Concurrent Systems

Nebenläufige Systeme

VI. Locks

Wolfgang Schröder-Preikschat

November 13, 2014

Subject Matter

- discussion on abstract concepts as to blocking synchronisation:
  - lock a critical section
  - shut simultaneous processes out of entrance
  - block (delay) interacting processes
  - unlock a critical section
  - give a simultaneous process the chance of entrance
  - unblock one or several interacting processes
- treatment of basic characteristics and common variants of locking
  - hierarchic placement of lock/unlock implementations ~ ISA level
  - standby position, control mode, properties, computational burden
  - relying on atomic read/write, with and without special instructions
- explanation of benefits, limits, shallows, drawbacks, but also myths

Spin-Lock (Ger. Umlaufsperre)

Blocking synchronisation under prevention of context switches and by active waiting, including processor halt, for unlocking.
Lockout [3, p.147]

A provision whereby two processes may negotiate access to common data is a necessary feature of an MCS.

already this original reference foreshadows two levels of abstraction at which an implementation may be organisationally attached to:

i by means of a program at instruction set architecture level (i.e., level 2)
   - busy waiting until success of a TAS-like instruction [3, p.147, Fig.3a]
   - the TAS-like instruction—was and still—is an unprivileged operation

ii by means of a program at operating system machine level (i.e., level 3)
   [To prevent hangup, ] inhibit interruption of a process between execution of a lock and execution of the following unlock. [3, p.147]
   - inhibit interruption beyond a hardware timeout is a privileged operation

note: (ii) takes a logical view as to hierarchic placement of lockout

Inhibit of Interruption/Preemption

in order that the mechanism is suited to pattern a hardware ELOP:¹

lock  ■ disables interrupts and acquires a (memory) bus lock
   ■ turns time monitoring on, i.e., arms some timeout mechanism
           - predefined worst-case execution time (WCET) or
           - upper limit of the number of processor instructions or cycles, resp.
   ↩ raises an exception or issues an instruction trap [7] upon timeout

unlock ■ turns time monitoring off
   ■ releases the (memory) bus lock and re-enables interrupts

for integrity reasons, the processor must enforce an absolute timeout

the instruction trap must be unmaskable at the level of lock/unlock
the instruction-trap handler must be indispensable
   - a necessary part that needs to be provided by the operating system

the lock/unlock pair does not have to be system calls to this end

it does have to “use” [11] an operating system and

it may benefit from an operating system as to problem-specific timeouts
   - in which case the lock/unlock pair does have to be system calls, yet

¹As indicated by [3, p.147], to prevent hangup of processes interrogating the lock indicator, and once supported by the Intel i860 [7, p.7-24].

Indivisibility Revisited

critical section considered as logical or physical ELOP, referred to [3]

logical

physical

process lock, only
   - passage is vulnerable to delays
   - blocking time is two-dimensional
   - WCET² of critical section and interrupt/preemption latency
   - hinges predictability
   - irrelevant for time-sharing mode
   - enables concurrent processes

interrupt and bus lock
   - passage is without delays
   - blocking time is one-dimensional
   - WCET² of critical section
   - eases predictability
   - relevant for real-time mode
   - disables concurrent processes

²abbr. worst-case execution time
Process Locks

Critical Section as ELOP in Logical Terms

Hint (Lockout)

Contemporary (real) processors do no longer offer a means to pattern a hardware ELOP. Instead, locking falls back on algorithmic solutions.

- the standby position of a process may be either active or passive
  - active
    - a spin-lock (Ger. Umlaufsperre), busy waiting
    - lock holder interruption/preemption is crucial to performance
    - periods out of processor increase latency for competing processes
    - extends the point in time until execution of unlock
  - passive
    - a sleeping lock (Ger. Schlafspere), idle waiting
    - lock/unlock entail system calls, thus are crucial to granularity
    - impact of system-call overhead depends on the critical sections
    - number, frequency of execution, and best-case execution time

- “passive waiting” for unlock is untypical for conventional locking
  - a sleeping lock typically falls back on a binary semaphore or mutex, resp.\(^3\)
  - a conventional lock manages on instruction set architecture level, only

\(^3\)Operating system machine level concepts are discussed in LEC 7.

