Mini-Ulix: Übungsblatt 9 (Aufgabe 4)

Hans-Georg Eßer, FOM Hochschule

26. Mai 2015

Inhaltsverzeichnis

1 Kernel-Layout

Das Layout der C-Quellcode-Datei `ulix.c` unseres Mini-Kernels sieht schon – grob – wie beim richtigen ULIX aus:

3a ⟨*ulix.c* 3a⟩≡
```
⟨constants 8b⟩
⟨type definitions 3d⟩
⟨global variables 3c⟩
⟨macros 8a⟩
⟨function prototypes 4b⟩
⟨function implementations 4e⟩
⟨kernel main 3b⟩
```

Auf einen Abdruck der Assembler-Datei `start.asm` verzichten wir in dieser Ausarbeitung.

Die `main`-Funktion hat den folgenden Aufbau:

3b ⟨*kernel main* 3b⟩≡ (3a)
```
int main () {
  ⟨kernel main: initialize variables 15e⟩
  printf ("[1] entering main()\n");

  // Speicherverwaltung
  ⟨kernel main: setup paging 8c⟩
  ⟨kernel main: install GDT 5a⟩  // replace "trick GDT" with regular GDT
  paging_ready = true; printf ("[4] regular GDT is active\n");
  ⟨kernel main: map physical RAM 14c⟩
  ⟨kernel main: setup frame table 15a⟩

  ⟨kernel main: initialize system 22c⟩   // <-  !!
  ⟨kernel main: user-defined tests 32e⟩
  for (;;);   // inifinite loop
}
```

Die ersten Code-Chunks befassen sich mit dem Einrichten der Speicherverwaltung; dieses Thema haben wir in der Vorlesung nicht behandelt. Für uns interessant wird es erst im Code Chunk ⟨*kernel main: initialize system* 22c⟩. Springen Sie zu Kapitel 7 (ab Seite 19).

3c ⟨*global variables* 3c⟩≡ (3a) 4a▷
```
int paging_ready = false;
```

2 Speicher einrichten: Die GDT

Zunächst brauchen wir die Datenstrukturen. Die GDT-Einträge haben die in der Vorlesung vorgestellte Struktur, bei der die Werte `base` und `limit` umständlich auf mehrere Komponenten verteilt sind.

Für den Zugriff auf eine GDT braucht es zudem immer eine Art Zeiger auf den Tabellenanfang, der auch die Größe der GDT festlegt (Typ `struct gdt_ptr`).

3d ⟨*type definitions* 3d⟩≡ (3a) 5b▷

```
struct gdt_entry {
  unsigned int limit_low    : 16;
  unsigned int base_low     : 16;
  unsigned int base_middle  :  8;
  unsigned int access       :  8;
  unsigned int flags        :  4;
  unsigned int limit_high   :  4;
  unsigned int base_high    :  8;
};

struct gdt_ptr {
  unsigned int limit        : 16;
  unsigned int base         : 32;
} __attribute__((packed));
```

Hier kommt die GDT in den Speicher:

4a ⟨*global variables* 3c⟩+≡ (3a) ◁3c 6b▷

```
struct gdt_entry gdt[6];
struct gdt_ptr gp;
```

Die GDT zu "flushen" ist eine Aufgabe für eine Funktion, die in der Assembler-Datei definiert ist. Wir definieren Sie hier als extern:

4b ⟨*function prototypes* 4b⟩≡ (3a) 4c▷

```
extern void gdt_flush();
```

Um einen GDT-Eintrag mit Werten zu füllen, verwenden wir die Funktion

4c ⟨*function prototypes* 4b⟩+≡ (3a) ◁4b 4d▷

```
void gdt_set_gate (int num, unsigned long base,
  unsigned long limit, unsigned char access, unsigned char gran);
```

die `base` und `limit` aufteilt und zusammen mit den Zugriffsrechten (`access`) und den Flags (in `gran`) an den richtigen Stellen einträgt. Wir benutzen Sie bei der Initialisierung mit der Funktion

4d ⟨*function prototypes* 4b⟩+≡ (3a) ◁4c 7a▷

```
void gdt_install ();
```

welche die drei benötigten Segmentdeskriptoren erzeugt (einen Nulldeskriptor und die beiden echten Deskriptoren für das Code- und das Daten-Segment).

Hier sind die Implementierungen:

4e ⟨*function implementations* 4e⟩≡ (3a) 7c▷

```
void gdt_set_gate(int num, unsigned long base, unsigned long limit,
                  unsigned char access, unsigned char gran) {
  /* Setup the descriptor base address */
  gdt[num].base_low = (base & 0xFFFF);            // 16 bits
  gdt[num].base_middle = (base >> 16) & 0xFF;     //  8 bits
  gdt[num].base_high = (base >> 24) & 0xFF;       //  8 bits

  /* Setup the descriptor limits */
  gdt[num].limit_low  = (limit & 0xFFFF);         // 16 bits
  gdt[num].limit_high = ((limit >> 16) & 0x0F);   //  4 bits
```

```
  /* Finally, set up the granularity and access flags */
  gdt[num].flags = gran & 0xF;
  gdt[num].access = access;
}

void gdt_install() {
  gp.limit = (sizeof(struct gdt_entry) * 6) - 1;
  gp.base = (int) &gdt;

  gdt_set_gate(0, 0, 0, 0, 0);    // NULL descriptor

  // code segment
  gdt_set_gate(1, 0, 0xFFFFFFFF, 0b10011010, 0b1100 /* 0xCF */);

  // data segment
  gdt_set_gate(2, 0, 0xFFFFFFFF, 0b10010010, 0b1100 /* 0xCF */);

  gdt_flush();
}
```

Im Hauptprogramm (`main`) rufen wir `gdt_install()` dann später auf:

5a ⟨*kernel main: install GDT* 5a⟩≡ (3b)

```
gdt_install ();
```

3 Paging

Wir beginnen wieder mit den benötigten Datenstrukturen.

3.1 Page Directory und Page Table

Wir definieren zwei Datenstrukturen:

- Auf der obersten Ebene gibt es das *Page Directory* (`page_directory`), das aus 1024 *Page Table Descriptors* (`page_table_desc`) besteht:

5b ⟨*type definitions* 3d⟩+≡ (3a) ◁3d 6a▷

```
typedef struct {
  unsigned int present         : 1;  //  0
  unsigned int writeable       : 1;  //  1
  unsigned int user_accessible : 1;  //  2
  unsigned int pwt             :  1;  //  3
  unsigned int pcd             :  1;  //  4
  unsigned int accessed        :  1;  //  5
  unsigned int undocumented    :  1;  //  6
  unsigned int zeroes          :  2;  //  8.. 7
  unsigned int unused_bits     :  3;  // 11.. 9
  unsigned int frame_addr      : 20;  // 31..12
} page_table_desc;

typedef struct { page_table_desc ptds[1024]; } page_directory;
```

Definiert:
page_directory, benutzt im Teil 6b.
page_table_desc, benutzt im Teil 7.

- Jeder Page Table Descriptor zeigt auf eine *Page Table* (page_table), welche wiederum aus 1024 *Page Descriptors* (page_desc) besteht:

6a ⟨*type definitions* 3d⟩+≡ (3a) ◁5b 11a▷

```
typedef struct {
  unsigned int present         : 1;  //  0
  unsigned int writeable       : 1;  //  1
  unsigned int user_accessible : 1;  //  2
  unsigned int pwt             :  1;  //  3
  unsigned int pcd             :  1;  //  4
  unsigned int accessed        :  1;  //  5
  unsigned int dirty           :  1;  //  6
  unsigned int zeroes          :  2;  //  8.. 7
  unsigned int unused_bits     :  3;  // 11.. 9
  unsigned int frame_addr      : 20;  // 31..12
} page_desc;

typedef struct { page_desc pds[1024]; } page_table;
```

Definiert:
page_desc, benutzt im Teil 7.
page_table, benutzt im Teils 6b, 11c, 13, und 14b.

- Die Page Descriptors zeigen schließlich auf die *Page Frames*, also 4 KByte große Blöcke im Hauptspeicher, welche unsere Seiten aufnehmen.

In dieser Miniversion von ULIX verwenden wir nur jeweils ein einziges Page Directory und 17 Page Tables, für die wir zunächst Speicher reservieren und dann noch Pointer darauf einrichten:

6b ⟨*global variables* 3c⟩+≡ (3a) ◁4a 9b▷

```
page_directory kernel_pd     __attribute__ ((aligned (4096)));
page_table kernel_pt         __attribute__ ((aligned (4096)));
page_table kernel_pt_ram[16] __attribute__ ((aligned (4096)));

page_directory* current_pd = &kernel_pd;
page_table*     current_pt = &kernel_pt;
```

Definiert:
current_pd, benutzt im Teils 8, 9a, 11c, 13, und 14.
current_pt, benutzt im Teil 8.
kernel_pd, nicht benutzt.
kernel_pt, nicht benutzt.
Benutzt page_directory 5b und page_table 6a.

