

FH Aachen

Fachbereich Elektrotechnik und Informationstechnik

Embedded Systems SS2022

Referat

Linux Tracing

Marcel Ochsendorf, Muhammed Parlak

Inhalt

1	Einleitung	1
1.1	Relevanz	1
1.2	Tracing	1
1.3	Ursprung	1
2	Grundlagen	2
2.1	Ringbuffer	2
2.2	Debug-Filesystem	2
2.3	Trace Events	2
2.4	Abfangen von Events	4
2.4.1	kprobes	4
2.4.2	kretprobes	5
2.4.3	CPU-Traps	5
2.4.4	uprobes	5
2.5	Nachteile / verschärfung	5
3	Tools	6
3.1	Trace-Log Aufzeichnung	6
3.1.1	trace-cmd	7
3.1.2	bpftrace	10
3.1.3	Kernelshark	10
4	Interpretation des Kernel-Trace Ergebnisses	12
5	Beispiel - TCP Paketanalyse Beispiel	13
6	Beispiel - Identifikation von Laufzeitproblemen	14
6.1	Ausgangsszenario	14
6.2	Aufzeichnung Trace-Log	15
6.2.1	Aktivierung der Events	15
6.3	Visualisierung und Beurteilung des Trace-Logs mittels kernelshark	16
	Literaturverzeichnis	17

1 Einleitung

Tracing ist die spezielle Verwendung der Protokollierung zur Aufzeichnung von Informationen über den Ausführungsablauf eines Programms. Oft werden mit eigenständig hinzugefügte Print-Messages der Code debuggt. Somit verfolgt man die Anweisungen mit einem eigenem tracing-System. Linux bringt einige eigenständige Tools mit, mit denen es möglich ist Vorgänge innerhalb von einem Embedded-System nachzuvollziehen und analysieren zu können. Die Linux-Tracing Funktionalität und die bestehenden Tools, welche im Linux-Kernel integriert sind, helfen so dabei bei der Identifikation von Laufzeiten, Nebenläufigkeiten und der Untersuchung von Latenzproblemen,

1.1 Relevanz

- multicore systeme unterscheid mikotrkontroller

1.2 Tracing

1.3 Ursprung

2 Grundlagen

2.1 Ringbuffer

2.2 Debug-Filesystem

- möglichkeiten

```
1 #ENABLE DEBUG FS
2 $ sudo mount -t debugfs debugfs /sys/kernel/debug
```

2.3 Trace Events

Durch das Debug-Filesystem ist jetzt der Zugriff auf die Debug und insbesondere auf die Tracing-Daten möglich. Im Debug-Filesystem ist nach aktivierung der `tracing`-Ordner vorhanden. In diesem werden die verfügbaren Events in Gruppen (Ordnern) dargestellt, auf welche im späteren Verlauf reagiert werden können.

- was sind tracers

```
1 # GET TRACERS
2 $ cat /sys/kernel/debug/tracing/available_tracers
3 hwlat blk mmiotrace function_graph wakeup_dl wakeup_rt wakeup
    function nop
```

In der `ls` Ausgabe des `events`-Ordners des Debug-Filesystems ist zu sehen, welche Events abgefangen werden können und mittels der Linux-Tracing-Tools protokolliert werden können.

```
1 # GET AVAILABLE EVENT LIST
2 $ cd /sys/kernel/debug/tracing/events
3 $ ls -lah | awk '{print $9}'
4 alarmtimer
5 clk
6 cpuhp
7 drm
8 exceptions
9 ext4
10 # => EXT4_READPAGE , EXT4_WRITEPAGE , EXT4_ERROR ,
11     EXT4_FREE_BLOCKS
12 filelock
13 filemap
14 fs_dax
15 ftrace
16 gpio
17 # => GPIO_DIRECTION , GPIO_VALUE
18 hda
19 i2c
20 irq
21 net
22 smbus
23 # => READ , WRITE , REPLY
24 sock
25 # => SOCKET_STATE_CHANGED , SOCK_EXCEED_BUFFER_LIMIT ,
26     SOCK_REC_QUEUE_FULL
27 spi
28 tcp
29 timer
30 # => TIMER_STOP , TIMER_INIT , TIMER_EXPIRED
```