Coordinating Cooperation

- enforcement of sequential execution of any critical section always goes according to one and the same pattern:
  - entry protocol
    - acquire exclusive right to run through the critical section
    - refuse other processes entrance to the critical section
    - as a function of the lock operation
  - exit protocol
    - release exclusive right to run through the critical section
    - provide a process entrance to the critical section
    - as a function of the unlock operation

- including the assurance of fundamental mandatory properties:
  - mutual exclusion: at any point in time, at most one process may “have a command of” (Ger. beherrschen) the critical section
  - deadlock freedom: if several processes simultaneously aim for entering the critical section, one of them will eventually succeed
  - starvation freedom: if a process aims for entering the critical section, it will eventually succeed
  - not least, desirable property is to not interfere with the scheduler

Lock Characteristics

- the control mode (Ger. Betriebsart, Prozessregelung) for a lockout may be either advisory or mandatory
  - advisory
    - locking is explicit, performed by cooperating processes
      - first-class object of the real processor, e.g. a critical section
      - assumes process-conformal protocol behaviour
      - a lock action must be followed by an unlock action
      - complies with a lower level of abstraction
  - mandatory
    - locking is implicit, as a side effect of a complex operation
      - first-class object of an operating system, e.g. a file
      - enables recognition of exceptional conditions
      - “extrinsic” access on a locked file by a simultaneous process
      - calls for a higher level of abstraction

- mandatory locks are implemented using advisory locks internally
  - the exception proves the rule...

Hint

Advisory locks are in the foreground of this lecture, mandatory locks (in its classical meaning) will not be covered.
Solutions Devoid of Dedicated Processor Instructions

- sole demand is the atomic read/write of one machine word from/to main memory by the real processor
- classical approaches are in the foreground
  - for \( N = 2 \) processes: Dekker (1965), Peterson (1981), and Kessels (1982)
  - more of Lamport (1974) and Peterson (1981) for \( N > 2 \) in the addendum
- all of them are more than an exercise to read, but significant even today
  - some are confined to two contenting processes, ideal for dual-core processors
  - others are computationally complex, but may result only in background noise
- they demonstrate what “coordination of cooperation” in detail means

- an additional and utmost important constraint of these approaches is related to the memory model of the real processor
- for sequential consistent memory only, less important in olden days
- but more recent, this changed dramatically and gives one a hard time
- mean to say: solutions for synchronisation that do not use specialised processor instructions are not necessarily portable!

The “state machine” approach will be picked up again later for non-blocking synchronisation (LEC 10), e.g. of a semaphore implementation (LEC 11).

### Lock Type I

Algorithms of Dekker, Peterson, and Kessel

```c
#ifndef NPROC
#define NPROC 2
#endif

#ifdef __FAME_LOCK_KESSEL__
#define NTURN NPROC
#else
#define NTURN NPROC - 1
#endif

typedef volatile struct
{
    bool want[NPROC]; /* initial : all false */
    char turn[NTURN]; /* initial : all 0 */
} lock_t;

inline unsigned earmark() {
    return /* hash of process ID for [0, NPROC - 1] */
}

void lock(lock_t *bolt) {
    unsigned self = earmark(); /* my process index */
    bolt->want[self] = true; /* I am interested */
    while (bolt->want[self - 1] /* you are interested */
        && (bolt->turn[0] != self)) { /* & inside CS */
        bolt->want[self] = false; /* I withdraw */
        while (bolt->turn[0] != self); /* & will wait */
        bolt->want[self] = true; /* & reconsider */
    }
}

void unlock(lock_t *bolt) {
    unsigned self = earmark(); /* my process index */
    bolt->turn[0] = self - 1; /* I defer to you */
    bolt->want[self] = false; /* I am uninterested */
}
```

