INF346: Distributed Systems

Theory and Practice

Telecom ParisTech, 2015

Administrivia

- Langue – français ou anglais?
- Lectures: Mon (8:30-11:45), Wed (8:30-11:45), Fri (13:30-16:45), check EOLE for exact time/place
- Exercises (TDs and TPs): 6.03, 23.03, 27.03
- Paper presentations: 22.04
- Exam: 24.04
- Two parts: algorithms (Petr Kuznetsov) and systems (Remi Sharrock)
- Web page: http://perso.telecom-paristech.fr/~kuznetso/INF346-2015/
- Office hours:
  - ✓ Petr Kuznetsov: C213-2, appointment by email: petr.kuznetsov@telecom-paristech.fr
- Credit = 2/3 exam + 1/3 bibliographic project
  - ✓ Papers to present on the course web page
  - ✓ Bonus for participation and lecture notes bugs

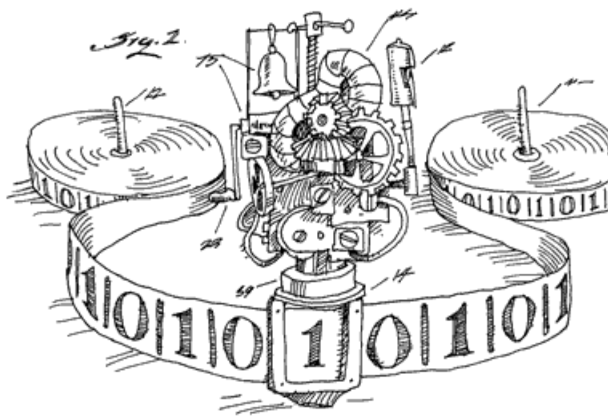
Literature

- Lecture notes: Concurrent computing
  http://perso.telecom-paristech.fr/~kuznetso/MPRI13/book-ln.pdf
- M. Herlihy and N. Shavit. The art of multiprocessor programming. Morgan Kaufman, 2008
- Lynch, N: Distributed Algorithms. Morgan Kaufmann Publishers, 1996.
- H. Attiya, J. Welch. Distributed Computing: Fundamentals, Simulations and Advanced Topics (2nd edition). Wiley. 2004

What is computing?

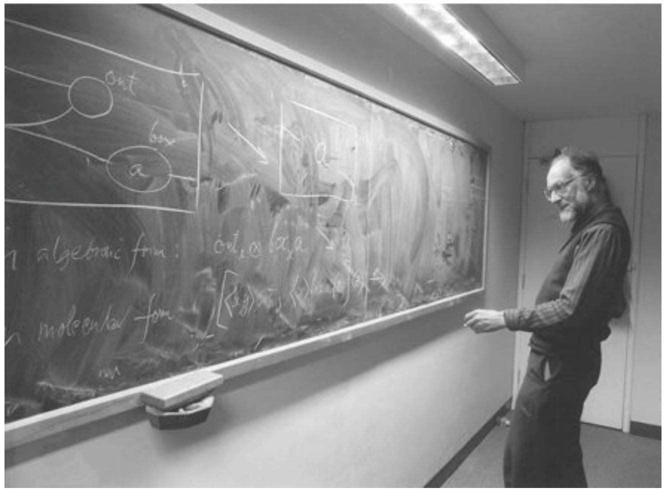
What is done by a Turing machine

Alan Turing
1912 – 1954



Not well adjusted to concurrency?

