

cs5460/6460: Operating Systems

Lecture 11: Context switch

Anton Burtsev
April, 2023

A bit of recap...

When OS context switches between processes?

When OS context switches between processes?

- Timer interrupt preempts the current process
- A process enters the kernel with a system call and has to wait on some resource
 - E.g., write to a pipe, but the pipe is full
- The process voluntarily yields CPU with the `yield()` system call

Lets look at timer interrupt

```
3351 trap(struct trapframe *tf)
3352 {
...
3363     switch(tf->trapno){
3364     case T_IRQ0 + IRQ_TIMER:
3365         if(cpu->id == 0){
3366             acquire(&tickslock);
3367             ticks++;
3368             wakeup(&ticks);
3369             release(&tickslock);
3370         }
3372     break;
...
3423     if(proc && proc->state == RUNNING
3424         && tf->trapno == T_IRQ0+IRQ_TIMER)
3424         yield();
```

trap()

```
3351 trap(struct trapframe *tf)
3352 {
...
3363     switch(tf->trapno){
3364     case T_IRQ0 + IRQ_TIMER:
3365         if(cpu->id == 0){
3366             acquire(&tickslock);
3367             ticks++;
3368             wakeup(&ticks);
3369             release(&tickslock);
3370         }
3372     break;
...
3423     if(proc && proc->state == RUNNING
3424         && tf->trapno == T_IRQ0+IRQ_TIMER)
3424         yield();
```

trap()

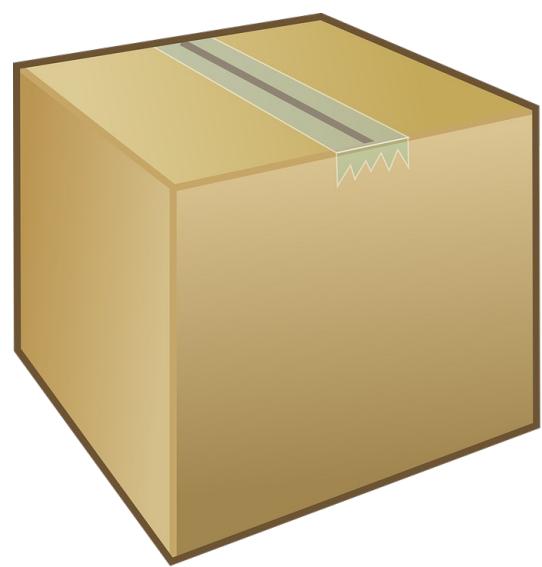
Invoke the scheduler

```
2777 yield(void)  
2778 {  
2779     acquire(&ptable.lock);  
2780     proc->state = RUNNABLE;  
2781     sched();  
2782     release(&ptable.lock);  
2783 }
```

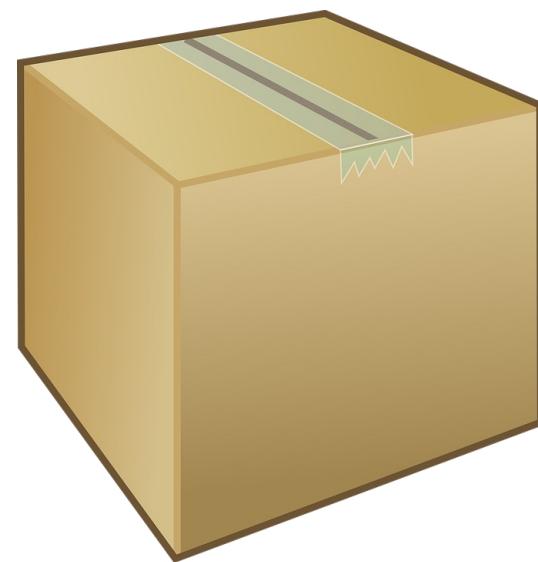
Start the context switch

```
2758 sched(void)
2759 {
...
2771     swtch(&proc->context,
...
2773 }
```

But what do you think needs to happen inside
`switch()`?



Process 1

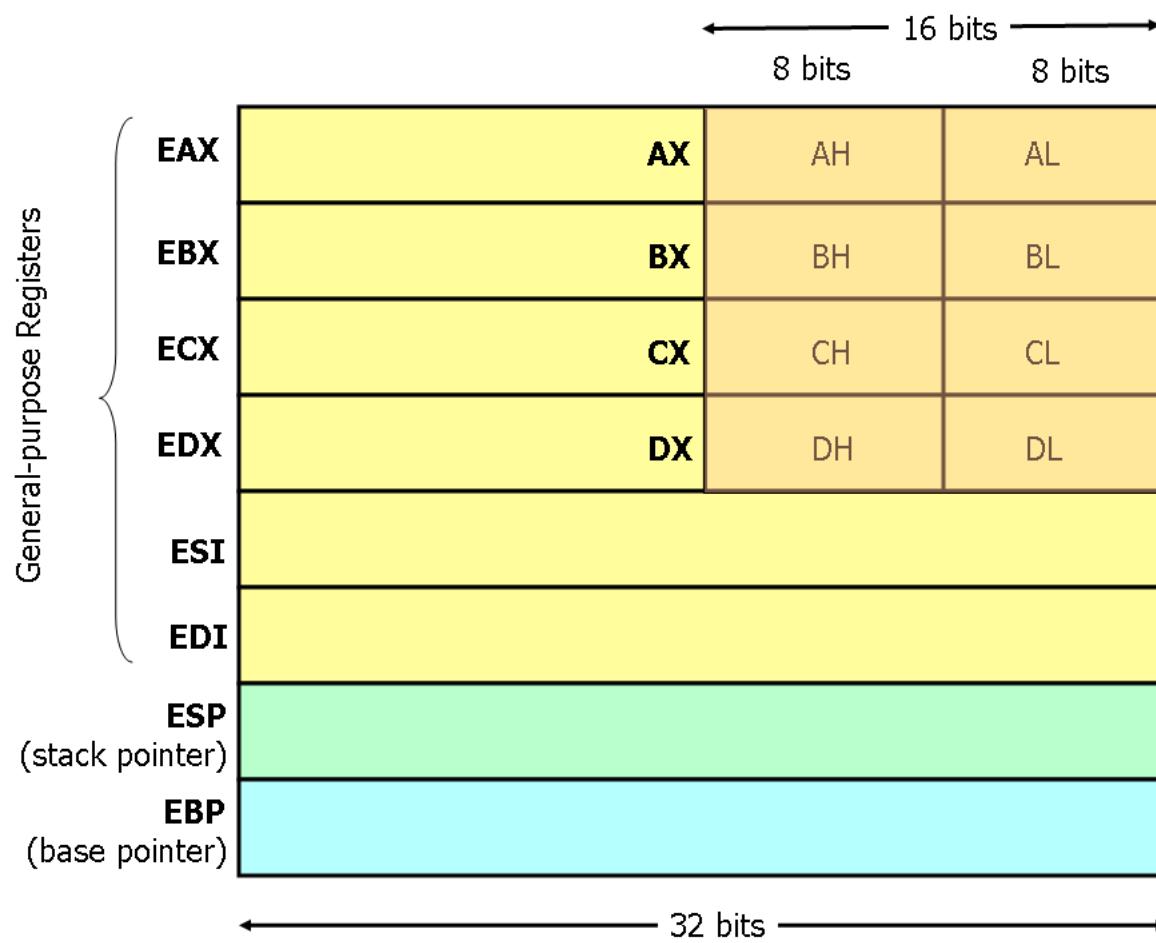


Process 2

What should be in the box?
i.e., what is the state of the process?

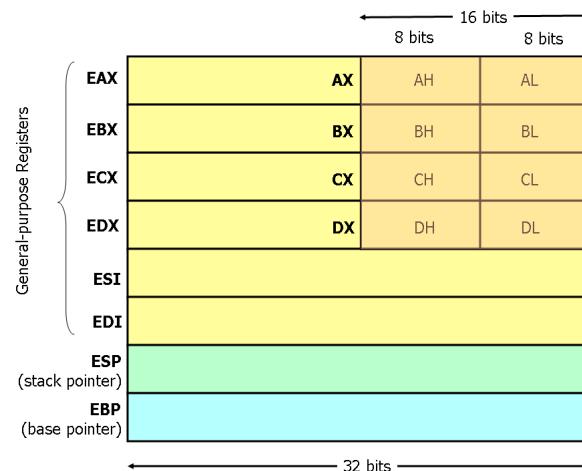
```
2103 struct proc {  
2104     uint sz; // Size of process memory (bytes)  
2105     pde_t* pgdir; // Page table  
2106     char *kstack; // Bottom of kernel stack for this process  
2107     enum procstate state; // Process state  
2108     volatile int pid; // Process ID  
2109     struct proc *parent; // Parent process  
2110     struct trapframe *tf; // Trap frame  
2111     struct context *context; // swtch() here to run  
2112     void *chan; // If non-zero, sleeping on chan  
2113     int killed; // If non-zero, have been killed  
2114     struct file *ofile[NFILE]; // Open files  
2115     struct inode *cwd; // Current directory  
2116     char name[16]; // Process name (debugging)  
2117 };
```

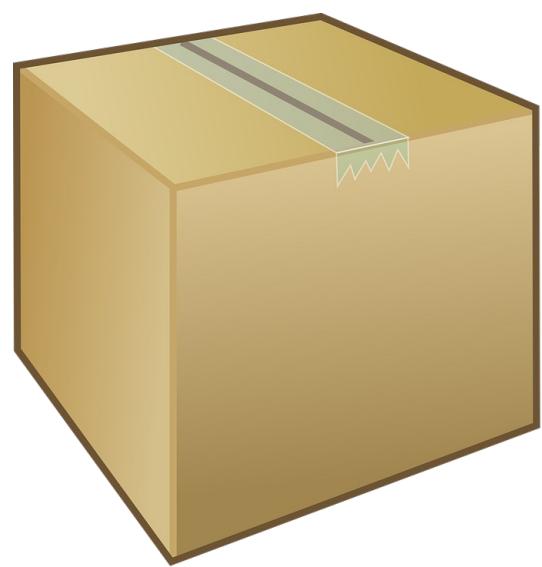
What about general registers?