For an interpretation, see also p. 38.
egoistic ("self-serving") entry protocol with no-passsing zone:

```c
void lock(lock_t *bolt) {
    unsigned self = earmark(); /* my process index */
    bolt->want[self] = true; /* I am interested */
    bolt->turn[0] = self; /* & you are interested */
    while (bolt->want[(self-1) % self])
        & (bolt->turn[0] == self)); /* & inside CS */
}
```

```c
void unlock(lock_t *bolt) {
    unsigned self = earmark(); /* my process index */
    bolt->want[self] = false; /* I am uninterested */
}
```

4–7 ■ compared to the entry protocol of Dekker’s algorithm, the interest in entering the critical section (l. 4) never disappears

---

**Kessel’s Algorithm for \( N = 2 \)** cf. [8]

refinement of Peterson’s solution, but a **mutable** entry protocol:

```c
#define __FAME_LOCK_KESSEL__
...
```

```c
void lock(lock_t *bolt) {
    unsigned self = earmark(); /* my process index */
    bolt->want[self] = true; /* I am interested */
    bolt->turn[self] = ((bolt->turn[self] + self) % 2);
    while (bolt->want[(self-1) % self])
        & (bolt->turn[self] == ((bolt->turn[self])+self)%2));
}
```

7 ■ who’s next uses feedback as to peer’s view on who’s turn was last

9 ■ in case of lock contention, gives only a single process precedence

**Hint (Progress)**

**Starvation Freedom** Question of Interpretation (cf. p. 11)

**Hint (Progress)**

A matter of interaction of processes by means of the entry and exit protocols, while abstracting away from potential delays caused by “external incidents” of the instruction set architecture (ISA) level.

in terms of the **lock callee** process: “bottom up” point of view of the level of abstraction of the entry protocol

■ the entry or exit, resp., protocol is shaped up as a logical ELOP (cf. p. 8)

■ depending on the solution, process delays are “accessory symptom” of:

**Dekker** ■ noncritical parts of the entry protocol (\( \text{want}_i = \text{false} \))

■ the critical section (\( \text{want}_i = \text{true} \))

■ in the terms of the **lock caller** process: “top down” point of view of the level of abstraction of the critical section

■ the entry or exit, resp., protocol appears to be instantaneous

---

Kessel’s Algorithm for **\( N = 2 \)**

**Solutions Based on Dedicated Processor Instructions**

**fundamental aspect common to all the solutions discussed before:**

■ processes rely on plain—but atomic—read/write operations, only

■ there is no read-modify-write cycle w.r.t. the same shared variable

■ as a consequence, arbitration at ISA level is less overhead-prone

→ solutions for \( N = 2 \) are “simple”, compared to \( N > 2 \) (cf. p. 40ff.)

**solutions for \( N > 2 \) processes benefit from special CPU instructions**

■ atomic read-modify-write instructions such as TAS, CAS, or FAA

■ but also load/store instructions that can be interlinked such as LL/SC

not only the memory model but in particular the **caching behaviour** of the real processor have a big impact on the solutions

■ most of the special instructions are considered harmful for data caches

■ unopt use breeds **interference** with all sorts of simultaneous processes

■ in case of high contention, this unwanted property is even more critical

mean to say: solutions for synchronisation making use of specialised processor instructions are not necessarily straightforward!
Lock Type II

in its simplest form, a binary variable indicating the lock status:

```c
#include <stdbool.h>

typedef volatile struct lock {
    bool busy; /* initial: false */
} lock_t;
```

true ■ occupied critical section, processes seeking entry will block
false ■ unoccupied critical section, unblocked processes will retry to enter

just as simple the exit protocol for a number of lock variants

```c
void unlock(lock_t *bolt) {
    bolt->busy = false; /* release lock */
}
```

more distinct is variant diversity of the entry protocol (p. 22 ff.)...

