Concurrency with Shared Variables

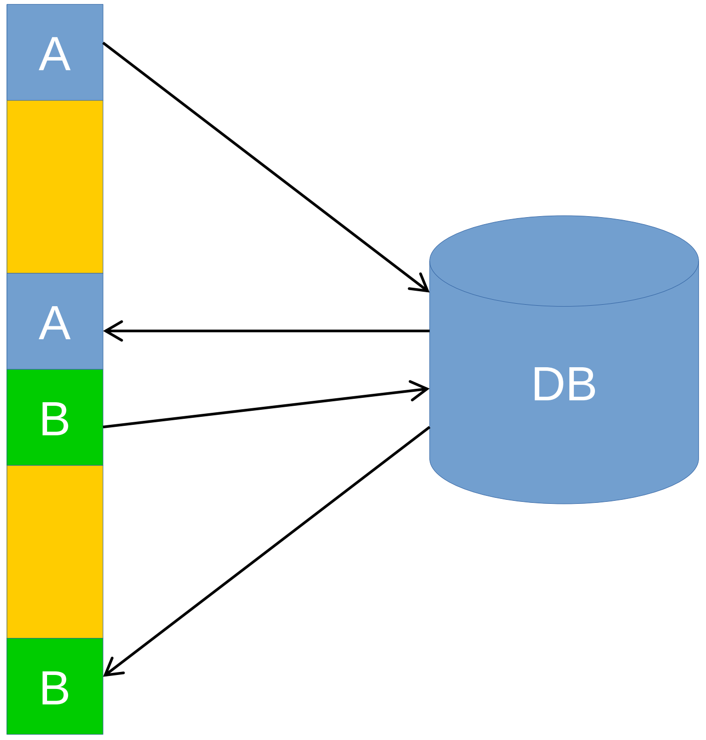
Goroutines and threads, dataraces, mutual exclusion, sync.Mutex, sync.RWMutex, sync.Once, sync.atomic, Go race detector, examples

Where to Find The Code and Materials?

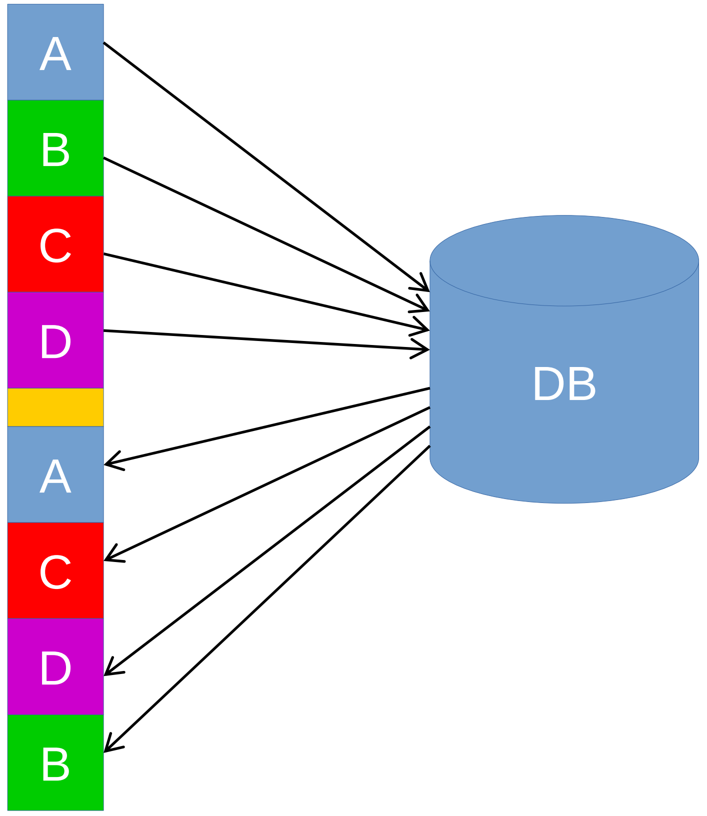
https://github.com/iproduct/coursego

Synchronous vs. Asynchronous IO

Synchronous



Asynchronous



Blocking vs. Non-blocking

- Blocking concurrency – uses **Mut**ual **Ex**clusion primitives (aka **Locks**) to prevent threads from simultaneously accessing/modifying the same resource
- Non-blocking concurrency does not make use of locks.
- One of the most advantageous feature of non-blocking vs. blocking is that, threads does not have to be suspended/waken up by the OS. Such overhead can amount to 1ms to a few 10ms, so removing this can be a big performance gain.

