



TECHNISCHE
UNIVERSITÄT
DRESDEN

Fakultät Informatik Institut für Systemarchitektur, Professur für Betriebssysteme

OPERATING-SYSTEM CONSTRUCTION

Material based on slides by Olaf
Spinczyk, Universität Osnabrück

Operating-System Development 101

<https://tud.de/inf/os/studium/vorlesungen/betriebssystembau>

HORST SCHIRMEIER



<https://pingo.coactum.de/816711>

OS Development (Not Always Comfy)

- **First Steps**

How to get your OS onto the target hardware?

- Compilation/Linking
- Boot process

- **Testing and Debugging**

What to do if your system doesn't respond?

- "printf debugging"
- Emulators, virtual machines
- Debuggers
- Remote Debugging
- Hardware support

OS Development (Not Always Comfy)

- **First Steps**

How to get your OS onto the target hardware?

- Compilation/Linking
- Boot process

- **Testing and Debugging**

What to do if your system doesn't respond?

- "printf debugging"
- Emulators, virtual machines
- Debuggers
- Remote Debugging
- Hardware support

Compilation/Linking – *Hello, World*

```
#include <iostream>

int main () {
    std::cout << "Hello, World" << std::endl;
}
```

```
$ g++ -o hello hello.cc
```

- Assumption:
 - Development system runs an x86 Linux
 - Target system also is a PC
- Does this program also run on **bare metal**?
- Is OS development in a **high-level programming language** possible at all?

Compilation/Linking – Problems and Solutions

- No dynamic linker available
 - link all necessary libraries statically
- libstdc++ and libc use Linux system calls (e.g., write)
 - We **cannot use** regular C/C++ runtime libraries. (We usually don't have alternatives either.)
- Generated addresses refer to *virtual* memory ("nm hello | grep main" yields "0000000000404745 T main")
 - We cannot use standard linker settings but **need a custom linker config**.
- High-level language code: environment expectations (CPU-register usage, address mapping, runtime environment, stack, ...)
 - Own **startup code** (written in assembler) must prepare high-level language code execution.



Booting

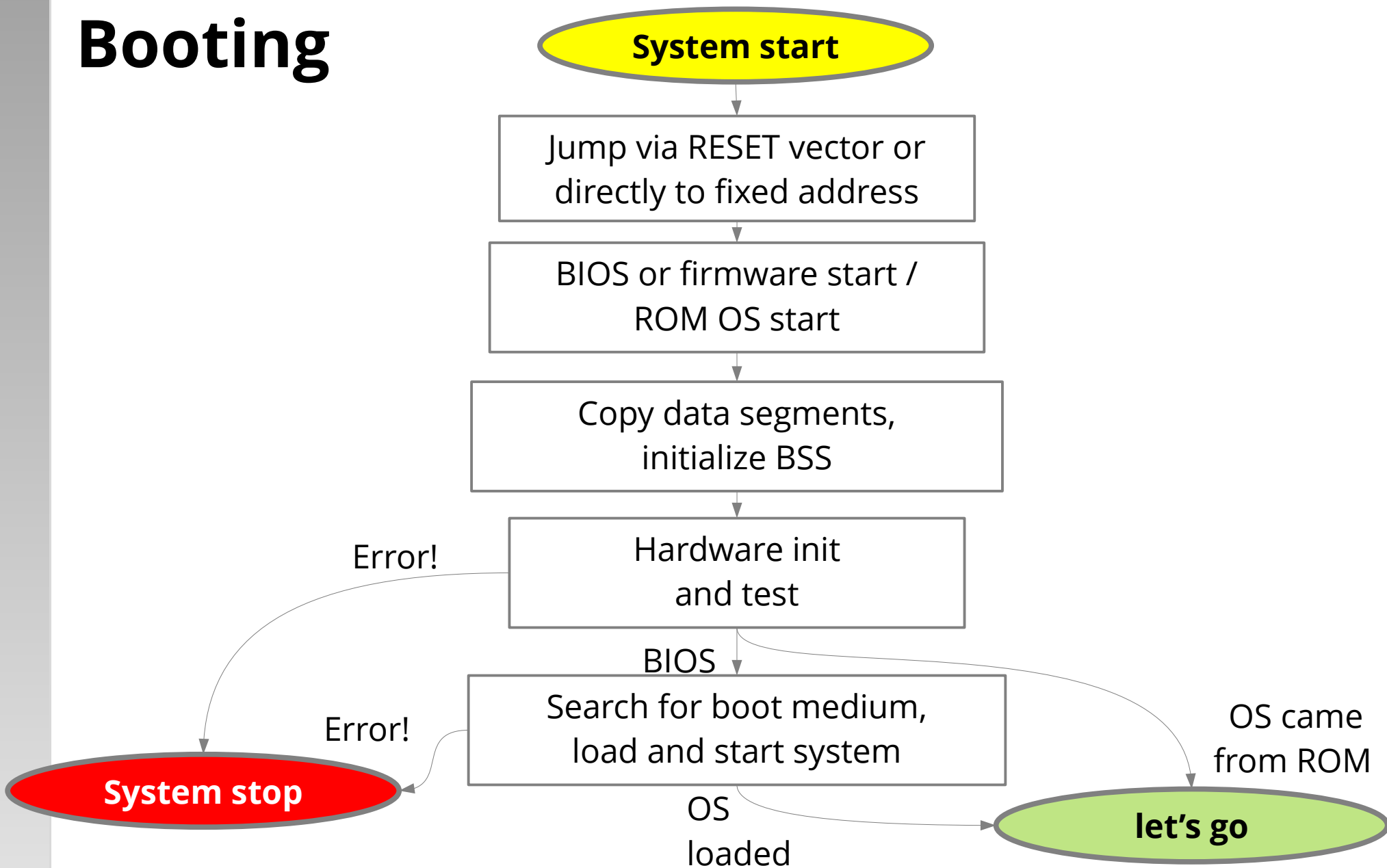
*“**Boot** is short for **bootstrap** or **bootstrap load** and derives from the phrase **to pull oneself up by one's bootstraps.**”*

