Systemnahe Software I
WS 2011/2012

Andreas F. Borchert
Universität Ulm

18. Oktober 2011

Widmung

Ausschnitt eines Fotos von Denise Panyik-Dale, CC-BY-2.0

Diese Vorlesung ist Dennis M. Ritchie gewidmet, der am 12. Oktober 2011 verstorben ist und ohne dessen Beiträge diese Vorlesung in dieser Form nicht denkbar wäre. Dennis M. Ritchie hat nicht nur die Programmiersprache C entworfen und implementiert, sondern auch zusammen mit Ken Thompson Unix entwickelt. Mit C wurde erstmals eine höhere Programmiersprache statt Assembler für die Systemprogrammierung verwendet. Dennis Ritchie und seinen Kollegen gelang es bei Unix, die Betriebsystemsschnittstellen und die Systemwerkzeuge in einem bis dahin unbekanntem Maß zu vereinfachen. Auf sie geht die Umsetzung des wichtigen Grundsatzes zurück, dass jedes Werkzeug genau eine Aufgabe zu erfüllen habe und das möglichst gut.

Syllabus

Inhalte:

- Einführung in die Programmiersprache C
- Dynamische Speicherverwaltung
- Entwicklungswerkzeuge im Umfeld von C
- Dateisysteme
- Systemnahe Programmierung

Syllabus: Ziele

- Erwerb von Grundkenntnissen der Programmiersprache C, wobei ein besonderer Wert gelegt wird auf den Umgang mit der dynamischen Speicherverwaltung und mit den Zeigern in C. Ziel ist es auch, den versehentlichen Einbau von Sicherheitslücken zu vermeiden.
- Erlernen des Umgangs mit den klassischen Entwicklungswerkzeugen unter UNIX wie beispielsweise make.
- Verständnis der Abstraktion eines Dateisystems, einiger Implementierungen und praktische Erfahrungen mit der zugehörigen System-Schnittstelle.

Syllabus: Prüfung

- Es gibt am Ende des Semesters eine schriftliche Prüfung (Termin steht noch nicht fest).
- Zur Teilnahme an der Prüfung ist eine Vorleistung erforderlich. Diese ist bei einer erfolgreichen Teilnahme an den Übungen gegeben (50% der Übungspunkte).
- Der Umfang beträgt 6 Leistungspunkte.
- Studiengänge:
  - **Bachelor:** Mathematik, Wirtschaftsmathematik, Informatik, Medieninformatik, Software Engineering, Physik, Wirtschaftsphysik und Elektrotechnik.
  - **Master:** Informatik und Medieninformatik (Technische und Systemnahe Informatik).
  - **Diplom:** Mündliche Prüfung. Dazu bitte mit mir Kontakt aufnehmen.

Syllabus: Voraussetzungen

- Grundkenntnisse in Informatik. Insbesondere sollte keine Scheu davor bestehen, etwas zu programmieren.
- Freude daran, etwas auch an einem Rechner auszuprobieren und genügend Ausdauer, dass nicht beim ersten Fehlversuch aufgegeben wird.

Syllabus: Struktur

- Jede Woche gibt es zwei Vorlesungsstunden an jedem Dienstag von 16-18 Uhr im H3.
- Die Übungen finden am Donnerstag von 16-18 Uhr im H12 statt.
- Organisatorische Feinheiten werden in der ersten Übungsstunde erläutert.
- Webseite: http://www.mathematik.uni-ulm.de/sai/ws11/soft1/

Syllabus: Übungen

- Wer an den Übungen teilnehmen möchte (zwingende Voraussetzung für eine Teilnahme an der schriftlichen Prüfung), der muss sich über SLC für die Vorlesung registrieren.
- Ebenso sollten alle über einen Shell-Zugang zu unseren Servern (wie z.B. die Theseus) verfügen.
- Für die Übungen werden Gruppen gebildet und die Gruppe trifft sich dann mit dem Tutor.
- Alle Gruppenmitglieder sollten jeweils an der Lösung mitwirken und mit der gesamten Lösung vertraut sein. Die Tutoren dürfen durchaus die einzelnen Gruppenmitglieder individuell mit Punkten bewerten.
- Einzelheiten werden in der ersten Übungsstunde am Donnerstag, den 20. Oktober, um 16 Uhr im H12 vorgestellt.