Alle Events sind in Gruppen gebündelt. Alle Events, welche das `ext4`-Filesystem betreffen, befinden sich im `ext4`-Ordner. Die Auflistung zeigt einige der für das `ext4` zur Verfügung stehenden Events. Zudem befinden sich zwei zusätzliche Dateien `enable`, `filter` in diesem Ordner. Durch diese ist es später möglich anzugeben, ob dieses Event aufgezeichnet werden soll.

```
1 $ cd /sys/kernel/debug/tracing/events/ext4
2 $ ls -lah | awk '{print $9}'
3
4 # EVENTS FOR EXT4
5 ext4_write_end
6 ext4_writepage
7 ext4_readpage
8 ext4_error
9
10 # INTERFACE FOR EVENT SETUP
11 enable
12 filter
```

13 format

Die optionale `format`-Datei kann zusätzliche Informationen bereitstellen über das, durch das Event bereitgestellt Format der Ausgabe.

```
1 $ cat /sys/kernel/debug/tracing/events/sched/sched_wakeup/
      format
2 name: sched_wakeup
3 ID: 60
4 format:
5     field:unsigned short common_type;          offset:0;
             size:2;
6     field:unsigned char common_flags;         offset:2;
             size:1;
7     field:unsigned char common_preempt_count; offset
             :3;           size:1;
8     field:int common_pid;        offset:4;       size:4;
9     field:int common_tgid;       offset:8;       size:4;
10
11    field:char comm[TASK_COMM_LEN]; offset:12;       size
             :16;
12    field:pid_t pid;            offset:28;       size:4;
13    field:int prio;            offset:32;       size:4;
14    field:int success;         offset:36;       size:4;
15    field:int cpu;             offset:40;       size:4;
```

2.4 Auffangen von Events

```
1 $ cd /sys/kernel/debug/tracing
2 $ echo 1 > events/sched/enable
```

2.4.1 kprobes

Kprobes können dazu verwendet werden, Laufzeit und Performance-Daten des Kernels zu sammeln. Der Vorteil an diesen ist, dass diese Daten ohne Unterbrechnung der Ausführung auf CPU-Instruktions-Ebene aggregiert werden können, anders wie bei dem Debuggen eines Programms mittels Breakpoints. Ein weiterer Vorteil ist, dass das Registrieren der Kprobes dynamisch zur Laufzeit und ohne Änderungen des Programmcodes geschieht.

2.4.2 kretprobes

2.4.3 CPU-Traps

2.4.4 uprobes

- für anwendung
- system libs

2.5 Nachteile / verschärfung

- welche effekte können entstehen
- tracing bracuth ressoucen
- last minimieren auf traget minimieren
- nur aufzeichnen und später analysieren z.B. auf einem anderen system
- wie verhindern

3 Tools

Allgemein sind keine speziellen Programme notwendig um die Laufzeiteigenschaften eines Programms aufzuzeichnen. Der Linux-Kernel bringt bereits alle nötigen Funktionalitäten mit. Jedoch gibt es Tools die eine visuelle Darstellung der aufgezeichneten Events ermöglichen.

3.1 Trace-Log Aufzeichnung

Für die Log-Aufzeichnung wird ein Ringbuffer genutzt. Das Aufzeichnen in den Ringpuffer ist Standardmäßig aktiviert.

```
1 # Disable the Recording on the ringbuffer
2 $ echo 0 > tracing on
```

Mit dem folgenden Befehl kann der Inhalt des Ringpuffers, auch während einer Aufzeichnung, ausgegeben werden:

```
1 $ less trace
```

Das Lesen während einer Aufzeichnung mit trace hat keinerlei Einfluss auf den Inhalt des Ringpuffers. Die Ausgabe des letzten Kommandos wird dabei in einem menschenlesbaren Format dargestellt:

Die bisherigen Aufzeichnungen der Ereignisse können mit einem einfachen Befehl entfernt werden:

```
1 $ echo > trace
```

Um einen Überlauf an Informationen zu verhindern kann die Aufzeichnung auch konsumierend gelesen werden. Somit werden beim Lesen zeitgleich diese aus dem Ringbuffer