Hier ist wichtig, dass die Datenstrukturen über das Attribut `((aligned (4096)))` jeweils ab einem Vielfachen von 4096 (der Seitengröße) beginnen; sie dürfen nicht "quer" im Speicher liegen.

Das Page Directory kernel_pd und die Page Table kernel_pt werden wir gleich nutzen, um das Paging an sich zu aktivieren; die 16 Page Tables im Array kernel_pt_ram brauchen wir später, um den Hauptspeicher (64 MByte) nach 0xD0000000... zu mappen.

3.2 Funktionen für den Zugriff

Wir definieren zwei Funktionen

7a ⟨*function prototypes* 4b⟩+≡ (3a) ◁4d 7b▷
```
page_table_desc* fill_page_table_desc (page_table_desc *ptd,
  unsigned int present, unsigned int writeable,
  unsigned int user_accessible, unsigned int frame_addr);
```
Benutzt page_table_desc 5b.

und

7b ⟨*function prototypes* 4b⟩+≡ (3a) ◁7a 10a▷
```
page_desc* fill_page_desc (page_desc *pd, unsigned int present,
  unsigned int writeable, unsigned int user_accessible,
  unsigned int dirty, unsigned int frame_addr);
```
Benutzt page_desc 6a.

mit denen wir einen Page Table Descriptor (oberste Ebene) bzw. einen Page Descriptor (zweite Ebene) mit Inhalten füllen können. Die beiden Funktionen sind ähnlich aufgebaut und unterscheiden sich nur darin, dass `fill_page_desc()` ein zusätzliches Argument akzeptiert, mit dem das `dirty`-Flag gesetzt werden kann: Das gibt es nur bei Page Descriptors, der Page Table Descriptor hat an derselben Stelle einen undokumentierten Eintrag, den wir immer auf 0 setzen.

Hier sind die Implementierungen:

7c ⟨*function implementations* 4e⟩+≡ (3a) ◁4e 10c▷
```
page_desc* fill_page_desc (page_desc *pd, unsigned int present,
  unsigned int writeable, unsigned int user_accessible,
  unsigned int dirty, unsigned int frame_addr) {

  memset (pd, 0, sizeof(pd));

  pd->present = present;
  pd->writeable = writeable;
  pd->user_accessible = user_accessible;
  pd->dirty = dirty;
  pd->frame_addr = frame_addr >> 12;   // right shift, 12 bits
  return pd;
};

page_table_desc* fill_page_table_desc (page_table_desc *ptd,
  unsigned int present, unsigned int writeable,
  unsigned int user_accessible, unsigned int frame_addr) {

  memset (ptd, 0, sizeof(ptd));

  ptd->present = present;
  ptd->writeable = writeable;
  ptd->user_accessible = user_accessible;
  ptd->frame_addr = frame_addr >> 12;   // right shift, 12 bits
  return ptd;
};
```
Benutzt page_desc 6a und page_table_desc 5b.

Die hier verwendete Funktion `memset` beschreiben wir am Ende dieses Textes.

Um beide Funktionen leichter mit Standardparametern aufrufen zu können, führen wir noch zwei Makros

8a ⟨*macros* 8a⟩≡ (3a) 9d▷
```
#define KMAP(pd,frame) \
  fill_page_desc (pd, true, true, false, false, frame)
#define KMAPD(ptd, frame) \
  fill_page_table_desc (ptd, true, true, false, frame)
```

ein, die Einträge für die Nutzung durch den Kernel erzeugen. Später, wenn wir den User Mode einführen, wird es ganz ähnlich aussehende Makros `UMAP` und `UMAPD` geben, die sich nur darin von `KMAP` und `KMAPD` unterscheiden, dass sie `user_accessible` jeweils auf `true` statt `false` setzen. Übrigens müssen wir diese beiden booleschen Konstanten auch erst definieren:

8b ⟨*constants* 8b⟩≡ (3a) 9c▷
```
#define false 0
#define true 1
```

3.3 Initialisierung des Paging

Beim Paging richten wir die Seitentabellen zunächst so ein, dass sie ordentlich mit unserer Trick-GDT zusammen spielen, welche über einen Base-Wert von `0x40000000` alle Adressen ab `0xC0000000` in Adressen ab `0` umrechnet.

Zunächst bereiten wir das Page Directory `current_pd` vor, indem wir es mit Null-Einträgen füllen:

8c ⟨*kernel main: setup paging* 8c⟩≡ (3b) 8d▷
```
for (int i=1; i<1024; i++) {
  fill_page_table_desc (&(current_pd->ptds[i]), false, false, false, 0);
};
```
Benutzt `current_pd` 6b.

Dabei lassen wir den ersten Eintrag aus, den wir separat bearbeiten: Wir lassen den Eintrag `0` und den Eintrag `768` auf die Seitentabelle `current_pt` zeigen, damit wir zwei Mappings auf die ersten 4 MB RAM haben, einmal ab `0` und einmal ab `0xC0000000`:

8d ⟨*kernel main: setup paging* 8c⟩+≡ (3b) ◁8c 8e▷
```
KMAPD ( &(current_pd->ptds[  0]), (unsigned int)(current_pt)-0xC0000000 );
KMAPD ( &(current_pd->ptds[768]), (unsigned int)(current_pt)-0xC0000000 );
```
Benutzt `current_pd` 6b und `current_pt` 6b.

In der Seitentabelle `current_pt` tragen wir für die ersten 1023 Rahmen das Mapping ein:

8e ⟨*kernel main: setup paging* 8c⟩+≡ (3b) ◁8d 9a▷
```
for (int i=0; i<1023; i++) {
  KMAP ( &(current_pt->pds[i]), i*4096 );
};
printf ("[2] page directory setup, with identity mapping\n");
```
Benutzt `current_pt` 6b.

Schließlich aktivieren wir das Paging, indem wir die Control Registers `cr0` und `cr3` mit passenden Werten füllen bzw. aktualisieren:

9a ⟨*kernel main: setup paging* 8c⟩+≡ (3b) ◁8e

```
unsigned int cr0;
char *kernel_pd_address;
kernel_pd_address = (char*)(current_pd) - 0xC0000000;
asm volatile ("mov %0, %%cr3" : : "r"(kernel_pd_address));
  // write CR3
asm volatile ("mov %%cr0, %0" : "=r"(cr0) : );  // read  CR0
cr0 |= (1<<31);      // Enable paging by setting PG bit 31 of CR0
asm volatile ("mov %0, %%cr0" : : "r"(cr0) );   // write CR0
printf ("[3] paging activated.\n");
```

Benutzt `current_pd` 6b.

4 Verwaltung der Rahmen

Bisher haben wir nur gezeigt, wie das System das Paging vorbereitet und aktiviert – im regulären Betrieb müssen wir aber auch dynamisch neuen Speicher allozieren. Es wird also nötig sein, Seiten anzufordern und wieder freizugeben. Als ersten Schritt in diese Richtung brauchen wir eine Verwaltung des physischen Speichers: Hier geht es um freie Seitenrahmen (*Frames*).

4.1 Datenstruturen

Wir legen eine Rahmentabelle `ftable` an, die für jeden der 64 MByte / 4 KByte = 16 K Rahmen des Hauptspeichers ein Bit enthält: Ist es `0`, ist der Rahmen frei; anderenfalls ist er belegt. Die Anzahl der noch verfügbaren Rahmen merken wir uns in einer Variablen `free_frames`:

9b ⟨*global variables* 3c⟩+≡ (3a) ◁6b 17a▷

```
unsigned int free_frames = NUMBER_OF_FRAMES;
char place_for_ftable[NUMBER_OF_FRAMES/8];
unsigned int* ftable = (unsigned int*)(&place_for_ftable);
```

Hier benutzen wir gleich zweimal die Konstante `NUMBER_OF_FRAMES`, die wir bisher nicht erwähnt haben. Sie berechnet sich, wie oben beschrieben, aus der Speichergröße `MEM_SIZE` und der Größe einer Seite `PAGE_SIZE`:

9c ⟨*constants* 8b⟩+≡ (3a) ◁8b 18b▷

```
#define MEM_SIZE 1024*1024*64        // 64 MByte
#define MAX_ADDRESS MEM_SIZE-1       // last valid physical address
#define PAGE_SIZE 4096               // Intel: 4K pages
#define NUMBER_OF_FRAMES MEM_SIZE/PAGE_SIZE
```

4.2 Funktionen für den Zugriff

Um auf die einzelnen Bits in `ftable` zugreifen zu können, definieren wir zunächs zwei Helfer-Makros:

9d ⟨*macros* 8a⟩+≡ (3a) ◁8a 12a▷

```
#define INDEX_FROM_BIT(b) (b/32)   // 32 bits in an unsigned int
#define OFFSET_FROM_BIT(b) (b%32)
```

`INDEX_FROM_BIT` findet für eine Rahmennummer zunächst heraus, in welchem der je 32 Bit breiten Integers der Tabelle sich das Bit befindet; im zweiten Schritt berechnet `OFFSET_FROM_BIT` dann die Position innerhalb dieses Integers. Mit diesen beiden Helfer-Makros können wir nun drei Funktionen

