We wrote a kernel once ...

... and it was ChaOS

Unleash the ChaOS

- Team effort with three members
  - Tim Scheuermann, Julian Holzwarth, Adrian Herrmann
  - 100% collaborative VS Code Live Share

- Free choice of design, code style, coding practices, …

- Developed over 14 weeks, 7 milestones
  1. Bare metal program (application → hardware)
  3. Driver with interrupts (application → driver → hardware)
  4. Multithreading (application → kernel/driver → hardware)
  5. Syscalls (application → library → kernel/driver → hardware)
  6. Memory protection
  7. Virtual memory

Target platform

- taskit Portux MiniPC
- AT91RM9200 SoC
- ARM920T CPU (ARMv4T)

- 16 MiB Flash, 64 MiB RAM
- Serial, Ethernet, USB host/target
- LEDs and 4x16 text display

- U-Boot bootloader in Flash

- Emulated with a patched QEMU

What is an operating system?

- Management & operation of physical resources
- Abstraction of physical with logical resources
- Process management (scheduling, switching)
- Inter-process communication
- Interface for programs

What can ChaOS do?

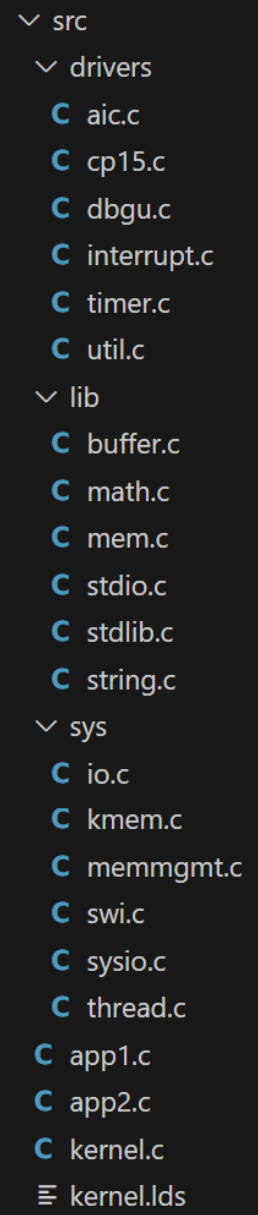
- Serial driver (DBGU) via MMIO (read/write, interrupts)
- Dynamic kernel memory management
- Processor modes, stacks
  - svc, und, abt, irq, fiq
- Interrupt handlers
  - Undefined Instruction, SWI, Prefetch Abort, Data Abort, IRQ, FIQ
- System timer & scheduling
- Processes/threads, preemptive context switching
- Memory protection, address spaces (MMU)
- User/kernel interface (syscalls, utility library)

Limitations

- No filesystem
  - Everything happens in RAM!
- No dynamic loading of code
  - Everything is statically linked into kernel binary

Project structure

- drivers: device drivers and functions that interact with hardware directly
- sys: OS libraries that are for internal use by the kernel and drivers – hardware-independent
- lib: application libraries, utility functions – hardware-independent



Kernel entry function

- System initialization
  - Stacks (per CPU mode), serial port, Interrupt Vector Table, interrupts, MMU, thread management, timer
  - CP15: Coprocessor for control of cache, TLB, MMU
- Thread creation: Program code is statically linked!
- Kernel now only runs for interrupts!

```
28  __attribute__((naked, section(".init")))
29  void _start() {
30      init_stacks();
31
32      // Init the kernel memory management
33      // kmem_init((void*) KMEM_START, KMEM_SIZE);
34
35      io_dbgu_init();
36
37      interrupt_enable();
38
39      dbgu_enable();
40      printf_isr("DBGU has been enabled.\n");
41
42      dbgu_rxrdy_interrupt_enable();
43      printf_isr("DBGU RXRDY Interrupt has been enabled.\n");
44
45      printf_isr("Create Interrupt Vector Table and initialize system.\n");
46      init_ivt();
47
48      printf_isr("Initializing Advanced Interrupt Controller.\n");
49      aic_enable_system_peripherals();
50
51      printf_isr("Initializing allocation table.\n");
52      memmgmt_init_allocation_table();
53
54      printf_isr("Initializing thread management.\n");
55      thread_init_management();
56
57      printf_isr("Initializing CP15 domains.\n");
58      cp15_init_domains();
59
60      printf_isr("Welcome to ChaOS.\n");
61
62      struct thread_tcb* thread = thread_create(&main, 0, 0, 0);
63      if (thread) {
64          thread_activate(thread->id);
65      }
66
67      timer_init_real_time(32);
68      timer_init_periodical(32);
69      // Nothing should be executed after this line
70
71      while(1);
72  }
```