*“**Booting** is the process of starting a computer, specifically with regard to starting its software. The process involves a chain of stages, in which at each stage, **a smaller, simpler program** loads and then executes the **larger, more complicated program** of the next stage.”*

The term is sometimes attributed to a story in Rudolf Erich Raspe's The Surprising Adventures of Baron Munchausen, but in that story Baron Munchausen pulls himself (and his horse) out of a swamp by his hair (specifically, his pigtail), not by his bootstraps – and no explicit reference to bootstraps has been found elsewhere in the various versions of the Munchausen tales.

en.wikipedia.org

Booting



PC Booting – Boot Sector

- PC BIOS loads 1st block (512 bytes) of boot drive at address 0x7c00 and jumps there (“blindly”)

- Boot-sector layout

FAT disk (DOS/Windows)

Offset	Inhalt
0x0000	jmp boot; nop; (ebx90)
0x0003	System name and version
0x000b	Bytes per sector
0x000d	Sectors per cluster
0x000e	reserved sectors (for boot record)
0x0010	number of FATs
0x0011	number of root-directory entries
0x0013	number of logical sectors
0x0015	media descriptor byte
0x0016	sectors per FAT
0x001a	number of heads
0x001c	number of hidden sectors
0x001e	boot : ...
0x01fe	0xaa55

PC Booting – Boot Sector

- PC BIOS loads 1st block (512 bytes) of boot drive at address 0x7c00 and jumps there (“blindly”)

- Boot-sector layout

Alternative (OOStuBS)

Offset	Inhalt
0x0000	<code>jmp boot; nop; (ebx90)</code>
0x0003	System name and version
0x000b	Bytes per sector
0x000d	Sectors per cluster
0x000e	reserved sectors (for boot record)
0x0010	number of FATs
0x0011	number of root-directory entries
0x0013	number of logical sectors
0x0015	media descriptor byte
0x0016	sectors per FAT
0x001a	number of heads
0x001c	number of hidden sectors
0x001e	<code>boot :</code> ...
0x01fe	<code>0xaa55</code>

In fact, only the beginning and the **“signature” (0xaa55)** at the end matters. Everything else is used by the **boot loader** to load the actual system.

PC Booting – Boot Loader

- Simple, **system-specific** boot loaders
 - Define hardware/software state
 - If necessary: Load further blocks with boot-loader code
 - Pinpoint the actual system on the boot media
 - Load the system (via BIOS functions)
 - Jump into loaded system
- Boot loader on disks not flagged as “bootable”
 - Error message, halt / reboot
- Boot loader with **boot menu** (e.g., GRUB)
(for example in the **Master Boot Record** of a HDD)
 - Display a menu
 - Emulate BIOS when booting the selected system
(load boot block to 0x7c00, jump)

OS Development (Not Always Comfy)

- **First Steps**

How to get your OS onto the target hardware?

- Compilation/Linking
- Boot process

- **Testing and Debugging**

What to do if your system doesn't respond?

- "printf debugging"
- Emulators, virtual machines
- Debuggers
- Remote Debugging
- Hardware support

Debugging

1947

9/9

0800 Antam started
 1000 " stopped - antam ✓
 13⁰⁰ (032) MP - MC ~~1.552647000~~ { 1.2700 9.037 847 025
 (033) PRO 2 2.130476415 } 9.037 846 995 connect
 connect 2.130676415 4.615925059(-2)

Relays 6-2 in 033 failed special speed test
 in relay " 10.000 test.

Relay 2145
 Relay 3376

1100 Started Cosine Tape (Sine check)
 1525 Started Multi Adder Test.

1545



Relay #70 Panel F
 (moth) in relay.

First actual case of bug being found.

1630 Antam started.
 1700 closed down.



Admiral Grace Hopper

Source: Wikipedia

“printf Debugging”

- Not that simple – if you don’t have a (working) printf
 - Often you don’t even have a display.
- printf() often changes the debuggee’s behavior
 - Problem vanishes / changes symptoms
 - Unfortunately particularly true for OS development
- Last resort:
 - blinking LED
 - serial interface

(Software) Emulators

- Emulate real hardware in software
 - Simplifies debugging
(Emulation software usually more communicative than real HW)
 - Shorter development cycles
- **Careful:** In the end, the system must run on real hardware!
 - Emulator and real hardware may differ in details!
 - Harder to find bugs in a complete system than during incremental development
- Emulation: a special case of **virtualization**
 - Provides a **virtual resource Y** (e.g., an Arm CPU)
based on a resource X (e.g., the systems x86-64 host CPU)

Emulators – Example “Bochs”

- Emulates i386, ..., Pentium, x86-64 (interpreter loop)
 - plus MMX, SSE–SSE4, 3DNow! instructions
 - Multiprocessor emulation
- Emulates a complete PC
 - Memory, devices (including sound, networking, ...)
 - Capable to run Windows, Linux
- Implemented in C++
- Development support
 - Logs helpful info, e.g. from crash
 - Built-in debugger (*GDB stub*)



Bochs in Bochs

Debugging

- Debugger helps locating software bugs by tracing/controlling the debuggee:
 - **Single-step mode**
 - **Breakpoints:** trigger when reaching a particular machine instruction
 - **Watchpoints:** trigger when a particular data element is accessed
- **Careful:** Bug-hunting might take *longer* when using a debugger
 - Taking a break and thinking about the problem can be more time-efficient
 - Single-stepping costs a lot of time
 - Often no way back in case you miss the problematic instruction
 - “printf debugging” allows better control over output format
 - Synchronization / race-condition bugs are impractical to debug with a debugger
- helpful: “Core dump” analysis
 - but of little relevance during OS development :-)

Debugging – Example Session

Setting a
breakpoint

Running the
program

Single-stepping

Continuing

```
$ g++ -static -g -o hello hello.cc
$ gdb hello
GNU gdb (Ubuntu 11.1-0ubuntu2) 11.1
...
(gdb) break main
Breakpoint 1 at 0x40474d: file hello.cc, line 4.
(gdb) run
Starting program: hello

Breakpoint 1, main () at hello.cc:4
4          std::cout << "Hello, World" << std::endl;
(gdb) next
Hello, World
5      }
(gdb) next
0x00000000004a7f4a in __libc_start_call_main ()
(gdb) continue
Continuing.
[Inferior 1 (process 663394) exited normally]
(gdb) quit
```