Spin with TAS
cf. p. 44

```c
void lock(lock_t *bolt) {
    while (!TAS(&bolt->busy)); /* loop if door closed */
}
```

be aware of the conventional implementation of TAS [13, p. 10 & 35]:

```c
atomic word TAS(word *ref) {
    word aux = *ref;
    *ref = 1;
    return aux;
}
```

the unconditional store has a deleterious effect for the cache
as to the cache operation (write invalidate or update, resp.), the cache line holding the main memory operand causes high bus traffic
for \( N \) contending processes, either \( N-1 \) cache misses or update requests

further problem dimension is non-stop instruction of TAS in the loop
blocks other processors from using the shared bus to access memory or other devices that are attached to access contention
thereby interfering in particular with processes that are unrelated to the spinning process, thus constraining concurrency

in non-functional terms, a solution that scales baddish...

Spin with CAS
cf. p. 44

```c
void lock(lock_t *bolt) {
    while (!CAS(&bolt->busy, false, true));
}
```

overcomes the problem of an “unconditional store”-prone TAS

the cache protocol runs write invalidate or update, resp., conditionally

but the problem of access contention at the shared bus remains
the processor is instructed to repeatedly run atomic “read-modify-write” cycles with only very short periods of leaving the bus unlocked
all sorts of simultaneous processes will have to suffer for bandwidth loss

in non-functional terms, a solution that scales bad...
Spin on Read

```c
void lock(lock_t *bolt) {
    do {
        while (bolt->busy);
    } while (!CAS(&bolt->busy, false, true));
}
```

Risk of degeneration to spin on CAS if the CSEC is too short and, thus, the cycle time of the entry/exit protocol possibly becomes shorter than the start-up time of the CPU for the next cycle within the cache (line 3): in the case of an x86, e.g., a handful (2–6) of processor instructions.

---

Backoff

### Definition

Static or dynamic **holding time**, stepped on a per-process(or) basis, that must elapse until resumption of a formerly contentious action.

- originally from telecommunications to facilitate **congestion control** (Ger. *Blockierungskontrolle*) by avoiding channel oversubscription:
  - statically (ALOHA [1]) or dynamically (Ethernet [10]) assigned delays
  - practised at broadcasting/sending time or to resolve contention, resp.
- adopted for parallel computing systems to reduce the probability\(^9\) of contention in case of conflicting accesses to shared resources
- common are dynamic approaches: exponential and proportional backoff

### Avoidance of Bus Lock Bursts

Allocation of stepped holding times on a per-process basis rivals with planning decisions of the process scheduler.

---

Lock Type III and IV

- for possibly lock-specific static/exponential backoff:
  - extended by a pointer to an open array of backoff values
  - typically, the array size complies with the number of processors

```c
typedef volatile struct lock {
    bool busy; /* initial: false */
    long (*rest)[]; /* initial: null */
} lock_t;
```

- for lock-specific proportional backoff: ticket-based
  - not dissimilar to a wait ticket dispenser (Ger. *Wartemarkenspender*) for a passenger paging system (Ger. *Personenaufnachlager*).

```c
typedef volatile struct lock {
    long next; /* number being served next */
    long this; /* number being currently served */
} lock_t;
```

---

Spin with Backoff I

- principle is to **pause** execution after a **collision** has been detected:
  - attenuate lock contention amongst known “wranglers” for the next trial

```c
void lock(lock_t *bolt) {
    while (!CAS(&bolt->busy, false, true))
        backoff(bolt, 1);
}
```

- combined with “**spin on read**” before (re-) sampling the lock flag:
  - combat lock contention for the next trial by assuming that “wranglers” could be overtaken by another simultaneous process

```c
void lock(lock_t *bolt) {
    do {
        while (bolt->busy);
        if (CAS(&bolt->busy, false, true)) break;
        backoff(bolt, 1);
    } while (true);
}
```

---

\(^8\)Note that the spinning processes may have been passed by a process.