The idle thread

- If the kernel only runs for interrupts, what to do if there are no threads? → idle thread
- Threads can yield processing time before scheduler kicks in
  - Idle thread does “nothing”, for as little time as possible

```
36  /**
37   * The main function to be executed by the idle thread.
38   */
39  __attribute__((section(".lib")))
40  void thread_idle_text(void) {
41      // The idle thread makes a repeated SWI_THREAD_YIELD call to
42      // ensure it only runs when there is no other thread to yield to.
43      while(1) {
44          asm volatile(
45              "swi 0x20"
46              ::
47          );
48      }
49  }
```

```
17  #define SWI_STR_WRITE       0x10
18  #define SWI_STR_READ        0x11
19  #define SWI_STR_READ_FLUSH  0x12
20  #define SWI_GETC            0x1A
21
22  #define SWI_THREAD_YIELD    0x20
23  #define SWI_THREAD_EXIT     0x21
24  #define SWI_THREAD_CREATE   0x22
25  #define SWI_THREAD_SLEEP    0x23
26
27  #define SWI_MEM_MAP         0x30
```

```
91  void swi_thread_yield(struct thread_tcb* tcb) {
92      UNUSED(tcb);
93      thread_select();
94  }
```

Thread switching

- Part of normal IRQ service routine
- Detect if IRQ came from system timer
  - Therefore, preemptive multitasking
- If so, call thread switching routine

```
133  /**
134   * Interrupt Request (IRQ)
135   */
136  __attribute__ ((interrupt ("IRQ")))
137  void isr_interrupt_request(void) {
138      char c;
139      struct thread_tcb* thread;
140
141      // Interrupt from the Period Interval Timer
142      if (timer_read_PIT_status()) {
143          thread_unblock_for_timer();
144          thread_switch();
145          return;
146      }
147  ...
```

Thread switching

- Every thread gets a time slot (round-robin)
- Save thread context in Thread Control Block
- Restore context of next thread

```
292  /**
293   * Switches the running thread.
294   *
295   * Use only in the IRQ Interrupt Service Routine!
296   */
297  __attribute__((always_inline))
298  inline void thread_switch(void) {
299
300      // Check that there is a thread currently running
301      if (thread_tcb_list[thread_sched_cur_idx].status == THREAD_STATUS_RUNNING) {
302          // Do not switch if the thread has not worked through its time slot yet
303          if (thread_switch_counter++ < THREAD_ROUND_ROBIN_TIME_SLOT) {
304              return;
305          }
306          thread_switch_counter = 0;
307
308          // Save the current thread's context and set its status
309          thread_save_context(&thread_tcb_list[thread_sched_cur_idx]);
310          thread_tcb_list[thread_sched_cur_idx].status = THREAD_STATUS_READY;
311      }
312
313      thread_select();
314
315      thread_restore_context(&thread_tcb_list[thread_sched_cur_idx]);
316      thread_tcb_list[thread_sched_cur_idx].status = THREAD_STATUS_RUNNING;
317
318  }
```

Thread context saving

```
116  __attribute__((always_inline))
117  inline uint32_t* thread_get_fp(void) {
118      uint32_t* fp;
119      asm volatile (
120          "mov %[fp], fp \n\t"
121          : [fp] "=r" (fp)
122      );
123      return fp;
124  }
```

- Some registers automatically pushed on the thread's stack by CPU when IRQ
- Others plainly available or shadowed (r13-14)