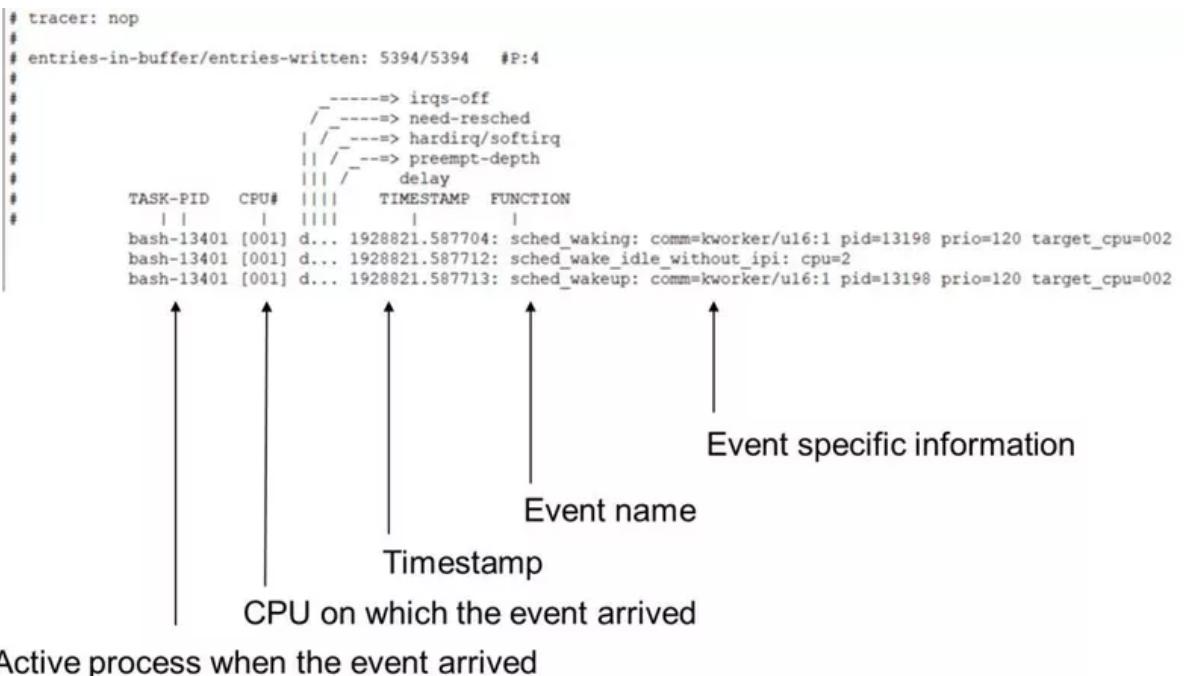


Bild 3-1: Trace-Log

entfernt.

Eine weitere Kernpunkt ist, dass in Mehrkernsystemen für jeden einzelnen Core ein separater Ringbuffer existiert. Damit die Analyse von verschiedenen Events getrennt werden kann, kann mit jeder weiteren Instanz pro Core ein weiterer Ringbuffer angelgt werden. Dies erfolgt im Unterverzeichnis `instances/`.

```

1 $ cd instances
2 $ mkdir inst0
3 $ mkdir inst1
4
5 # Remove if the instances is not needed anymore
6 # rmdir inst0

```

3.1.1 trace-cmd

Das Tool `trace-cmd` ist das bekannteste und am meisten genutzte Hilfmittel zur Aufzeichnung. Dies ist ein kommandozeilenwerkzeug, das auf den meisten gängigen Linux Distrubitionen bereits vorinstalliert ist.

```

1 # CHECK IF TRACING IS ENABLED
2 # RUN AS SUDO
3 $ mount | grep tracefs