10a ⟨*function prototypes* 4b⟩+≡ (3a) ◁7b 10d▷

```
static void set_frame (unsigned int frame_addr);
static void clear_frame (unsigned int frame_addr);
static unsigned int test_frame (unsigned int frame_addr);
```

implementieren, mit denen wir einzelne Bits setzen, zurücksetzen oder abfragen können. Alle nutzen zunächst die Makros, um aus der Frame-Adresse `frame_addr` die zwei Werte `index` und `offset` zu berechnen

10b ⟨*calculate index and offset* 10b⟩≡ (10c)

```
unsigned int frame = frame_addr / PAGE_SIZE;
unsigned int index  = INDEX_FROM_BIT  (frame);
unsigned int offset = OFFSET_FROM_BIT (frame);
```

und dann das jeweilige Bit zu verändern oder auszulesen:

10c ⟨*function implementations* 4e⟩+≡ (3a) ◁7c 10e▷

```
static void set_frame (unsigned int frame_addr) {
  ⟨calculate index and offset 10b⟩
  ftable[index] |= (1 << offset);
}

static void clear_frame (unsigned int frame_addr) {
  ⟨calculate index and offset 10b⟩
  ftable[index] &= ~(1 << offset);
}

static unsigned int test_frame (unsigned int frame_addr) {
  // returns true if frame is in use (false if frame is free)
  ⟨calculate index and offset 10b⟩
  return ((ftable[index] & (1 << offset)) >> offset);
}
```

Damit ist der Zugriff auf die Frame-Tabelle geregelt, jetzt können wir Funktionen entwickeln, mit denen der Kernel explizit einen neuen Frame (für die eigene Nutzung) anfordert oder wieder freigibt. Wir nennen diese Funktionen

10d ⟨*function prototypes* 4b⟩+≡ (3a) ◁10a 11b▷

```
int request_new_frame ();
void release_frame (unsigned int frameaddr);
```

– die letzte der beiden nimmt eine physische Speicheradresse als Argument, wie sie von der ersten zurück gegeben wird. Der Aufruf

```
release_frame ( request_new_frame () );
```

sollte also neutral sein. Zur Implementierung ist nur bei `request_new_frame()` ein wenig Arbeit nötig, weil hier die Frame-Tabelle durchsucht wird:

10e ⟨*function implementations* 4e⟩+≡ (3a) ◁10c 11c▷

```
int request_new_frame () {
  unsigned int frameid;
  boolean found=false;
  for (frameid = 0; frameid < NUMBER_OF_FRAMES; frameid++) {
    if ( !test_frame (frameid*4096) ) {
      found=true;
      break;   // frame found
    };
  }
  if (found) {
    memset ((void*)PHYSICAL(frameid << 12), 0, PAGE_SIZE);
    set_frame (frameid*4096);
    free_frames--;
    return frameid;
  } else {
    return -1;
  }
};

void release_frame (unsigned int frameaddr) {
  if ( test_frame (frameaddr) ) {
    // only do work if frame is marked as used
    clear_frame (frameaddr);
    free_frames++;
  };
};
```

Den hier benutzten Typ `boolean` müssen wir noch deklarieren:

11a ⟨*type definitions* 3d⟩+≡ (3a) ◁6a 20d▷

```
typedef unsigned int boolean;
```

5 Verwaltung der Seiten

Ohne Rahmen keine Seiten: Nachdem nun die Verwaltung der Frames funktioniert, können wir uns der Vergabe von Seiten zuwenden.

Die Datenstrukturen für das Paging sind schon vorhanden, die haben wir bei der Initialisierung des Pagings besprochen. Hier geht es nun darum, im laufenden Betrieb dynamisch neue Seiten anzufordern und diese wieder zurückzugeben – beides ist nur möglich, indem auch Frames angefordert und freigegeben werden, und wir müssen dazu auch das Page Directory und die Page Tables überarbeiten sowie ggf. neue Page Tables erzeugen.

Wir starten mit der einfachen Funktion

11b ⟨*function prototypes* 4b⟩+≡ (3a) ◁10d 12b▷

```
unsigned int pageno_to_frameno (unsigned int pageno);
```

welche für bereits "gemappten" virtuellen Speicher eine Seitennummer in die zugehörige Rahmennummer umrechnet. Das funktioniert genauso wie in der MMU (Memory Management Unit):

11c ⟨*function implementations* 4e⟩+≡ (3a) ◁10e 12c▷

```
unsigned int pageno_to_frameno (unsigned int pageno) {
  unsigned int pdindex = pageno/1024;
  unsigned int ptindex = pageno%1024;
  if ( ! current_pd->ptds[pdindex].present ) {
    return -1;       // we don't have that page table
  } else {
    // get the page table
    page_table* pt = (page_table*)
      ( PHYSICAL(current_pd->ptds[pdindex].frame_addr << 12) );
    if ( pt->pds[ptindex].present ) {
      return pt->pds[ptindex].frame_addr;
    } else {
      return -1;     // we don't have that page
    };
  };
};
```

Benutzt `current_pd` 6b und `page_table` 6a.

Die Funktion verwendet das Makro `PHYSICAL`, das einfach zu jeder Adresse den Wert `0xD0000000` addiert, um über das Mapping des physischen Speichers in den virtuellen Adressraum (ab `0xD0000000`) auf den Hauptspeicher zuzugreifen:

12a ⟨*macros* 8a⟩+≡ (3a) ◁9d 18a▷

```
#define PHYSICAL(x) ((x)+0xd0000000)
```

Wenn kein zugeordneter Frame gefunden wird (entweder weil es schon im Page Directory oder in der richtigen Page Table keinen Eintrag gibt), gibt die Funktion `-1` zurück.

Jetzt wird es kompliziert: Wir implementieren nun die Funktion

12b ⟨*function prototypes* 4b⟩+≡ (3a) ◁11b 14a▷

```
unsigned int* request_new_page (int need_more_pages);
```

die eine neue Seite anfordert. Der Parameter `need_more_pages` wird in der aktuellen Version des Codes noch nicht ausgewertet; er dient später dazu, mehrere zusammenhängende Seiten anzufordern.

Die Funktion besorgt sich zunächst einen frischen Frame und sucht dann nach einer freien Seitennummer:

12c ⟨*function implementations* 4e⟩+≡ (3a) ◁11c 14b▷

```
unsigned int* request_new_page (int need_more_pages) {
  ⟨page request implementation 12d⟩
}
```

12d ⟨*page request implementation* 12d⟩≡ (12c) 13a▷

```
unsigned int newframeid = request_new_frame ();
if (newframeid == -1) { return NULL; }  // exit if no frame was found
unsigned int pageno = -1;
for (unsigned int i=0xc0000; i<1024*1024; i++) {
  if ( pageno_to_frameno (i) == -1 ) {
    pageno = i;
    break;       // end loop, unmapped page was found
  };
};
```

```
if ( pageno == -1 ) {
  return NULL;   // we found no page -- whole 4 GB are mapped???
};
```

An dieser Stelle haben wir den Frame und die Seitennummer. Jetzt müssen wir das neue Mapping eintragen. Dazu berechnen wir zunächst die Positionen im Page Directory und der jeweiligen Page Table:

13a ⟨*page request implementation* 12d⟩+≡ (12c) ◁12d 13b▷

```
unsigned int pdindex = pageno/1024;
unsigned int ptindex = pageno%1024;
page_table* pt;
```

Benutzt page_table 6a.

Wenn `ptindex == 0` gilt, müssen wir eine neue Seitentabelle "anbrechen". Wir gehen hier davon aus, dass wir diese zunächst erstellen müssen. (Tatsächlich müssten wir prüfen, ob sie schon existiert – das könnte passieren, wenn wir Speicher wieder freigeben und dann erneut anfordern.) Für die neue Seitentabelle verwenden wir dann den bereits angeforderten Frame und benötigen danach einen neuen.

13b ⟨*page request implementation* 12d⟩+≡ (12c) ◁13a 13c▷

```
if (ptindex == 0) {
  // last entry! // create a new page table in the reserved frame
  page_table* pt = (page_table*) PHYSICAL(newframeid<<12);
  memset (pt, 0, PAGE_SIZE);
  KMAPD ( &(current_pd->ptds[pdindex]), newframeid << 12 );

  newframeid = request_new_frame ();  // get yet another frame
  if (newframeid == -1) {
    return NULL;                      // exit if no frame was found
    // note: we're not removing the new page table since we assume
    // it will be used soon anyway
  }
};
```

Benutzt current_pd 6b und page_table 6a.

Weiter geht es mit dem Eintragen des Mappings, dazu suchen wir zunächst die richtige Seitentabelle heraus (die über den Index `pdindex` im Page Directory festgelegt ist) und tragen an deren Position `ptindex` den Frame ein, wobei uns das Makro `KMAP` hilft:

13c ⟨*page request implementation* 12d⟩+≡ (12c) ◁13b

```
pt = (page_table*)( PHYSICAL(current_pd->ptds[pdindex].frame_addr << 12) );
// finally: enter the frame address
KMAP ( &(pt->pds[ptindex]), newframeid * PAGE_SIZE );

// invalidate cache entry
asm volatile ("invlpg %0" : : "m"(*(char*)(pageno<<12)) );

memset ((unsigned int*) (pageno*4096), 0, 4096);
return ((unsigned int*) (pageno*4096));
```

Benutzt current_pd 6b und page_table 6a.