- Saving frame pointer essential
  - Lots of inline code for thread switching to not mess with fp

```
126  /**
127   * Saves the context of the last running thread into its TCB.
128   * Use only in the IRQ Interrupt Service Routine!
129   *
130   * @param tcb               A pointer to the thread's TCB
131   */
132  __attribute__((always_inline))
133  inline void thread_save_context(struct thread_tcb* tcb) {
134
135      uint32_t* ptr;
136      uint8_t i;
137      uint32_t* b;
138      uint32_t s;
139
140      ptr = thread_get_fp() - 6;
141      for (i = 0; i < 4; i++) { // r0-r3
142          tcb->r[i] = ptr[i];
143      }
144
145      tcb->r[11] = ptr[4]; // fp r11
146      tcb->r[12] = ptr[5]; // ip r12
147      tcb->r[15] = ptr[6]; // pc r15
148
149      b = &tcb->r[4];
150      asm volatile ( // r4-r10
151          "stm %[rs], {r4-r10} \n\t"
152          : [rs] "=r" (b)
153      );
154
155      b = &tcb->r[13];
156      asm volatile ( // r13-r14
157          "stm %[rs], {r13-r14}^ \n\t"
158          : [rs] "=r" (b)
159      );
160
161      s = tcb->r[THREAD_REG_CPSR];
162      asm volatile ( // cpsr
163          "mrs %[rs], SPSR \n\t"
164          : [rs] "=r" (s)
165      );
166
167  }
```

Software interrupt handler

- Run desired swi function within software interrupt handler
- Thread selection outside scheduling!
  - E.g., swi_getc() blocks thread and calls to select next one
  - Don't forget to save the context

```c
/**
 * Software Interrupt (SWI)
 */
__attribute__ ((interrupt ("SWI")))
void isr_software_interrupt(void) {

    void* iptr = read_link_register() - 4;
    uint32_t inst = *(uint32_t*)iptr & 0xFF;
    uint8_t i = 0;

    struct thread_tcb* tcb = thread_get_current();
    thread_save_context(tcb);

    while (swi_types[i++]) {
        if (inst == swi_types[i-1]) {
            ((func)swi_functions[i-1])(tcb);

            tcb = thread_get_current();
            tcb->status = THREAD_STATUS_RUNNING;
            thread_restore_context(tcb);

            return;
        }
    }

    printf_isr("Unknown software interrupt 0x%x detected at address 0x%p.\n", inst, iptr);

}
```

```c
#define SWI_STR_WRITE       0x10
#define SWI_STR_READ        0x11
#define SWI_STR_READ_FLUSH  0x12
#define SWI_GETC            0x1A

#define SWI_THREAD_YIELD    0x20
#define SWI_THREAD_EXIT     0x21
#define SWI_THREAD_CREATE   0x22
#define SWI_THREAD_SLEEP    0x23

#define SWI_MEM_MAP         0x30
```

```c
void swi_getc(struct thread_tcb* tcb) {
    thread_block_for_char(tcb);
    thread_select();
}
```

Assembling the thing

- Quite some asm code
  - Part of the fun and challenge

```c
/**
 * Enables the IRQ signal.
 */
void interrupt_enable_irq(void) {

    asm volatile (
        "mrs r3, CPSR \n\t"
        "and r3, r3, #0xFFFFFF7F \n\t"
        "msr CPSR, r3 \n\t"
    );

}

/**
 * Enables the FIQ signal.
 */
void interrupt_enable_fiq(void) {

    asm volatile (
        "mrs r3, CPSR \n\t"
        "and r3, r3, #0xFFFFFFBF \n\t"
        "msr CPSR, r3 \n\t"
    );

}

/**
 * Disables the IRQ signal.
 */
void interrupt_disable_irq(void) {

    asm volatile (
        "mrs r3, CPSR \n\t"
        "orr r3, r3, #0x80 \n\t"
        "msr CPSR, r3 \n\t"
    );

}

/**
 * Disables the FIQ signal.
 */
void interrupt_disable_fiq(void) {

    asm volatile (
        "mrs r3, CPSR \n\t"
        "orr r3, r3, #0x40 \n\t"
        "msr CPSR, r3 \n\t"
    );

}
```

```c
/**
 * Writes a given number of bytes into the standard output.
 *
 * @param source    Pointer to the buffer to write bytes from
 * @param size      The number of bytes to write
 *
 * @return          The number of bytes written
 */
__attribute__((section(".lib")))
size_t write_string(char* source, size_t size) {
    size_t len;
    asm volatile(
        "mov r7, %[source] \n"
        "mov r8, %[size] \n"
        "swi 0x10 \n"
        "mov %[len], r7"
        : [len] "=r" (len)
        : [source] "r" (source), [size] "r" (size)
        : "r7", "r8"
    );
    return len;
}
```

```c
/**
 * Writes the address of the translation table base to the MMU.
 *
 * @param ptr       The address of the TTB
 */
void cp15_write_translation_table_base(uint32_t* ptr) {

    ptr = (uint32_t*) ((uint32_t)ptr & 0xFFFFC000);
    asm volatile (
        "mov r7, %[ptr] \n"
        "mcr p15, 0, r7, c2, c0, 0 \n"
        :
        : [ptr] "r" (ptr)
        : "r7"
    );

}
```

