal number of steps. See Section 5.7.3 in the `keystats` section.

items
: number of items in the hash table

buckets
: number of buckets in the hash table

mc
: the maximum chain length found in the hash table (uthash usually tries to keep fewer than 10 items in each bucket, or in some cases a multiple of 10)

fl
: flags (either `ok`, or `NX` if the expansion-inhibited flag is set)

bloom/sat
if the hash table uses a Bloom filter, this is the size (as a power of two) of the filter (e.g. 16 means the filter is 2^16 bits in size). The second number is the "saturation" of the bits expressed as a percentage. The lower the percentage, the more potential benefit to identify cache misses quickly.

fcn
symbolic name of hash function

keys saved to
file to which keys were saved, if any

How hashscan works When hashscan runs, it attaches itself to the target process, which suspends the target process momentarily. During this brief suspension, it scans the target's virtual memory for the signature of a uthash hash table. It then checks if a valid hash table structure accompanies the signature and reports what it finds. When it detaches, the target process resumes running normally. The hashscan is performed "read-only"-- the target process is not modified. Since hashscan is analyzing a momentary snapshot of a running process, it may return different results from one run to another.

5.9 Expansion internals

Internally this hash manages the number of buckets, with the goal of having enough buckets so that each one contains only a small number of items.

Why does the number of buckets matter? When looking up an item by its key, this hash scans linearly through the items in the appropriate bucket. In order for the linear scan to run in constant time, the number of items in each bucket must be bounded. This is accomplished by increasing the number of buckets as needed.

5.9.1 Normal expansion

This hash attempts to keep fewer than 10 items in each bucket. When an item is added that would cause a bucket to exceed this number, the number of buckets in the hash is doubled and the items are redistributed into the new buckets. In an ideal world, each bucket will then contain half as many items as it did before.

Bucket expansion occurs automatically and invisibly as needed. There is no need for the application to know when it occurs.

Per-bucket expansion threshold

Normally all buckets share the same threshold (10 items) at which point bucket expansion is triggered. During the process of bucket expansion, uthash can adjust this expansion-trigger threshold on a per-bucket basis if it sees that certain buckets are over-utilized.

When this threshold is adjusted, it goes from 10 to a multiple of 10 (for that particular bucket). The multiple is based on how many times greater the actual chain length is than the ideal length. It is a practical measure to reduce excess bucket expansion in the case where a hash function over-utilizes a few buckets but has good overall distribution. However, if the overall distribution gets too bad, uthash changes tactics.

5.9.2 Inhibited expansion

You usually don't need to know or worry about this, particularly if you used the `keystats` utility during development to select a good hash for your keys.

A hash function may yield an uneven distribution of items across the buckets. In moderation this is not a problem. Normal bucket expansion takes place as the chain lengths grow. But when significant imbalance occurs (because the hash function is not well suited to the key domain), bucket expansion may be ineffective at reducing the chain lengths.

Imagine a very bad hash function which always puts every item in bucket 0. No matter how many times the number of buckets is doubled, the chain length of bucket 0 stays the same. In a situation like this, the best behavior is to stop expanding, and accept O(n) lookup performance. This is what uthash does. It degrades gracefully if the hash function is ill-suited to the keys.

If two consecutive bucket expansions yield `ideal%` values below 50%, uthash inhibits expansion for that hash table. Once set, the *bucket expansion inhibited* flag remains in effect as long as the hash has items in it. Inhibited expansion may cause `HASH_FIND` to exhibit worse than constant-time performance.

5.10 Hooks

You don't need to use these hooks- they are only here if you want to modify the behavior of uthash. Hooks can be used to change how uthash allocates memory, and to run code in response to certain internal events.

5.10.1 malloc/free

By default this hash implementation uses `malloc` and `free` to manage memory. If your application uses its own custom allocator, this hash can use them too.

Example 5.5 Specifying alternate memory management functions

```
#include "uthash.h"

/* undefine the defaults */
#undef uthash_malloc
#undef uthash_free

/* re-define, specifying alternate functions */
#define uthash_malloc(sz) my_malloc(sz)
#define uthash_free(ptr) my_free(ptr)

...
```

5.10.2 Out of memory

If memory allocation fails (i.e., the malloc function returned `NULL`), the default behavior is to terminate the process by calling `exit(-1)`. This can be modified by re-defining the `uthash_fatal` macro.

```
#undef uthash_fatal
#define uthash_fatal(msg) my_fatal_function(msg);
```

The fatal function should terminate the process or `longjmp` back to a safe place. Uthash does not support "returning a failure" if memory cannot be allocated.

5.10.3 Internal events

There is no need for the application to set these hooks or take action in response to these events. They are mainly for diagnostic purposes.

These two hooks are "notification" hooks which get executed if uthash is expanding buckets, or setting the *bucket expansion inhibited* flag. Normally both of these hooks are undefined and thus compile away to nothing.

Expansion

There is a hook for the bucket expansion event.

Example 5.6 Bucket expansion hook

```
#include "uthash.h"

#undef uthash_expand_fyi
#define uthash_expand_fyi(tbl) printf("expanded to %d buckets\n", tbl->num_buckets)

...
```

Expansion-inhibition

This hook can be defined to code to execute in the event that uthash decides to set the *bucket expansion inhibited* flag.

Example 5.7 Bucket expansion inhibited hook

```
#include "uthash.h"

#undef uthash_noexpand_fyi
#define uthash_noexpand_fyi printf("warning: bucket expansion inhibited\n");

...
```

5.11 Debug mode

If a program that uses this hash is compiled with `-DHASH_DEBUG=1`, a special internal consistency-checking mode is activated. In this mode, the integrity of the whole hash is checked following every add or delete operation. This is for debugging the uthash software only, not for use in production code.