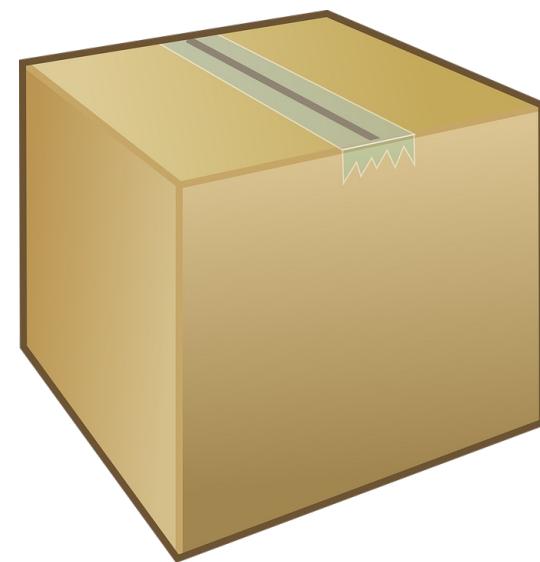
Save them on the stack

```
2093 struct context {  
2094     uint edi;  
2095     uint esi;  
2096     uint ebx;  
2097     uint ebp;  
2098     uint eip;  
2099 };
```



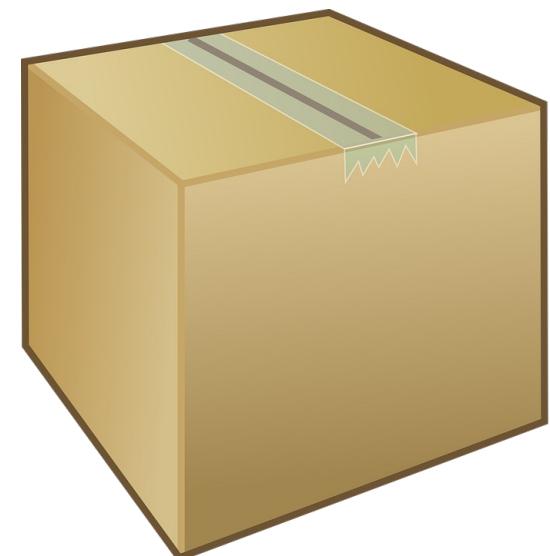


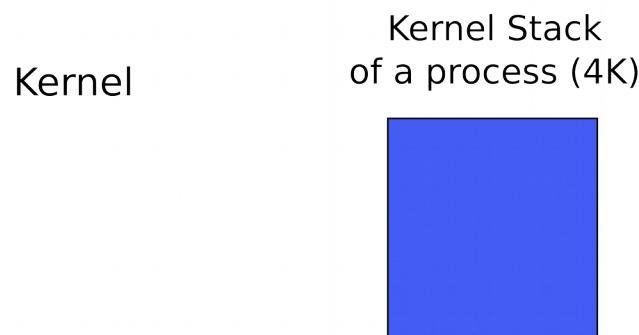
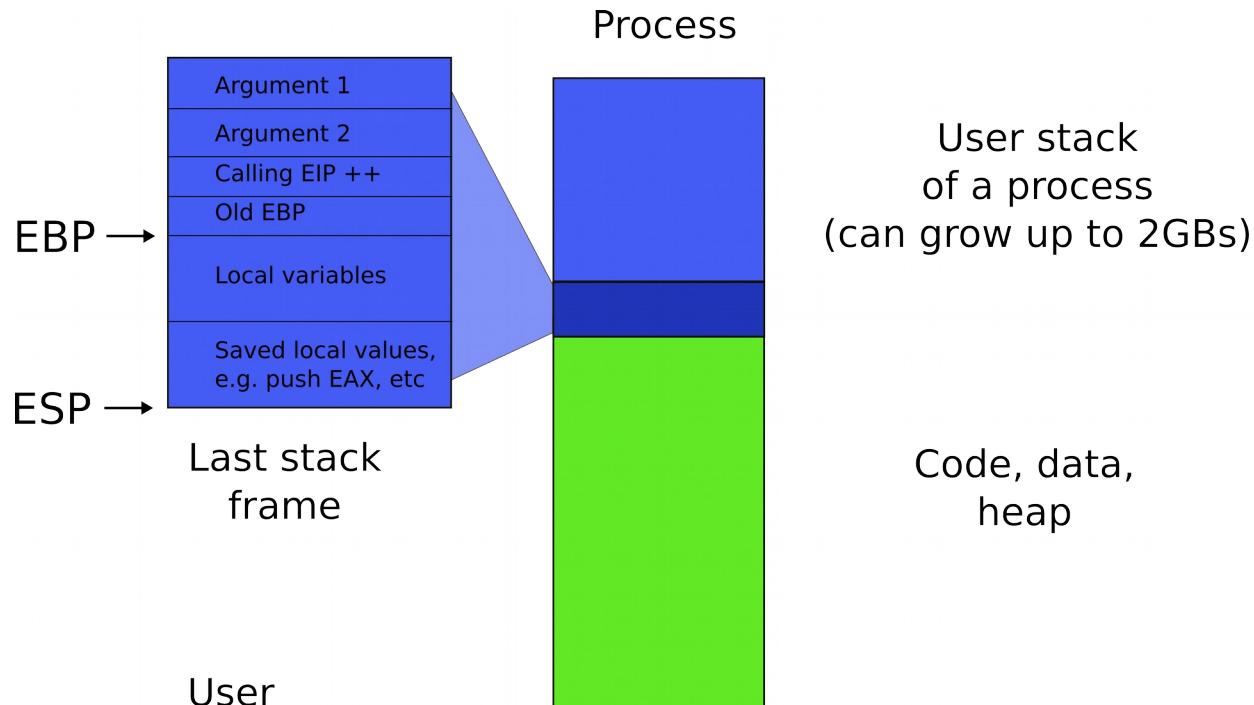
Process 1



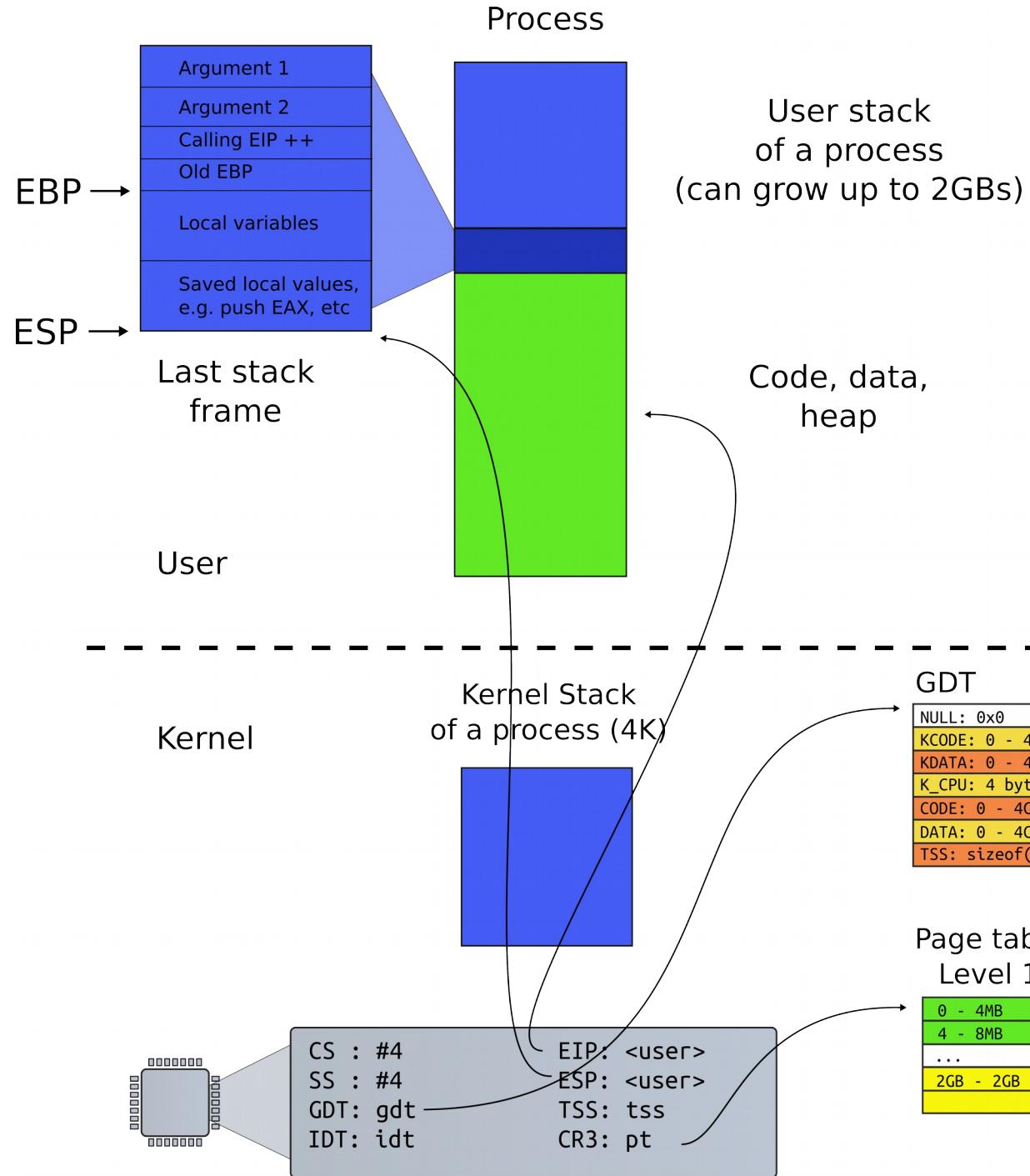
Process 2

Back to the timer interrupt path
(keep track of what happens to the stack, lets see
how everything gets packed in the “box”!)