Computation as interaction



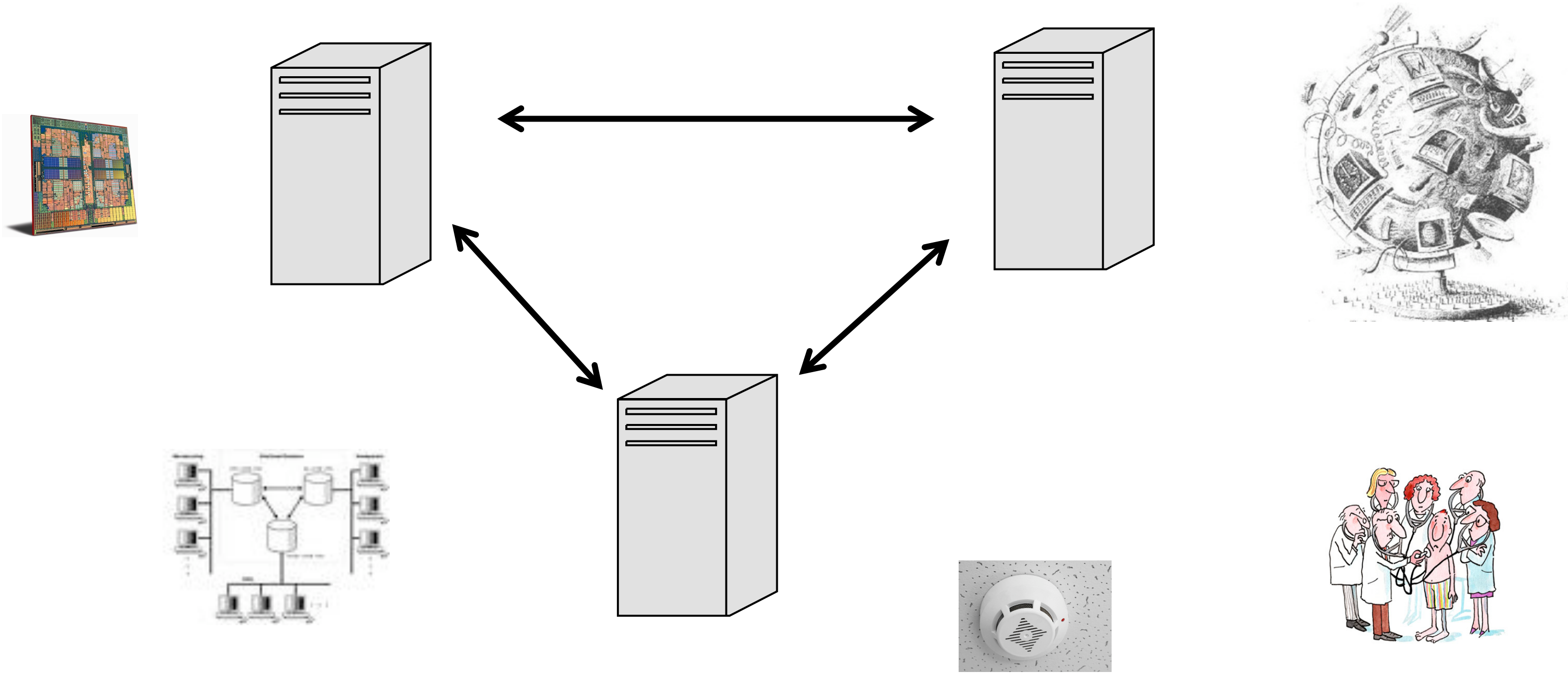
Robin Milner

1934-2010

This course is about distributed computing:

independent sequential processes that communicate

Concurrency is everywhere!