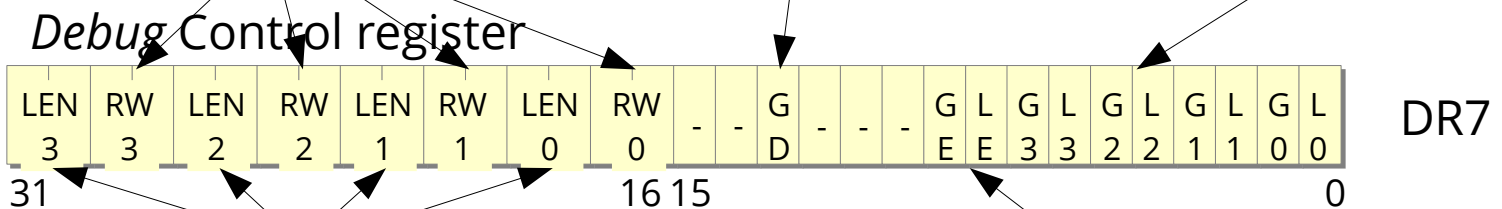
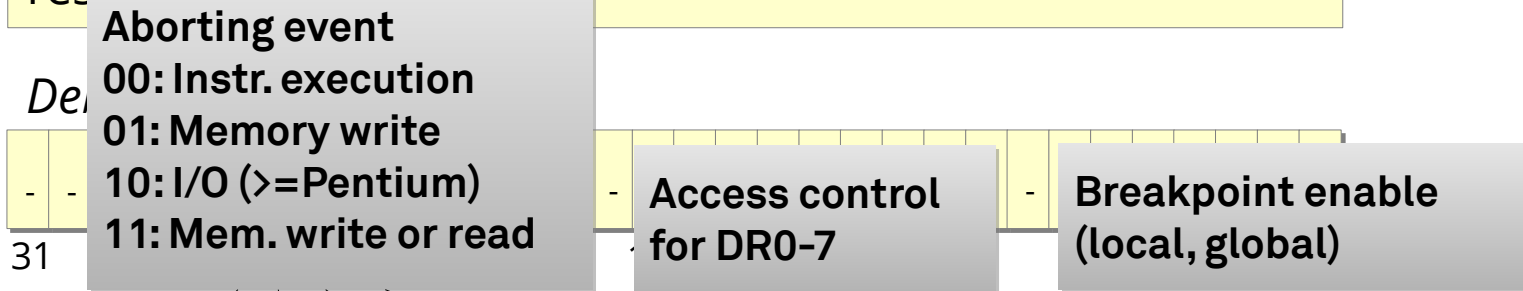
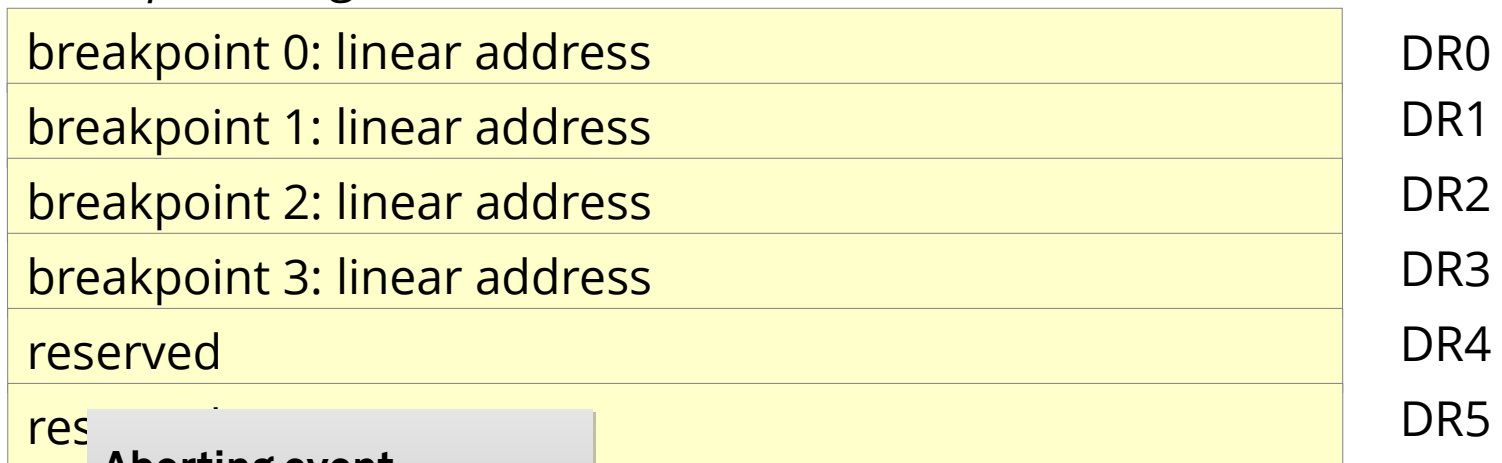
Debugging – Technical Background (1)

- Practically all CPUs support debugging
- Example: Intel x86
 - **INT3** instruction triggers a “*breakpoint interrupt*” (in fact a *trap*)
 - User “sets breakpoint”, debugger (at runtime) places INT3 in program code
 - Trap handler redirects control flow to debugger
 - **enabled Trap Flag (TF)** in status register (EFLAGS / RFLAGS): trigger “*debug interrupt*” after every instruction
 - Can be used for implementing single-stepping in the debugger
 - Trap handler is *not* executed in single-stepping mode
 - **Debug Registers DR0–DR7** can monitor up to 4 breakpoints or watchpoints
 - No code manipulation necessary: breakpoints in ROM/FLASH or read-only memory segments (e.g. *shared libraries!*)
 - Efficient watchpoints only possible through this mechanism

Debugging - Technical Background (2)

80386 Debug Registers

Breakpoint Register



Debugging – Technical Background (3)

- For debugging **regular user-space applications**, the OS must provide an interface
 - e.g. Linux: **ptrace (2)**

```
#include <sys/ptrace.h>
long ptrace(enum __ptrace_request request, pid_t pid,
            void *addr, void *data);
```