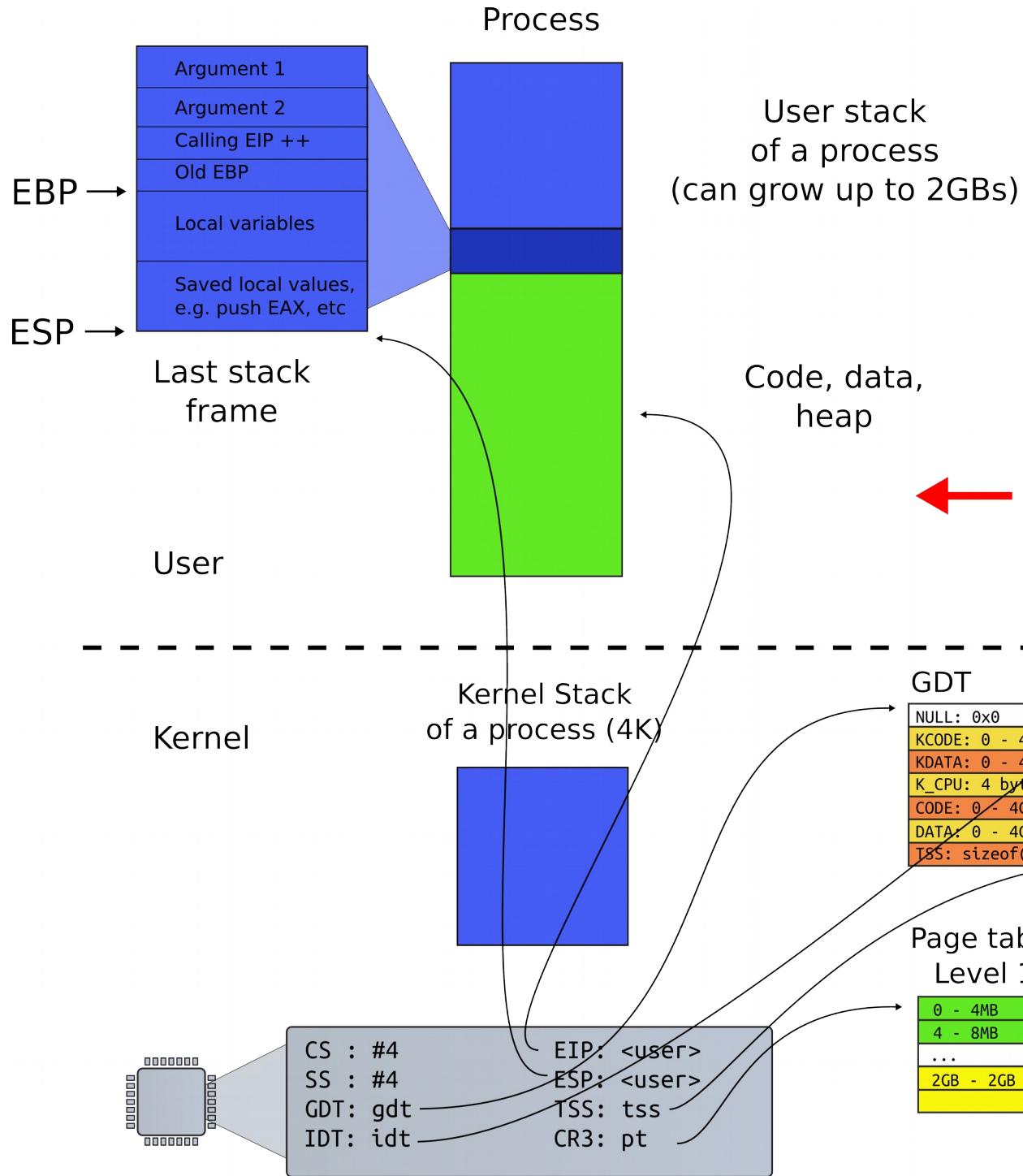


- User mode
- Two stacks
 - Kernel and user
 - Kernel stack is empty

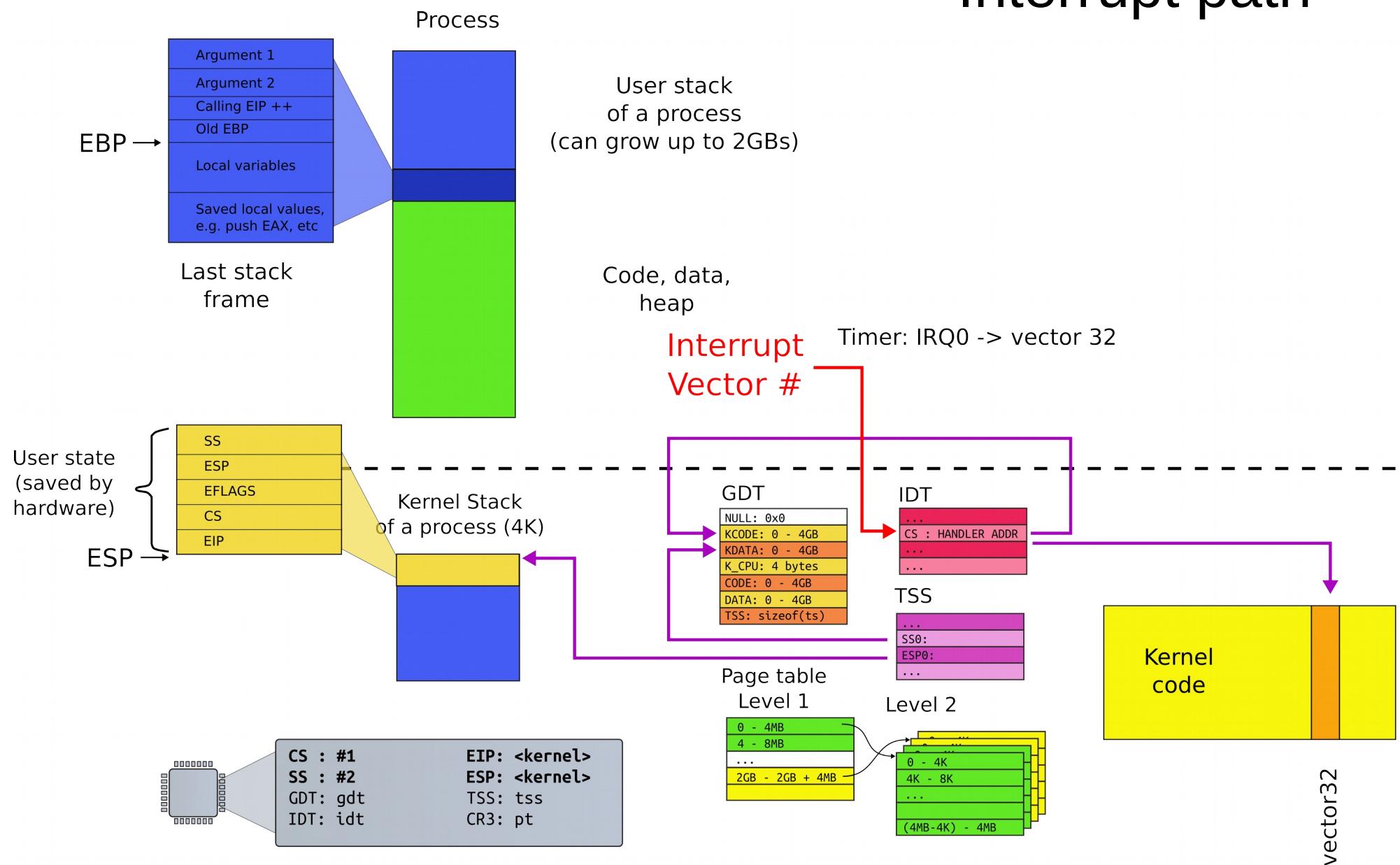


- Page table
- GDT

Timer interrupt



Interrupt path



Where does IDT (entry 32) point to?

vector32:

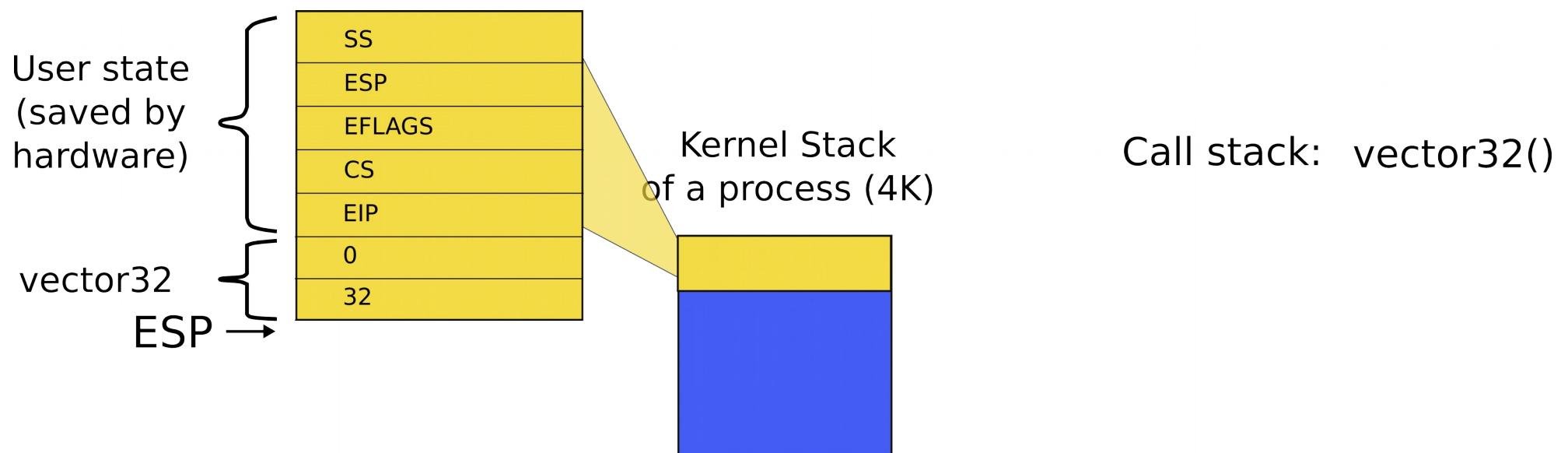
```
pushl $0      // error code
```

```
pushl $32      // vector #
```

```
jmp alltraps
```

- Automatically generated
- From vectors.pl
 - vector.S

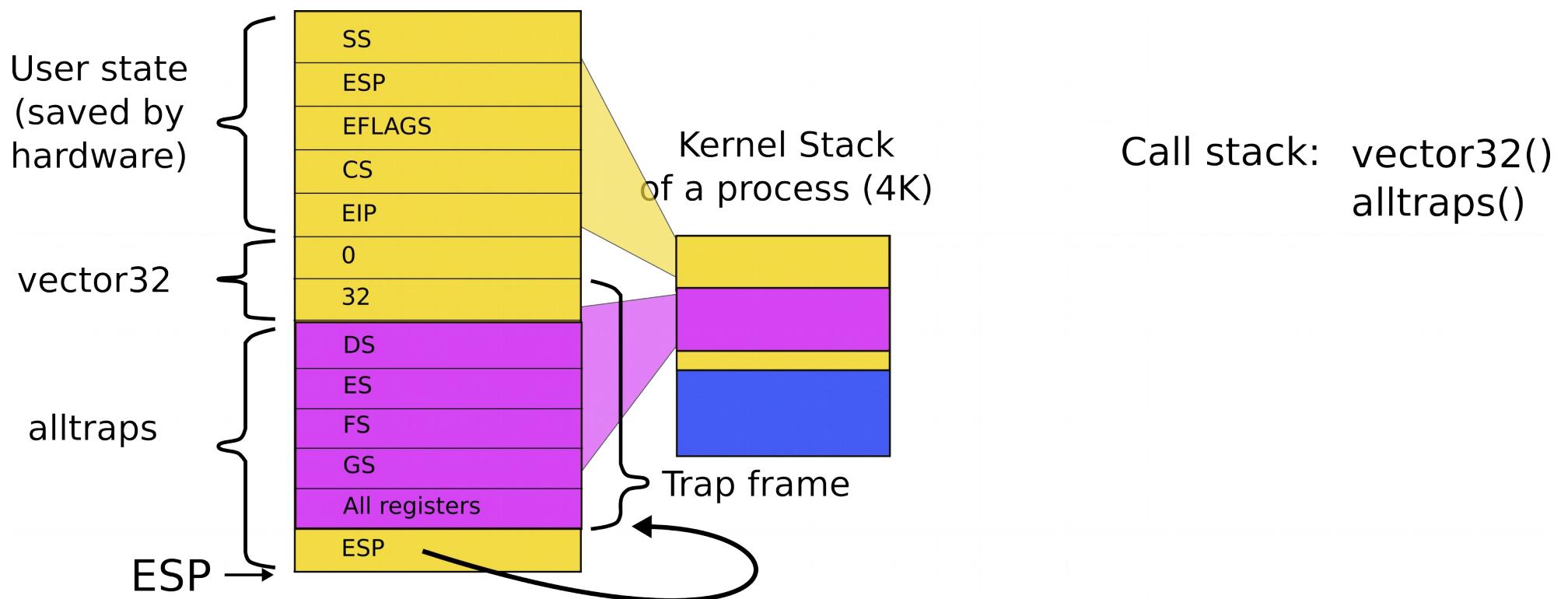
Kernel stack after interrupt



alltraps()

```
3254 alltraps:  
3255     # Build trap frame.  
3256     pushl %ds  
3257     pushl %es  
3258     pushl %fs  
3259     pushl %gs  
3260     pushal  
3261  
3262     # Set up data and per-cpu segments.  
3263     movw $(SEG_KDATA<<3), %ax  
3264     movw %ax, %ds  
3265     movw %ax, %es  
3266     movw $(SEG_KCPU<<3), %ax  
3267     movw %ax, %fs  
3268     movw %ax, %gs  
3269  
3270     # Call trap(tf), where tf=%esp  
3271     pushl %esp  
3272     call trap
```

Kernel stack after interrupt



```
3254 alltraps:  
3255 # Build trap frame.  
3256 pushl %ds  
3257 pushl %es  
3258 pushl %fs  
3259 pushl %gs  
3260 pushal  
3261  
3262 # Set up data and per-cpu segments.  
3263 movw $(SEG_KDATA<<3), %ax  
3264 movw %ax, %ds  
3265 movw %ax, %es  
3266 movw $(SEG_KCPU<<3), %ax  
3267 movw %ax, %fs  
3268 movw %ax, %gs  
3269  
3270 # Call trap(tf), where tf=%esp  
3271 pushl %esp  
3272 call trap
```

alltraps()

```
3351 trap(struct trapframe *tf)
3352 {
...
3363     switch(tf->trapno){
3364     case T_IRQ0 + IRQ_TIMER:
3365         if(cpu->id == 0){
3366             acquire(&tickslock);
3367             ticks++;
3368             wakeup(&ticks);
3369             release(&tickslock);
3370         }
3372     break;
...
3423     if(proc && proc->state == RUNNING
3424         && tf->trapno == T_IRQ0+IRQ_TIMER)
3424         yield();
```

trap()

```
3351 trap(struct trapframe *tf)
3352 {
...
3363     switch(tf->trapno){
3364     case T_IRQ0 + IRQ_TIMER:
3365         if(cpu->id == 0){
3366             acquire(&tickslock);
3367             ticks++;
3368             wakeup(&ticks);
3369             release(&tickslock);
3370         }
3372     break;
...
3423     if(proc && proc->state == RUNNING
3424         && tf->trapno == T_IRQ0+IRQ_TIMER)
3424         yield();
```

trap()

Invoke the scheduler

```
2777 yield(void)  
2778 {  
2779     acquire(&ptable.lock);  
2780     proc->state = RUNNABLE;  
2781     sched();  
2782     release(&ptable.lock);  
2783 }
```

