TECHNISCHE UNIVERSITÄT DRESDEN

Faculty of Computer Science Institute for System Architecture, Operating Systems Group

Betriebssysteme und Sicherheit, WS 2012/13

2. Aufgabenblatt – Threads

Geplante Bearbeitungszeit: drei Wochen

Aufgabe 2.1 Threads und Prozesse

Prozess (nach NEHMER): dynamisches Objekt, das selbständige, voneinander isolierte sequentielle Aktivitäten bezüglich Anfordern und Besitz von Betriebsmitteln in einem Rechensystem repräsentiert.

Ein Prozess ist gegeben durch:

Adressraum: Behälter für die Aufnahme abgegrenzter Programme mit zugeordneten Daten, Abstraktion des physischen Speichers

Handlungsvorschrift: im Adressraum in Form eines sequentiellen Programms gespeichert

Aktivitätsträger oder Thread: führt Handlungsvorschrift aus, Abstraktion des physischen Prozessors

(a) Diskutieren Sie die oben stehende Definition eines Prozesses. Mit welchem Ziel wurde der Prozessbegriff für Betriebssysteme eingeführt, warum hat man dieses Ziel angestrebt? Welche grundlegenden Konsequenzen ergeben sich daraus? Warum unterscheidet man zwischen Single-Thread-Prozessen und Multi-Thread-Prozessen?

(b) Erklären Sie Aufgabe und Struktur eines Threadsteuerblocks (TCB). Welche Informationen sind für einen Prozesssteuerblock (PCB) erforderlich?

(c) Eine Technik zur Leistungssteigerung eines Rechensystems ist das sogenannte Auslagern eines Prozesses bzw. Threads („swapping"), d.h. das Speichern der im Hauptspeicher befindlichen prozessrelevanten Informationen auf einer Platte und das anschließende Freigeben dieses Speicherbereichs.

- Welche Vor- und Nachteile hat diese Technik?
- Entwerfen Sie ein Thread-Zustandsmodell (analog zu Folie 5 des Foliensatzes „Threads: Einführung"), in dem diese Technik durch einen weiteren Zustand „ausgelagert" berücksichtigt wird. Welche Zustandsübergänge sind notwendig, welche sind darüber hinaus sinnvoll, welche sollten ausgeschlossen sein? Erläutern Sie dabei auch die verwendeten Zustände und Übergänge.

(d) Betrachtet werde ein Mail-Server, der die in einer Liste abgelegte Nachrichten an die jeweiligen Empfänger versenden soll (jede Nachricht habe nur einen einzigen Empfänger). Der Server arbeitet folgendermaßen: er entnimmt der Liste eine Nachricht, extrahiert den Empfänger, ermittelt dessen IP-Adresse, baut die entsprechende Verbindung auf, sendet die Nachricht und schließt die Verbindung.

1. Beschreiben Sie (in Pseudocode) eine Implementation des Servers durch ein Programm, das als sequentieller Prozess ausgeführt wird. Dabei werde angenommen, dass stets zu versendende Nachrichten vorliegen.
2. Warum ist diese Lösung ineffizient? Welche Möglichkeiten gibt es, sie zu verbessern? Hat sie dennoch Vorteile?
3. Beschreiben Sei eine Implementation, bei der mehrere Threads innerhalb eines Prozesses genutzt werden können. Welche Vor- und Nachteile hat dieses Vorgehen?
4. Welche Konsequenzen hat es, wenn anstelle der Threads selbst wieder sequentielle Prozesse (mit nur jeweils einem Thread) verwendet werden?

(e) In einem System paralleler Prozesse erfolge das Ausdrucken von Dateien folgendermaßen: Ein Prozess, der eine Datei ausgedruckt haben möchte, schreibt den Dateinamen in ein (genügend großes) Feld, jedes Feldelement enthält also genau einen Dateinamen. Die Einträge erfolgen fortlaufend. Eine Variable `frei` zeigt den nächsten freien Platz an, eine weitere Variable `drucke` die nächste zu druckende Datei. Die Variablen `frei` und `druckliste` können von allen Prozessen genutzt werden. Ein Prozess `druckprozess` ermittelt periodisch, ob eine Datei auszudrucken ist, druckt sie gegebenenfalls, streicht sie aus `druckliste` und aktualisiert die Variable `drucke`.

Erklären Sie an diesem Beispiel die Begriffe Wettlaufsituation, kritischer Abschnitt und wechselseitiger Ausschluss.

Aufgabe 2.2 Wechselseitiger Ausschluss

Gegeben: P: System paralleler Prozesse, B: Menge globaler Betriebsmittel

Wettlaufsituation (race condition): Mehrere Prozesse aus P bewerben sich unabhängig voneinander um die zeitweise exklusive Nutzung derselben Betriebsmittel aus B.