(Am Ende invalidieren wir mit der CPU-Instruktion `invlpg` einen eventuell vorhandenen Eintrag im Cache der MMU, initialisieren die neue Seite, so dass sie nur Nullen enthält, und geben ihre (virtuelle) Adresse zurück.

Eine Seite wieder freizugeben, ist leichter: Die Funktion

14a ⟨*function prototypes* 4b⟩+≡ (3a) ◁12b 15b▷
```
void release_page (unsigned int pageno);
```

die eine Seitennummer als Argument erwartet, ist schnell implementiert: Sie ersetzt den von `request_new_frame` eingetragenen Page Descriptor wieder durch einen Null-Deskriptor und gibt auch den zugeordneten Frame frei.

14b ⟨*function implementations* 4e⟩+≡ (3a) ◁12c 15c▷
```
void release_page (unsigned int pageno) {
  int frameno = pageno_to_frameno (pageno);  // we will need this later
  if ( frameno == -1 )  { return; }          // exit if no such page
  unsigned int pdindex = pageno/1024;
  unsigned int ptindex = pageno%1024;
  page_table* pt;
  pt = (page_table*)
    ( PHYSICAL(current_pd->ptds[pdindex].frame_addr << 12) );
  // write null page descriptor
  memset (&(pt->pds[ptindex]), 0, 4);
  fill_page_desc (&(pt->pds[ptindex]), false, false, false, false, 0);
  release_frame (frameno<<12);   // expects an address, not an ID
  asm volatile ("invlpg %0" : : "m"(*(char*)(pageno<<12)) );
  // gdt_flush ();
};
```
Benutzt `current_pd` 6b und `page_table` 6a.

Auch hier wird am Ende wieder `invlpg` verwendet, um eventuelle Informationen über diese Seite im MMU-Cache zu löschen.

5.1 Den physischen Speicher mappen

Das Mapping des physischen Speichers in den Adressbereich ab `0xD0000000` funkioniert so, dass wir jeweils für die virtuelle Seite `x + 0xD0000` den physischen Frame `x` eintragen. Dafür benötigen wir die bereits oben erwähnten 16 Seitentabellen:

14c ⟨*kernel main: map physical RAM* 14c⟩≡ (3b)
```
memset (kernel_pt_ram, 0, 4);

for (unsigned int fid=0; fid<NUMBER_OF_FRAMES; fid++) {
  KMAP ( &(kernel_pt_ram[fid/1024].pds[fid%1024]), fid*PAGE_SIZE );
}
unsigned int physaddr;
for (int i=0; i<16; i++) {
  // get physical address of kernel_pt_ram[i]
  physaddr = (unsigned int)(&(kernel_pt_ram[i])) - 0xc0000000;
  KMAPD ( &(current_pd->ptds[832+i]), physaddr );
};

gdt_flush ();
```
Benutzt `current_pd` 6b.

Das Einrichten der Frame-Tabelle besteht nur darin, passende `0`- und `1`-Bits hineinzuschreiben. Dafür können wir zweimal `memset` verwenden: Der erste Aufruf füllt die Tabelle mit Nullen, und der zweite setzt die vordersten $128 \times 8 = 1024$ Bits auf `1`, weil die ersten 1024 Rahmen (also die ersten 4 MByte) nicht von der Speicherverwaltung verwendet werden sollen.

15a ⟨*kernel main: setup frame table* 15a⟩≡ (3b)

```
memset (ftable, 0, NUMBER_OF_FRAMES/8);  // all frames are free
memset (ftable, 0xff, 128);
free_frames -= 1024;
```

6 Hilfsfunktionen

Hier finden Sie alle Funktionen, die eher uninteressant sind – z. B., weil sie klassische Hilfsfunktionen sind, die sonst von Bibliotheken bereitgestellt werden.

6.1 Speicher initialisieren mit `memset`

Die Funktionen

15b ⟨*function prototypes* 4b⟩+≡ (3a) ◁14a 15d▷

```
void *memset  (void *dest, char val,  int count);
void *memsetw (void *dest, short val, int count);
```

erwarten als Argumente eine Startadresse, ein Füll-Byte oder Füll-Wort und die Anzahl der zu füllenden Bytes. Sie ist schnell geschrieben:

15c ⟨*function implementations* 4e⟩+≡ (3a) ◁14b 16b▷

```
void *memset (void *dest, char val, int count) {
  char *temp = (char *)dest;
  for( ; count != 0; count--) *temp++ = val;
  return dest;
}

void *memsetw (void *dest, short val, int count) {
  short *temp = (short *)dest;
  for( ; count != 0; count--) *temp++ = val;
  return dest;
}
```

6.2 Text mit `printf` ausgeben

Die Funktion

15d ⟨*function prototypes* 4b⟩+≡ (3a) ◁15b 16a▷

```
extern int printf(const char *format, ...);
```

stellen wir über eine separate C-Datei bereit, deren Inhalt wir hier nicht weiter erklären. Damit sie funktioniert, müssen wir zwei Variablen initialisieren:

15e ⟨*kernel main: initialize variables* 15e⟩≡ (3b)

```
posx = 0; posy = 8;  // set cursor
```

Die `printf`-Funktion nutzt die einfachere Funktion

16a ⟨*function prototypes* 4b⟩+≡ (3a) ◁15d 16c▷
```
void kputch (char c);
```

welche ein einzelnes Zeichen auf dem Bildschirm ausgeben kann. Sie schreibt direkt in den Textmodus-Framebuffer der Grafikkarte, welcher in den physischen Adressraum eingeblendet ist. Um sich die aktuelle Cursorposititon zu merken, verwendet sie die Variablen `posx` und `posy`.

16b ⟨*function implementations* 4e⟩+≡ (3a) ◁15c 16e▷
```
void kputch (char c) {
  char *screen;

  if (c=='\n') {
    posy ++;
    posx = 0;
    uartputc ('\n');
    return;
  }

  if (paging_ready)
    screen = (char*) 0xb8000 + posy*160 + posx*2;
  else
    screen = (char*) 0xc0000000 + 0xb8000 + posy*160 + posx*2;
  *screen = c;
  posx++;
  if (posx == 80) {
    posy++; posx = 0;
  }

  // auf serielle Konsole schreiben; ohne Erklärung
  if (c == 0x100) {  //  backspace
    uartputc('\b'); uartputc(' '); uartputc('\b');
  } else uartputc(c);
}
```

Die Funktion erzeugt oben für Zeilenumbrüche und auch am Ende auch eine Ausgabe über die serielle Konsole, was wir hier nicht weiter erklären; die Funktion `uartputc` müssen wir allerdings als extern deklarieren:

16c ⟨*function prototypes* 4b⟩+≡ (3a) ◁16a 16d▷
```
extern void uartputc (int c);
```

6.3 Den Bildschirm mit `clrscr` löschen

Wenn Sie mehr Platz auf dem Bildschirm (für weitere Ausgaben) brauchen, können Sie durch einen Aufruf der Funktion

16d ⟨*function prototypes* 4b⟩+≡ (3a) ◁16c 17b▷
```
void clrscr ();
```

den Bildschirm löschen; das setzt auch den Cursor automatisch nach links oben:

16e ⟨*function implementations* 4e⟩+≡ (3a) ◁16b 17c▷

```
void clrscr () {
  posx = posy = 0;
  unsigned blank = 0x20 + (0x0f<<8);   // blank character (word)
  char *screen;
  if (paging_ready)
    screen = (char*) 0xb8000;
  else
    screen = (char*) 0xc0000000 + 0xb8000;
  memsetw (screen, blank, 80*25);
}
```

Die beiden hier verwendeten Variablen für die Cursor-Position müssen wir noch deklarieren:

17a ⟨*global variables* 3c⟩+≡ (3a) ◁9b 21a▷
```
int posx, posy;
```

6.4 String kopieren mit `strncpy`

17b ⟨*function prototypes* 4b⟩+≡ (3a) ◁16d 17d▷
```
void *strncpy(void *dest, const void *src, int count);
```

17c ⟨*function implementations* 4e⟩+≡ (3a) ◁16e 17e▷
```
void *strncpy (void *dest, const void *src, int count) {
  // like memcpy, but copies only until first \0 character
  const char *sp = (const char *)src;
  char *dp = (char *)dest;
  for (; count != 0; count--) {
    *dp = *sp;
    if (*dp == 0) break;
    dp++; sp++;
  }
  return dest;
}
```

6.5 Hexdump

Zum Bearbeiten der Übungsaufgabe stellen wir noch die folgende Funktion

17d ⟨*function prototypes* 4b⟩+≡ (3a) ◁17b 19b▷
```
void hexdump (unsigned int start, unsigned int end);
```

bereit, welche einen Hexdump des virtuellen Speichers zwischen `start` und `end` erzeugt:

17e ⟨*function implementations* 4e⟩+≡ (3a) ◁17c 19c▷
```
void hexdump (unsigned int start, unsigned int end) {
  char z;
  for (unsigned int i=start; i < end; i+=16) {
    printf ("%x  ", i);  // address
    // hex values
    for (int j=i; j<i+16; j++) {
      printf ("%02x ", (unsigned char)PEEK(j));
```

```
      if (j==i+7) kputch (' ');
    };
    kputch (' ');
    // char values
    for (int j=i; j<i+16; j++) {
      z = PEEK(j);
      if ((z>32)&&(z<127)) {
        kputch (PEEK(j));
      } else {
        kputch ('.');
      }
    }

    kputch ('\n');
  }
}
```

Sie verwendet das Makro `PEEK` zum Auslesen des Speichers:

18a ⟨*macros* 8a⟩+≡ (3a) ◁12a 25e▷

```
#define PEEK(addr) (*(unsigned char *)(addr))
```

Zum Abschluss noch drei Makros, die wir im Code verwendet haben:

18b ⟨*constants* 8b⟩+≡ (3a) ◁9c 19a▷

```
#define asm __asm__
#define volatile __volatile__
#define NULL ((void*) 0)
```

7 Interrupts

Wir implementieren nun die Interrupt- (und im nächsten Kapitel die Fault-) Handler und starten dabei mit den Interrupt-Nummern:

19a ⟨*constants* 8b⟩+≡ (3a) ◁18b 19d▷

```
#define IRQ_TIMER      0
#define IRQ_KBD        1
#define IRQ_SLAVE      2     // Here the slave PIC connects to master
#define IRQ_COM2       3
#define IRQ_COM1       4
#define IRQ_FDC        6
#define IRQ_IDE       14     // primary IDE controller; secondary has IRQ 15
```

Definiert:
IRQ_COM1, nicht benutzt.
IRQ_COM2, nicht benutzt.
IRQ_FDC, nicht benutzt.
IRQ_IDE, nicht benutzt.
IRQ_KBD, benutzt im Teil 28a.
IRQ_SLAVE, benutzt im Teil 21e.
IRQ_TIMER, benutzt im Teil 28a.

Weiter geht es mit den `in`- und `out`-Befehlen

19b ⟨*function prototypes* 4b⟩+≡ (3a) ◁17d 21b▷

```
unsigned char  inportb (unsigned short port);
unsigned short inportw (unsigned short port);
void outportb (unsigned short port, unsigned char data);
void outportw (unsigned short port, unsigned short data);
```

Benutzt inportb 19c, inportw 19c, outportb 19c, und outportw 19c.

die wir wir in der Vorlesung implementieren; der Code für `inportb` und `outportb` steht bereits in der Datei `printf.c`, weswegen wir ihn hier weglassen.

19c ⟨*function implementations* 4e⟩+≡ (3a) ◁17e 21c▷

```
unsigned short inportw (unsigned short port) {
  unsigned short retval;
  asm volatile ("inw %%dx, %%ax" : "=a" (retval) : "d" (port));
  return retval;
}

void outportw (unsigned short port, unsigned short data) {
  asm volatile ("outw %%ax, %%dx" : : "d" (port), "a" (data));
}
```

Definiert:
inportb, benutzt im Teils 19b, 22b, und 28d.
inportw, benutzt im Teil 19b.
outportb, benutzt im Teils 19, 20, 22b, und 24a.
outportw, benutzt im Teil 19b.

Damit können wir nun die PICs einrichten. Ihre Ports haben die folgenden Adressen:

19d ⟨*constants* 8b⟩+≡ (3a) ◁19a 24b▷

```
// I/O Addresses of the two programmable interrupt controllers
#define IO_PIC_MASTER_CMD   0x20  // Master (IRQs 0-7), command register
#define IO_PIC_MASTER_DATA  0x21  // Master, control register
```

```
#define IO_PIC_SLAVE_CMD    0xA0  // Slave (IRQs 8-15), command register
#define IO_PIC_SLAVE_DATA   0xA1  // Slave, control register
```

Definiert:
IO_PIC_MASTER_CMD, benutzt im Teils 20b und 24a.
IO_PIC_MASTER_DATA, benutzt im Teils 20 und 22b.
IO_PIC_SLAVE_CMD, benutzt im Teils 20b und 24a.
IO_PIC_SLAVE_DATA, benutzt im Teils 20 und 22b.

Wir müssen bei der Initialisierung die Interrupt-Nummern umbiegen; detaillierte Erläuterungen dazu finden Sie im Buch-Kapitel 12 auf der Webseite.

20a ⟨*remap the interrupts to 32..47* 20a⟩≡ (21e)
```
⟨PIC: program/initialize the PICs 20b⟩
⟨PIC: set the initial interrupt mask 20c⟩
```

20b ⟨*PIC: program/initialize the PICs* 20b⟩≡ (20a)
```
outportb (IO_PIC_MASTER_CMD,  0x11);  // ICW1: initialize; begin programming
outportb (IO_PIC_SLAVE_CMD,   0x11);  // ICW1: dito, for PIC2
outportb (IO_PIC_MASTER_DATA, 0x20);  // ICW2 for PIC1: offset 0x20
                                      // (remaps 0x00..0x07 -> 0x20..0x27)
outportb (IO_PIC_SLAVE_DATA,  0x28);  // ICW2 for PIC2: offset 0x28
                                      // (remaps 0x08..0x0f -> 0x28..0x2f)
outportb (IO_PIC_MASTER_DATA, 0x04);  // ICW3 for PIC1: there's a slave on IRQ 2
                                      // (0b00000100 = 0x04)
outportb (IO_PIC_SLAVE_DATA,  0x02);  // ICW3 for PIC2: your slave ID is 2
outportb (IO_PIC_MASTER_DATA, 0x01);  // ICW4 for PIC1 and PIC2: 8086 mode
outportb (IO_PIC_SLAVE_DATA,  0x01);
```

Benutzt IO_PIC_MASTER_CMD 19d, IO_PIC_MASTER_DATA 19d, IO_PIC_SLAVE_CMD 19d, IO_PIC_SLAVE_DATA 19d, und outportb 19c.

Im zweiten Schritt setzen wir die Interrupt-Maske (und schalten alle Interrupts aus).

20c ⟨*PIC: set the initial interrupt mask* 20c⟩≡ (20a)
```
outportb (IO_PIC_MASTER_DATA, 0x00);  // PIC1: mask 0
outportb (IO_PIC_SLAVE_DATA,  0x00);  // PIC2: mask 0
```

Benutzt IO_PIC_MASTER_DATA 19d, IO_PIC_SLAVE_DATA 19d, und outportb 19c.

Weiter geht es mit der Interrupt Descriptor Table. Ihre Einträge haben folgenden Aufbau:

20d ⟨*type definitions* 3d⟩+≡ (3a) ◁11a 20e▷
```
struct idt_entry {
    unsigned int addr_low  : 16;   // lower 16 bits of address
    unsigned int gdtsel    : 16;   // use which GDT entry?
    unsigned int zeroes    :  8;   // must be set to 0
    unsigned int type      :  4;   // type of descriptor
    unsigned int flags     :  4;
    unsigned int addr_high : 16;   // higher 16 bits of address
} __attribute__((packed));
```

und es gibt (wie bei der GDT) zusätzlich einen Pointer auf den Anfang der Tabelle:

20e ⟨*type definitions* 3d⟩+≡ (3a) ◁20d 23a▷

```
struct idt_ptr {
    unsigned int limit   : 16;
    unsigned int base    : 32;
} __attribute__((packed));
```

Wir verwenden zwei globale Variablen für die Tabelle und den Pointer:

21a ⟨*global variables* 3c⟩+≡ (3a) ◁17a 24c▷

```
struct idt_entry idt[256] = { 0 };
struct idt_ptr idtp;
```

Wir können einen Deskriptor mit

21b ⟨*function prototypes* 4b⟩+≡ (3a) ◁19b 21d▷

```
void fill_idt_entry (unsigned char num, unsigned long address,
    unsigned short gdtsel, unsigned char flags, unsigned char type);
```

Benutzt `fill_idt_entry` 21c.

füllen:

21c ⟨*function implementations* 4e⟩+≡ (3a) ◁19c 22b▷

```
void fill_idt_entry (unsigned char num, unsigned long address,
    unsigned short gdtsel, unsigned char flags, unsigned char type) {
  if (num >= 0 && num < 256) {
    idt[num].addr_low  = address & 0xFFFF; // address is the handler address
    idt[num].addr_high = (address >> 16) & 0xFFFF;
    idt[num].gdtsel    = gdtsel;           // GDT sel.: user mode or kernel mode?
    idt[num].zeroes    = 0;
    idt[num].flags     = flags;
    idt[num].type      = type;
  }
}
```

Definiert:
`fill_idt_entry`, benutzt im Teils 21 und 25e.