Start the context switch

2758 sched(void)

2759 {

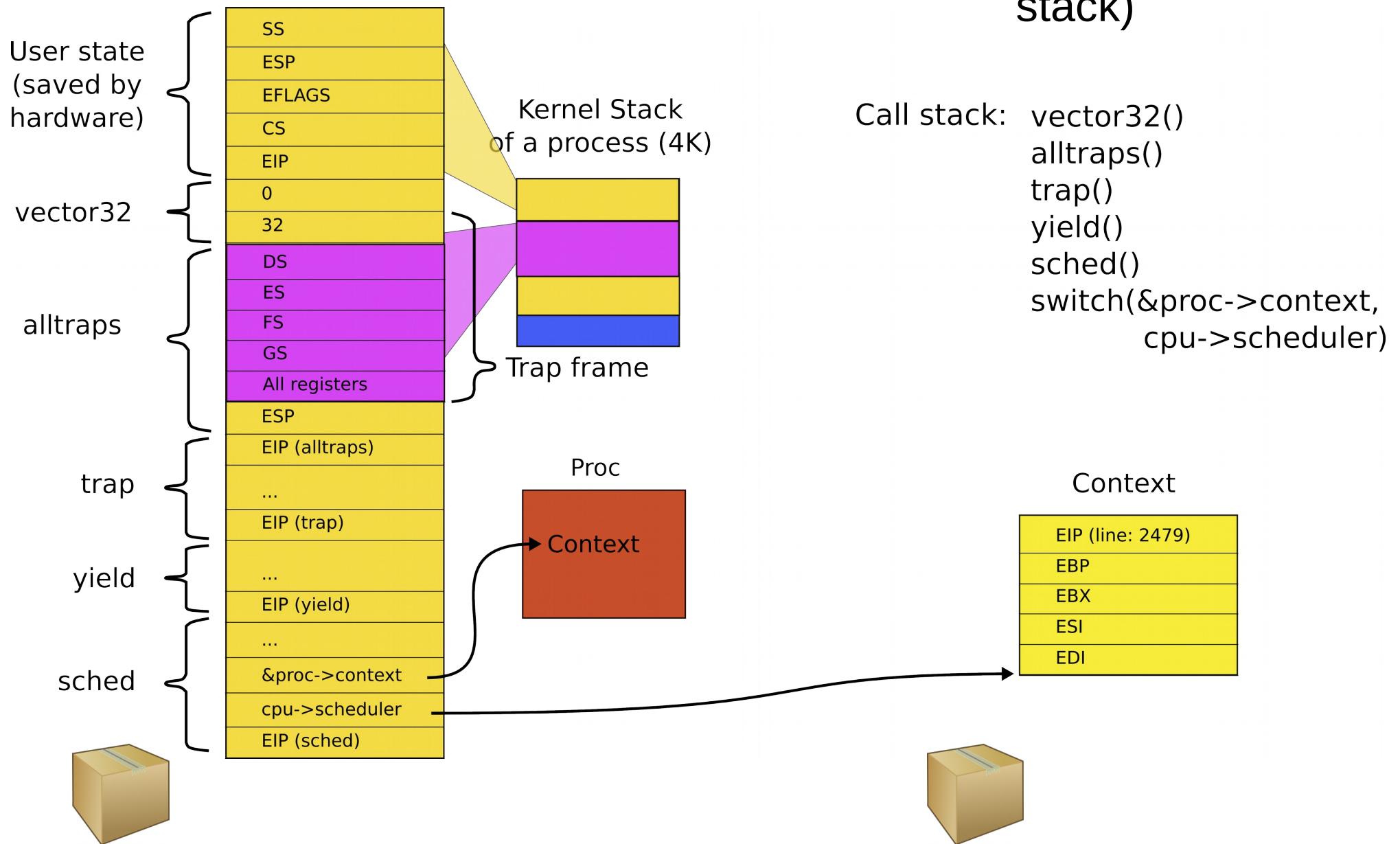
• • •

```
2771     swtch(&proc->context,  
                  cpu->scheduler);
```

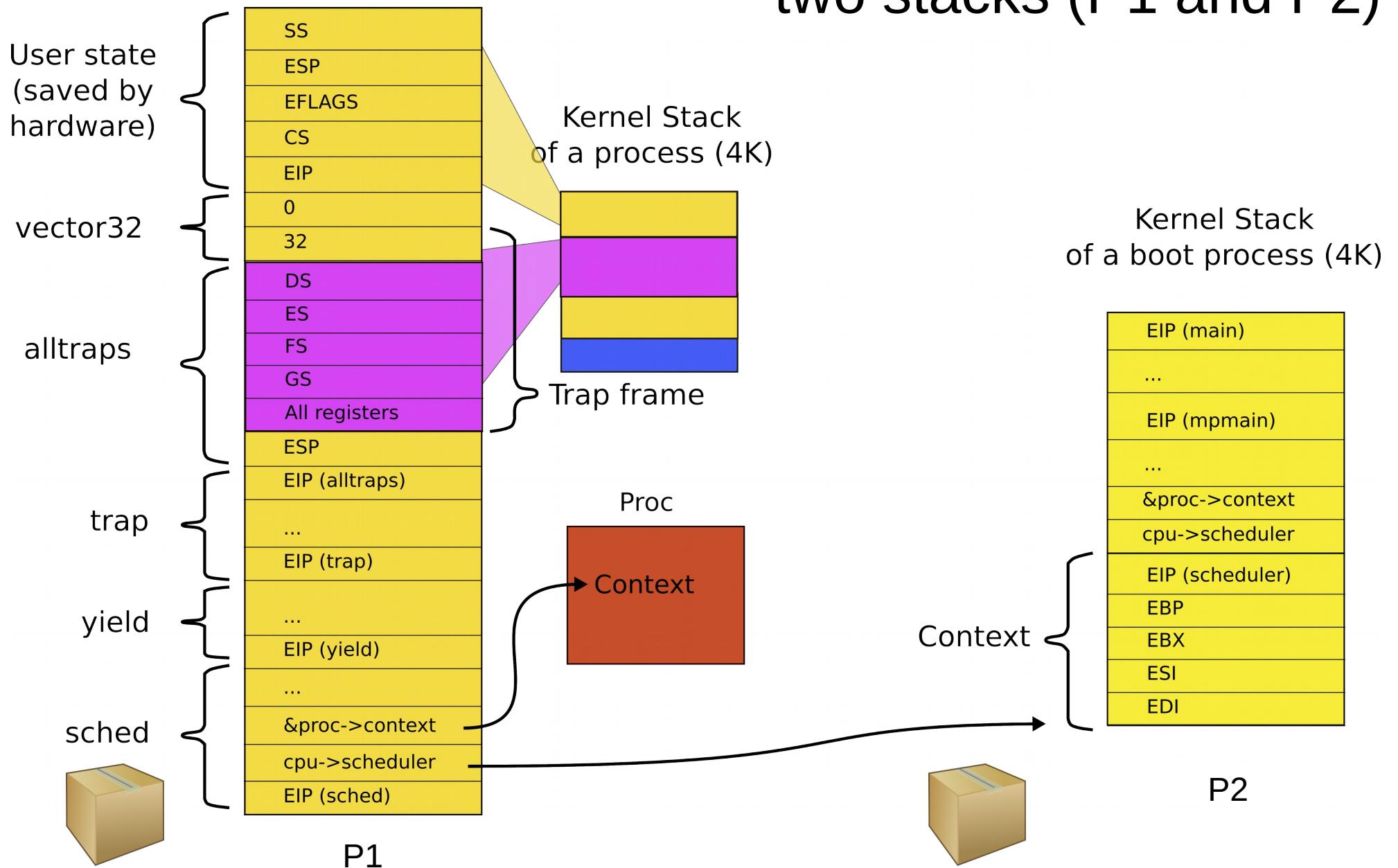
10 of 10

2773 }

Stack inside swtch() and its two arguments (passed on the stack)



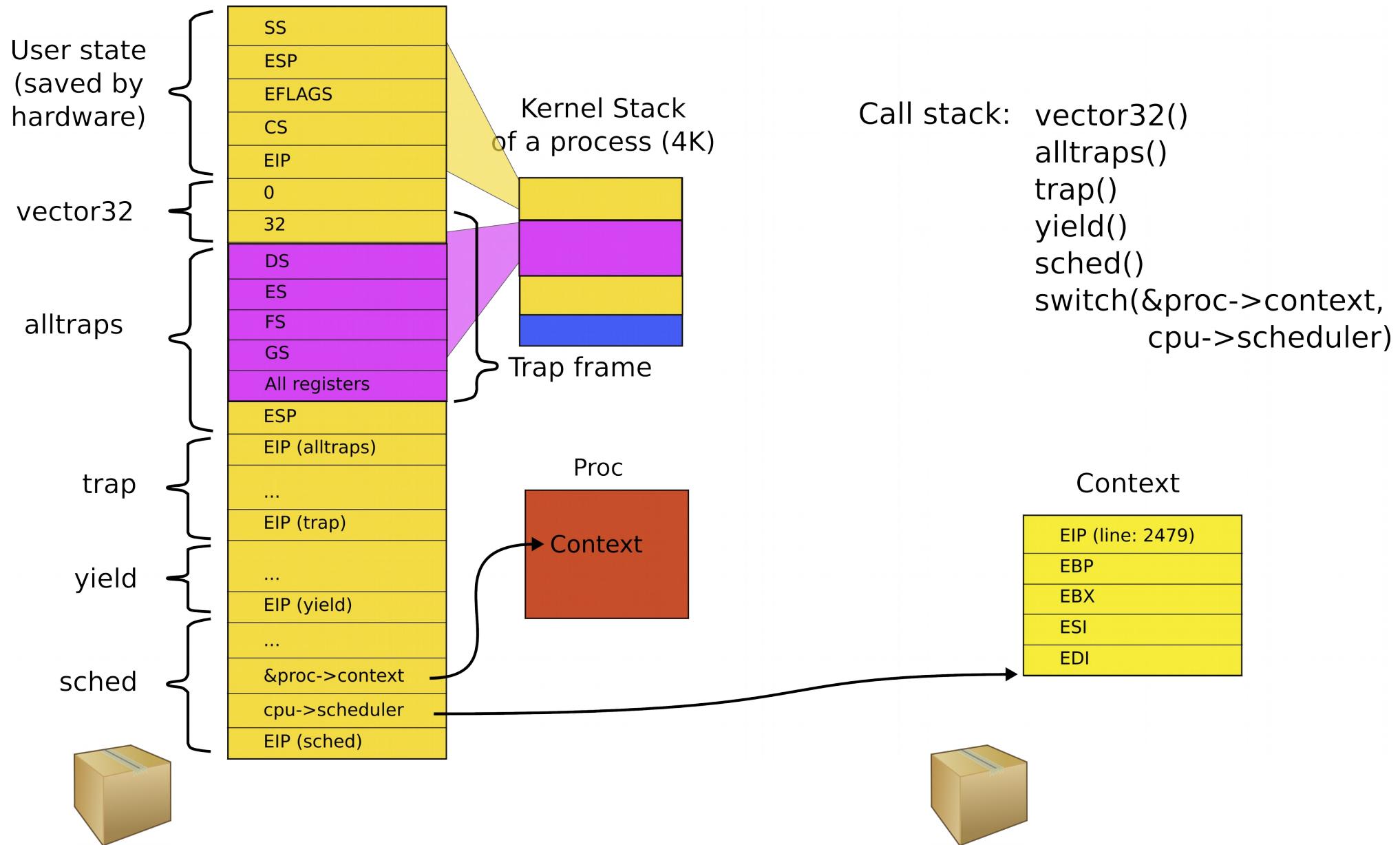
Remember you have
two stacks (P1 and P2)



The context switch function should pack everything what is left in the box and switch stacks, i.e., save the pointer to the old stack (P1) and load the new stack (P2)

What is left unsaved?