Kritischer Abschnitt (critical section) bezüglich einer Teilmenge $B' \subseteq B$:

Abschnitt eines zu einem Prozess aus P gehörenden Programms, in dem Werte von Betriebsmitteln aus B' durch andere Prozesse aus P geändert werden können.

Wechselseitiger Ausschluss (mutual exclusion): Koordinierung der Abläufe der Prozesse aus P so, dass die kritischen Abschnitte jeweils nur von einem Prozess betreten werden können.

(a) Wie wird wechselseitiger Ausschluss prinzipiell realisiert? Welche allgemeinen Anforderungen muss jede Realisierung erfüllen?

(b) Welche Möglichkeiten zur Realisierung des wechselseitigen Ausschlusses gibt es auf Maschinenebene? Welche Vor- und Nachteile haben sie?

(c) Erklären Sie, warum die in der Vorlesung angegebene erste Lösung zum wechselseitigen Ausschluss mittels Sperrvariablen (Folie 66) das Problem nicht löst. Worin liegt die prinzipielle Ursache?

(d) Im Gegensatz zu Aufgabe (c) ist der wechselseitige Ausschluss mehrerer Threads bezüglich eines kritischen Abschnitts mittels einer einfachen Schleife beispielsweise dann korrekt möglich, wenn in der zugrundeliegenden Hardware ein Maschinenbefehl der Form `xchg R adr` bereitgestellt wird, durch den atomar der Inhalt eines Registers `R` und einer Hauptspeicherzelle `adr` miteinander ausgetauscht werden. Diskutieren Sie unter diesem Gesichtspunkt die nachstehende Implementation auf der Basis eines i386-Systems.

```
        .globl  lock, unlock /* Funktionen extern zugänglich machen */
mutex:  .long   0            /* Speicherstelle namens mutex, Inhalt 0 */
lock:   movl    $1, %eax
wait:   xchg    %eax, mutex
        cmp     $0, %eax
        jne     wait
        ret
unlock: movl    $0, mutex
        ret
```

(e) Die für die Lösung des Wettlaufproblems geforderte Bedingung der Lebendigkeit kann folgendermaßen untergliedert werden: Lebendigkeit ist das *Nicht*-Vorliegen von

- Fernwirkung: Ein Thread außerhalb seines kritischen Abschnitts (und außerhalb der Dienste `entersection`, `leavesection`) behindert den Thread-Ablauf.
- Ausgrenzung: Ein Thread wird ständig am Eintritt in seinen kritischen Abschnitt gehindert.
- Verklemmung: Zwei Threads behindern sich gegenseitig am Eintritt in ihren kritischen Abschnitt.

Diskutieren Sie unter diesem Gesichtspunkt, inwieweit die folgenden fünf Versuche das Wettlaufproblem zwischen zwei Threads T_1, T_2 lösen. (Die Threads sollen die angegebenen Befehlsfolgen jeweils im Zyklus „ewig“ durchlaufen, sofern möglich.)

- Voraussetzung für 1. Versuch:

```
int s = 1;
```

- Voraussetzung für 2. – 5. Versuch:

```
#define false 0
#define true 1
int s1 = s2 = true;
```

- weitere Voraussetzung für 5. Versuch:

```
int next = 1;
```

	T_1	T_2

1. Versuch	`while (s == 2) {`
	`}`
	`/* kritischer Abschnitt */`
	`s = 2;`
		`while (s == 1) {`
		`}`
		`/* kritischer Abschnitt */`
		`s = 1;`

2. Versuch	`while (s2 == false) {`
	`}`
	`s1 = false;`
	`/* kritischer Abschnitt */`
	`s1 = true;`
		`while (s1 == false) {`
		`}`
		`s2 = false;`
		`/* kritischer Abschnitt */`
		`s2 = true;`

3. Versuch	`s1 = false;`
	`while (s2 == false) {`
	`}`
	`/* kritischer Abschnitt */`
	`s1 = true;`
		`s2 = false;`
		`while (s1 == false) {`
		`}`
		`/* kritischer Abschnitt */`
		`s2 = true;`

4. Versuch	`s1 = false;`
	`while (s2 == false) {`
	`  s1 = true;`
	`  s1 = false;`
	`}`
	`/* kritischer Abschnitt */`
	`s1 = true;`
		`s2 = false;`
		`while (s1 == false) {`
		`  s2 = true;`
		`  s2 = false;`
		`}`
		`/* kritischer Abschnitt */`
		`s2 = true;`

5. Versuch	`s1 = false;`
	`while (s2 == false) {`
	`  if (next == 2) {`
	`  s1 = true;`
	`  while (next == 2) {`
	`  }`
	`  s1 = false;`
	`  }`
	`}`
	`/* kritischer Abschnitt */`
	`next = 2;`
	`s1 = true;`
		`s2 = false;`
		`while (s1 == false) {`
		`  if (next == 1) {`
		`  s2 = true;`
		`  while (next == 1) {`
		`  }`
		`  s2 = false;`
		`  }`
		`}`
		`/* kritischer Abschnitt */`
		`next = 1;`
		`s2 = true;`