Syllabus: Skript

- Es gibt ein Skript (entwickelt von mehreren Mitgliedern unseres Instituts), das auf der Vorlesungswebseite zur Verfügung steht.
- Parallel gibt es gelegentlich Präsentationen (wie diese), die ebenfalls als PDF zur Verfügung gestellt werden.
- Wenn Sie das Skript oder die Präsentationen ausdrucken möchten, nutzen Sie dazu bitte die entsprechenden Einrichtungen des KIZ. Im Prinzip können Sie dort beliebig viel drucken, wenn Sie genügend Punkte dafür erworben haben.
- Das Druck-Kontingent, das Sie bei uns kostenfrei erhalten (das ist ein Privileg und kein natürliches Recht), darf für die Übungen genutzt werden, jedoch nicht für das Ausdrucken von Skripten oder Präsentationen.

Syllabus: Sprechstunden

- Sie sind eingeladen, mich jederzeit per E-Mail zu kontaktieren:
  E-Mail: `andreas.borchert@uni-ulm.de`
- Meine reguläre Sprechzeit ist am Mittwoch 14-16 Uhr. Zu finden bin ich in der Helmholtzstraße 18, Zimmer E02.
- Zu anderen Zeiten können Sie auch gerne vorbeischauen, aber es ist dann nicht immer garantiert, daß ich Zeit habe. Gegebenenfalls lohnt sich vorher ein Telefonanruf: 23572.

Syllabus: Nicht verzweifeln!

- Immer wieder kann es mal vorkommen, dass es zu scheinbar unlösbaren Problemen bei einer Übungsaufgabe kommt.
- Geben Sie dann bitte nicht auf. Nutzen Sie unsere Hilfsangebote.
- Sie können (und sollen) dazu gerne Ihren Tutor oder Tutorin kontaktieren oder den Übungsleiter Markus Schnalke oder bei Bedarf gerne auch mich.
- Schicken Sie bitte in so einem Fall alle Quellen zu und vergessen Sie nicht, eine präzise Beschreibung des Problems mitzuliefern.
- Das kann auch am Wochenende funktionieren.

Syllabus: Feedback

- Feedback ist ausdrücklich erwünscht.
- Es besteht insbesondere auch immer die Möglichkeit, auf Punkte noch einmal einzugehen, die zunächst noch nicht klar geworden sind.
- Vertiefende Fragen und Anregungen sind auch willkommen.
- Wir spulen hier nicht immer das gleiche Programm ab. Jede Vorlesung und jedes Semester verläuft anders und das hängt auch von Ihnen ab!

Syllabus: Wie geht es weiter?

- Im SS 2012 folgt der zweite Teil, der u.a. Interprozesskommunikation und Netzwerke behandelt. (Dies ist auch für an Numerik interessierte Hörer relevant.)
- Voraussichtlich im WS 2012/2013 biete ich wieder eine vertiefende Vorlesung mit C++ an.
- Im SS 2013 werde ich wahrscheinlich wieder Parallele Programmierung mit C++ anbieten.
- Gelegentlich wird auch in Zusammenarbeit mit Prof. Urban und Prof. Funken Scientific Computing gelesen. Dies eröffnet den Weg zur parallen Programmierung mit C++ für numerische Anwendungen. Hierzu empfehlen sich Systemnahe Software I+II und C++ als solide Grundlage.

Was ist Systemnahe Software?

- Der Begriff »System« bezieht sich hier auf den Kern eines Betriebssystems.
- Betriebssysteme (bzw. deren Kerne) erfüllen drei Funktionen:
  - Sie greifen direkt auf die Hardware zu,
  - sie verwalten all die Hardware-Ressourcen wie beispielsweise Speicherplatz, Plattenplatz und CPU-Zeit und
  - sie bieten eine Schnittstelle für Anwendungsprogramme.
- Systemnahe Software ist Software, die direkt mit der Betriebssystems-Schnittstelle zu tun hat.

Wie sehen die Schnittstellen aus?