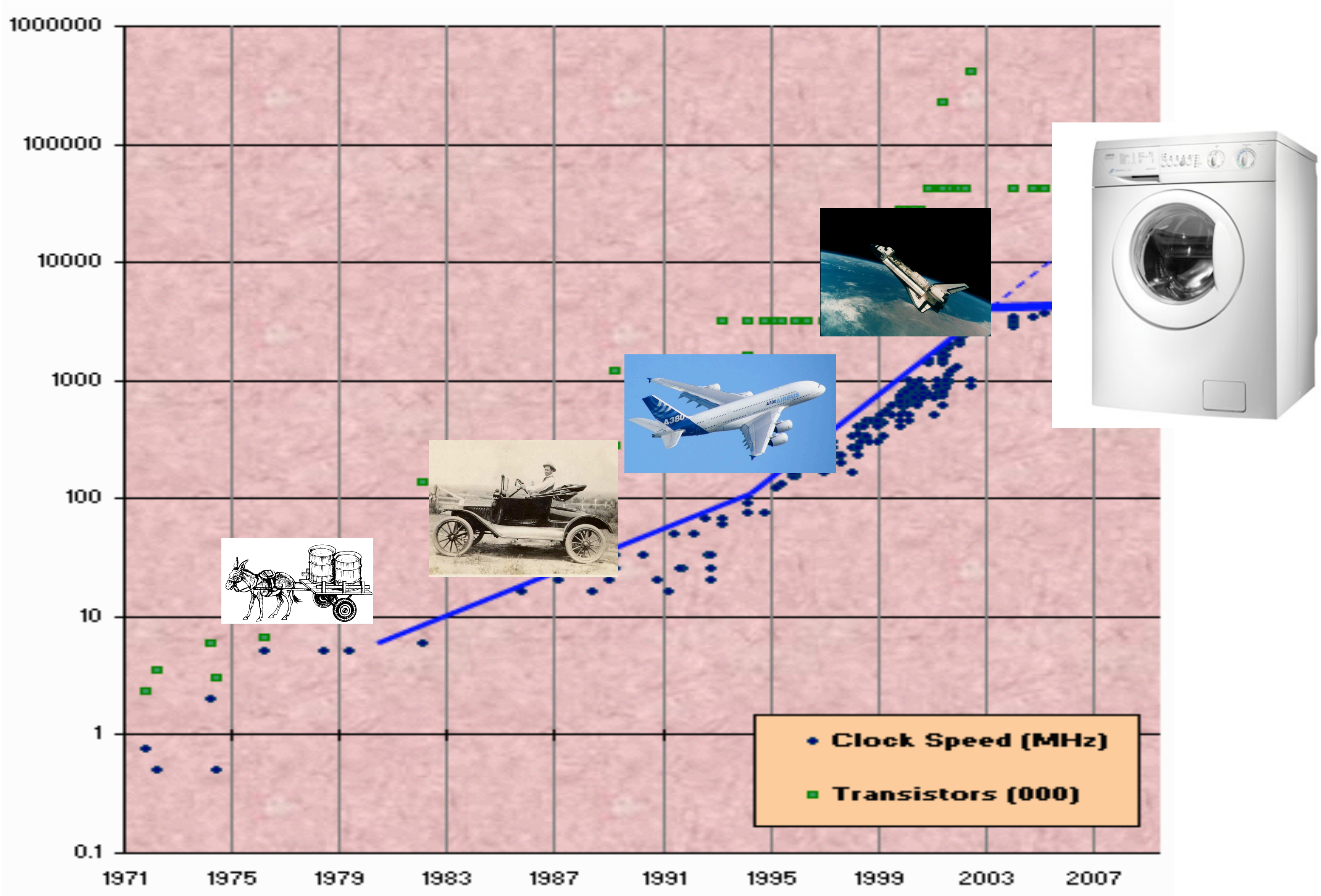
- Multi-core processors
- Sensor networks
- Internet
- Basically everything related computing

Communication models

- Shared memory
  - ✓Processes apply (read–write) operations on shared variables
  - ✓Failures and asynchrony

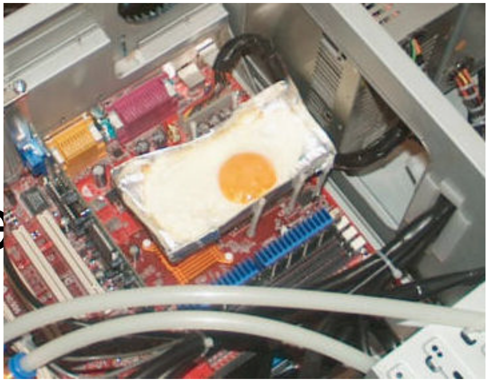
- Message passing
  - ✓Processes send and receive messages
  - ✓Communication graphs
  - ✓Message delays

Moore's Law and CPU speed



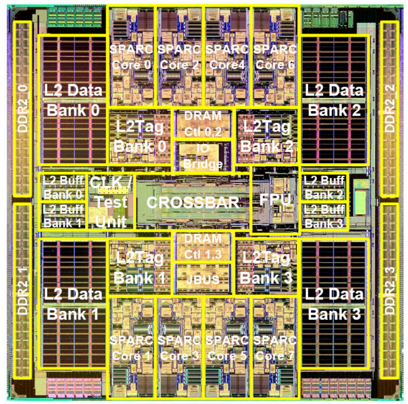
Clock speed deadend

- Memory wall
  - ✓Performance gap between memory and CPU
- ILP wall
  - ✓Not enough work to spend the cycles
- Power wall
  - ✓Thermal problems caused by higher clock speeds



The case against the “washing machine science”