Callee saved registers



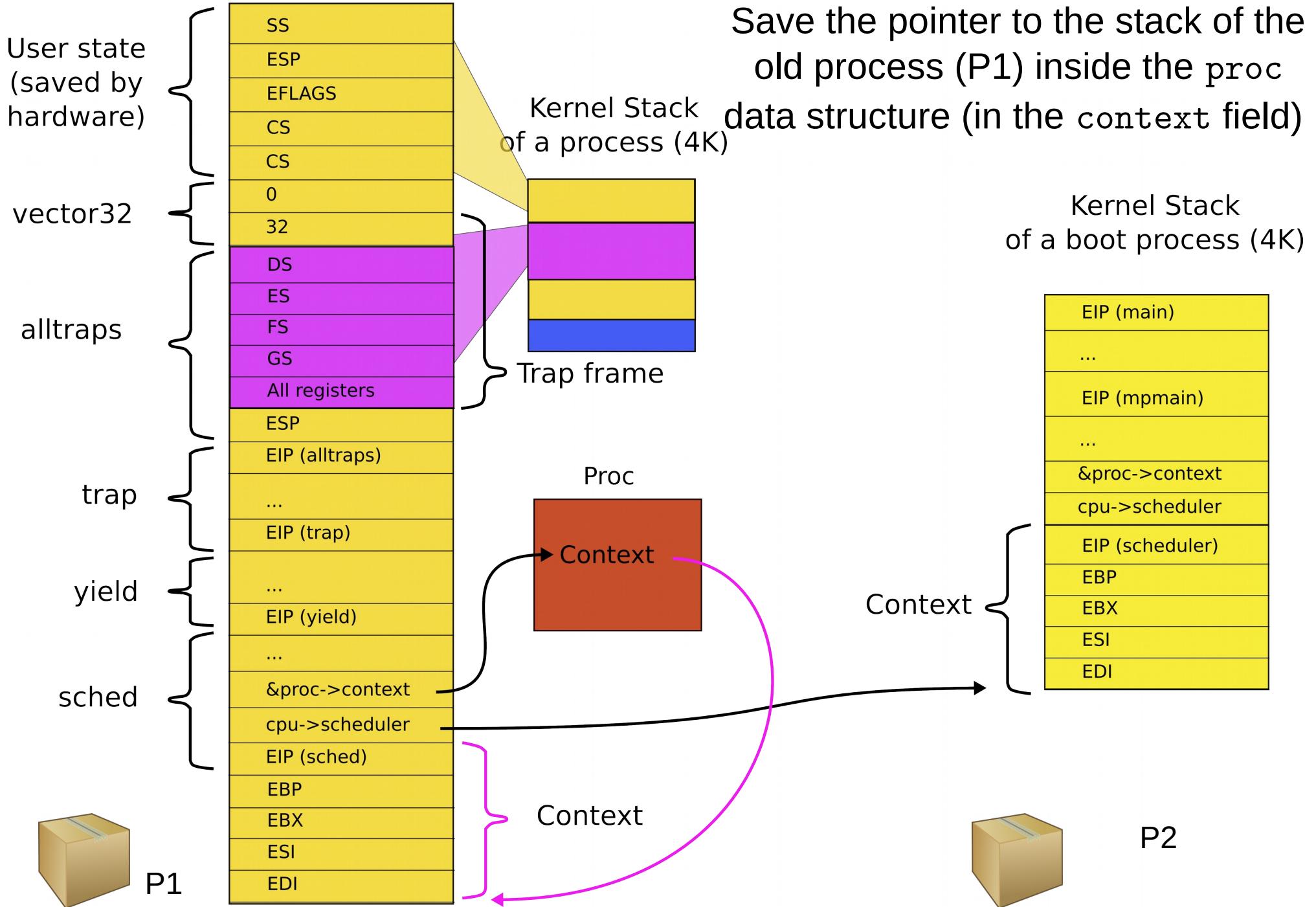
swtch(): save registers on the stack

```
2958 swtch:  
2959    movl 4(%esp), %eax  
2960    movl 8(%esp), %edx  
2961  
2962    # Save old callee-save registers  
2963    pushl %ebp  
2964    pushl %ebx  
2965    pushl %esi  
2966    pushl %edi  
2967  
2968    # Switch stacks  
2969    movl %esp, (%eax)  
2970    movl %edx, %esp  
2971  
2972    # Load new callee-save registers  
2973    popl %edi  
2974    popl %esi  
2975    popl %ebx  
2976    popl %ebp  
2977    ret
```

```
2093 struct context {  
2094     uint edi;  
2095     uint esi;  
2096     uint ebx;  
2097     uint ebp;  
2098     uint eip;  
2099 };
```

Main trick: context is always saved on the top of a stack

And the context switch just saves the old context
and loads the new

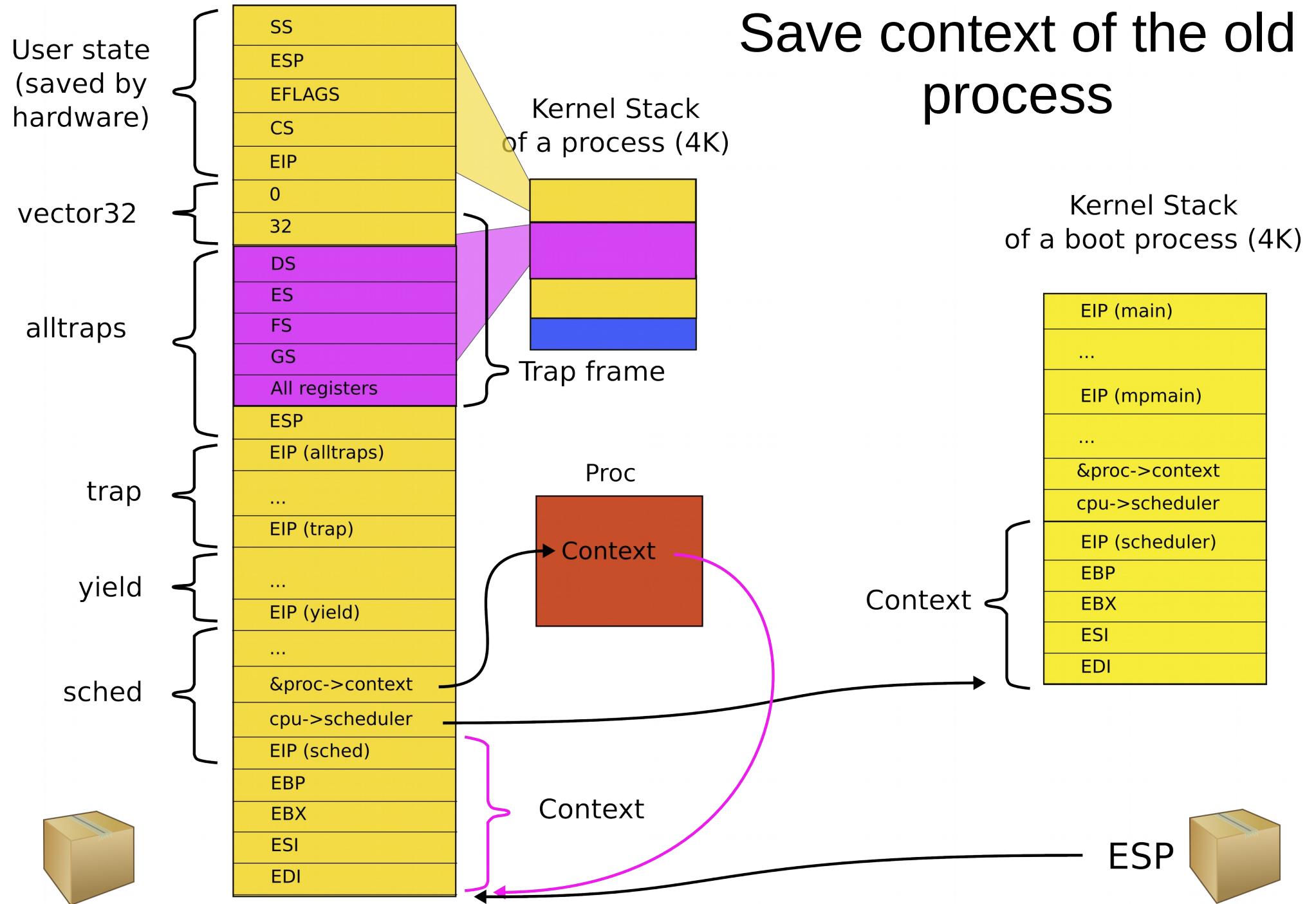


```
2958 swtch:  
2959    movl 4(%esp), %eax      // struct context **old  
2960    movl 8(%esp), %edx      // struct context *new  
2961  
2962 # Save old callee-save registers  
2963    pushl %ebp  
2964    pushl %ebx  
2965    pushl %esi  
2966    pushl %edi  
2967  
2968 # Switch stacks  
2969    movl %esp, (%eax)    // load current context (top of current stack) into  
                  // the memory location pointed by *old  
2970    movl %edx, %esp      // set stack to be equal to *new (the top of the new context)  
2971  
2972 # Load new callee-save registers  
2973    popl %edi  
2974    popl %esi  
2975    popl %ebx  
2976    popl %ebp  
2977    ret
```

swtch()

Now you can simply load the new context (P2) into ESP and continue returning on that new stack

Save context of the old process

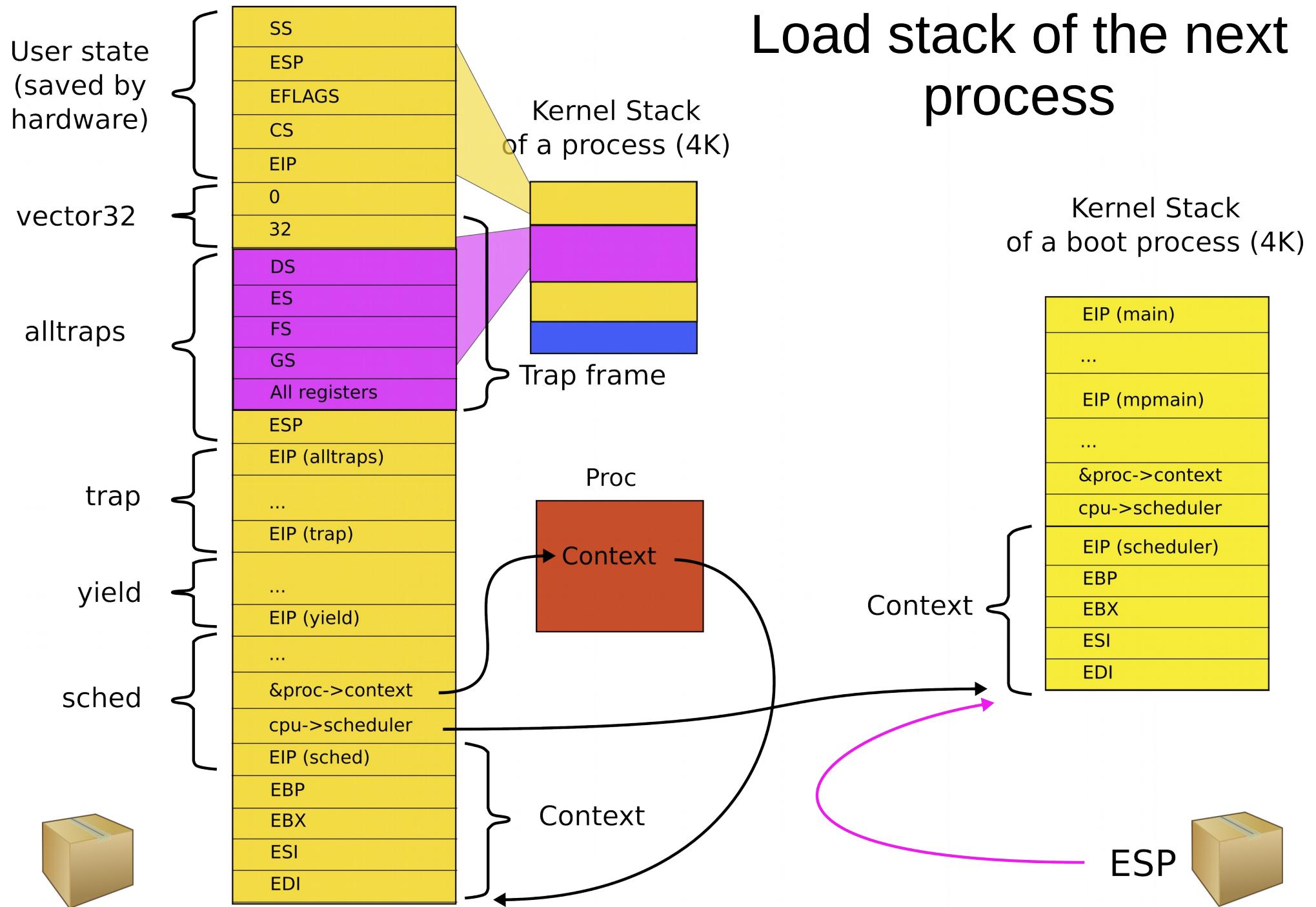


```
2958 swtch:  
2959    movl 4(%esp), %eax      // struct context **old  
2960    movl 8(%esp), %edx      // struct context *new  
2961  
2962 # Save old callee-save registers  
2963    pushl %ebp  
2964    pushl %ebx  
2965    pushl %esi  
2966    pushl %edi  
2967  
2968 # Switch stacks  
2969    movl %esp, (%eax)      // save current context (top of current stack) into  
                  // the memory location pointed by *old  
2970    movl %edx, %esp       // set stack to be equal to *new (the top of the new context)  
2971  
2972 # Load new callee-save registers  
2973    popl %edi  
2974    popl %esi  
2975    popl %ebx  
2976    popl %ebp  
2977    ret
```