```

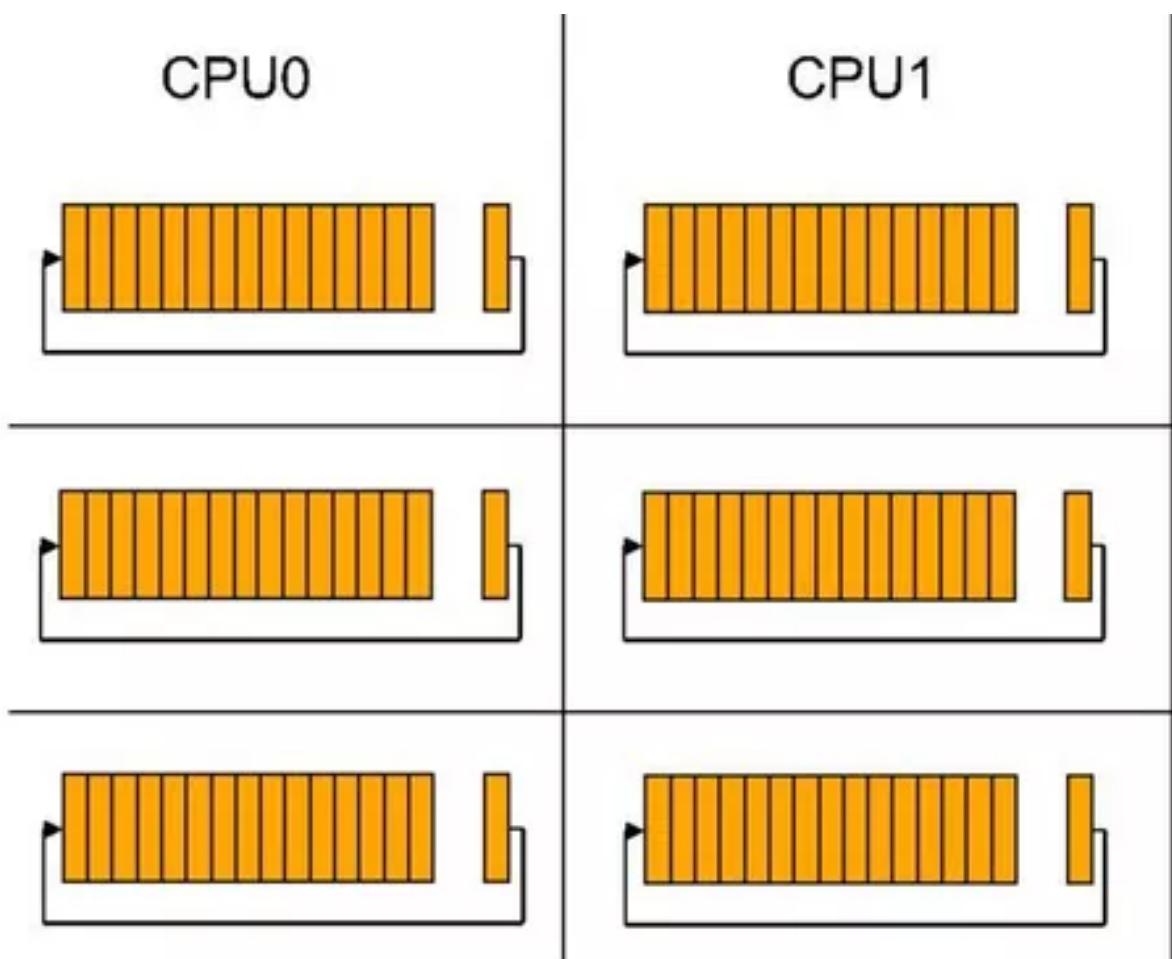


Bild 3-2: Ringbuffer

```

cpus=4
sleep-19405 [000] 915.755341: sched_stat_runtime: comm=trace-cmd pid=19405 runtime=168772 [ns] vruntime=124673167878 [ns]
sleep-19405 [000] 915.755344: sched_switch: trace-cmd:19405 [120] D ==> swapper/0:0 [120]
<idle>-0 [003] 915.759205: sched_waking: comm=rcu_sched pid=10 prio=120 target_cpu=003
<idle>-0 [003] 915.759210: sched_wakeup: rcu_sched:10 [120] success=1 CPU:003
<idle>-0 [003] 915.759212: sched_waking: comm=kworker/3:1 pid=2850 prio=120 target_cpu=003
<idle>-0 [003] 915.759214: sched_wakeup: kworker/3:1:2850 [120] success=1 CPU:003
<idle>-0 [003] 915.759226: sched_switch: swapper/3:0 [120] R ==> kworker/3:1:2850 [120]
kworker/3:1-2850 [003] 915.759231: sched_stat_runtime: comm=kworker/3:1 pid=2850 runtime=14167 [ns] vruntime=60565680765 [ns]
kworker/3:1-2850 [003] 915.759233: sched_switch: kworker/3:1:2850 [120] W ==> rcu_sched:10 [120]
rcu_sched-10 [003] 915.759238: sched_stat_runtime: comm=rcu_sched pid=10 runtime=7532 [ns] vruntime=60565679935 [ns]
rcu_sched-10 [003] 915.759246: sched_switch: rcu_sched:10 [120] W ==> swapper/3:0 [120]
<idle>-0 [002] 915.762699: sched_waking: comm=apt-get pid=2554 prio=120 target_cpu=002
<idle>-0 [002] 915.762706: sched_wakeup: apt-get:2554 [120] success=1 CPU:002
<idle>-0 [002] 915.762709: sched_waking: comm=kworker/2:3 pid=389 prio=120 target_cpu=002

```

Bild 3-3: trace-cmd Report

```

4 ## => none on /sys/kernel/tracing type tracefs (rw,relatime,
      seclabel)
5
6 # IF TRACING IS NOT ACTIVE, ENABLE IT FIRST
7
8 ## ONLY SCHEDULER EVENTS
9 $ echo sched_wakeup >> /sys/kernel/debug/tracing/set_event
10 ## ALL EVENTS
11 $ echo *:* > /sys/kernel/debug/tracing/set_event
12
13
14 # RECORD
15 $ trace-cmd record -e sched ./program_executable

```

- output erklärung