- Teilweise bieten die Betriebssystems-Schnittstellen (auch Systemaufrufe genannt) ein sehr hohes Abstraktions-Niveau.
- So kann beispielsweise aus der Sicht einer Anwendung eine Netzwerk-Kommunikation abgewickelt werden, ohne darüber nachzudenken, was für Netzwerk-Hardware konkret genutzt wird, wie die Pakete geroutet werden oder wann Pakete erneut zu senden sind, wenn der erste Versuch nicht geklappt hat.
- Zwar gibt es teilweise große Unterschied bei den Schnittstellen, jedoch steht erfreulicherweise ein Standard zur Verfügung, kurz POSIX genannt oder ausführlicher IEEE Standard 1003.1.
- Dieser Standard entspricht weitgehend einer gemeinsamen Schnittmenge von Unix, Linux und den BSD-Varianten. Dank Cygwin gibt es auch weitgehend eine POSIX-Schnittstelle unter Windows.

Warum ist C relevant?

- Durch den Erfolg von Unix erreichte C eine gewisse Monopolstellung für die systemnahe Software.
- Da POSIX (und auch die einzelnen Betriebssysteme) die Schnittstelle nur auf der Ebene von C definieren, führt an C kaum ein Weg vorbei.
- Praktisch alle anderen Sprach-Implementierungen (wie beispielsweise C++, Java, Fortran oder Ada) basieren letztenendes auf C bzw. benötigen die C-Bibliothek, die die Schnittstelle zum Betriebssystem liefert.

Warum überhaupt etwas anderes als C?

- Ursprünglich wurde systemnahe Software in Assembler geschrieben. Das ist sehr umständlich und überhaupt nicht portabel, da jede Prozessor-Architektur anders zu programmieren ist.
- C entstand Mitte der 70er Jahre als Alternative zu Assembler. Manche bezeichnen C deswegen bis heute als »portablen Assembler«.
- C liefert Portabilität, ist aber immer noch sehr maschinennah.
- Wir verlassen mit C die gewohnte »heile Welt« von Java (oder anderer ähnlicher moderner Programmiersprachen).
- C setzt maschinennahes Denken voraus und bietet viele Fallstricke, die wir in der »heilen Welt« nicht kennen.
- Insofern wird heute C bevorzugt nur im systemnahen Bereich eingesetzt.

Wie funktionieren Systemaufrufe?

- Programme laufen auf modernen Betriebssystemen in ihrer eigenen virtuellen Welt ab, d.h. sie sehen in ihrem Adressraum weder das Betriebssystem noch die anderen parallel laufenden Programme.
- Die virtuelle Welt wird nur durch besondere Ereignisse verlassen, wenn z.B. durch 0 geteilt wird, ein `null`-Zeiger dereferenziert wird, die Uhr sagt, dass ein anderer Prozess mal an der Reihe ist, sich die Platte meldet, weil ein gewünschter Datenblock endlich da ist, irgendeine Taste auf der Tastatur gedrückt wurde oder ...
- ... ein Programm mit dem Betriebssytem kommunizieren möchte.
- All diese Ereignisse unterbrechen den regulären Betrieb und führen dazu, dass das Benutzerprogramm zu arbeiten aufhört und der Betriebssystems-Kern die Kontrolle übernimmt, um festzustellen, was zur Unterbrechung geführt hat.
- Im Falle eines Systemaufrufs werden die Parameter aus der Welt des Benutzerprogramms mühsam herausgeholt, der Aufruf bearbeitet und die Resultate in die Benutzer-Welt überführt.
- In Wirklichkeit ist das noch viel komplizierter ...

Wie wird eine Unterbrechung initiiert?

- Für absichtliche Unterbrechungen gibt es spezielle Maschinen-Instruktionen.
- Diese gehören nicht zum Vokabular eines C-Compilers, so dass in jeder C-Bibliothek die Systemaufrufe in Assembler geschrieben sind.
- Die Aufrufsyntax in C ist portabel, die jeweilige Implementierung ist es nicht, da sie in Assembler geschrieben ist.