\(^9\)Note that in interference-prone environments of unknown frequency, periods, and lengths of delays it is hardly feasible to prevent lock contention.
Spin with Backoff II

Truncated Exponential Backoff

- rely on feedback to decrease the rate of simultaneous processes:
  - gradual doubling of the per-process holding time when allocation failed
  - increasing lock-retry timeout with "ceiling value" (most significant bit)

```c
void lock(lock_t *bolt) {
    int hold = 1;
    do {
        while (bolt->busy);
        if (CAS(&bolt->busy, false, true)) break;
        backoff(bolt, hold);
        if ((hold << 1) != 0) hold <<= 1;
    } while (true);
}
```

in non-functional terms, solutions that scale to some extent...
- including the solutions of static backoff as shown before

Spin with Ticket

Proportional Backoff

```c
void lock(lock_t *bolt, long cset) {
    long self = FAA(&bolt->next, 1);
    if (self != bolt->this) {
        rest((self - bolt->this) * cset);
        while (self < bolt->this);
    }
}
void unlock(lock_t *bolt) {
    bolt->this += 1; // register next one's turn */
}
```

note that self – this gives the number of waiting processes that will be served first in order to run the critical section

knowing the critical section execution time (CSET) would be great
- a choice of best-, average-, or worst-case execution time (B/A/WCET)
- depends on the structure of critical sections as well as “background noise”

Interference by Ticket-Lock

Entry policy is first-come, first-served (FCFS), which rarely complies with the process scheduler policy.

Outline

Preface

Fundamentals
- Bifocal Perspective
- Basic Attributes

Avenues of Approach
- Atomic Memory Read/Write
- Specialised Instructions

Summary


© wosch CS (WS 2014, LEC 6) Avenues of Approach–Specialised Instructions 30–44

© wosch CS (WS 2014, LEC 6) Summary 32–44
Résumé

- conventional locking under prevention of context switches
- hierarchic placement of lock/unlock implementations ~ ISA level
- standby position, control mode, properties, computational burden
- approaches with atomic read/write or added specialised instructions
- algorithms falling back on TAS, CAS, FAA, and backoff procedures
- although simple in structure, potential deleterious cache effects
- lock contention when processes try to acquire a lock simultaneously
- bus lock bursts when processes run the entry protocol in common mode

Critical Section Execution Time (CSEC)

That locks are suitable for a short CSEC is computer-science folklore, but by far too flat. Much more important is to have a bounded and, even better, constant CSEC. Above all, this makes high demands on the design of critical sections and non-sequential programs.

Reference List I


Reference List II


Reference List III


Peterson's Solution for $N = 2$: Transformation

- the construct of the busy wait loop in the entry protocol originally described in [12] is to be read as follows:

  \[
  \text{wait until condition} = \text{repeat nothing until condition} = \text{do nothing while } \neg \text{condition}
  \]

  applied to C = while (\neg \text{condition});

  with condition = \neg Q_i \text{ or turn } = i

  inserted and factored out = while (\neg (\neg Q_i \text{ or turn } = i));

  = while (Q_i \text{ and turn } \neq i);

  = while (Q_i \text{ and turn } = j);

  with $j \neq i$

- this results in a code structure of the entry protocol that is different from the many examples as can be found in the Web

Original Dekker's Algorithm for $N = 2$ Interpretation

- let self be the current process, peer be the counterpart, and bolt be the lock variable used to protect some critical section CS

- a first glance at the entry protocol reveals:

  4. self shows interest in entering CS, maybe simultaneously to peer's intend to enter the same CS as well

  5–9. if applicable, self hence waits on peer to yield CS and appoint self being candidate to run CS next

- upon a closer look, the entry protocol takes care of the following:

  5–6. as the case my be, self contends with peer for entrance but retries if it should be self's turn to enter

  7–8. in that case, while preventing potential deadlock[11] of the processes, self waits on peer for being appointed to enter CS

- 9. reconsider entering of the critical section...

---

Disregarding the original reference, EWD is also renowned for a pamphlet that argues for abolishment of goto from high-level programming languages [5].