Documentation

```c
/**
 * Initializes the Domain Access Control Register.
 */
void cp15_init_domains(void) {

    /*
     * The CP 15 Register 3, or Domain Access Control Register, defines the domain's access
     * permission.
     * MMU accesses are controlled through the use of 16 domains.
     * Each field of register 3 is associated with one domain.
     *
     * The 2-bit field value allows domain access as described in the table below.
     *
     * Value    Access      Description
     * ------------------------------------------------------------
     * 00       No access   Any access generates a domain fault
     * 01       Client      Accesses are checked against the access permission
     *                      bits in the section or page descriptor
     * 10       Reserved    Reserved. Currently behaves like the no access mode
     * 11       Manager     Accesses are not checked against the access permission
     *                      bits so a permission fault cannot be generated
     */

    // Currently we are not interested in a case where we wouldn't want to check accesses
    // against the access permission, so we define one single domain where this is the case.
    asm volatile (
        "mov r1, #1 \n"
        "mcr p15, 0, r1, c3, c0, 0 \n"
        : : : "r1"
    );

}
```

```c
#define DBGUB       0xFFFFF200  // Base address

#define DBGU_CR     0x0000      // Control Register                (WRITE ONLY)
#define DBGU_MR     0x0004      // Mode Register                   (READ/WRITE)    (RESET VALUE 0x00)
#define DBGU_IER    0x0008      // Interrupt Enable Register       (WRITE ONLY)
#define DBGU_IDR    0x000C      // Interrupt Disable Register      (WRITE ONLY)
#define DBGU_IMR    0x0010      // Interrupt Mask Register         (READ ONLY)     (RESET VALUE 0x00)
#define DBGU_SR     0x0014      // Status Register                 (READ ONLY)
#define DBGU_RHR    0x0018      // Receive Holding Register        (READ ONLY)     (RESET VALUE 0x00)
                                //
                                // This holds only one value inside the first eight bits, the rest is always empty.
                                //
                                // RXCHR: Received Character
                                // Last received character if RXRDY is set.
                                //
                                // We just need to address one byte beginning in DBGU_RHR and don't need any defines
                                // for this register.
#define DBGU_THR    0x001C      // Transmit Holding Register       (WRITE ONLY)
                                //
                                // This holds only one value inside the first eight bits, the rest is always empty.
                                //
                                // TXCHR: Character to be Transmitted
                                // Next character to be transmitted after the current character if TXRDY is not set.
                                //
                                // We just need to address one byte beginning in DBGU_THR and don't need any defines
                                // for this register.
#define DBGU_BRGR   0x0020      // Baud Rate Generator Register    (READ/WRITE)    (RESET VALUE 0x00)
                                //
                                // This holds only one value inside the first 16 bits, the rest is always empty.
                                //
                                // CD: Clock Divisor
                                //
                                // CD           Baud Rate Clock
                                // ----------------------------
                                // 0            Disabled
                                // 1            MCK
                                // 2 to 65535   MCK / (CD * 16)
                                //
                                // We just need to address two bytes beginning in DBGU_BRGR and don't need any defines
                                // for this register.

// Reserved memory space from 0x0024 - 0x003C

#define DBGU_CIDR   0x0040      // Chip ID Register                (READ ONLY)
#define DBGU_EXID   0x0044      // Chip ID Extension Register      (READ ONLY)
                                // Adding further specifications for these two registers would add a tremendous
                                // overhead to this file for something that is most definitely not going to be
                                // needed.
                                // Therefore, we have left it out. If needed, it can still be added in.
```

- Thorough manual work – no shortcuts, no AI

🪄 Demo 🪄