Wie sehen Systemaufrufe konkret aus?

hello.s

```
/*
   Hello world demo in Assembler
   for the SPARCv8/Solaris platform
*/
        .section ".text"
        .globl  _start
_start:
/* write(1, msg, 13); */
        or      %g0,4,%g1
        or      %g0,1,%o0
        sethi   %hi(msg),%o1
        add     %o1,%lo(msg),%o1
        or      %g0,13,%o2
        ta      8
/* exit(0) */
        or      %g0,1,%g1
        or      %g0,0,%o0
        ta      8
msg:    .ascii  "Hello world!\012"
```

Wie sehen Systemaufrufe konkret aus?

- Das Beispiel wurde für die SPARC-Architektur geschrieben.
- Das Assembler-Programm besteht 9 Instruktionen, die jeweils 4 Bytes benötigen und 13 Bytes Text.
- `%g1`, `%o0`, `%o1` und `%o2` sind alles sogenannte Register, die im 32-Bit-Modus jeweils 32 Bit aufnehmen können.
- `%g0` ist ein spezielles Register, das immer den Wert 0 hat.
- Instruktionen haben bei der SPARC-Architektur normalerweise drei Operanden, wobei der dritte Operand das Ziel ist. Beispiel: `or %g0,4,%g1`
  Das ist eine binäre Oder-Operation mit `%g0` (also dem Wert 0) und der Zahl 4, dessen Resultat in `%g1` abgelegt wird. Kurz gefasst wird damit dem Register `%g1` der Wert 4 zugewiesen.

Wie sehen Systemaufrufe konkret aus?

- Die spezielle Instruktion `ta` (*trap always*) unterbricht die Programmausführung, bis der Prozess vom Betriebssystem wieder zum Leben erweckt wird.
- Die Parameter des Systemaufrufs werden bei der SPARC/Solaris-Plattform in den Registern %o0 bis %o5 abgelegt (bis zu 6 Parameter, die allerdings auf irgendwelche Speicherflächen mit mehr Parametern verweisen können).
- Im Register %g1 wird eine Nummer abgelegt, die den Systemaufruf selektiert. So steht beispielsweise die 1 für `exit()` und 4 für `write()`.
- Die Nummer 8, die bei `ta` angegeben wird, dient als Index in die Trap-Tabelle ...

Was passiert nach einem Systemaufruf?

usr/src/uts/sun4u/ml/trap_table.s

```
trap_table:
    /* hardware traps */
    NOT;                        /* 000  reserved */
    RED;                        /* 001  power on reset */
    RED;                        /* 002  watchdog reset */
    RED;                        /* 003  externally initiated reset */
/* ... */
    /* user traps */
    GOTO(syscall_trap_4x);      /* 100  old system call */
    TRAP(T_BREAKPOINT);         /* 101  user breakpoint */
    TRAP(T_DIV0);               /* 102  user divide by zero */
    FLUSHW();                   /* 103  flush windows */
    GOTO(.clean_windows);       /* 104  clean windows */
    BAD;                        /* 105  range check ?? */
    GOTO(.fix_alignment);       /* 106  do unaligned references */
    BAD;                        /* 107  unused */
    SYSCALL(syscall_trap32)     /* 108  ILP32 system call on LP64 */
/* ... */
```

- (Der Programmtext stammt aus den OpenSolaris-Quellen, siehe `http://www.opensolaris.org/`).

Was passiert nach einem Systemaufruf?

usr/src/uts/sun4u/ml/trap_table.s

```
#define SYSCALL(which)                          \
        TT_TRACE(trace_gen)                     ;\
        set     (which), %g1                    ;\
        ba,pt   %xcc, sys_trap                  ;\
        sub     %g0, 1, %g4                     ;\
        .align  32
```

- Zunächst werden alle Register gesichert und es findet ein Wechsel in den privilegierten Prozessor-Modus statt.
- Zu jeder Unterbrechungsart gibt es eine Nummer, wobei Unterbrechungen durch Benutzerprogramme von Hardware-Unterbrechungen unterschieden werden. Aus `ta 8` wird die Nummer $256 + 8 = 0x108$.
- Zu jedem der 512 verschiedenen Unterbrechungsmöglichkeiten sind in der Trap-Tabelle 32 Bytes Code vorgesehen, die den Trap behandeln, indem sie typischerweise eine entsprechende Routine aufrufen.