---

Original Dekker's Algorithm for $N = 2$ cf. [4, p.17–18]

```c
void lock(lock_t *bolt) {
  unsigned self = earmark();

  A: bolt->want[self] = true;
  L: if (bolt->want[self-1]) {
    if (bolt->turn[0] == self) goto L;
    bolt->want[self] = false;
  }
  B: if (bolt->turn[0] == (self-1)) goto B;
     goto A;
}
```

---

Peterson's Solution for $N > 2$ cf. [6] or [12], resp.

```c
void lock(lock_t *lock) {
  unsigned rank, next, self = earmark();

  for (rank = 0; rank < NPROC - 1; rank++) {
    lock->want[self] = rank;
    lock->turn[rank] = self;

    for (next = 0; next < NPROC; next++)
      if (next != self)
        while ((lock->want[next] >= rank) && (lock->turn[rank] == self));
  }

  void unlock(lock_t *lock) {
    unsigned self = earmark();

    lock->want[self] = -1;
  }
```

---

Memory Barriers/Fences

Beware of dynamic ordering of read/write operations.
### Peterson's Solution for $N > 2$

**Interpretation**

**Hint**

Every process must have proved oneself for $n-1$ ranks to be eligible for entering the critical section.

- Basic idea is to apply the two-process solution at each rank repeatedly
  - at least one process is eliminated, stepwise, until only one remains
- Let $want[p]$ be the rank of process $p$, let $turn[r]$ be the process that entered rank $r$ last, and let $CS$ be a critical section:
  - in attempting to enter $CS$, indicate interest to reach the next rank
  - for it, check all other processes for their particular rank and
  - busy wait if there are still higher ranked processes and the current process is still designed to be promoted
- Often also labelled as filter or tournament algorithm:
  - deters one out of $N$ simultaneous processes from entering $CS$
  - repeated for $N-1$ times, only one process will be granted access finally

---

### Lamport’s Bakery Algorithm I

**Lock Type and Ticket Dispenser**

```c
#include <stdbool.h>

typedef volatile struct lock {
  bool want[NPROC];  /* initial: all false */
  long turn[NPROC];  /* initial: all 0 */
} lock_t;

void ticketing(lock_t *bolt, unsigned slot) {
  unsigned next, high = 0;
  bolt->want[slot] = true;  /* enter choosing */
  for (next = 0; next < NPROC; next++)
    if (bolt->turn[next] > high)
      high = bolt->turn[next];
  bolt->turn[slot] = high + 1;  /* state number */
  bolt->want[slot] = false;  /* leave choosing */
}
```

---

### Lamport’s Bakery Algorithm II

**cf. [9]**

```c
void lock(lock_t *bolt) {
  unsigned next, self = earmark();
  ticketing(bolt, self);  /* take a number */
  for (next = 0; next < NPROC; next++) {
    while (bolt->want[next]);  /* next chooses */
    while ((bolt->turn[next] != 0)
      && ((bolt->turn[next] < bolt->turn[self])
        || ((bolt->turn[next] == bolt->turn[self])
          && (next < self))));
    }
}

void unlock(lock_t *bolt) {
  unsigned self = earmark();
  bolt->turn[self] = 0;
}
```

---

### Spin with TAS or CAS, resp.

**cf. p.23 and p.24**

```asm
_MEMORY_BARRIERS/FENCES

number of "busy wait" loop actions with bus locked and unlocked:

_line (5) v. lines (4, 6, 7)_

_lines (14, 15) v. lines (13, 16, 17)_

in case of x86, there is no difference as to the number of actions
but there is still the difference as to the frequency of cache interference
the ratio depends on the code generator (compiler) and the CPU

_line (5) v. lines (4, 6, 7)_

_lines (14, 15) v. lines (13, 16, 17)_

in case of x86, there is no difference as to the number of actions
but there is still the difference as to the frequency of cache interference
the ratio depends on the code generator (compiler) and the CPU

```