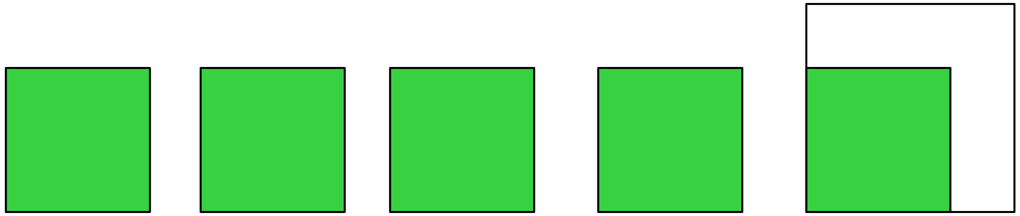
- Single-processor performance does not improve
- But we can add more cores
- Run concurrent code on multiple processors



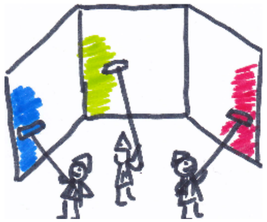
Can we expect a proportional speedup?
(ratio between sequential time and parallel time for executing a job)

Example: painting in parallel

- 5 friends want to paint 5 equal-size rooms, one friend per room
  - ✓Speedup = 5



- What if one room is twice as big?



Amdahl's Law

- p – fraction of the work that can be done in parallel (no synchronization)
- n - the number of processors
- Time one processor needs to complete the job = 1

$$S = \frac{1}{1 - p + p / n}$$

Painting in parallel

- Assigning one painter to one room, 5/6 of the work can be performed in parallel.
- Parallel execution time = 1-5/6+1/6 = 1/6+1/6 = 2/6 = 1/3

$$S = 1/(1/3) = 3$$

- Can be worse: 10 rooms, 10 painters, one room twice bigger

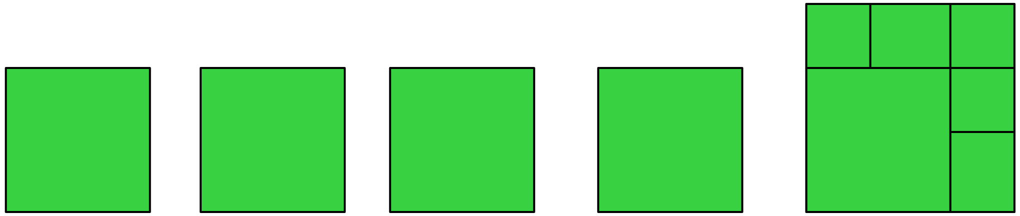
$$S = 1/\ (1-10/11+1/11) = 11/2 = 5.5$$

- But >90% of the work can be parallelized!

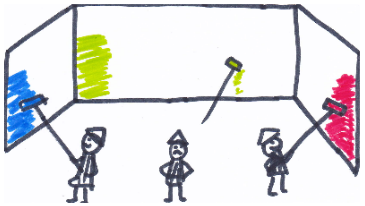
Cannot be better than 11, regardless of the number of processors!

A better solution

- When done, help the others
  - ✓All 5 paint the remaining half-room in parallel
- Communication and agreement is required!
- **This is a hard task**



- And this is exactly what synchronization algorithms try to achieve!



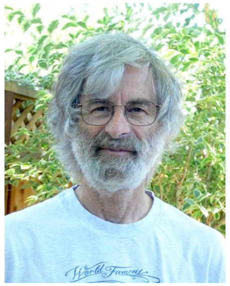
Challenges

- What is a correct implementation?
  - ✓Safety and liveness
- What is the cost of synchronization?
  - ✓Time and space lower bounds
- Failures/asynchrony
  - ✓Fault-tolerant concurrency?
- How to distinguish possible from impossible?
  - ✓Impossibility results

Distributed ≠ Parallel

- The main challenge is synchronization

- "you know you have a distributed system when the crash of a computer you've never heard of stops you from getting any work done" (Leslie Lamport)



History

- Dining philosophers, mutual exclusion (Dijkstra)~60's
- Distributed computing, logical clocks (Lamport), distributed transactions (Gray) ~70's
- Consensus (Lynch) ~80's
- Distributed programming models, since ~90's
- Multicores now

Why theory of distributed computing?

- Every computing system is distributed
- Computing getting mission-critical
  - ✓Understanding fundamentals is crucial
- Intellectual challenge
  - ✓A distinct math domain?