Wir definieren Interrupt-Handler-Funktionen `irq0` bis `irq15` in der Assembler-Datei. Im C-Programm müssen wir die Funktionen als extern deklarieren:

21d ⟨*function prototypes* 4b⟩+≡ (3a) ◁21b 22a▷

```
extern void irq0(), irq1(), irq2(),  irq3(),  irq4(),  irq5(),  irq6(),  irq7();
extern void irq8(), irq9(), irq10(), irq11(), irq12(), irq13(), irq14(), irq15();
```

Dann können wir die Handler auf die einzelnen IDT-Einträge verteilen:

21e ⟨*install the interrupt handlers* 21e⟩≡ (22c)

```
⟨install the IDT 25a⟩
⟨install the fault handlers 25f⟩
⟨remap the interrupts to 32..47 20a⟩
set_irqmask (0xFFFF);           // initialize IRQ mask
enable_interrupt (IRQ_SLAVE);   // IRQ slave

// flags: 1 (present), 11 (DPL 3), 0; type: 1110 (32 bit interrupt gate)
fill_idt_entry (32, (unsigned int)irq0,  0x08, 0b1110, 0b1110);
fill_idt_entry (33, (unsigned int)irq1,  0x08, 0b1110, 0b1110);
fill_idt_entry (34, (unsigned int)irq2,  0x08, 0b1110, 0b1110);
fill_idt_entry (35, (unsigned int)irq3,  0x08, 0b1110, 0b1110);
```

```
fill_idt_entry (36, (unsigned int)irq4,  0x08, 0b1110, 0b1110);
fill_idt_entry (37, (unsigned int)irq5,  0x08, 0b1110, 0b1110);
fill_idt_entry (38, (unsigned int)irq6,  0x08, 0b1110, 0b1110);
fill_idt_entry (39, (unsigned int)irq7,  0x08, 0b1110, 0b1110);
fill_idt_entry (40, (unsigned int)irq8,  0x08, 0b1110, 0b1110);
fill_idt_entry (41, (unsigned int)irq9,  0x08, 0b1110, 0b1110);
fill_idt_entry (42, (unsigned int)irq10, 0x08, 0b1110, 0b1110);
fill_idt_entry (43, (unsigned int)irq11, 0x08, 0b1110, 0b1110);
fill_idt_entry (44, (unsigned int)irq12, 0x08, 0b1110, 0b1110);
fill_idt_entry (45, (unsigned int)irq13, 0x08, 0b1110, 0b1110);
fill_idt_entry (46, (unsigned int)irq14, 0x08, 0b1110, 0b1110);
fill_idt_entry (47, (unsigned int)irq15, 0x08, 0b1110, 0b1110);
```

Benutzt `enable_interrupt` 22b, `fill_idt_entry` 21c, `IRQ_SLAVE` 19a, und `set_irqmask` 22b.

wobei wir noch

22a ⟨*function prototypes* 4b⟩+≡ (3a) ◁21d 24d▷

```
static void set_irqmask (unsigned short mask);
static void enable_interrupt (int number);
unsigned short get_irqmask ();
```

Benutzt `enable_interrupt` 22b, `get_irqmask` 22b, und `set_irqmask` 22b.

implementieren:

22b ⟨*function implementations* 4e⟩+≡ (3a) ◁21c 24a▷

```
static void set_irqmask (unsigned short mask) {
  outportb (IO_PIC_MASTER_DATA, (char)(mask % 256) );
  outportb (IO_PIC_SLAVE_DATA,  (char)(mask >> 8)  );
}

unsigned short get_irqmask () {
  return inportb (IO_PIC_MASTER_DATA)
      + (inportb (IO_PIC_SLAVE_DATA) << 8);
}

static void enable_interrupt (int number) {
  set_irqmask (
    get_irqmask ()        // the current value
    & ~(1 << number)      // 16 one-bits, but bit "number" cleared
  );
}
```

Definiert:
`enable_interrupt`, benutzt im Teils 21e, 22a, und 28a.
`get_irqmask`, benutzt im Teil 22a.
`set_irqmask`, benutzt im Teils 21e und 22a.
Benutzt `inportb` 19c, `IO_PIC_MASTER_DATA` 19d, `IO_PIC_SLAVE_DATA` 19d, und `outportb` 19c.

Das Installieren der Interrupt Handler im Code Chunk ⟨*install the interrupt handlers* 21e⟩ muss bei der Systeminitialisierung erfolgen, also:

22c ⟨*kernel main: initialize system* 22c⟩≡ (3b) 28a▷

```
⟨install the interrupt handlers 21e⟩
```

Unsere Interrupt-Handler erhalten die Registerinhalte, dafür legen wir einen eigenen Datentyp an:

23a ⟨*type definitions* 3d⟩+≡ (3a) ◁20e

```
typedef struct {
  unsigned int gs, fs, es, ds;
  unsigned int edi, esi, ebp, esp, ebx, edx, ecx, eax;
  unsigned int int_no, err_code;
  unsigned int eip, cs, eflags, useresp, ss;
} context_t;
```

In der Assembler-Datei liegen die Anfangsstücke der Interrupt-Handler:

23b ⟨*start.asm* 23b⟩≡ 25c▷

```
global irq0, irq1, irq2,  irq3,  irq4,  irq5,  irq6,  irq7
global irq8, irq9, irq10, irq11, irq12, irq13, irq14, irq15

%macro irq_macro 1
       cli                  ; disable interrupts
       push byte 0          ; error code (none)
       push byte %1         ; interrupt number
       jmp irq_common_stub  ; rest is identical for all handlers
%endmacro

irq0:  irq_macro 32
irq1:  irq_macro 33
irq2:  irq_macro 34
irq3:  irq_macro 35
irq4:  irq_macro 36
irq5:  irq_macro 37
irq6:  irq_macro 38
irq7:  irq_macro 39
irq8:  irq_macro 40
irq9:  irq_macro 41
irq10: irq_macro 42
irq11: irq_macro 43
irq12: irq_macro 44
irq13: irq_macro 45
irq14: irq_macro 46
irq15: irq_macro 47

extern irq_handler          ; defined in the C source file

irq_common_stub:            ; this is the identical part
       pusha
       push ds
       push es
       push fs
       push gs
       push esp  ; pointer to the context_t
       call irq_handler     ; call C function
       pop esp
       pop gs
       pop fs
       pop es
       pop ds
       popa
       add esp, 8
```

```
        iret
```
Benutzt `irq_handler` 24a.

– wir haben in der Vorlesung besprochen, warum die Register in dieser Reihenfolge auf den Stack gelegt (und später wieder ausgelesen) werden.

(*Achtung: Sie können diesen Code Chunk nicht in dieser NoWeb-Datei verändern; er wird nicht exportiert.*)

Es fehlt nun noch die C-Funktion `irq_handler()`:

24a ⟨*function implementations* 4e⟩+≡ (3a) ◁22b 24e▷
```
void irq_handler (context_t *r) {
  int number = r->int_no - 32;                    // interrupt number
  void (*handler)(context_t *r);                  // type of handler functions

  if (number >= 8)
    outportb (IO_PIC_SLAVE_CMD, END_OF_INTERRUPT);  // notify slave  PIC
  outportb (IO_PIC_MASTER_CMD, END_OF_INTERRUPT);   // notify master PIC

  handler = interrupt_handlers[number];
  if (handler != NULL)  handler (r);
}
```
Definiert:
`irq_handler`, benutzt im Teil 23b.
Benutzt `interrupt_handlers` 24c, `IO_PIC_MASTER_CMD` 19d, `IO_PIC_SLAVE_CMD` 19d, und `outportb` 19c.

Sie verwendet die Konstante

24b ⟨*constants* 8b⟩+≡ (3a) ◁19d 29b▷
```
#define END_OF_INTERRUPT  0x20
```

und das Array

24c ⟨*global variables* 3c⟩+≡ (3a) ◁21a 26b▷
```
void *interrupt_handlers[16] = { 0 };
```
Definiert:
`interrupt_handlers`, benutzt im Teil 24.

Um einen Interrupt-Handler zu installieren (also seine Adresse in das Array einzutragen), verwenden wir

24d ⟨*function prototypes* 4b⟩+≡ (3a) ◁22a 25b▷
```
void install_interrupt_handler (int irq, void (*handler)(context_t *r));
```
Benutzt `install_interrupt_handler` 24e.

mit folgender Implementierung:

24e ⟨*function implementations* 4e⟩+≡ (3a) ◁24a 26d▷
```
void install_interrupt_handler (int irq, void (*handler)(context_t *r)) {
  if (irq >= 0 && irq < 16)
    interrupt_handlers[irq] = handler;
}
```
Definiert:
`install_interrupt_handler`, benutzt im Teils 24d und 28a.
Benutzt `interrupt_handlers` 24c.