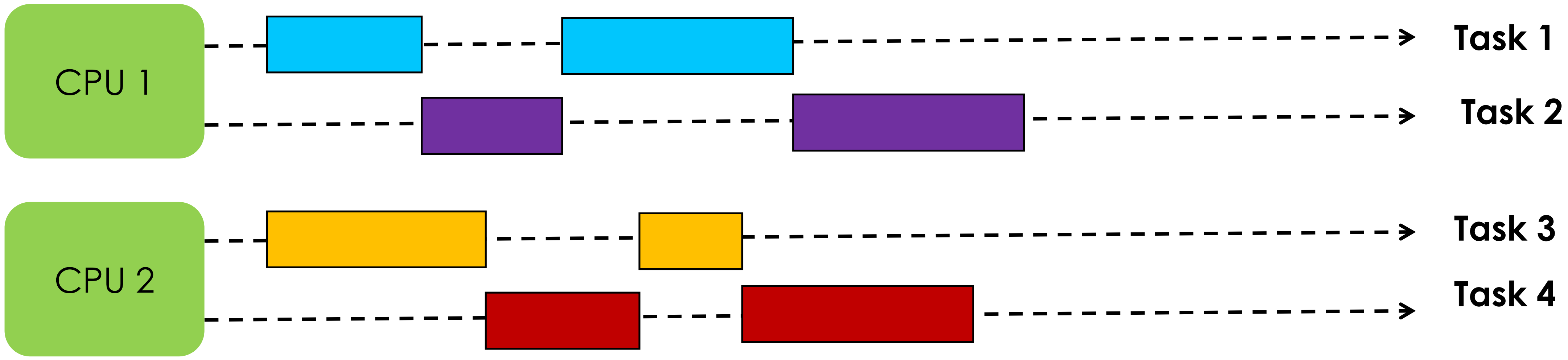
Non-blocking Concurrency

- In computer science, an algorithm is called **non-blocking** if failure or suspension of any thread cannot cause failure or suspension of another thread;[1] for some operations, these algorithms provide a useful alternative to traditional blocking implementations. A non-blocking algorithm is **lock-free** if there is guaranteed system-wide progress, and **wait-free** if there is also guaranteed per-thread progress. "Non-blocking" was used as a synonym for "lock-free" in the literature until the introduction of obstruction-freedom in 2003.
- It has been shown that widely available atomic *conditional* primitives, CAS and LL/SC, cannot provide starvation-free implementations of many common data structures without memory costs growing linearly in the number of threads.

[Wikipedia]

Concurrency vs. Parallelism

- Concurrency refers to how a single CPU can make progress on multiple tasks seemingly at the same time (AKA concurrently).
- Parallelism allows an application to parallelize the execution of a single task - typically by splitting the task up into subtasks which can be completed in parallel.



Scalability Problem

- Scalability is the ability of a program to handle growing workloads.
- One way in which programs can scale is parallelism: if we want to process a large chunk of data, we describe its processing as a sequence of transforms on a stream, and by setting it to parallel we ask multiple processing cores to process the parts of the task simultaneously.
- The problem is that the processes and threads, the OS supported units of concurrency, cannot match the scale of the application domain's natural units of concurrency — a session, an HTTP request, or a database transactional operation.
- A server can handle upward of a million concurrent open sockets, yet the operating system cannot efficiently handle more than a few thousand active threads. So it becomes a mapping problem - M:N

Why are OS Threads Heavy?

- Universal – represent all languages and types of workloads
- Can be suspended and resumed - this requires preserving its state, which includes the instruction pointer, as well as all of the local computation data, stored on the stack.
- The stack should be quite large, because we can not assume constraints in advance.
- Because the OS kernel must schedule all types of threads that behave very differently it terms of processing and blocking — some serving HTTP requests, others playing videos => its scheduler must be all-encompassing, and not optimized.

Solutions To Thread Scalability Problem

- Because threads are costly to create, we pool them -> but we must pay the price: leaking thread-local data and a complex cancellation protocol.
- Thread pooling is coarse grained – not enough threads for all tasks.
- So instead of blocking the thread, the task should return the thread to the pool while it is waiting for some external event, such as a response from a database or a service, or any other activity that would block it.
- The task is no longer bound to a single thread for its entire execution.
- Proliferation of asynchronous APIs, from Noide.js to NIO in Java, to the many “reactive” libraries (Reactive Extensions - Rx, etc.) => intrusive, all-encompassing frameworks, even basic control flow, like loops and try/catch, need to be reconstructed in “reactive” DSLs, supporting classes with hundreds of methods.

What Color is Your Function?
[http://journal.stuffwithstuff.com/2015/02/01/what-color-is-your-function/]

- Synchronous functions return values, async ones do not and instead invoke callbacks.
- Synchronous functions give their result as a return value, async functions give it by invoking a callback you pass to it.
- You can't call an async function from a synchronous one because you won't be able to determine the result until the async one completes later.
- Async functions don't compose in expressions because of the callbacks, have different error-handling.
- Node's whole idea is that the core libs are all asynchronous. (Though they did dial that back and start adding ___Sync() versions of a lot functions.)

Async/Await in JS (and some other languages)

- Cooperative scheduling points are marked explicitly with await => scalable synchronous code – but we mark it as async – a bit of confusing!
- Solves the context issue by introducing a new kind of context that is like thread but is incompatible with threads – one blocks and the other returns some sort of Promise - you can not easily mix sync and async code.

Right-Sized Concurrency

- If we could make threads lighter, we could have more of them, and can use them as intended:
  1. to directly represent domain units of concurrency;
  2. by virtualizing scarce computational resources;
  3. hiding the complexity of managing those resources.
- Example: Goroutines, Erlang
- creating and suspending goroutines is cheap
- Managed by the Golang runtime, and scheduled cooperatively and/or preemptively.