Outline

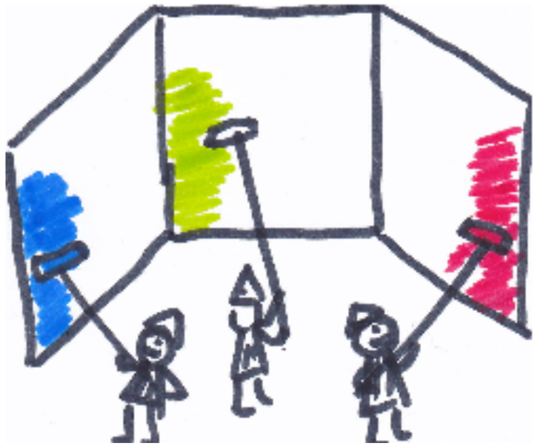
I. Synchronization problems
II. Correctness: safety and liveness
III. Read-write and snapshot memory
IV. Consensus
V. Transactional memory
VI. CAP theorem, synchrony assumptions
VII. Strong consistency: Paxos
VIII. Semantic-aware consistency: operational transformation
IX. Advanced topics in distributed computing (SDN,…)

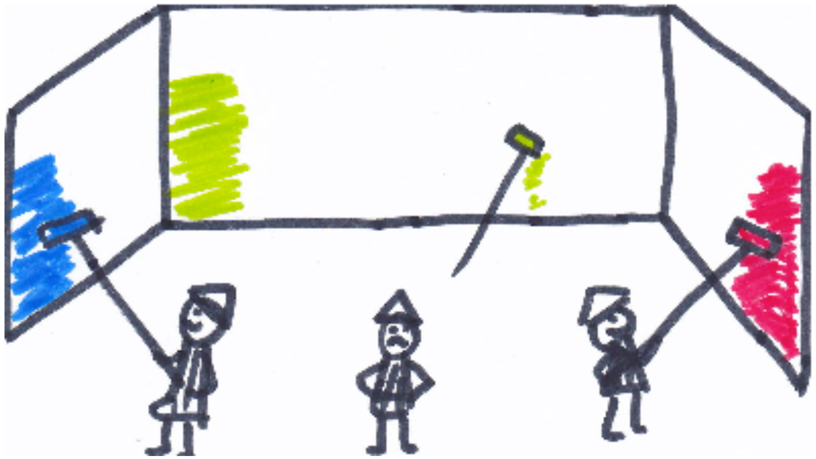
Real concurrency--in which one program actually continues to function while you call up and use another--is more amazing but of small use to the average person. How many programs do you have that take more than a few seconds to perform any task?

New York Times, 25 April 1989, in an article on new operating systems for IBM PC

Synchronization, blocking and non-blocking

INF346, 2015





$$S = \frac{1}{1 - p + p/n}$$

Why synchronize ?

- Concurrent access to a shared resource may lead to an inconsistent state
  - ✓ E. g., concurrent file editing
  - ✓ Non-deterministic result (**race condition**): the resulting state depends on the scheduling of processes

- Concurrent accesses need to be **synchronized**
  - ✓ E. g., decide who is allowed to update a given part of the file at a given time

- Code leading to a race condition is called **critical section**
  - ✓ Must be executed sequentially

- **Synchronization problems**: mutual exclusion, readers-writers, producer-consumer, …

Edsger Dijkstra
1930-2002

Dining philosophers (Dijkstra, 1965)

- To make progress (to eat) each process (philosopher) needs two resources (forks)
- Mutual exclusion: no fork can be shared
- Progress conditions:
  - ✓Some philosopher does not starve (deadlock-freedom)
  - ✓No philosopher starves (starvation-freedom)

Mutual exclusion

- No two processes are in their critical sections (CS) at the same time

\+

- Deadlock-freedom: **at least one** process eventually enters its CS
- Starvation-freedom: **every** process eventually enters its CS
  - ✓ Assuming no process **blocks** in **CS** or **Entry section**

- Originally: implemented by reading and writing
  - ✓ Peterson's lock, Lamport's bakery algorithm
- Currently: in hardware (mutex, semaphores)