swtch(): load next context

- Load address of the next context (it's in **%edx**) into **%esp**

Load stack of the next process



Remember: The context switch function should just save the pointer to the old stack (P1) and load the new stack (P2)

```
2958 swtch:  
2959    movl 4(%esp), %eax  
2960    movl 8(%esp), %edx  
2961  
2962 # Save old callee-save registers  
2963    pushl %ebp  
2964    pushl %ebx  
2965    pushl %esi  
2966    pushl %edi  
2967  
2968 # Switch stacks  
2969    movl %esp, (%eax)  
2970    movl %edx, %esp  
2971  
2972 # Load new callee-save registers  
2973    popl %edi  
2974    popl %esi  
2975    popl %ebx  
2976    popl %ebp  
2977    ret
```

Now: exit from swtch()

Where does this switch() return?

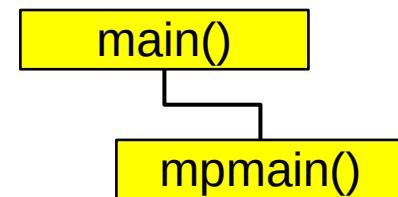
Context is always top of some stack...ok, but how?

- How does initialization of each CPU end?

```
1317 main(void)
1318 {
1319     kinit1(end, P2V(4*1024*1024)); // phys page allocator
1320     kvmalloc(); // kernel page table
1321     mpinit(); // detect other processors
1322     ...
1323     seginit(); // segment descriptors
1324     ...
1325     tvinit(); // trap vectors
1326     ...
1327     userinit(); // first user process
1328     mpmain(); // finish this processor's setup
1329 }
1330 }
```

main()

```
1260 // Common CPU setup code.  
1261 static void  
1262 mpmain(void)  
1263 {  
1264     cprintf("cpu%d: starting\n", cpu->id);  
1265     idtinit(); // load idt register  
1266     xchg(&cpu->started, 1);  
1267     scheduler(); // start running processes  
1268 }
```



We ended boot by starting the scheduler

```
2458 scheduler(void)
2459 {
2462     for(;;){
2468         for(p = ptable.proc; p < &ptable.proc[NPROC] ; p++){
2469             if(p->state != RUNNABLE)
2470                 continue;
2475             proc = p;
2476             switchuvm(p);
2477             p->state = RUNNING;
2478             swtch(&cpu->scheduler, proc->context);
2479             switchkvm();
2483             proc = 0;
2484         }
2487     }
2488 }
```

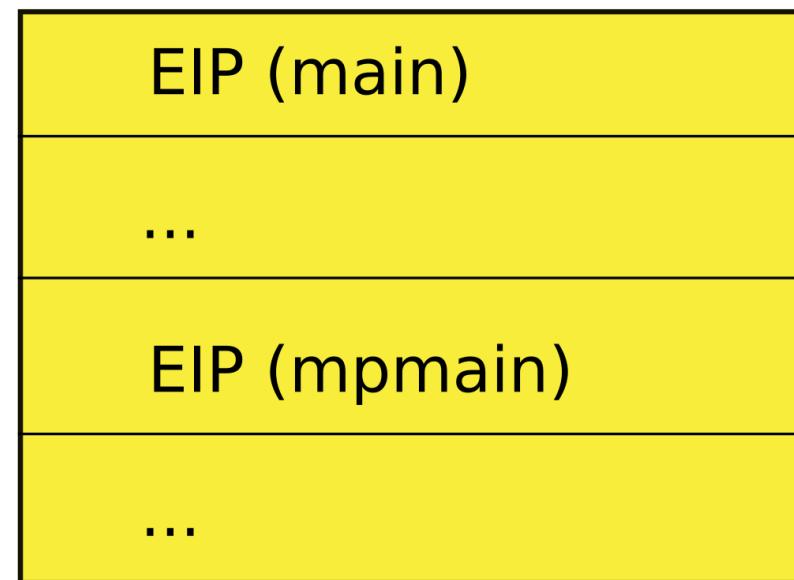
Scheduler()

- Chooses next process to run
- Switches to it
 - From the current context

```
2301 struct cpu {  
2302     uchar apicid;           // Local APIC  
2303     struct context *scheduler; // swtch() here to enter scheduler  
2304     struct taskstate ts;      // TSS  
2305     struct segdesc gdt[NSEGS]; // x86 global descriptor table  
2306     volatile uint started;    // Has the CPU started?  
2307     int ncli;                // Depth of pushcli nesting.  
2308     int intena;              // Were interrupts enabled ...  
2309     struct proc *proc;        // The process running on this cpu  
2310 };  
2311  
2312 extern struct cpu cpus[NCPU];  
2313 extern int ncpu;
```

This is how the stack looked after boot finished, i.e., inside `mpmain()`

Kernel Stack of a boot process (4K)



- So when the scheduler context switched the first time

```
2478 swtch(&cpu->scheduler,  
           proc->context);
```

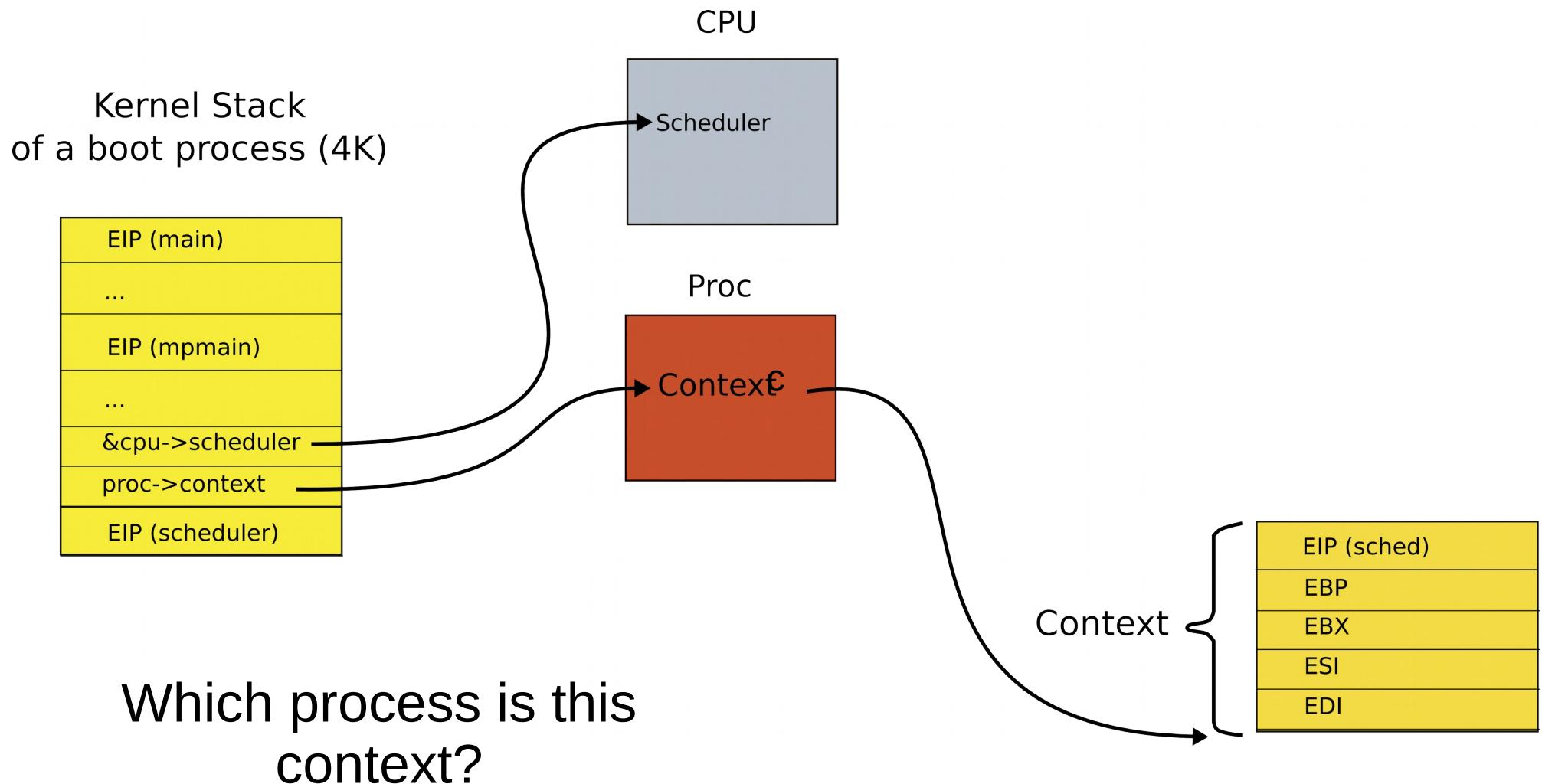
- We save the current context of the scheduler into:

```
&cpu->scheduler
```