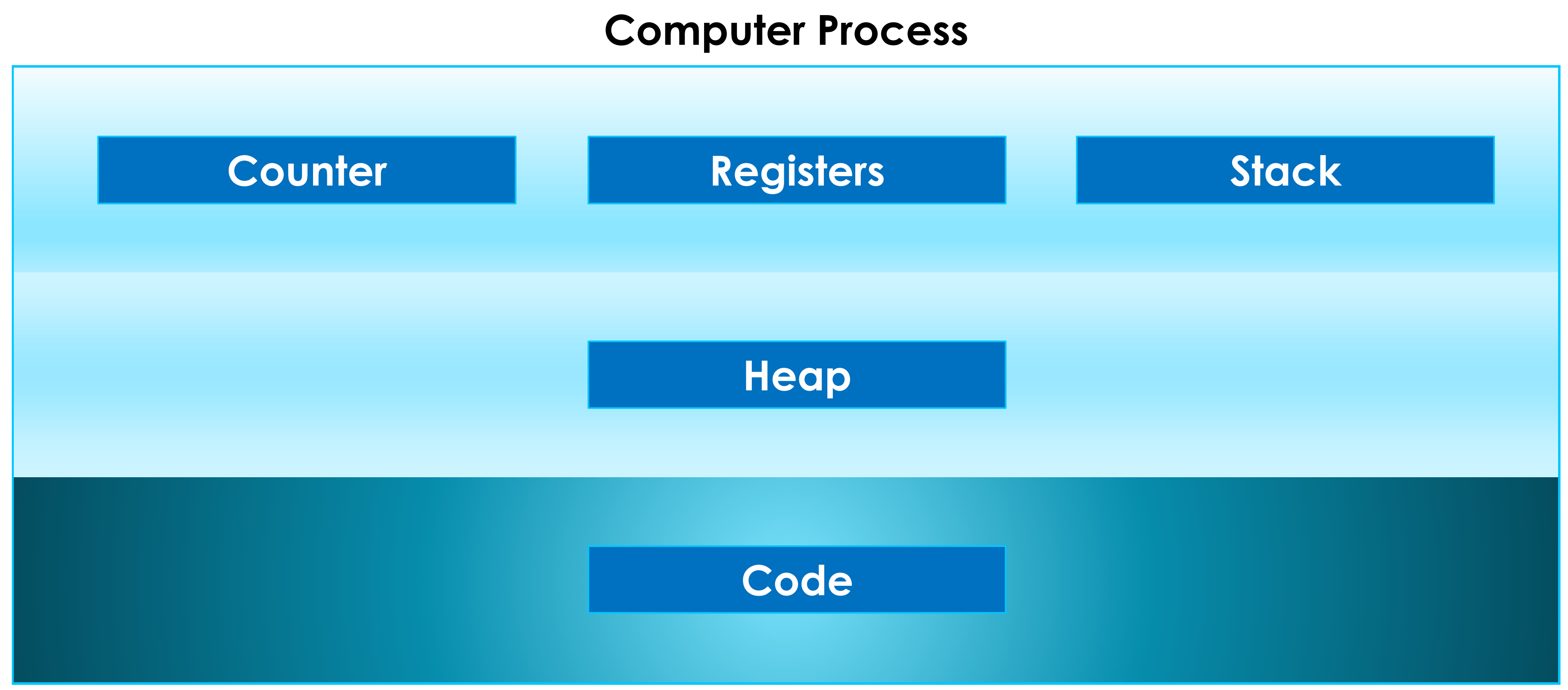
How Are Goroutines Better?

- The goroutines make use of a stack, so it can represent execution state more compactly.
- Control over execution by optimized Dispatcher;
- Millions of goroutines => every unit of concurrency in the application domain can be represented by its own goroutine
- Just spawn a new goroutine, one per task.
- Example: HTTP request - a new goroutine is already spawned to handle it, but now, in the course of handling the request, you want to simultaneously query a database, and issue outgoing requests to three other services? No problem: spawn more goroutines.

How Are Goroutines Better?