Mit dem letzten Befehl werden die ganzen Events zu Scheduler aufgezeichnet. Dabei werden während der Aufzeichnung kontinuierlich die Ringbuffer in konsumierender Form ausgelesen und in die Datei `trace.dat` geschrieben, falls mit dem `-o` keine eigene Datei eingegeben wurde. Als Informationen werden zu dem Inhalt des Ringbuffers auch zusätzlich notwendige Informationen über das Target, für die Auswertung auf beliebigen System gespeichert.

Die `trace-cmd` Konsolenanwendung dient nicht nur zur Aufzeichnung der Trace-Events, sondern bietet auch die Möglichkeit augezeichnet Reports visuell darzustellen. Die Ausgabe erfolgt mit dem Befehl `trace-cmd report [-i <Dateiname>]` als Tabelle in der Konsole und ist somit rein Textbasiert, siehe dazu die nächste Abbildung.

Auf diese Aufzeichnung zusätzlich ein Filter angewendet werden, um die Suche auf bestimmten Ereignissen einzuschränken.

Mit dem Tool ist es einfach die Teilschritte zu automatisieren.

3.1.2 bpftrace

Seit der Kernelversion >4.x, kann ein weiteres Tool mit dem Namen `bpftrace` verwendet werden. Dieses bietet jedoch zusätzlich eine eigene Skriptsprache mit der nicht nur Aggregation, sondern auch die Eventfilter und die Verarbeitung der Ergebnisse automatisiert werden können.

```
1 # Block I/O latency as a histogram EXAMPLE
2 $ wget https://raw.githubusercontent.com/iovisor/bpftrace/
   master/tools/biolatency.bt
3 $ bpftrace ./biolatency.bt
4 # @usecs:
5 # [512, 1K)           10 | @
6 # [ 1K, 2K)          426 | 000000000000000000000000
7 # [2K, 4K)           230 | 000000000000000000000000
8 # [4K, 8K)            9 | @
9 # [8K, 16K)          128 | 000000000000000000000000
10 # [16K, 32K)         68 | 000000000000000000000000
11 # ...
```

3.1.3 Kernelshark

Das zuvor erklärte `trace-cmd` ist wie oben erwähnt nur ein textbasiertes Analysetool. Das kommende Kernelshark Tool bietet dem Anwender die Möglichkeit die Traceaufzeichnungen grafisch zu analysieren. Dabei sind die beiden Tools aufeinander abgestimmt und werden gemeinsam entwickelt. Auch dieses Tool ist in den meisten Linux Distributionen vorinstalliert.