Request (PTRACE_...)	Semantics
TRACEME	Indicate that this process is to be traced by its parent
ATTACH, DETACH	Seize control over another process (alt. to TRACEME)
PEEKTEXT, PEEKDATA, PEEKUSER	Read data from debuggee's address space
POKETEXT, POKEDATA, POKEUSER	Change data in debuggee's address space
SYSCALL, CONT	Monitor system calls and continue
SINGLESTEP	Single-stepping mode (machine instruction granularity)
KILL	Abort debuggee

Debugging – Technical Background (4)

```
int main(void) {
    long long counter = 0; /* machine instruction counter */
    int wait_val;          /* child's return value */
    int pid;              /* child's process id */

    puts("Please wait");
    pid = fork();         /* create child process */
    if (pid == -1)        /* failed to create child process */
        perror("fork");
    else if (pid == 0) { /* child process starts */
        ptrace(PTRACE_TRACEME, 0, 0, 0); /* allow parent to control child */
        execl("/bin/ls", "ls", NULL); /* run child program (ls) and terminate*/
    }
    else {                /* parent process starts */
        /* wait for SIGTRAP */
        while (wait(&wait_val) != 1 && WIFSTOPPED(wait_val) && WSTOPSIG(wait_val)) {
            counter++;
            if (ptrace(PTRACE_SINGLESTEP, pid, 0, 0) != 0) { /* enable single step mode */
                perror("ptrace");
                break;
            }
        }
    }
    printf("Number of machine instructions : %lld\n", counter);
    return 0;
} }
```

**ptrace(2)
example**

Debugging – Technical Background (5)

- User expects **source-code** visualization: *source-level debugging*
 - **Prerequisites:** access to sources, (compiler-generated) *debug information*

```
$ g++ -g -o hello hello.cc
$ objdump --section-headers hello
hello:      file format elf64-x86-64
Sections:
Idx Name          Size      VMA           LMA           File off  Algn
...
 24 .data           00000010  0000000000004000  0000000000004000  00003000  2**3
                CONTENTS, ALLOC, LOAD, DATA
 25 .bss            00000118  0000000000004040  0000000000004040  00003010  2**6
                ALLOC
 26 .comment       00000025  0000000000000000  0000000000000000  00003010  2**0
                CONTENTS, READONLY
 27 .debug_aranges  00000030  0000000000000000  0000000000000000  00003035  2**0
                CONTENTS, READONLY, DEBUGGING, OCTETS
 28 .debug_info    000023bb  0000000000000000  0000000000000000  00003065  2**0
                CONTENTS, READONLY, DEBUGGING, OCTETS
 29 .debug_abbrev  0000059b  0000000000000000  0000000000000000  00005420  2**0
                CONTENTS, READONLY, DEBUGGING, OCTETS
 30 .debug_line    0000014a  0000000000000000  0000000000000000  000059bb  2**0
                CONTENTS, READONLY, DEBUGGING, OCTETS
 31 .debug_str     0000120b  0000000000000000  0000000000000000  00005b05  2**0
                CONTENTS, READONLY, DEBUGGING, OCTETS
 32 .debug_line_str 0000028b  0000000000000000  0000000000000000  00006d10  2**0
                CONTENTS, READONLY, DEBUGGING, OCTETS
```


Remote Debugging

- Allows debugging programs on platforms we cannot (yet) work on interactively
 - Requires **communications link** (serial, Ethernet, ...)
 - ... which in turn necessitates a **device driver**
 - Target “device” can also be an emulator (e.g., QEMU)
- Debugging component on the target system (“stub”) should be as simple as possible



Remote Debugging – Example GDB (1)

- Communication protocol
(“GDB Remote Serial Protocol” – RSP)
 - Reflects requirements on GDB *stub*
 - Based on transferring ASCII strings
 - Message format: **\$**<command or reply>**#**<checksum>
 - Messages are directly acknowledged with **+** (OK) or **-** (error)
- Examples:
 - **\$g#67** ▶ Read contents of all registers
 - Reply: **+\$123456789abcdef0...#...** ▶ Reg. 1 = 0x12345678, 2 = 0x9...
 - **\$G123456789abcdef0...#...** ▶ Set register contents
 - Reply: **+\$OK#9a** ▶ Success
 - **\$m4015bc,2#5a** ▶ Read 2 bytes starting at address 0x4015bc
 - Reply: **+\$2f86#06** ▶ Value 0x2f86

Remote Debugging – Example GDB (2)

- Communication protocol – all command categories:
 - Register and memory commands
 - read/write all registers
 - **read/write single register**
 - **read/write memory area**
 - Controlling program execution
 - request reason for latest interruption
 - single-step
 - **continue execution**
 - Miscellaneous
 - Print to debug console
 - Error messages

Minimum stub functionality

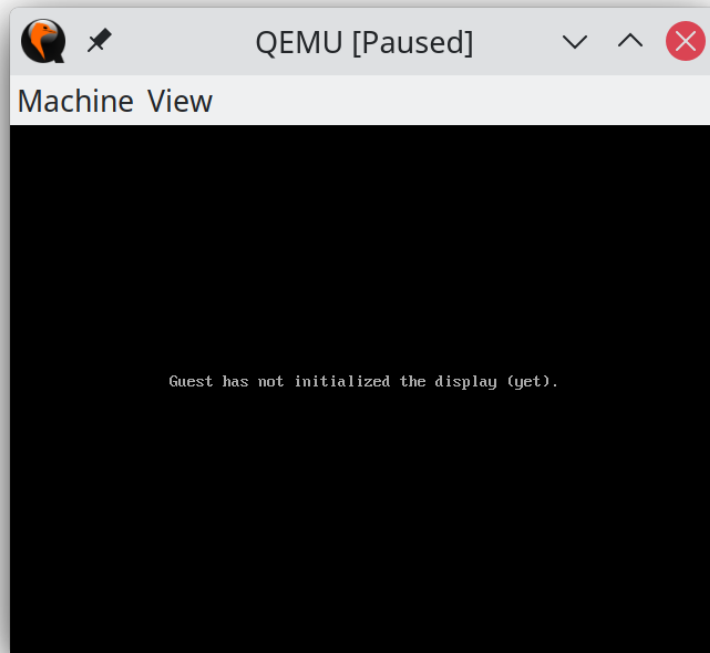


Remote Debugging – with QEMU

- With the right command-line parameters, QEMU offers a GDB stub communicating via TCP

```
$ make qemu-gdb
```

```
...
```