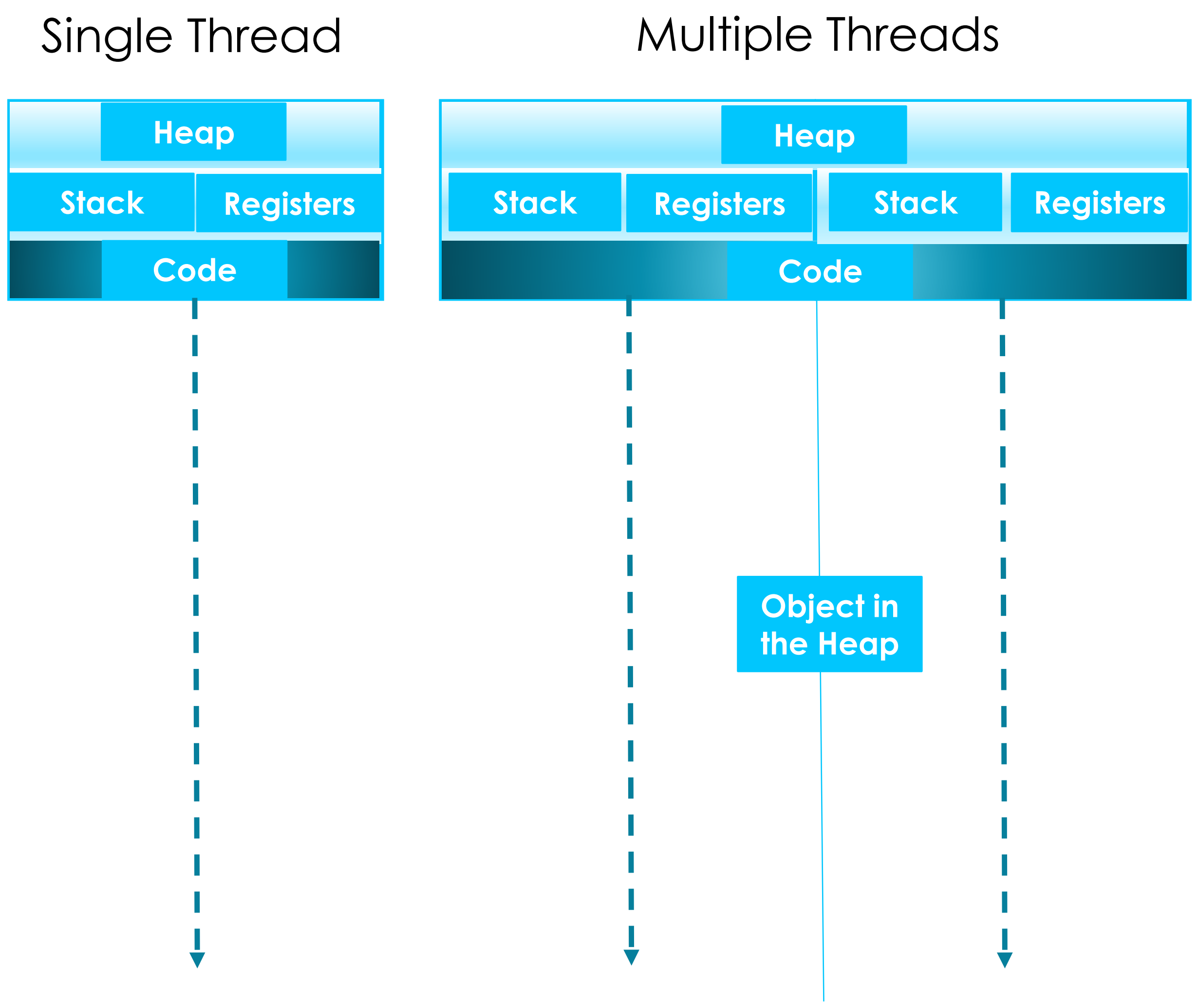
- You need to wait for something to happen without wasting precious resources – forget about callbacks or reactive stream chaining – just block!
- Write straightforward, boring code.
- Goroutines preserve all the benefits threads give only the runtime cost in footprint and performance is gone!

Concurrency Approaches: Processes



Threads

- There can be many threads in the same process
- The threads can access the shared memory
- This means that the global objects can be accessed by all threads
- Provided by the OS
- Cheaper to create than processes
- Some languages expose them directly other hide them behind a level of abstraction