Peterson's lock: 2 processes

```
bool flag[0]   = false;
bool flag[1]   = false;
int turn;
```

```
P0:

flag[0] = true;
turn = 1;
while (flag[1] and turn==1)
{
          // busy wait
}
// critical section
…
// end of critical section
flag[0] = false;
```

```
P1:

flag[1] = true;
turn = 0;
while (flag[0] and turn==0)
{
          // busy wait
}
// critical section
…
// end of critical section
flag[1] = false;
```

Peterson's lock: N ≥ 2 processes

```
// initialization
level[N] = -1;     // current level of processes 0...N-1
waiting[N-1] = -1; // the waiting process of each level 0...N-2

// code for process i that wishes to enter CS
for (m = 0; m < N-1; ++m) {
    level[i] = m;
    waiting[m] = i;
    while(waiting[m] == i &&(exists k ≠ i: level[k] ≥ m)) {
        // busy wait
    }
}
// critical section
level[i] = -1; // exit section
```

Bakery [Lamport'74,simplified]

```
// initialization
flag: array [1..N] of bool = {false};
label: array [1..N] of integer = {0}; //assume no bound

// code for process i that wishes to enter CS

flag[i] = true; //enter the "doorway"
label[i] = 1 + max(label[1], ..., lebel[N]); //pick a ticket
while (for some k ≠ i: flag[k] and (label[k],k)<<(label[i],i));
// wait until all processes "ahead" are served
…
// critical section
…
flag[i] = false; // exit section
```

Processes are served in the "ticket order": first-come-first-serve

Readers-writers problem

- Writer updates a file
- Reader keeps itself up-to-date
- Reads and writes are non-atomic!

Why synchronization? Inconsistent values might be read

```
        Writer                                Reader
T=0: write("sell the cat")

                                   T=1: read("sell …")

T=2: write("wash the dog")

                                   T=3: read("… the dog")
```

Sell the dog?

Producer-consumer (bounded buffer) problem

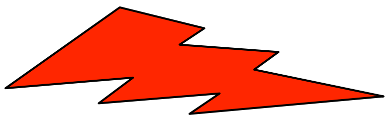
- Producers **put** items in the buffer (of bounded size)
- Consumers **get** items from the buffer
- Every item is consumed, no item is consumed twice

(Client-server, multi-threaded web servers, pipes, …)

Why synchronization? Items can get lost or consumed twice:

Producer

```
/* produce item */
while (counter==MAX);
buffer[in] = item;
in = (in+1) % MAX;
counter++;
```



Race!

Consumer

```
/* to consume item */
while (counter==0);
item=buffer[out];
out=(out+1) % MAX;
counter--;
/* consume item */
```

Synchronization tools

- Busy-waiting (TAS)
- Semaphores (locks), monitors
- Nonblocking synchronization
- Transactional memory

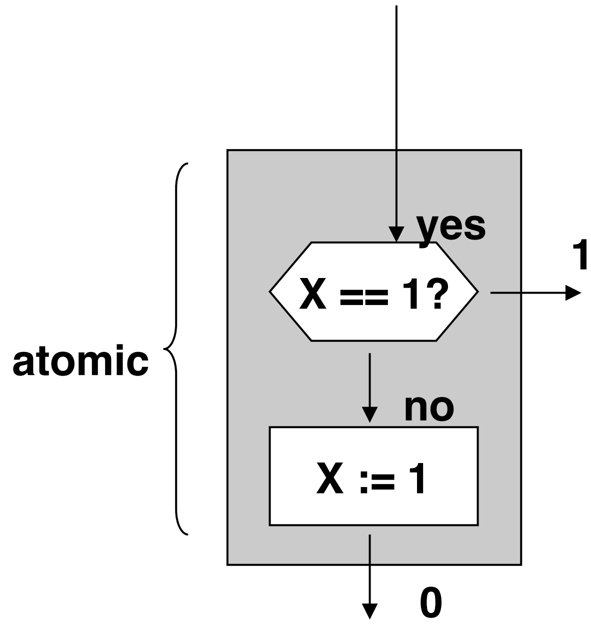
Busy-wait: Test and Set

- TAS(X) **tests** if X = 1, **sets** X to 1 if not, and returns the old value of X
  - ✓ Instruction available on almost all processors