(f) Untersuchen Sie, ob der folgende Algorithmus die Bedingungen zum Schutz kritischer Abschnitte erfüllt.

```
int s1 = s2 = next = 1;
```

Thread 1	Thread 2

`  s1 = 0;`
`M1:`
`  if (s2 == 0) {`
`  if (next == 1)`
`  goto M1;`
`  s1 = 1;`
`  while (next == 2) {`
`  }`
`  }`
`  /* kritischer Abschnitt */`
`  s1 = 1;`
`  next = 2;`
	`  s2 = 0;`
	`M2:`
	`  if (s1 == 0) {`
	`  if (next == 2)`
	`  goto M2;`
	`  s2 = 1;`
	`  while (next == 1) {`
	`  }`
	`  }`
	`  /* kritischer Abschnitt */`
	`  s2 = 1;`
	`  next = 1;`

Aufgabe 2.3 Semaphore

(a) Was versteht man unter einem Semaphor? Warum wurde der Begriff eingeführt? Erläutern Sie die in der Vorlesung angegebene Implementation verbal. Welche Vor- und Nachteile hat diese Implementation? Welche Vor- und Nachteile haben Semaphore generell?

(b) Die Semaphor-Operationen `P` und `V` müssen als unteilbare Operationen implementiert werden. Skizzieren Sie einen Ablauf, der andernfalls zu falschen Ergebnissen führt.

(c) Geben Sie in C eine Implementation für einen zählenden Semaphor an, dessen Zählvariable `count` auch negative Werte annehmen kann mit folgender Bedeutung:

- Ist `count` $\geqslant 0$, so gibt `count` die Anzahl der Prozesse an, die noch den kritischen Abschnitt betreten können.
- Ist `count` $\leqslant 0$, so gibt `-count` die Anzahl der Prozesse an, die auf das Betreten des kritischen Abschnitts warten.

Stellen Sie zunächst einen Programmablaufplan für die Funktionen `P` und `V` auf.

(d) Beschreiben Sie die Steuerung für ein System, bei dem Fahrzeuge über eine Brücke wollen, für folgende Fälle der maximalen Belastbarkeit der Brücke:

1. ein Fahrzeug, unabhängig von der Richtung (1 Fahrspur)
2. ein Fahrzeug je Richtung gleichzeitig (2 Fahrspuren)
3. drei Fahrzeuge je Richtung gleichzeitig (2 Fahrspuren)
4. drei Fahrzeuge, unabhängig von der Richtung (1 Fahrspur; die Fahrzeuge aus einer Richtung müssen so lange warten, bis alle Fahrzeuge der Gegenrichtung die Brücke passiert haben)

Verwenden Sie Semaphore. Welches Problem tritt in 4. auf, wie lässt es sich lösen?

(e) Erläutern Sie das Erzeuger-Verbraucher-Problem und seine prinzipiellen Lösungsmöglichkeiten.

Der Puffer sei einelementig. Geben Sie eine Lösung dieses speziellen Problems mittels Semaphoren an. Verdeutlichen Sie die Korrektheit Ihrer Lösung anhand eines repräsentativen Ablaufbeispiels.

Inwieweit lässt sich das spezielle Problem auch ohne Semaphore (allein auf Maschinenebene) lösen?

(f) Ein Ringpuffer werde durch einen Thread *A* mit Elementen vom Typ `elem_t` gefüllt und durch einen Thread *B* geleert (gemäß FIFO); der Puffer soll dabei als *n*-elementiges Feld (*n* konstant) implementiert werden.

- In welchen Fällen können Konflikte auftreten?
- Beschreiben und diskutieren Sie eine Lösung des Problems in C mittels Semaphoren.

(g) Erläutern Sie das Philosophenproblem. Worin liegt seine Bedeutung?

Betrachtet werde der in der Vorlesung angegebene Lösungsversuch. Zeigen Sie, dass:

- weder diese Implementation
- noch das „Klammern" der gesamten Anweisungsfolge vom Aufnehmen der ersten bis zum Ablegen der zweiten Gabel
- noch das einzelne „Klammern" des Aufnehmens und des Ablegens beider Gabeln

mit Semaphoren das Problem lösen.

Welches ist der entscheidende Gedanke, der zu einer korrekten Lösung führt?

(h) Was versteht man unter dem Leser-Schreiber-Problem? Skizzieren Sie mindestens zwei Lösungsansätze. Welches sind die wesentlichen Gedanken der in der Vorlesung angegebenen Lösung?

(i) Erklären Sie den Begriff „Prioritätsumkehr" (priority inversion) in einem System ohne beziehungsweise mit Semaphoren.