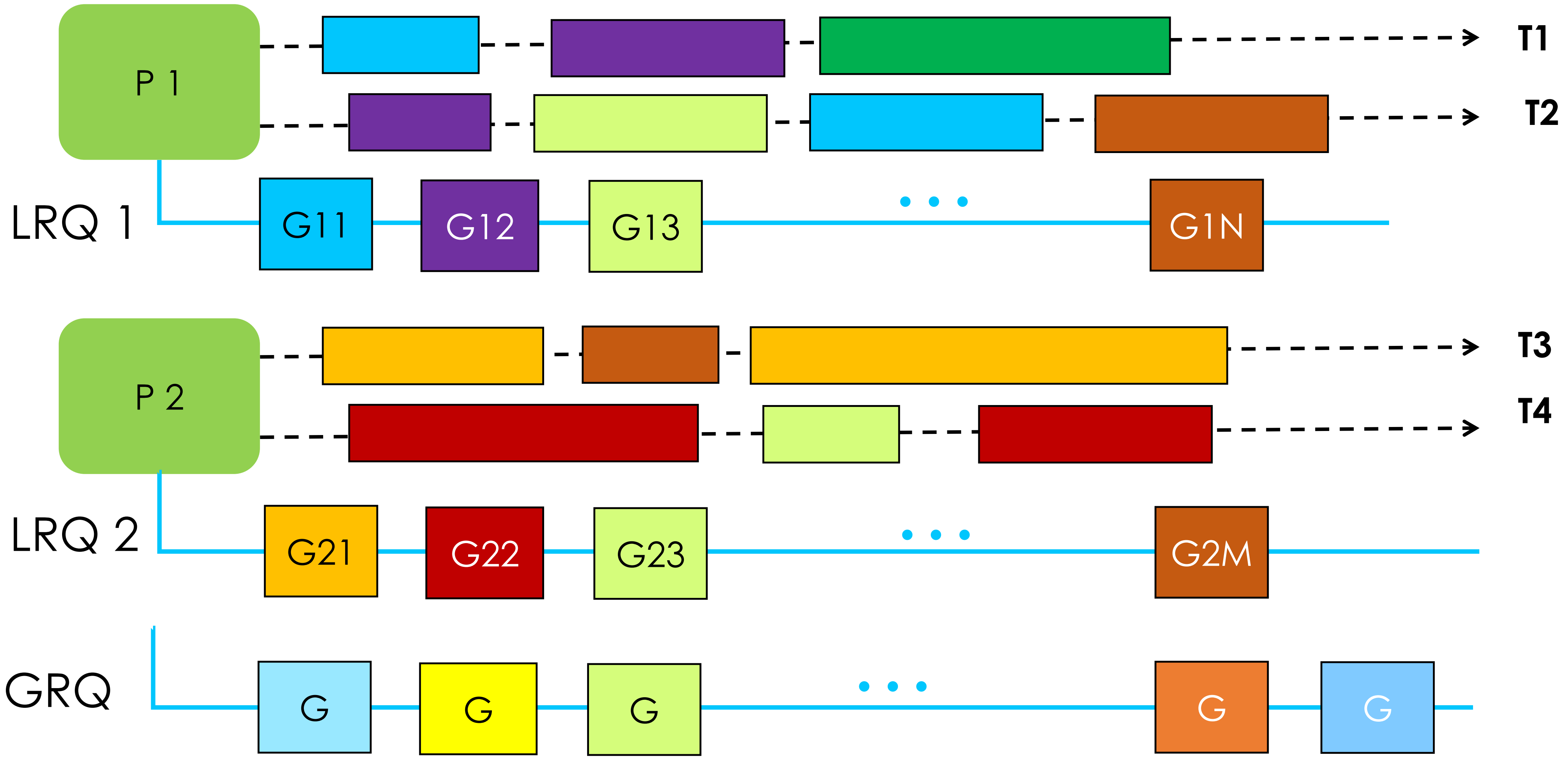
Goroutines

- They are executed independently from the main function
- Can be hundreds of thousands of them – the initial stack can be as low as 2KB
- The goroutines stack can grow dynamically as needed
- In Golang there is a smart scheduler that can map goroutines to OS threads
- Goroutines follow the idea of Communicating sequential processes of Hoare

Goroutines Scheduling

GRQ – Global Run Queue **LRQ – Local Run Queue**

P – Processor T – Thread G – Goroutine



But why should we bother: concurrency problems

If we access the same memory from two threads/ goroutines, than we have a race! Lets see this pseudo-code:

```
int i = 0

thread1 { i++ }
thread2 { i++ }

wait { thread1 } { thread2 }
print i
```

What will be the result of the computation?

Critical Sections

- We provide Mutual Exclusion between different threads accessing the same resource concurrently
- There are many ways to implement Mutual Exclusion
- In Golang there are sync.Mutex, sync.RWMutex, atomic operations, semaphores, error groups (structured concurrency), concurrent hash maps, etc.
- And the message passing using Channels of course :)

Share by communicating vs. Communicate by sharing

We can use Mutex / Atomic primitives for mutual exclusion between goroutines as in other languages, but in most cases it happens to be simpler and more obvious to handle data to other goroutines using channels.

Example Problem: Concurrent Counter Update

```
import ("fmt")
const NUM_GOROUTINES = 100
var n int64 = 0
func increment(id int, complete chan<- struct{}) {
        for i := 0; i < 10000; i++ { n++ }
        complete <- struct{} {}
}

func main() {
        complete := make(chan struct{})
        for i := 0; i < NUM_GOROUTINES; i++ {
                go increment(i, complete)
        }
        for i := 0; i < NUM_GOROUTINES; i++ {
                <-complete
        }
        fmt.Println(n)
}
```

Concurrent Counter Update Using WaitGroup

```
import ("fmt"; "sync")
const NUM_GOROUTINES = 100
var n int64 = 0
func inrement(id int, wg *sync.WaitGroup) {
        for i := 0; i < 10000; i++ {
                n++
        }
        wg.Done()
}
func main() {
        var wg sync.WaitGroup
        for i := 0; i < NUM_GOROUTINES; i++ {
                wg.Add(1)
                go inrement(i, &wg)
        }
        wg.Wait()
        fmt.Println(n)
}
```

Concurrent Counter Update Solution: Atomic

```
import ("fmt"; "sync"; "sync/atomic")
var n int64 = 0
func increment(id int, wg *sync.WaitGroup) {
        for i := 0; i < 10000; i++ {
                atomic.AddInt64(&n, 1)
        }
        wg.Done()
}
func main() {
        var wg sync.WaitGroup
        for i := 0; i < 100; i++ {
                wg.Add(1)
                go increment(i, &wg)
        }
        wg.Wait()
        fmt.Println(n)
}
```

Concurrent Counter Update Solution: Mutex

```
var n int64 = 0
var mu sync.Mutex
func increment(id int, wg *sync.WaitGroup) {
        for i := 0; i < 10000; i++ {
                mu.Lock()
                n++
                mu.Unlock()
        }
        wg.Done()
}
func main() {
        var wg sync.WaitGroup
        for i := 0; i < 100; i++ {
                wg.Add(1)
                go increment(i, &wg)
        }
        wg.Wait()
        fmt.Println(n)
}
```