Bei der Initialisierung des Kernels müssen wir noch das IDTR-Register laden:

25a ⟨*install the IDT* 25a⟩≡ (21e)
```
idtp.limit = (sizeof (struct idt_entry) * 256) - 1;   // must do -1
idtp.base  = (int) &idt;
idt_load ();
```

wobei der letzte Befehl wieder als Assembler-Code vorliegt:

25b ⟨*function prototypes* 4b⟩+≡ (3a) ◁24d 25d▷
```
extern void idt_load ();
```

25c ⟨*start.asm* 23b⟩+≡ ◁23b 26a▷
```
extern idtp  ; defined in the C file
global idt_load
idt_load:    lidt [idtp]
             ret
```

8 Faults

Das Fault-Handling funktioniert ganz ähnlich wie die Interrupt-Behandlung, darum gehen wir hier auf die Details nur kurz ein.

Die Funktionen `isr0()`, ..., `isr31()` und `isr128()` (neu, für die System Calls) implementieren wir wieder als Assembler-Funktionen; im C-Programm müssen wir ihre Namen bekannt machen:

25d ⟨*function prototypes* 4b⟩+≡ (3a) ◁25b 26c▷
```
extern void isr0(),  isr1(),  isr2(),  isr3(),  isr4(),  isr5(),
   isr6(),  isr7(),  isr8(),  isr9(),  isr10(), isr11(), isr12(),
   isr13(), isr14(), isr15(), isr16(), isr17(), isr18(), isr19(),
   isr20(), isr21(), isr22(), isr23(), isr24(), isr25(), isr26(),
   isr27(), isr28(), isr29(), isr30(), isr31(), isr128();
```

Um die IDT-Einträge für die 33 Fault-Handler anzulegen, benutzen wir ein kleines Makro:

25e ⟨*macros* 8a⟩+≡ (3a) ◁18a
```
#define FILL_IDT(i) \
  fill_idt_entry (i, (unsigned int)isr##i, 0x08, 0b1110, 0b1110)
```
Definiert:
FILL_IDT, benutzt im Teil 25f.
Benutzt fill_idt_entry 21c.

und können dann mit wenigen Zeilen die 32 Eintragungen vornehmen:

25f ⟨*install the fault handlers* 25f⟩≡ (21e)
```
FILL_IDT( 0); FILL_IDT( 1); FILL_IDT( 2); FILL_IDT( 3); FILL_IDT( 4);
FILL_IDT( 5); FILL_IDT( 6); FILL_IDT( 7); FILL_IDT( 8); FILL_IDT( 9);
FILL_IDT(10); FILL_IDT(11); FILL_IDT(12); FILL_IDT(13); FILL_IDT(14);
FILL_IDT(15); FILL_IDT(16); FILL_IDT(17); FILL_IDT(18); FILL_IDT(19);
FILL_IDT(20); FILL_IDT(21); FILL_IDT(22); FILL_IDT(23); FILL_IDT(24);
FILL_IDT(25); FILL_IDT(26); FILL_IDT(27); FILL_IDT(28); FILL_IDT(29);
FILL_IDT(30); FILL_IDT(31); FILL_IDT(128);
```
Benutzt FILL_IDT 25e.

In der Assembler-Datei geben wir an, dass die Symbole exportiert werden sollen:

26a ⟨*start.asm* 23b⟩+≡ ◁25c

```
global isr0,  isr1,  isr2,  isr3,  isr4,  isr5,  isr6,  isr7,  isr8
global isr9,  isr10, isr11, isr12, isr13, isr14, isr15, isr16, isr17
global isr18, isr19, isr20, isr21, isr22, isr23, isr24, isr25, isr26
global isr27, isr28, isr29, isr30, isr31, isr128
```

Dann brauchen wir ein Array mit Fehlermeldungen; hinter Fault-Nummer 18 kommt nichts mehr: Die restlichen Werte sind reserviert.

26b ⟨*global variables* 3c⟩+≡ (3a) ◁24c 28c▷

```
char *exception_messages[] = {
  "Division By Zero",       "Debug",                          //  0,  1
  "Non Maskable Interrupt", "Breakpoint",                     //  2,  3
  "Into Detected Overflow", "Out of Bounds",                  //  4,  5
  "Invalid Opcode",         "No Coprocessor",                 //  6,  7
  "Double Fault",           "Coprocessor Segment Overrun",    //  8,  9
  "Bad TSS",                "Segment Not Present",            // 10, 11
  "Stack Fault",            "General Protection Fault",       // 12, 13
  "Page Fault",             "Unknown Interrupt",              // 14, 15
  "Coprocessor Fault",      "Alignment Check",                // 16, 17
  "Machine Check",                                            // 18
  "Reserved", "Reserved", "Reserved", "Reserved", "Reserved",
  "Reserved", "Reserved", "Reserved", "Reserved", "Reserved",
  "Reserved", "Reserved", "Reserved"                          // 19..31
};
```

Definiert:
`exception_messages`, benutzt im Teil 26e.

Den eigentlichen Fault-Handler

26c ⟨*function prototypes* 4b⟩+≡ (3a) ◁25d 28b▷

```
void fault_handler (context_t *r);
```

Benutzt `fault_handler` 26d.

präsentieren wir hier in der vereinfachten Variante (die noch keine Prozesse berücksichtigt). Der Handler gibt einige Informationen aus und hält dann das System an.

26d ⟨*function implementations* 4e⟩+≡ (3a) ◁24e 28d▷

```
void fault_handler (context_t *r) {
  if (r->int_no >= 0 && r->int_no < 32) {
    ⟨fault handler: display status information 26e⟩
    printf ("System Stops\n");
    asm ("cli; \n hlt;");
  }
}
```

Definiert:
`fault_handler`, benutzt im Teil 26c.

In der Ausgabe erscheinen Name und Nummer des Faults sowie die Inhalte einiger Register.

26e ⟨*fault handler: display status information* 26e⟩≡ (26d)

```
printf ("'%s' (%d) Exception at 0x%08x.\n",
  exception_messages[r->int_no], r->int_no, r->eip);
```

```
  printf ("eflags: 0x%08x  errcode: 0x%08x\n", r->eflags, r->err_code);
  printf ("eax: %08x  ebx: %08x  ecx: %08x  edx: %08x \n",
    r->eax, r->ebx, r->ecx, r->edx);
  printf ("eip: %08x  esp: %08x  int: %8d  err: %8d \n",
    r->eip, r->esp, r->int_no, r->err_code);
  printf ("ebp: %08x  cs: %d  ds: %d  es: %d  fs: %d  ss: %x \n",
    r->ebp, r->cs, r->ds, r->es, r->fs, r->ss);
```

Benutzt `exception_messages` 26b.

9 Tastatur-Treiber

Wir legen hier fest, wie das System auf Tastendrücke reagiert. Dazu installieren wir einen Interrupt-Handler für den Tastatur-Interrupt `IRQ_KBD`: Die Funktion `keyboard_handler` müssen wir im nächsten Schritt schreiben.

28a ⟨*kernel main: initialize system* 22c⟩+≡ (3b) ◁22c

```
install_interrupt_handler (IRQ_KBD, keyboard_handler);
enable_interrupt (IRQ_KBD);
// install_interrupt_handler (IRQ_TIMER, timer_handler);
// enable_interrupt (IRQ_TIMER);
printf ("ENABLE INTERRUPTS\n");
asm ("sti");
```

Benutzt `enable_interrupt` 22b, `install_interrupt_handler` 24e, `IRQ_KBD` 19a, und `IRQ_TIMER` 19a.

Wie alle Handler-Funktionen hat auch `keyboard_handler` die Signatur `void handler (context_t *r)`. Bei der Gelegenheit definieren wir außerdem einen Timer-Handler.

28b ⟨*function prototypes* 4b⟩+≡ (3a) ◁26c 30a▷

```
void keyboard_handler (context_t *r);
void timer_handler (context_t *r);
```

28c ⟨*global variables* 3c⟩+≡ (3a) ◁26b 29c▷

```
unsigned long int ticks = 0;
```

Hier folgt nun die Implementation des eigentlichen Keyboard-Handlers. Er muss als erstes den aktuellen Scan-Code von der Tastatur lesen – wenn man das vergisst, erzeugt die Tastatur keine weiteren Interrupts.

Der Handler prüft zunächst, ob der Scan-Code einen Wert kleiner als 128 hat: Das bedeutet, dass eine Taste gedrückt wurde. (Scan-Codes mit gesetztem 8. Bit, also einem Wert ≥ 128, stehen für das Loslassen von Tasten.)

Erkannte Zeichen schreibt der Handler auf die Konsole und außerdem in den Puffer `buffer`.

28d ⟨*function implementations* 4e⟩+≡ (3a) ◁26d 29a▷

```
void keyboard_handler (context_t *r) {
  unsigned char s = inportb (0x60);
  char c;
  if (s < 128)
    c = scancode_table[s];
  else
    c = 0;

  if (c != 0 && pos < 256) {
    buffer[pos] = c;
    printf ("%c", c);
    pos++;
    unread++;
  }

  return;
};
```

Benutzt `inportb` 19c.

Unser Timer-Handler wird automatisch regelmäßig aufgerufen; wir zählen hier nur eine Variable `ticks` hoch. Wer mag, kann den auskommentierten Code aktivieren: Dann gibt der Handler bei jedem Aufruf ein “T” aus.