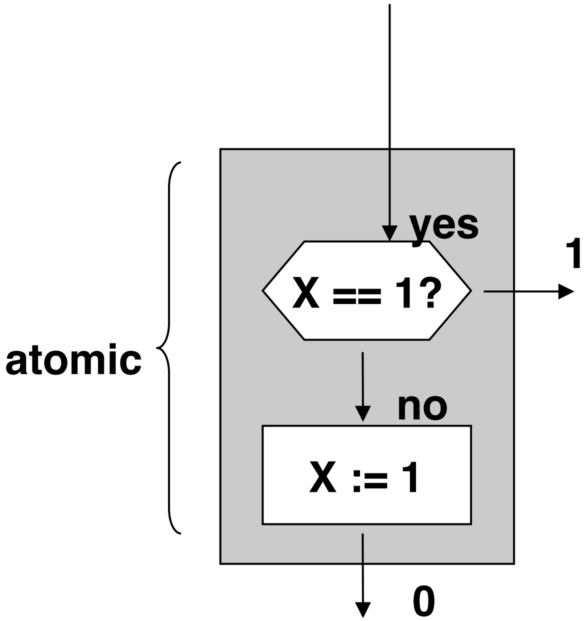
```
TAS(X):
```

atomic $\left\{ \right.$

```
if X == 1 return 1;
X = 1;
return 0;
```



Busy-wait: Test and Set



shared X:=0

Producer	Consumer
`while(counter==MAX);`	`while (counter==0);`
`. . .`	`. . .`
`buffer[in] = item;`	`item = buffer[out];`
`. . .`	`. . .`
`while TAS(X);`	`while TAS(X);`
`counter++;`	`counter--;`
`X:=0;`	`X:=0;`
`. . .`	`...`

Problems:

- busy waiting
- no record of request order (for multiple producers and consumers)

Semaphores [Dijkstra 1968]: specification

- A semaphore S is an integer variable accessed (apart from initialization) with two atomic operations P(S) and V(S)
  - ✓ Stands for “passeren” (to pass) and “vrijgeven” (to release) in Dutch

- The value of S indicates the number of resource elements available (if positive), or the number of processes waiting to acquire a resource element (if negative)

```
Init(S,v){ S := v; }

P(S){
        while S<=0;     /* wait until a resource is available */
        S--;            /* pass to a resource */
}

V(S){
        S++;            /* release a resource */
}
```

Semaphores: implementation

S is associated with a composite object:

- ✓ S.counter: the **value** of the semaphore
- ✓ S.wq: the **waiting queue,** memorizing the processes having requested a resource element

```
Init(S,R_nb) {
    S.counter=R_nb;
    S.wq=empty;
}
P(S) {
    S.counter--;
    if S.counter<0{
     put the process in S.wq and wait until
    READY;}
}
V(S) {
    S.counter++
    if S.counter>=0{
                mark 1st process in S.wq as
                READY;}
}
```

TELECOM ParisTech

Lock

- A semaphore initialized to 1, is called a **lock** (or **mutex)**
- When a process is in a critical section, no other process can come in

shared semaphore S := 1

Producer	Consumer
`while (counter==MAX);`	`while (counter==0);`
`. . .`	`. . .`
`buffer[in] = item;`	`item = buffer[out];`
`. . .`	`. . .`
`P(S);`	`P(S);`
`counter++;`	`counter--;`
`V(S)`	`V(S);`
`. . .`	`...`

Problem: still waiting until the buffer is ready

Semaphores for producer-consumer

- 2 semaphores used :
  - ✓ **empty:** indicates empty slots in the buffer (to be used by the producer)
  - ✓ **full:** indicates full slots in the buffer (to be read by the consumer)

```
shared  semaphores empty := MAX, full := 0;
```

Producer	Consumer
`P(empty)`	`P(full);`
`buffer[in] = item;`	`item = buffer[out];`
`in = (in+1) % MAX;`	`out=(out+1) % MAX;`
`V(full)`	`V(empty);`

Potential problems with semaphores/locks

- **Blocking**: progress of a process is conditional (depends on other processes)
- **Deadlock:** no progress ever made

```
X1:=1; X2:=1

      Process 1        |      Process 2
-----------------------+-----------------------
  ...                  |  ...
  P(X1)                |  P(X2)
  P(X2)                |  P(X1)
  critical section     |  critical section
  V(X2)                |  V(X1)
  V(X1)                |  V(X2)
  ...                  |  ...
```

- **Starvation**: waiting in the waiting queue forever

Other problems of blocking synchronization

- Priority inversion
  - ✓High-priority threads blocked
- No robustness
  - ✓Page faults, cache misses etc.
- Not composable

Can we think of anything else?