Remote Debugging - with Bochs

```
$ gdb build/system
GNU gdb (Ubuntu 11.1-0ubuntu2) 11.1
...
Reading symbols from build/system...
(gdb) break main
Breakpoint 1 at 0x10167f: file main.cc, line 11.
(gdb) target remote localhost:2024
Remote debugging using localhost:2024
0x0000000000000000fff0 in ?? ()
(gdb) continue
Continuing.

Breakpoint 1, main () at main.cc:4
4      {
(gdb) next
11          return 0;
(gdb) continue
Continuing.
```

**Automated in OOSTuBS
Makefile to prevent TCP-port
collisions:**

make gdb

(and skip the `target remote ...` step)

Debugging *Deluxe*

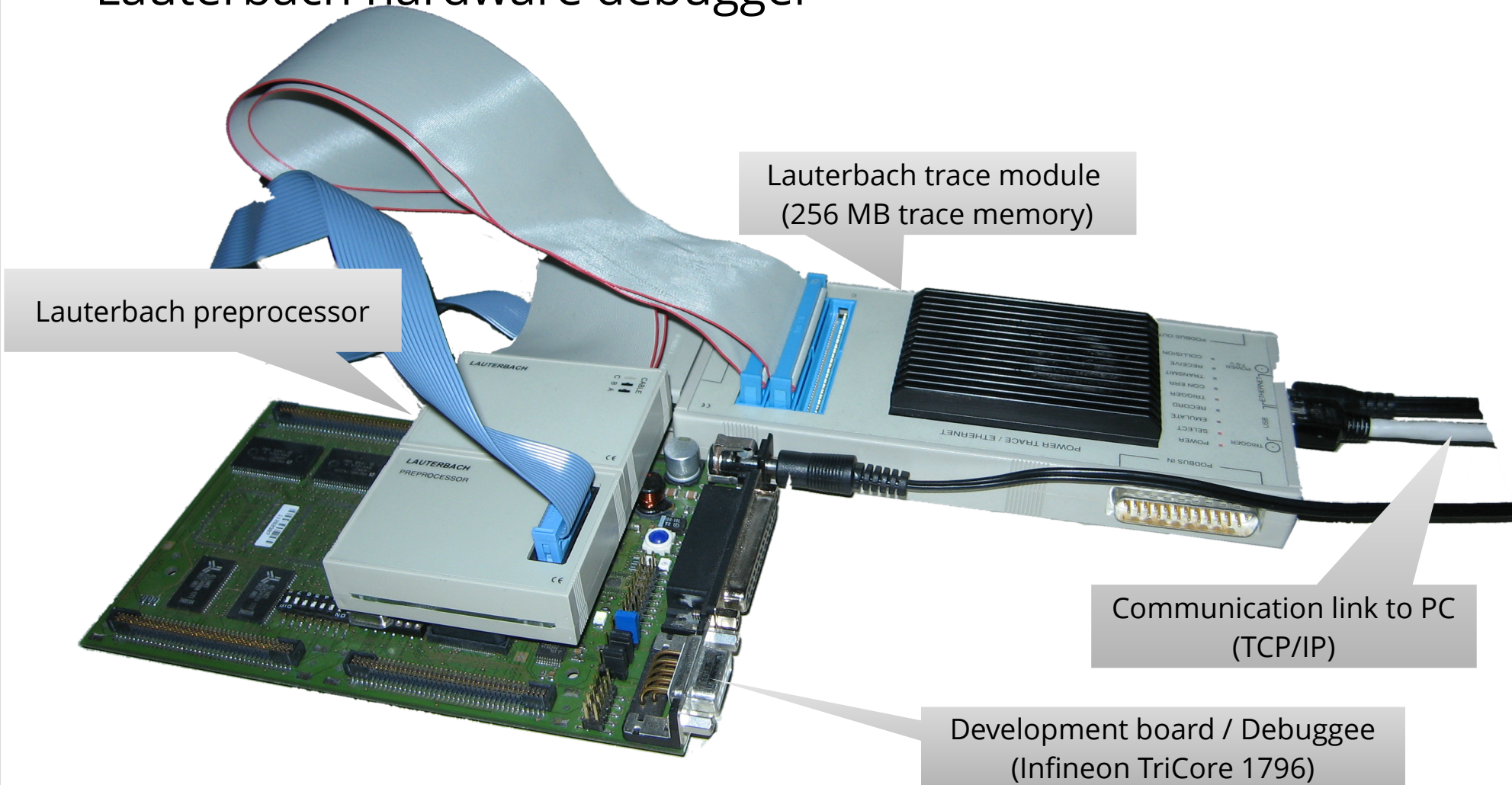
- Many chip manufacturers integrate **hardware support for debugging** (OCDS – *On Chip Debug System*)
 - BDM, OnCE, MPD, JTAG
- Usually simple serial protocols between debugging unit and external debugger (save chip pins!)
- Advantages:
 - *Debug Monitor* (e.g. gdb stub) does not use any application memory
 - Debug Monitor implementation unnecessary
 - ROM/FLASH breakpoints using hardware breakpoints
 - Concurrent access to memory and CPU registers
 - Specialized hardware partially allows to record a control-flow trace (ex post analysis)