Was passiert nach einem Systemaufruf?

usr/src/uts/sparc/v9/ml/syscall_trap.s

```
    ENTRY_NP(syscall_trap32)
    ldx      [THREAD_REG + T_CPU], %g1     ! get cpu pointer
    mov      %o7, %l0                      ! save return addr
/* ... */
    lduw     [%l1 + G1_OFF + 4], %g1       ! get 32-bit code
    set      sysent32, %g3                 ! load address of vector table
    cmp      %g1, NSYSCALL                 ! check range
    sth      %g1, [THREAD_REG + T_SYSNUM] ! save syscall code
    bgeu,pn %ncc, _syscall_ill32
      sll    %g1, SYSENT_SHIFT, %g4        ! delay - get index
    add      %g3, %g4, %g5                 ! g5 = addr of sysentry
    ldx      [%g5 + SY_CALLC], %g3         ! load system call handler
/* ... */
    call     %g3                           ! call system call handler
    nop
/* ... */
    jmp      %l0 + 8
     nop
```

- Danach werden die Parameter so kopiert, dass sie als Parameter einer C-Funktion übergeben werden können (nicht dargestellt) und dann wird passend zur Systemaufrufsnummer die entsprechende Funktion aus einer Tabelle ausgewählt.

Was passiert nach einem Systemaufruf?

usr/src/uts/common/os/sysent.c

```
struct sysent sysent[NSYSCALL] =
{
/* ONC_PLUS EXTRACT END */
    /*  0 */ IF_LP64(
                    SYSENT_NOSYS(),
                    SYSENT_C("indir",       indir,          1)),
    /*  1 */ SYSENT_CI("exit",              rexit,          1),
    /*  2 */ SYSENT_2CI("forkall",          forkall,        0),
    /*  3 */ SYSENT_CL("read",              read,           3),
    /*  4 */ SYSENT_CL("write",             write,          3),
    /*  5 */ SYSENT_CI("open",              open,           3),
```

- In der `sysent`-Tabelle finden sich alle Systemaufrufe zusammen mit einigen Infos zur Anzahl der Parameter und die Art der Rückgabewerte. Unter Solaris hat diese Tabelle inzwischen 256 Einträge.

Was passiert nach einem Systemaufruf?

usr/src/uts/common/syscall/rw.c

```
ssize_t write(int fdes, void *cbuf, size_t count) {
        struct uio auio;
        struct iovec aiov;
/* ... */
        if ((cnt = (ssize_t)count) < 0)
                return (set_errno(EINVAL));
        if ((fp = getf(fdes)) == NULL)
                return (set_errno(EBADF));
        if (((fflag = fp->f_flag) & FWRITE) == 0) {
                error = EBADF;
                goto out;
        }
/* ... */
        aiov.iov_base = cbuf;
        aiov.iov_len = cnt;
/* ... */
        auio.uio_loffset = fileoff;
        auio.uio_iov = &aiov;
/* ... */
        error = VOP_WRITE(vp, &auio, ioflag, fp->f_cred, NULL);
        cnt -= auio.uio_resid;
/* ... */
out:
/* ... */
        if (error)
                return (set_errno(error));
        return (cnt);
}
```

Zusammenfassung

- Die Schnittstelle zwischen Anwendungen und dem Betriebssystems-Kern besteht aus über 200 einzelnen Funktionen, die zu einem großen Teil standardisiert sind.
- Systemaufrufe sind sehr viel teurer als reguläre Funktionsaufrufe. Das liegt an dem Mechanismus der Unterbrechungsbehandlung, dem notwendigen Kontextwechsel, der Parameterschaufelei und auch der asynchronen Natur vieler Funktionen (wie beispielsweise beim I/O).
- Deswegen ist es wichtig, auf der Anwendungsseite Bibliotheken zu entwickeln, die die Zahl der Systemaufrufe bzw. deren Aufwand minimieren.
- Verbunden mit den einzelnen Systemaufrufen sind viele Abstraktionen und Objekte, die wir uns im Laufe der Vorlesung genauer ansehen werden wie etwa das Dateisystem, die Prozesse, die Signalbehandlung (Unterbrechungen auf der Benutzerseite) die Interprozess-Kommunikation und die allgemeine Netzwerk-Kommunikation.