In the `tests/` directory, running `make debug` will run all the tests in this mode.

In this mode, any internal errors in the hash data structure will cause a message to be printed to `stderr` and the program to exit.

The `UT_hash_handle` data structure includes `next`, `prev`, `hh_next` and `hh_prev` fields. The former two fields determine the "application" ordering (that is, insertion order-- the order the items were added). The latter two fields determine the "bucket chain" order. These link the `UT_hash_handles` together in a doubly-linked list that is a bucket chain.

Checks performed in `-DHASH_DEBUG=1` mode:

- the hash is walked in its entirety twice: once in *bucket* order and a second time in *application* order
- the total number of items encountered in both walks is checked against the stored number
- during the walk in *bucket* order, each item's `hh_prev` pointer is compared for equality with the last visited item
- during the walk in *application* order, each item's `prev` pointer is compared for equality with the last visited item

Macro debugging: Sometimes it's difficult to interpret a compiler warning on a line which contains a macro call. In the case of uthash, one macro can expand to dozens of lines. In this case, it is helpful to expand the macros and then recompile. By doing so, the warning message will refer to the exact line within the macro.

Here is an example of how to expand the macros and then recompile. This uses the `test1.c` program in the `tests/` subdirectory.

```
gcc -E -I../src test1.c > /tmp/a.c
egrep -v '^#' /tmp/a.c > /tmp/b.c
indent /tmp/b.c
gcc -o /tmp/b /tmp/b.c
```

The last line compiles the original program (test1.c) with all macros expanded. If there was a warning, the referenced line number can be checked in `/tmp/b.c`.

5.12 Thread safety

You can use uthash in a threaded program. But you must do the locking. Use a read-write lock to protect against concurrent writes. It is ok to have concurrent readers (since uthash 1.5).

For example using pthreads you can create an rwlock like this:

```
pthread_rwlock_t lock;
if (pthread_rwlock_init(&lock,NULL) != 0) fatal("can't create rwlock");
```

Then, readers must acquire the read lock before doing any `HASH_FIND` calls or before iterating over the hash elements:

```
if (pthread_rwlock_rdlock(&lock) != 0) fatal("can't get rdlock");
HASH_FIND_INT(elts, &i, e);
pthread_rwlock_unlock(&lock);
```

Writers must acquire the exclusive write lock before doing any update. Add, delete, and sort are all updates that must be locked.

```
if (pthread_rwlock_wrlock(&lock) != 0) fatal("can't get wrlock");
HASH_DEL(elts, e);
pthread_rwlock_unlock(&lock);
```

If you prefer, you can use a mutex instead of a read-write lock, but this will reduce reader concurrency to a single thread at a time.

An example program using uthash with a read-write lock is included in `tests/threads/test1.c`.

6 Macro reference

6.1 Convenience macros

The convenience macros do the same thing as the generalized macros, but require fewer arguments.

In order to use the convenience macros,

1. the structure's `UT_hash_handle` field must be named `hh`, and
2. for add or find, the key field must be of type `int` or `char[]`

Table 3: Convenience macros

macro	arguments
HASH_ADD_INT	(head, keyfield_name, item_ptr)
HASH_FIND_INT	(head, key_ptr, item_ptr)
HASH_ADD_STR	(head, keyfield_name, item_ptr)
HASH_FIND_STR	(head, key_ptr, item_ptr)
HASH_DEL	(head, item_ptr)
HASH_SORT	(head, cmp)
HASH_COUNT	(head)

6.2 General macros

These macros add, find, delete and sort the items in a hash. You need to use the general macros if your `UT_hash_handle` is named something other than `hh`, or if your key's data type isn't `int` or `char[]`.

Table 4: General macros

macro	arguments
HASH_ADD	(hh_name, head, keyfield_name, key_len, item_ptr)
HASH_ADD_KEYPTR	(hh_name, head, key_ptr, key_len, item_ptr)
HASH_FIND	(hh_name, head, key_ptr, key_len, item_ptr)
HASH_DELETE	(hh_name, head, item_ptr)
HASH_SRT	(hh_name, head, cmp)
HASH_CNT	(hh_name, head)
HASH_CLEAR	(hh_name, head)
HASH_SELECT	(dst_hh_name, dst_head, src_hh_name, src_head, condition)

Note
HASH_ADD_KEYPTR is used when the structure contains a pointer to the key, rather than the key itself.

6.2.1 Argument descriptions

hh_name
name of the `UT_hash_handle` field in the structure. Conventionally called `hh`.

head
the structure pointer variable which acts as the "head" of the hash. So named because it initially points to the first item that is added to the hash.

keyfield_name
the name of the key field in the structure. (In the case of a multi-field key, this is the first field of the key). If you're new to macros, it might seem strange to pass the name of a field as a parameter. See note.

key_len
the length of the key field in bytes. E.g. for an integer key, this is `sizeof(int)`, while for a string key it's `strlen(key)`. (For a multi-field key, see the notes in this guide on calculating key length).

key_ptr
for HASH_FIND, this is a pointer to the key to look up in the hash (since it's a pointer, you can't directly pass a literal value here). For HASH_ADD_KEYPTR, this is the address of the key of the item being added.

item_ptr
pointer to the structure being added, deleted or looked up. This is an input parameter for HASH_ADD and HASH_DELETE macros, and an output parameter for HASH_FIND.

cmp
pointer to comparison function which accepts two arguments (pointers to items to compare) and returns an int specifying whether the first item should sort before, equal to, or after the second item (like strcmp).

condition
a function or macro which accepts a single argument-- a void pointer to a structure, which needs to be cast to the appropriate structure type. The function or macro should return (or evaluate to) a non-zero value if the structure should be "selected" for addition to the destination hash.