Debugging *Deluxe* – Example BDM

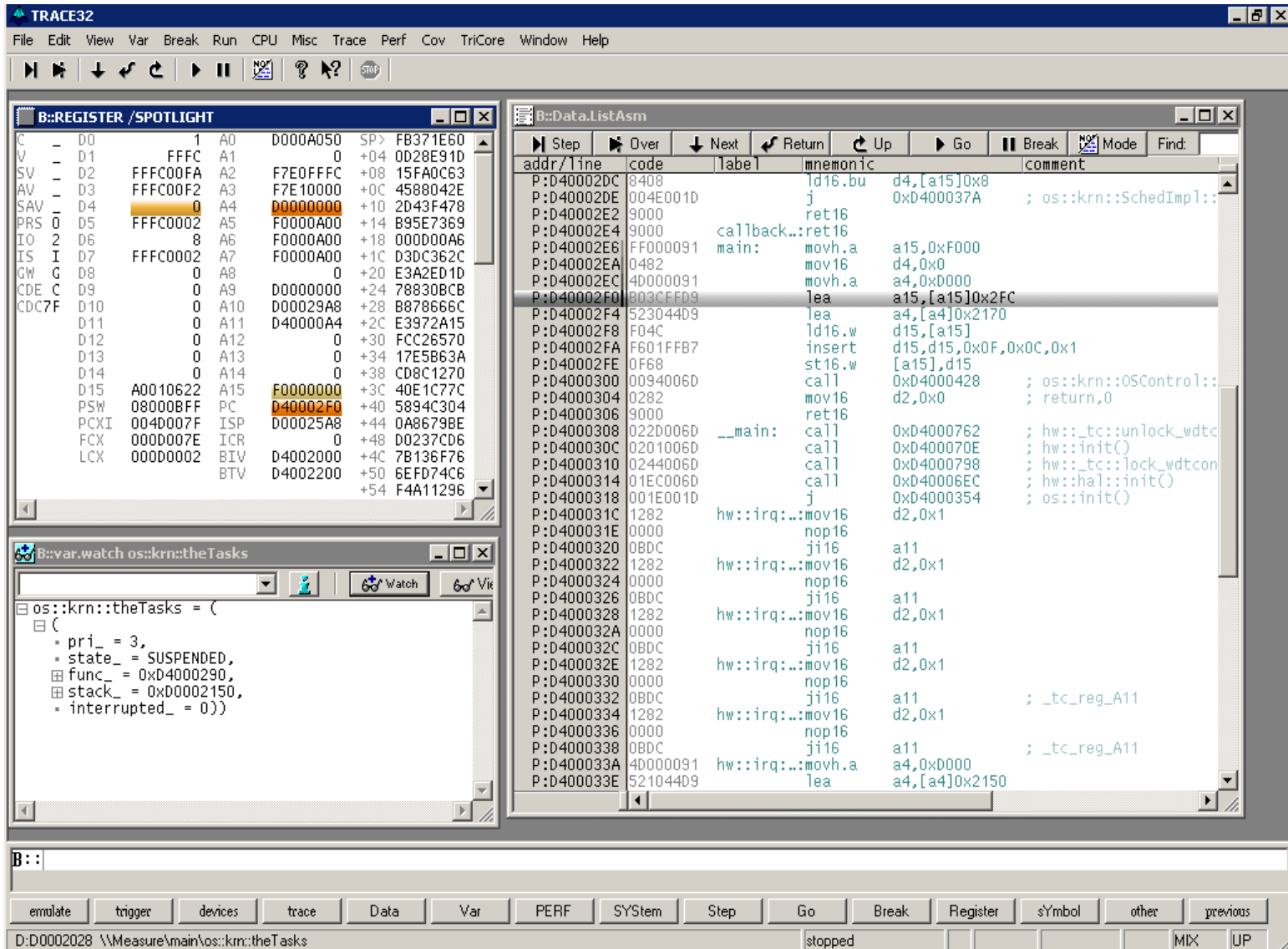
- “Background Debug Mode” – on-chip debug solution by Motorola
- Serial communication via 3 lines (DSI, DSO, DSCLK)
- BDM commands of 68k and ColdFire processors:
 - RAREG/RDREG – Read Register
 - read particular data or address register
 - WAREG/WDREG – Write Register
 - write particular data or address register
 - READ/WRITE – Read Memory/Write Memory
 - read/write specific memory location
 - DUMP/FILL – Dump Memory/Fill Memory
 - read/fill block of memory
 - BGND/GO – Enter BDM/Resume
 - stop/continue execution

Debugging *Deluxe* - Hardware Solution

- Lauterbach hardware debugger



Debugging *Deluxe* - Lauterbach Frontend



The screenshot displays the TRACE32 debugger interface with the following components:

- TRACE32 Title Bar:** File Edit View Var Break Run CPU Misc Trace Perf Cov TriCore Window Help
- Navigation Bar:** Play, Stop, Step Over, Step Into, Step Out, Step Return, Step Up, Step Down, Break, Mode, Find.
- B::REGISTER /SPOTLIGHT:**