Quiz 1: Data Race – Which 1 character should you change?

```
type RPC struct {
	result int
	done   chan struct{}
}
func (rpc *RPC) compute() {
	time.Sleep(time.Second) // strenuous computation intensifies
	rpc.result = 42
	close(rpc.done)
}
func (RPC) version() int {
	return 1 // never going to need to change this
}
func main() {
	rpc := &RPC{done: make(chan struct{})}
	go rpc.compute()        // kick off computation in the background
	version := rpc.version() // grab some other information while we're waiting
	<-rpc.done              // wait for computation to finish
	result := rpc.result
	fmt.Printf("RPC computation complete, result: %d, version: %d\n", result, version)
}
```

Deadlock

Deadlock happens if a group of goroutines can not continue to progress, because they are mutually waiting each other for something (e.g. to free a resource, or to write/read from channel). For example:

```
func main() {
		ch := make(chan int)
		ch <- 1
		fmt.Println(<-ch)
}
```

```
fatal error: all goroutines are asleep - deadlock!
goroutine 1 [chan send]:
main.main()
		D:/CourseGO/git/coursegopro/08-goroutines-channels/deadlock-simple/deadlock-
simple.go:7 +0x37
```

Deadlock

- A goroutine can get stuck
  - either because it's waiting for a channel or
  - because it is waiting for one of the locks in the sync package
- Common reasons are that
  - no other goroutine has access to the channel or the lock,
  - a group of goroutines are waiting for each other and none of them is able to proceed
- Currently Go only detects when the program as a whole freezes, NOT when a subset of goroutines get stuck.
- With channels it's often easy to figure out what caused a deadlock. Programs that make heavy use of mutexes can, on the other hand, be notoriously difficult to debug.

Nil channels

- Generally NOT vary useful!
- Nil channels have a few interesting behaviors:
  - Sends to them block forever
  - Receives from them block forever
  - Closing them leads to panic

```
func main() {
        var c chan string
        c <- "message" // Deadlock
}
func main() {
        var c chan string
        fmt.Println(<-c) // Deadlock
}
```

But sometimes NIL channels can be useful:

```
func sendTo(c chan<- int, iter int) {
   for i := 0; i <= iter; i++ {
      c <- i
   }

   close(c)
}
```

```
func main() {
   ch1 := make(chan int)
   ch2 := make(chan int)
   go sendTo(ch1, 5)
   go sendTo(ch2, 10)
   for {
      select {
      case x, ok := <-ch1:
         if ok {
            fmt.Println("Channel 1 sent", x)
         } else { ch1 = nil  }
      case y, ok:= <-ch2:
         if ok {
            fmt.Println("Channel 2 sent", y)
         } else {  ch2 = nil  }
      }
      if ch1 == nil && ch2 == nil { break }
   }
   fmt.Println("Program finished normally.")
}
```

Semaphors: limiting the number of concurrent goroutines

```
func DoWork(n int, jobs <-chan struct{}) {
        fmt.Println("doing", n)
        time.Sleep(500 * time.Millisecond)
        fmt.Println("finished", n)
        <-jobs // release the token
}
func main() {
        // concurrentJobs is a buffered channel implemeting semaphore that blocks
        // if more than 20 goroutines are started at once
        var concurrentJobs = make(chan struct{}, 5)
        for i := 0; i < 30; i++ {
                concurrentJobs <- struct{}{} // acquire a  token
                go DoWork(i, concurrentJobs)
                fmt.Printf("Current number of goroutines: %d\n", runtime.NumGoroutine())
        }
}
```

sync.WaitGroup

```
func main() {
	var wg sync.WaitGroup
	var urls = []string{
		"http://www.golang.org/",
		"http://www.google.com/",
		"http://www.somestupidname.com/",
	}
	for _, url := range urls {
		wg.Add(1) // Increment the WaitGroup counter.
		go func(url string) {  // Launch a goroutine to fetch the URL.
			defer wg.Done() // Decrement the counter when the goroutine completes.
			http.Get(url) // Fetch the URL
		}(url)
	}
	// Wait for all HTTP fetches to complete.
	wg.Wait()
}
```

sync.Once

```
type Worker struct {
        once sync.Once
}

func (w *Worker) Run() {
        w.once.Do(func() {
                fmt.Println("this will run only once")     // printed only once!
        })
}

func main() {
        w := Worker{}
        w.Run()
        w.Run()
        w.Run()
}
```

Mutexes: sync.Mutex

```
func main() {
	var wg sync.WaitGroup
	var mu sync.Mutex
	wg.Add(2)
	go func() {
		mu.Lock()
		time.Sleep(3 * time.Second)
		fmt.Println("go routine 1 releasing lock after 3s:", time.Now())
		mu.Unlock()
		wg.Done()
	}()
	go func() {
		fmt.Println("go routine 2 trying to acquire lock:", time.Now())
		mu.Lock()
		fmt.Println("go routine 2 acquired lock after 3s:", time.Now())
		mu.Unlock()
		wg.Done()
	}()
	wg.Wait()
}
```