Das vom trace-cmd erzeugte `trace.dat`-Format wird im Kernelshark als Eingabe erwartet. Wenn im folgendem ersten Befehl nichts eingegeben, dann wird nach der entsprechenden `trace.dat` im Verzeichnis gesucht.

```
1 $ kernelshark
2 $ kernelshark -i <Dateiname>
```

Im folgenden ist die grafische Darstellung zu sehen. Dabei besitzt jeder Task einen Farbton. Für jede CPU wird eine eigene Zeile dargestellt. Dieses Tool hat eine gewisse Ähnlichkeit mit der von uns genutzten Logic 2 Software.

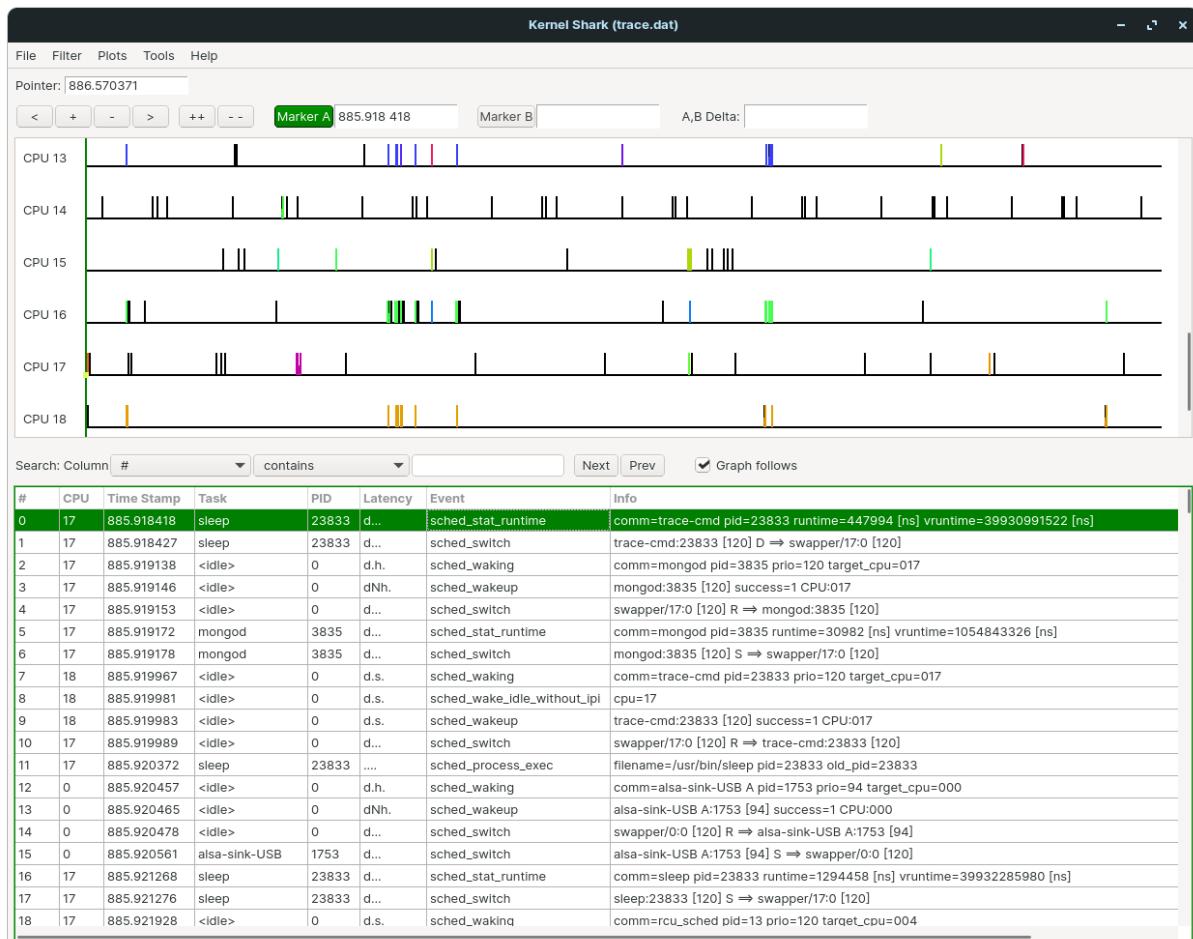


Bild 3-4: Kernelshark

4 Interpretation des Kernel-Trace Ergebnisses

5 Beispiel - TCP Paketanalyse

Beispiel

Dieses erste Beispiel

6 Beispiel - Identifikation von Laufzeitproblemen

In diesem Abschnitt soll an einem einfache Beispiel gezeigt werden, wie es mittels Tracing möglich ist, eine Laufzeitanalyse auf verschiedenen Systemen für eine Anwendung durchzuführen.

6.1 Ausgangsszenario

Als Ausgangspunkt dieses Beispiels, soll das Laufzeitverhalten eines Programms auf einem Linux-System analysiert werden. Die zugrunde liegende Software wurde bisher nur auf einem Linux-Realtime Kernel verwendet, jedoch erfordert die Implementation neuer Features eine neuere Kernel-Version, welche noch nicht als RT-Version auf dem System zur Verfügung steht. Somit soll ermittelt werden, ob die unmodifizierte Software eins zu eins auf dem neuen System lauffähig ist und die Laufzeitandorderungen erfüllt.

Das System besteht hier aus einem [RaspberryPi 4B](#) mit einer angeschlossenen LED am GPIO-Port 25 und zu testende Programm lässt diese dabei in 100ms Abständen Blinken.

```
1 #include <iostream>
2 #include <wiringPi.h>
3 #include <csignal>
4 using namespace std;
5
6 void signal_callback_handler(int signum) {
7
8 }
9
10 int main(int argc, char *argv[])
11 {
12     // REGISTER SIGNAL HANDLER