Non-blocking algorithms

A process makes progress, regardless of the other processes

shared buffer[MAX]:=empty; head:=0; tail:=0;

Producer `put(item)`	Consumer `get()`

`if (tail-head == MAX){`
`        return(full);`
`}`
`buffer[tail%MAX]=item;`
`tail++;`
`return(ok);`
	`if (tail-head == 0){`
	`        return(empty);`
	`}`
	`item=buffer[head%MAX];`
	`head++;`
	`return(item);`

Problems:

- works for 2 processes but hard to say why it works ☺
- multiple producers/consumers? Other synchronization pbs?

(stay in class to learn more)

Transactional memory

- Mark sequences of instructions as an **atomic transaction**, e.g., the resulting producer code:

```
atomic {
        if (tail-head == MAX){
        return full;
        }
        items[tail%MAX]=item;
        tail++;
}
return ok;
```

- A transaction can be either **committed** or **aborted**
  - ✓ Committed transactions are **serializable**
  - ✓ Let the transactional memory (TM) care about the conflicts
  - ✓ Easy to program, but performance may be problematic

Summary

- Concurrency is indispensable in programming:
  - ✓ Every system is now concurrent
  - ✓ Every parallel program needs to synchronize
  - ✓ Synchronization cost is high ("Amdahl's Law")

- Tools:
  - ✓ Synchronization primitives (e.g., monitors, TAS, CAS, LL/SC)
  - ✓ Synchronization libraries (e.g., java.util.concurrent)
  - ✓ Transactional memory, also in hardware (Intel Haswell, IBM Blue Gene,…)

- Coming next:
  - ✓ Nonblocking synchronization using read-write memory
  - ✓ Read-write transformations and snapshot memory

Quiz

- What if we reverse the order of the first two lines the 2-process Peterson's algorithm

```
P0:                          P1:
turn = 1;                    turn = 0;
flag[0] = true;              flag[1] = true;
…                            …
```

  Would it work?
- Prove that Peterson's N-process algorithm ensures:
  - ✓ mutual exclusion: no two processes are in the critical section at a time
  - ✓ starvation freedom: every process in the trying section eventually reaches the critical section (assuming no process fails in the trying, critical, or exit sections)
- Show that the bounded (black-white) Bakery algorithm in correct

Bakery [Lamport'74,original]

```
// initialization
flag: array [1..N] of bool = {false};
label: array [1..N] of integer = {0}; //assume no bound

// code for process i that wishes to enter CS
flag[i] = true; //enter the doorway
label[i] = 1 + max(label[1], ..., lebel[N]); //pick a ticket
flag[i] = false; //exit the doorway
for j=1 to N do
          while (flag[j]); //wait until j is not in the doorway
          while (label[j]≠0 and (label[j],j)<<(label[i],i));
          // wait until j is not "ahead"
…
// critical section
…
label[i] = 0; // exit section
```

Ticket withdrawal is "protected" with flags: a very useful trick

Black-White Bakery [Taubenfeld'06]

```
// initialization
color: {black,white};
flag: array [1..N] of bool = {false};
label[1..N]: array of type {0,…,n} = {0}  //bounded ticket numbers
mycolor[1..N]: array of type {black,white}

// code for process i that wishes to enter CS
flag[i] = true; //enter the "doorway"
mycolor[i] =color;
label[i] = 1 + max({label[j]| j=1,…,N: mycolor[i]=mycolor[j]});
flag[i] = false; //exit the "doorway"
for j=1 to N do
   while (flag[j]);
   if mycolor[j]=mycolor[i] then
      while (label[j]≠0 and (label[j],j)<<(label[i],i) and mycolor[j]=mycolor[i]  );
   else
      while (label[j]≠0 and mycolor[i]=color and mycolor[j] ≠ mycolor[i]);

// wait until all processes "ahead" of my color are served
…
// critical section
…
If mycolor[i]=black then color = white else color – black;
label[i] = 0; // exit section
```

Colored tickets => bounded variables!