Mutexes: sync.RWMutex

```
type RWMutex struct {
	mu    sync.RWMutex
	value int
}

func (m *RWMutex) Store(value int) {
	m.mu.Lock()
	defer m.mu.Unlock()
	m.value = value
}

func (m *RWMutex) Load() int {
	m.mu.RLock()
	time.Sleep(100 * time.Nanosecond)
	defer m.mu.RUnlock()
	return m.value
}
```

Mutexes: sync.RWMutex – benchmarking I

```
func BenchmarkMain(b *testing.B) {
	b.Run("BasicMutex-6-1", func(b *testing.B) {
		BasicMutex_Load(b, 6, 1)
	})
	b.Run("BasicMutex-1-6", func(b *testing.B) {
		BasicMutex_Load(b, 1, 6)
	})
	b.Run("BasicMutex-0-6", func(b *testing.B) {
		BasicMutex_Load(b, 0, 6)
	})
	b.Run("RWMutex-6-1", func(b *testing.B) {
		RWMutex_Load(b, 6, 1)
	})
	b.Run("RWMutex-1-6", func(b *testing.B) {
		RWMutex_Load(b, 1, 6)
	})
	b.Run("RWMutex-0-6", func(b *testing.B) {
		RWMutex_Load(b, 0, 6)
	})
}
```

Mutexes: sync.RWMutex – benchmarking II

```
func RWMutex_Load(b *testing.B, nLoads int, nStores int) {
        mu := RWMutex{}
        mu.Store(10)
        for i := 0; i < b.N; i++ {
                for j := 0; j < nLoads; j++ {
                        go mu.Load()
                }
                for j := 0; j < nStores; j++ {
                        go mu.Store(i)
                }
        }
}
```

Mutexes: sync.RWMutex – benchmarking III

```
goarch: amd64
cpu: Intel(R) Core(TM) i7-8700K CPU @ 3.70GHz
BenchmarkMain
BenchmarkMain/BasicMutex-3-0
BenchmarkMain/BasicMutex-3-0-12          109976     11434 ns/op
BenchmarkMain/BasicMutex-6-1
BenchmarkMain/BasicMutex-6-1-12           46353     25159 ns/op
BenchmarkMain/BasicMutex-3-3
BenchmarkMain/BasicMutex-3-3-12           58431     21415 ns/op
BenchmarkMain/BasicMutex-1-6
BenchmarkMain/BasicMutex-1-6-12           51291     24679 ns/op
BenchmarkMain/BasicMutex-0-6
BenchmarkMain/BasicMutex-0-6-12         1000000      1164 ns/op
BenchmarkMain/RWMutex-3-0
BenchmarkMain/RWMutex-3-0-12            1220876      1002 ns/op
BenchmarkMain/RWMutex-6-1
BenchmarkMain/RWMutex-6-1-12             397364      4849 ns/op
BenchmarkMain/RWMutex-3-3
BenchmarkMain/RWMutex-3-3-12             394406     18694 ns/op
BenchmarkMain/RWMutex-1-6
BenchmarkMain/RWMutex-1-6-12             295240     31376 ns/op
BenchmarkMain/RWMutex-0-6
BenchmarkMain/RWMutex-0-6-12             851794      1462 ns/op
PASS
```

Patterns: Worker Pool - I

```
jobs := make(chan job)
results := make(chan result)
ctx, cancel := context.WithTimeout(context.Background(), 5*time.Second)
runtime.GOMAXPROCS(runtime.NumCPU())
for i := 0; i < runtime.NumCPU(); i++ {
        go func(ctx context.Context, workerId int, jobs <-chan job, results chan<- result) {
                defer close(results)
                for job := range jobs {
                        var r result
                        if job.input > 0 {
                            r = result{job.id, workerId, process(job.input), nil}
                        } else {
                            r = result{job.id, workerId, 0, fmt.Errorf("invalid argument error: %f", job.input)}
                        }
                        select {
                        case results <- r:
                        case <-ctx.Done():
                                return
                        }
                }
        }(ctx, i, jobs, results)
}
```

Patterns: Worker Pool - II

```go
//  send out jobs
go func() {
		for i := 0; i < numJobs; i++ {
				jobs <- job{i, math.Abs(float64(i-14) * math.Pi)}
		}
		close(jobs)
}()

//  do something with results
for r := range results {
		if r.err != nil {
				// do something with error
				log.Printf("Error executing Job %d: %v\n", r.jobId, r.err.Error())
				cancel()
		} else {
				fmt.Printf("Job %d executed by worker %d -> Result: %v\n", r.jobId, r.workerId, r.output)
		}
}
```

Concurrent Datastructures: ConcurrentHashSet I

```
import "sync"
type ConcurrentHashSet struct {
            mu     sync.Mutex
            values map[string]struct{}
}
func New() *ConcurrentHashSet {
            return &ConcurrentHashSet{values: make(map[string]struct{})}
}
func (s *ConcurrentHashSet) Add(val string) {
            s.mu.Lock()
            s.values[val] = struct{}{}
            s.mu.Unlock()
}
func (s *ConcurrentHashSet) IsMember(val string) bool {
            s.mu.Lock()
            _, ok := s.values[val]
            s.mu.Unlock()
            return ok
}
```

Concurrent Datastructures: ConcurrentHashSet II

```
func (s *ConcurrentHashSet) Remove(val string) {
		s.mu.Lock()
		delete(s.values, val)
		s.mu.Unlock()
}
```