```

```
13     signal(SIGINT, signal_callback_handler);
14
15     wiringPiSetup();           // Setup the library
16     pinMode(0, OUTPUT);       // Configure GPIO0 as an output
17     pinMode(1, INPUT);        // Configure GPIO1 as an input
18     bool state = false;
19
20     while(1)
21     {
22         state = !state;
23         digitalWrite(0, state);
24         delay(500);
25     }
26     return 0;
27 }
```

Für den Test wurde als RT Kernel die Version 4.19.59-rt23-v7l+ verwendet, welche nicht alle Funktionalitäten des aktuellen 5.10 Kernel besitzt. In diesem fiktiven Beispiel, wird die [systemd-networking >V.248](#) Funktionalität für das Batman-Prokoll benötigt, welche den Grund für die Umstellung darstellt und nicht trivial in den 4.x Kernel integriert werden kann.

Die Messungen wurden zuerst auf dem aktuellen 5.10 LTS Kernel aufgezeichnet und im Anschluss wurde der RT-Kernel auf einem anderen System per Cross-Compilation aus dem rpi-4.19.y-rt Branch des raspberrypi/linux Repository gebaut. Dieser Schritt war notwendig, da es kein fertiges RT-Kernel Image zur Verfügung stand. Die erzeugten Dateien wurden dann auf die Boot-Partition der SD Karte geschrieben und in der /boot/config.txt Datei wurde der neue Kernel hinterlegt kernel=kernel7_rt.img.

6.2 Aufzeichnung Trace-Log

Zur Aufzeichnung des Trace-Logs wurde trace-cmd verwendet. Auf dem Zielsystem wurde dabei nur die Aufzeichnung vorgenommen und die Analyse der Logs erfolgte auf einem separaten System.

6.2.1 Aktivierung der Events

```
1 $ echo 1 > /sys/kernel/debug/tracing/tracing_on
2 $ cat /sys/kernel/debug/tracing/trace
```

```
3 $echo > /sys/kernel/debug/tracing/trace
```

```
1 trace-cmd record -e sched ./blink
```

6.3 Visualisierung und Beurteilung des Trace-Logs mittels kernelshark

Literaturverzeichnis