29a ⟨*function implementations* 4e⟩+≡ (3a) ◁28d 30b▷

```
void timer_handler (context_t *r) {
  ticks++;
  /*
  if (ticks % 10 == 0) {
    if (posy == 25) clrscr ();
    printf ("T");
  }
  */
  return;
};
```

Wir definieren, anders als in der Aufgabe gefordert, die vollständige Scan-Tabelle:

29b ⟨*constants* 8b⟩+≡ (3a) ◁24b 30c▷

```
#define KEY_UP     191
#define KEY_DOWN   192
#define KEY_LEFT   193
#define KEY_RIGHT  194
```

29c ⟨*global variables* 3c⟩+≡ (3a) ◁28c 30d▷

```
char scancode_table[128] = {
  /*  0.. 9 */    0,  27, '1', '2', '3', '4', '5', '6', '7', '8',
  /* 10..19 */   '9', '0', '-', '=', '\b',       /* Backspace */
                 '\t', /* Tab */   'q', 'w', 'e', 'r',
  /* 20..29 */   't', 'z', 'u', 'i', 'o', 'p', '[', ']',
                 '\n', /* Enter */  0, /* Control */
  /* 30..39 */   'a', 's', 'd', 'f', 'g', 'h', 'j', 'k', 'l', ';',
  /* 40..49 */   '\'', '`', 0, /* Left shift */  '\\', 'y', 'x',
                 'c', 'v', 'b', 'n',
  /* 50..59 */   'm', ',', '.', '/', 0, /* Right shift */
                 '*', 0, /* Alt */  ' ', /* Space bar */
                 0, /* CapsLock */  0, /* F1 */
  /* 60..69 */   0, 0, 0, 0, 0, 0, 0, 0, 0, /* F2..F10 */
                 0, /* NumLock */
  /* 70..79 */   0, /* Scroll Lock */   0, /* Home key */
                 KEY_UP, 0, /* Page Up */
                 '-', KEY_LEFT, 0, KEY_RIGHT,
                 '+', 0, /* End */
  /* 80..89 */   KEY_DOWN, 0, /* Page Down */
                 0, /* Insert Key */    0, /* Delete */
                 0, 0, 0, 0, /* F11 */  0, /* F12 */   0,
  /* 90..127     not defined */
};   // Scan Codes
short int pos = 0;                  // Puffer-Position
short int unread = 0;               // Anzahl "frischer" Zeichen
char buffer[256] = { 0 };           // Eingabe-Puffer
```

Die Funktion `kreadline` liest maximal `len` Zeichen in den String ein, dessen Speicheradresse über den Pointer `s` übergeben wird:

30a ⟨*function prototypes* 4b⟩+≡ (3a) ◁28b 30e▷
```
void kreadline (char *s, int len);
```

Die Funktion liest solange ein neues Zeichen aus dem Tastaturpuffer `buffer`, bis die Eingabetaste gedrückt wird:

30b ⟨*function implementations* 4e⟩+≡ (3a) ◁29a 30f▷
```
void kreadline (char *s, int len) {
  int read_chars = 0;
  for (;;) {
    if (pos>0 && buffer[pos-1]=='\n') {
      buffer[pos-1] = 0;  // String terminieren
      strncpy (s, buffer, len);
      pos = 0;
      return;
    }
  }
}
```

10 System Calls

Wir definieren zunächst die System-Call-Tabelle:

30c ⟨*constants* 8b⟩+≡ (3a) ◁29b
```
#define MAX_SYSCALLS 0x8000          // max syscall number: 0x7fff
```
Definiert:
`MAX_SYSCALLS`, benutzt im Teil 30.

30d ⟨*global variables* 3c⟩+≡ (3a) ◁29c
```
void *syscall_table[MAX_SYSCALLS];
```
Definiert:
`syscall_table`, benutzt im Teils 30f und 31a.
Benutzt `MAX_SYSCALLS` 30c.

Um neue System Calls anzulegen, richten wir eine Funktion

30e ⟨*function prototypes* 4b⟩+≡ (3a) ◁30a 31b▷
```
void install_syscall_handler (int syscallno, void *syscall_handler);
```
Benutzt `install_syscall_handler` 30f und `syscall_handler` 31a.

ein, die eine Syscall-Nummer und die Adresse einer Handler-Funktion als Argumente erwartet:

30f ⟨*function implementations* 4e⟩+≡ (3a) ◁30b 31a▷
```
void install_syscall_handler (int syscallno, void *syscall_handler) {
  if (syscallno < MAX_SYSCALLS)
    syscall_table[syscallno] = syscall_handler;
  return;
};
```
Definiert:
`install_syscall_handler`, benutzt im Teils 30e und 32e.
Benutzt `MAX_SYSCALLS` 30c, `syscall_handler` 31a, und `syscall_table` 30d.

Der generische System-Call-Handler, der dann den gewünschten Handler aufruft, sieht wie folgt aus:

31a ⟨*function implementations* 4e⟩+≡ (3a) ◁30f 31c▷

```
void syscall_handler (context_t *r) {
  void (*handler) (context_t*);   // handler is a function pointer
  int number = r->eax;
  handler = syscall_table[number];
  if (handler != 0) {
    handler (r);
  } else {
    printf ("Unknown syscall no. eax=0x%x; ebx=0x%x. eip=0x%x, esp=0x%x. "
            "Continuing.\n", r->eax, r->ebx, r->eip, r->esp);
  };
  return;
}
```

Definiert:
`syscall_handler`, benutzt im Teil 30.
Benutzt `syscall_table` 30d.

In die Assembler-Datei `start.asm` haben wir Code aufgenommen, der bei einem Interrupt mit der Nummer 128 (0x80) die Funktion `syscall_handler` anspringt; das ist vergleichbar mit den Aufrufen von `irq_handler` oder `fault_handler`.

Sie können einzelne System-Call-Handler in ⟨*syscall prototypes* 32a⟩ deklarieren ...

31b ⟨*function prototypes* 4b⟩+≡ (3a) ◁30e 31d▷
⟨*syscall prototypes* 32a⟩

... und dann später in ⟨*syscall functions* 32b⟩ implementieren:

31c ⟨*function implementations* 4e⟩+≡ (3a) ◁31a 31e▷
⟨*syscall functions* 32b⟩

10.1 Funktionen für den Aufruf

Mit den folgenden Funktionen können Sie System Calls standardisisert aufrufen. Das ist eigentlich ein Feature für den User Mode, es funktioniert aber auch jetzt, während es noch keine Prozesse gibt.

31d ⟨*function prototypes* 4b⟩+≡ (3a) ◁31b 32c▷

```
int syscall1 (int eax);
int syscall2 (int eax, int ebx);
int syscall3 (int eax, int ebx, int ecx);
int syscall4 (int eax, int ebx, int ecx, int edx);
```

Mit diesen Funktionen können Sie System Calls aufrufen, die 0 bis 3 Argumente haben.

31e ⟨*function implementations* 4e⟩+≡ (3a) ◁31c 32d▷

```
int syscall1 (int eax) {
  int result;
  asm ( "int $0x80" : "=a" (result) : "a" (eax) );
  return result ;
}
```

```
int syscall2 (int eax, int ebx) {
  int result;
  asm ( "int $0x80" : "=a" (result) : "a" (eax), "b" (ebx) );
  return result ;
}

int syscall3 (int eax, int ebx, int ecx) {
  int result;
  asm ( "int $0x80" : "=a" (result) : "a" (eax), "b" (ebx), "c" (ecx) );
  return result ;
}

int syscall4 (int eax, int ebx, int ecx, int edx) {
  int result;
  asm ( "int $0x80" : "=a" (result) : "a" (eax), "b" (ebx), "c" (ecx),
        "d" (edx) );
  return result ;
}
```

10.2 Beispiele

32a ⟨*syscall prototypes* 32a⟩≡ (31b)

```
// Hier fuegen Sie den Prototyp ein
```

32b ⟨*syscall functions* 32b⟩≡ (31c)

```
// Hier fuegen Sie die Implementierung ein
```

32c ⟨*function prototypes* 4b⟩+≡ (3a) ◁31d

```
// Hier ist Platz fuer den Prototyp von userprint()
```

32d ⟨*function implementations* 4e⟩+≡ (3a) ◁31e

```
// Hier ist Platz fuer die Implementierung von userprint()
```

32e ⟨*kernel main: user-defined tests* 32e⟩≡ (3b)

```
// Aufgabe 2 b) iv; diesen Teil koennen Sie loeschen
char input[41];
for (;;) {
  if (posy == 25) clrscr ();
  printf ("Eingabe: ");
  kreadline ((char*)input, 40);
  if (posy == 25) clrscr ();
  printf ("Ausgabe: %s\n", input);
}

// Tragen Sie den selbst definierten System Call in
// die Syscall-Tabelle ein:

// install_syscall_handler (nummer, funktion);

// Hier testen Sie den Syscall
```

Benutzt `install_syscall_handler` 30f.