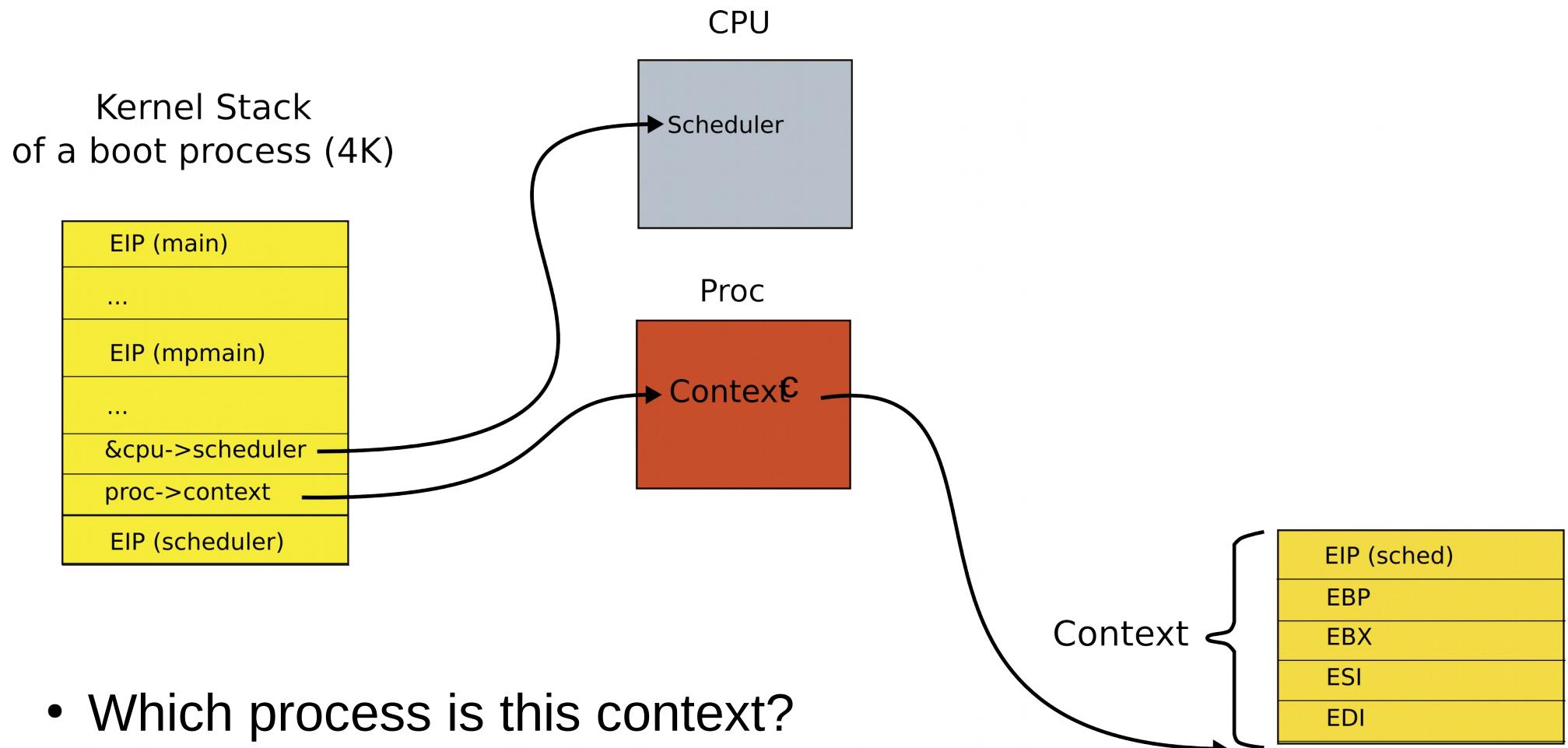
- And restore the context of the first process

```
proc->context
```

This is how stack looked like
when scheduler() invoked
switch() for the first time



This is how stack looked like
when scheduler() invoked
swtch() for the first time



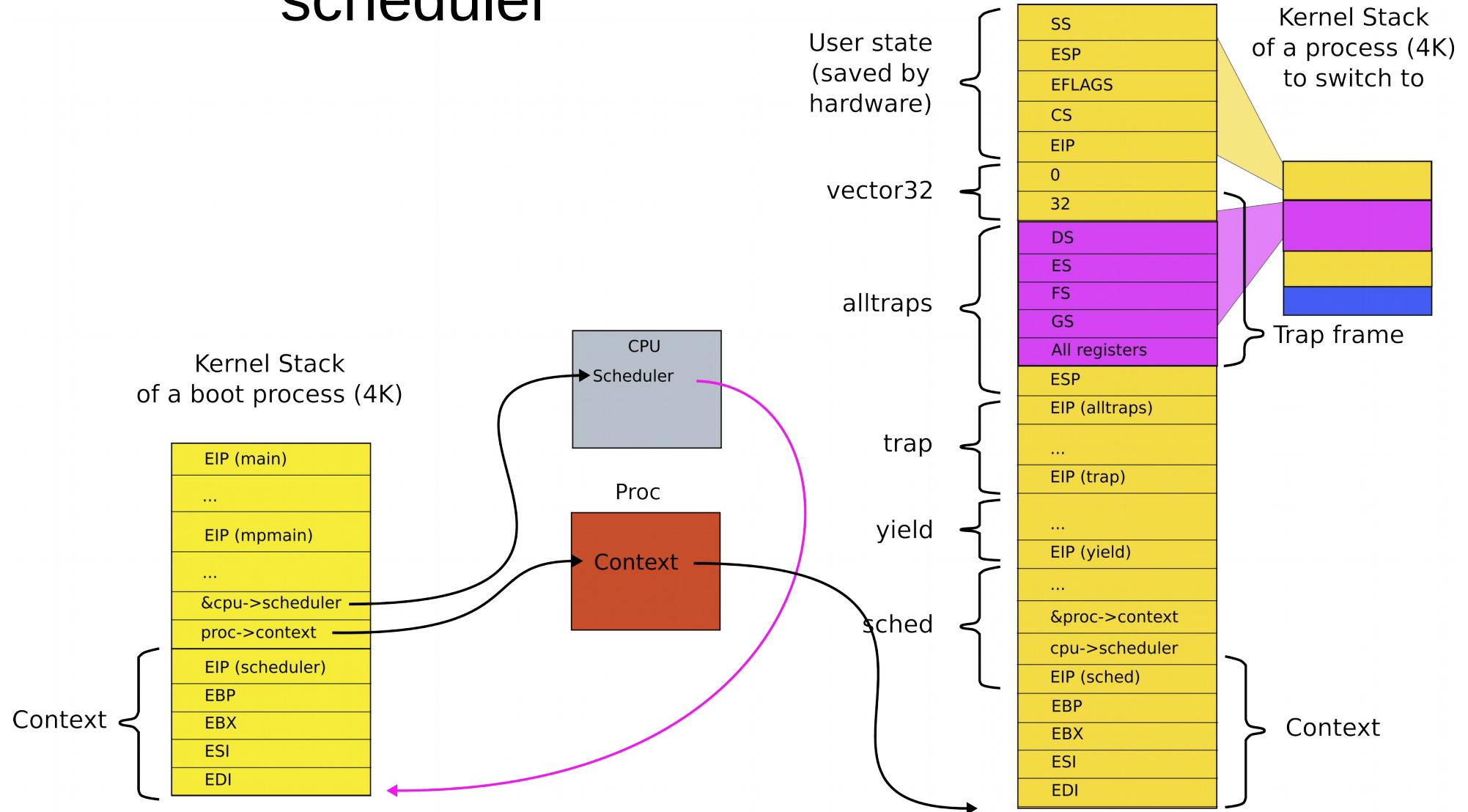
- Which process is this context?
- It's the context of the first process scheduler decides to run

```
2456 allocproc(void)
2457 {
...
2477 // Leave room for trap frame.
2478 sp -= sizeof *p->tf;
2479 p->tf = (struct trapframe*)sp;
2480
2481 // Set up new context to start executing at forkret,
2482 // which returns to trapret.
2483 sp -= 4;
2484 *(uint*)sp = (uint)trapret;
2485
2486 sp -= sizeof *p->context;
2487 p->context = (struct context*)sp;
2488 memset(p->context, 0, sizeof *p->context);
2489 p->context->eip = (uint)forkret;
...
2492 }
```

Context is configured as top of the stack when new process is created inside allocproc() function

- Remember `exec()`?

Save context of the scheduler



```
2958 swtch:  
2959    movl 4(%esp), %eax      // struct context **old  
2960    movl 8(%esp), %edx      // struct context *new  
2961  
2962 # Save old callee-save registers  
2963    pushl %ebp  
2964    pushl %ebx  
2965    pushl %esi  
2966    pushl %edi  
2967  
2968 # Switch stacks  
2969    movl %esp, (%eax)    // load current context (top of current stack) into  
                  // the memory location pointed by *old  
2970    movl %edx, %esp      // set stack to be equal to *new (the top of the new context)  
2971  
2972 # Load new callee-save registers  
2973    popl %edi  
2974    popl %esi  
2975    popl %ebx  
2976    popl %ebp  
2977    ret
```

swtch()

The context is the top of some stack

- Initially it was the stack of mpenter()
 - On which scheduler started
- Then first process...
 - Then scheduler again
 - And the next process...

Back to the context switch
(we now know we return into the scheduler)

Where does this `swtch()` return?

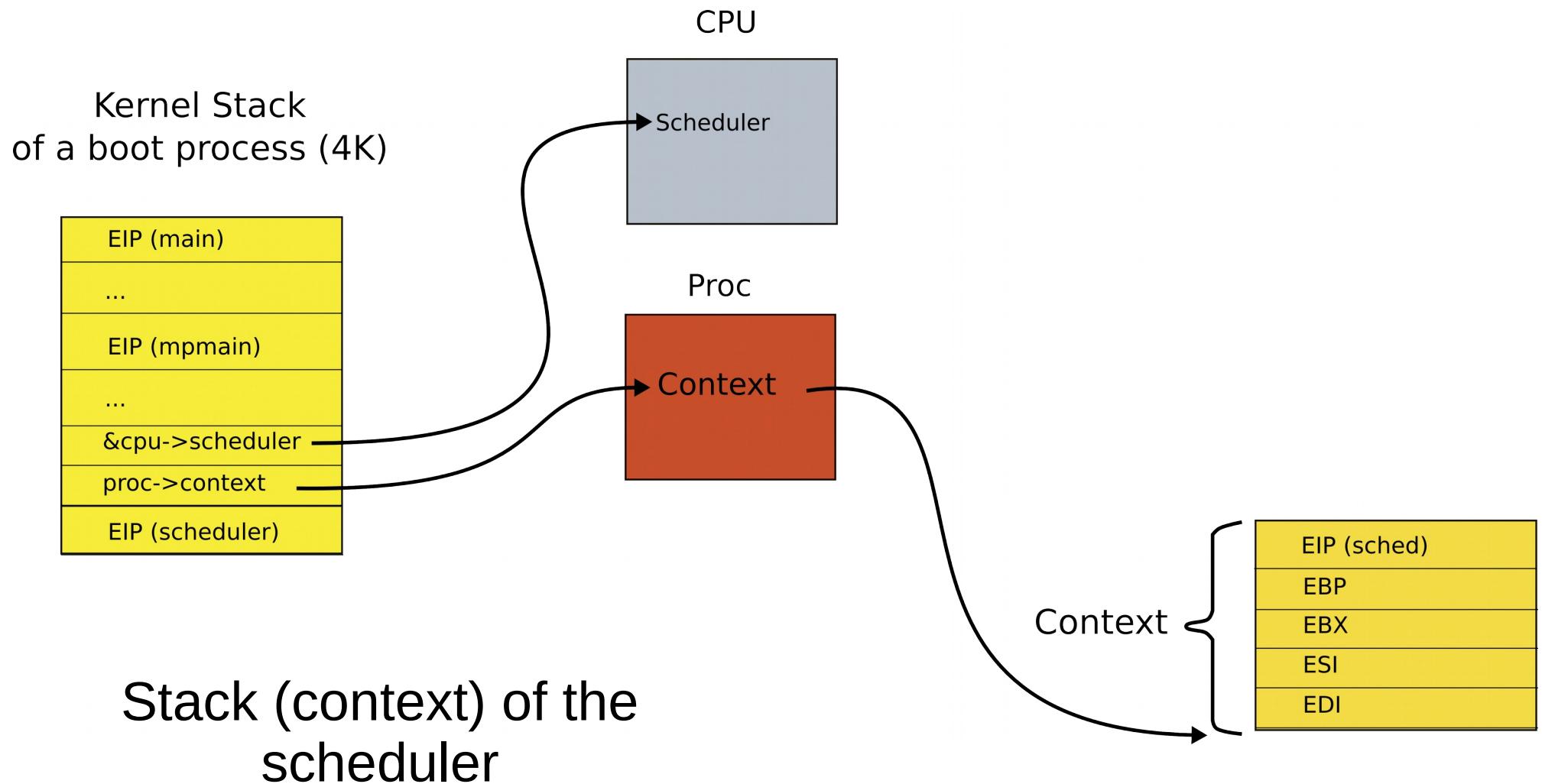
- Scheduler!
- After all remember
 - We started with timer interrupt
 - Entered the kernel
 - Entered `schedule()`
 - Entered `switch`
- And are currently on our way from the process into the scheduler

```
2458 scheduler(void)
2459 {
2462     for(;;){
2468         for(p = ptable.proc; p < &ptable.proc[NPROC] ; p++){
2469             if(p->state != RUNNABLE)
2470                 continue;
2475             proc = p;
2476             switchuvm(p);
2477             p->state = RUNNING;
2478             swtch(&cpu->scheduler, proc->context);
2479             switchkvm();
2483             proc = 0;
2484         }
2487     }
2488 }
```