Using ConcurrentHashSet – Prevent Duplicate URLs

```
func DoWork(url string,
        ctx context.Context,
        visited *concurrentset.ConcurrentHashSet,
        numUrls *uint64,
        urls chan<- string,
        jobs *semaphor.Semaphor,
        ) {
        defer jobs.Release() // release the token
        for i := 0; i < MAX_GOROUTINES; i++ {
                newUrl := fmt.Sprintf("%s/%d", url, i)
                num := atomic.AddUint64(numUrls, 1)
                if visited.IsMember(newUrl) || atomic.LoadUint64(numUrls) >= MAX_URLS {
                        continue
                }
                select {
                case urls <- newUrl:
                case <-ctx.Done():
                        return
                }
        }
}
```

sync.Condititon - I

The call to Wait does the following under the hood
1. Calls Unlock() on the condition Locker
2. Notifies the list wait
3. Calls Lock() on the condition Locker

The Cond type besides the Locker also has access to 2 important methods:
1. Signal - wakes up 1 go routine waiting on a condition (rendezvous point)
2. Broadcast - wakes up all go routines waiting on a condition (rendezvous point)

```
func condition(n *uint64) bool {
		return atomic.LoadUint64(n) % 2 == 0
}
```

sync.Condititon - II

```go
func main() {
        var wg sync.WaitGroup
        cond := sync.NewCond(&sync.Mutex{})
        var n uint64 = 0
        wg.Add(2)
        go func() {
                defer wg.Done()
                for i:= 0; i < 10; i++ {
                        cond.L.Lock()
                        for !condition(&n) {
                                cond.Wait()
                        }
                        atomic.AddUint64(&n, 1)   // ... make use of condition ...
                        fmt.Println("go routine 1")
                        cond.L.Unlock()
                        cond.Signal()  // OR cond.Broadcast()
                }
        }()
```

sync.Condititon - III

```
        go func() {
                defer wg.Done()
                for i:= 0; i < 10; i++ {
                        cond.L.Lock()
                        for condition(&n) {
                                cond.Wait()
                        }
                        // ... make use of condition ...
                        atomic.AddUint64(&n, 1)
                        fmt.Println("go routine 2")
                        cond.L.Unlock()
                        cond.Signal()   // OR cond.Broadcast()
                }
        }()
        wg.Wait()
}
```

atomic.Value - I

```
import ("sync/atomic"; "time")
func loadConfig() map[string]string { return make(map[string]string) }
func requests() chan int { return make(chan int) }
func main() {
        var config atomic.Value // holds current server configuration
        // Create initial config value and store into config.
        config.Store(loadConfig())
        for i := 0; i < 10; i++ {
                go func() {
                // Reload config every 10 seconds and update config value with the new version.
                        for {
                                time.Sleep(10 * time.Millisecond)
                                config.Store(loadConfig())
                        }
                }()
        }
```

atomic.Value - II

```
        // Create worker goroutines that handle incoming requests using the latest config value.
        for i := 0; i < 100; i++ {
                go func() {
                        for r := range requests() {
                                c := config.Load()
                                // Handle request r using config c.
                                _, _ = r, c
                        }
                }()
        }
}
```

Golang Concurrent Programming Advices - I

https://yourbasic.org/golang/concurrent-programming/ [CC BY 3.0 license.]

- **Goroutines are lightweight threads** - A goroutine is a lightweight thread of execution. All goroutines in a single program share the same address space.
- **Channels offer synchronized communication** - A channel is a mechanism for two goroutines to synchronize execution and communicate by passing values.
- **Select waits on a group of channels** - A select statement allows you to wait for multiple send or receive operations simultaneously.
- **Data races explained** - A data race is easily introduced by mistake and can lead to situations that are very hard to debug. This article explains how to avoid this headache.
- **How to detect data races** - By starting your application with the '-race' option, the Go runtime might be able to detect and inform you about data races.
- **How to debug deadlocks** - The Go runtime can often detect when a program freezes because of a deadlock. This article explains how to debug and solve such issues.

Golang Concurrent Programming Advices - II

https://yourbasic.org/golang/concurrent-programming/ [CC BY 3.0 license.]

Waiting for goroutines - A sync.WaitGroup waits for a group of goroutines to finish.

Broadcast a signal on a channel - When you read from a closed channel, you receive a zero value. This can be used to broadcast a signal to several goroutines on a single channel.

How to kill a goroutine - One goroutine can't forcibly stop another. To make a goroutine stoppable, let it listen for a stop signal on a channel.

Timer and Ticker: events in the future - Timers and Tickers are used to wait for, repeat, and cancel events in the future.

Mutual exclusion lock (mutex) - A sync.Mutex is used to synchronize data by explicit locking in Go.

3 rules for efficient parallel computation - To efficiently schedule parallel computation on separate CPUs is more of an art than a science. This article gives some rules of thumb.

Thank's for Your Attention!

Trayan Iliev

IPT – Intellectual Products & Technologies

http://iproduct.org/

http://robolearn.org/

https://github.com/iproduct

https://twitter.com/trayaniliev

https://www.facebook.com/IPT.EACAD