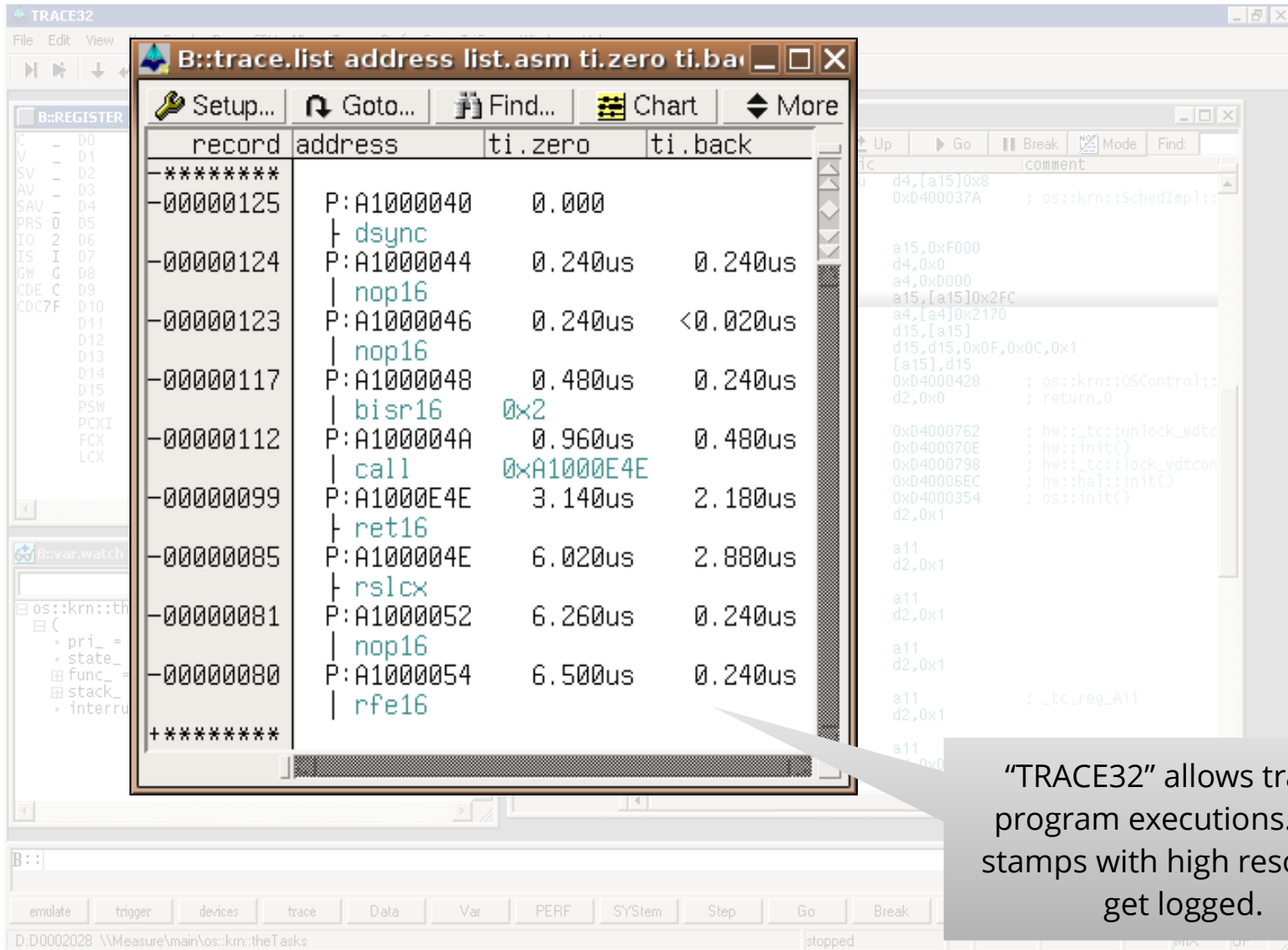
C	D0	1	A0	D000A050	SP>	FB371E60
V	D1	FFFC	A1	+04	0028E91D	
SV	D2	FFFC00FA	A2	F7E0FFFC	+08	15FA0C63
AV	D3	FFFC00F2	A3	F7E10000	+0C	4588042E
SAV	D4	0	A4	F0000000	+10	2D43F478
PRS	D5	FFFC0002	A5	F0000A00	+14	895E7369
IO	D6	8	A6	F0000A00	+18	000000A6
IS	D7	FFFC0002	A7	F0000A00	+1C	D3DC362C
GW	D8	0	A8	0	+20	E3A2ED1D
CDE	D9	0	A9	D0000000	+24	78830BCB
CDC7F	D10	0	A10	D00029A8	+28	B878666C
	D11	0	A11	D4000A4	+2C	E3972A15
	D12	0	A12	0	+30	FCC26570
	D13	0	A13	0	+34	17E5B63A
	D14	0	A14	0	+38	CD8C1270
	D15	A0010622	A15	F0000000	+3C	40E1C77C
PSW	08000BFF	PC		D40002F0	+40	5894C304
PXCI	004D007F	ISP		D00025A8	+44	0A8679BE
FCX	000D007E	ICR		0	+48	0D237CD6
LCX	000D0002	BIV		D4002000	+4C	7B136F76
		BTV		D4002200	+50	6EFD74C6
					+54	F4A11296
- B::Data.ListAsm:**

addr/line	code	label	mnemonic	comment
P:D40002DC	8408		ld16.bu	d4,[a15]0x8
P:D40002DE	004E001D		j	0xD400037A ; os::krn::SchedImpl::
P:D40002E2	9000		ret16	
P:D40002E4	9000	callback::	ret16	
P:D40002E6	FF000091	main:	movh.a	a15,0xF000
P:D40002EA	0482		mov16	d4,0x0
P:D40002EC	4D000091		movh.a	a4,0xD000
P:D40002F0	803CFE09		lea	a15,[a15]0x2FC
P:D40002F4	523044D9		lea	a4,[a1]0x2170
P:D40002F8	F04C		ld16.w	d15,[a15]
P:D40002FA	F601FFB7		insert	d15,d15,0x0F,0x0C,0x1
P:D40002FE	0F68		st16.w	[a15],d15
P:D4000300	0094006D		call	0xD4000428 ; os::krn::OSControl::
P:D4000304	0282		mov16	d2,0x0 ; return,0
P:D4000306	9000		ret16	
P:D4000308	022D006D	__main:	call	0xD4000762 ; hw::_tc::unlock_wdct
P:D400030C	02D1006D		call	0xD400070E ; hw::init()
P:D4000310	0244006D		call	0xD4000798 ; hw::_tc::lock_wdctcon
P:D4000314	01EC006D		call	0xD40006EC ; hw::hal::init()
P:D4000318	001E001D		j	0xD4000354 ; os::init()
P:D400031C	1282	hw::irq::	mov16	d2,0x1
P:D400031E	0000		nop16	
P:D4000320	0BDC		ji16	a11
P:D4000322	1282	hw::irq::	mov16	d2,0x1
P:D4000324	0000		nop16	
P:D4000326	0BDC		ji16	a11
P:D4000328	1282	hw::irq::	mov16	d2,0x1
P:D400032A	0000		nop16	
P:D400032C	0BDC		ji16	a11
P:D400032E	1282	hw::irq::	mov16	d2,0x1
P:D4000330	0000		nop16	
P:D4000332	0BDC		ji16	a11 ; _tc_reg_A11
P:D4000334	1282	hw::irq::	mov16	d2,0x1
P:D4000336	0000		nop16	
P:D4000338	0BDC		ji16	a11 ; _tc_reg_A11
P:D400033A	4D000091	hw::irq::	movh.a	a4,0xD000
P:D400033E	521044D9		lea	a4,[a4]0x2150
- B::var.watch os::krn::theTasks:**

```

os::krn::theTasks = (
  (
    * pri_ = 3,
    * state_ = SUSPENDED,
    * func_ = 0xD4000290,
    * stack_ = 0x00002150,
    * interrupted_ = 0))
  
```
- Bottom Bar:** emulate trigger devices trace Data Var PERF SYStem Step Go Break Register sYmbol other previous
- Status Bar:** D:D0002028 \Measure\main\os::krn::theTasks | stopped | MIX | UP

Debugging *Deluxe* - Lauterbach Frontend



The screenshot shows the TRACE32 interface with a window titled "B::trace.list address list.asm ti.zero ti.ba". The window contains a table of execution records:

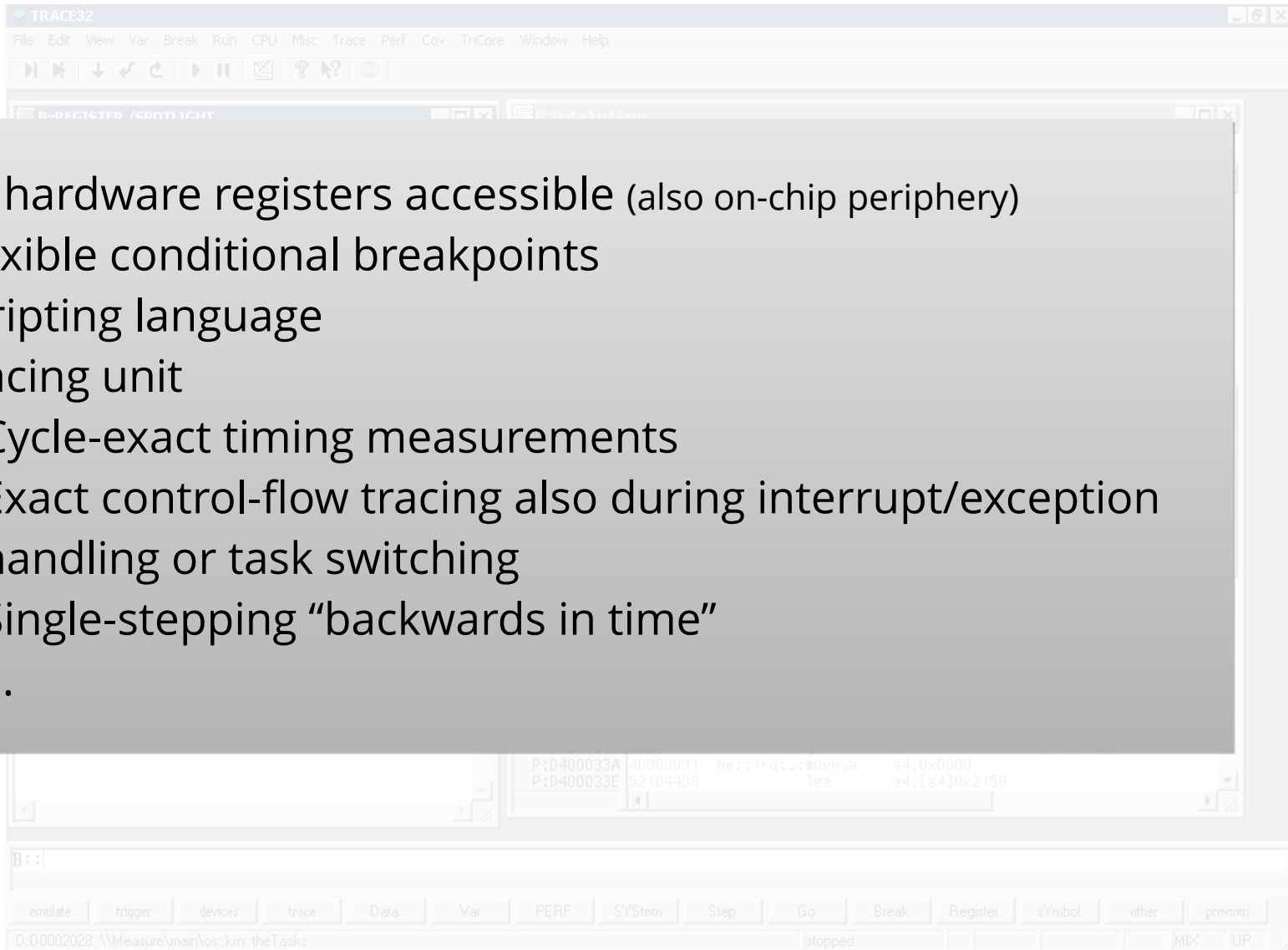
record	address	ti.zero	ti.back

-00000125	P: A1000040 dsync	0.000	
-00000124	P: A1000044 nop16	0.240us	0.240us
-00000123	P: A1000046 nop16	0.240us	<0.020us
-00000117	P: A1000048 bisr16 0x2	0.480us	0.240us
-00000112	P: A100004A call 0xA1000E4E	0.960us	0.480us
-00000099	P: A1000E4E ret16	3.140us	2.180us
-00000085	P: A100004E rslcx	6.020us	2.880us
-00000081	P: A1000052 nop16	6.260us	0.240us
-00000080	P: A1000054 rfe16	6.500us	0.240us

A callout box points to the table with the text: "TRACE32" allows tracing program executions. Time stamps with high resolution get logged.

Debugging *Deluxe* - Lauterbach Frontend

- All hardware registers accessible (also on-chip periphery)
- Flexible conditional breakpoints
- Scripting language
- Tracing unit
 - Cycle-exact timing measurements
 - Exact control-flow tracing also during interrupt/exception handling or task switching
 - Single-stepping “backwards in time”
 - ...



Summary

- Operating-system development differs significantly from regular application development:
 - No libraries
 - Bare metal is the basis we build upon
- The first steps are often the hardest
 - Compilation/linking, booting, system initialization
- Comfortable bug hunting necessitates infrastructure
 - Device drivers for “printf debugging”
 - Stub and communication link/driver for remote debugging
 - Hardware debugging support like with BDM
 - Ideal: Professional hardware debuggers (e.g. Lauterbach)