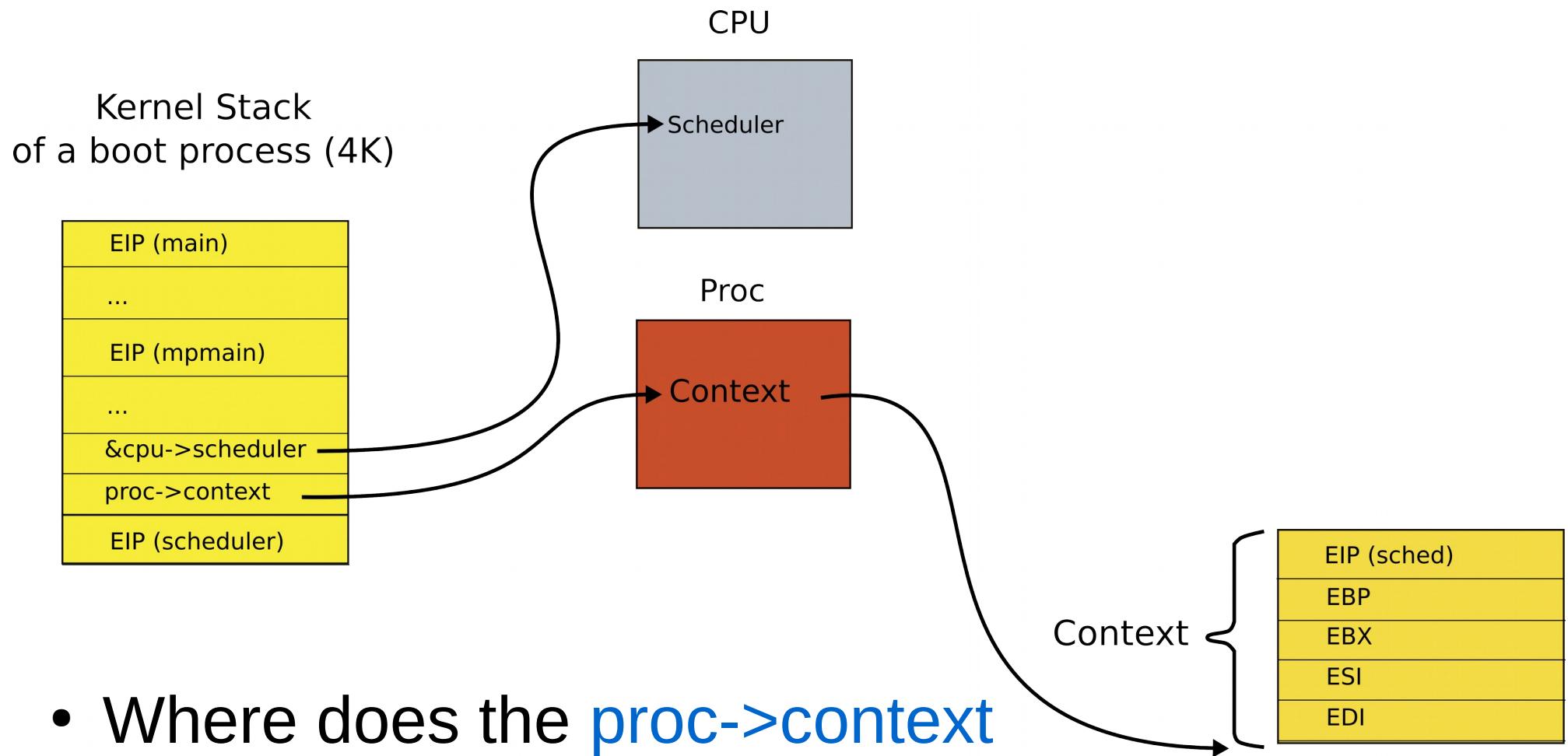
What does scheduler do?

- Chooses next process to run
- Switches to it

What does stack look like when
scheduler() invokes swtch()?

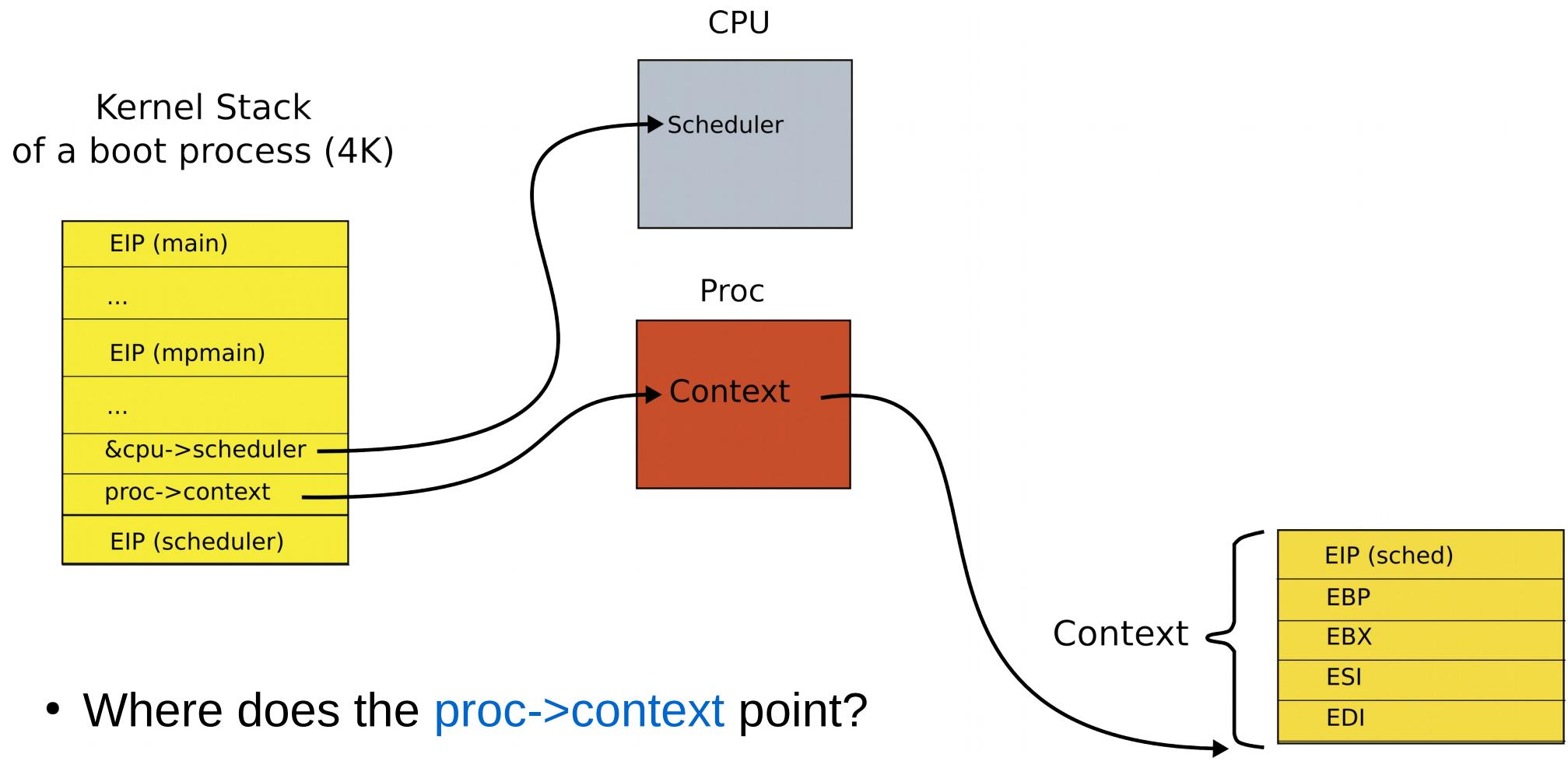


What does stack look like when scheduler() invokes swtch()?



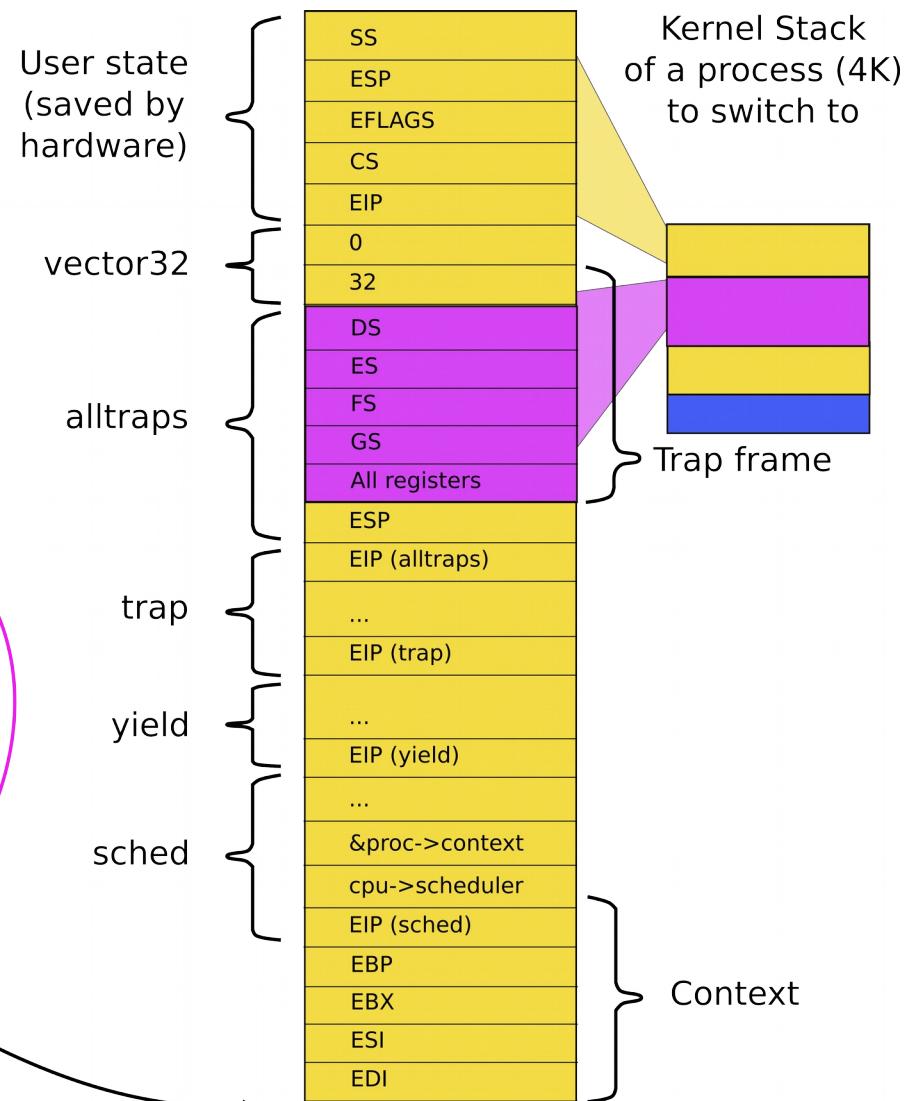
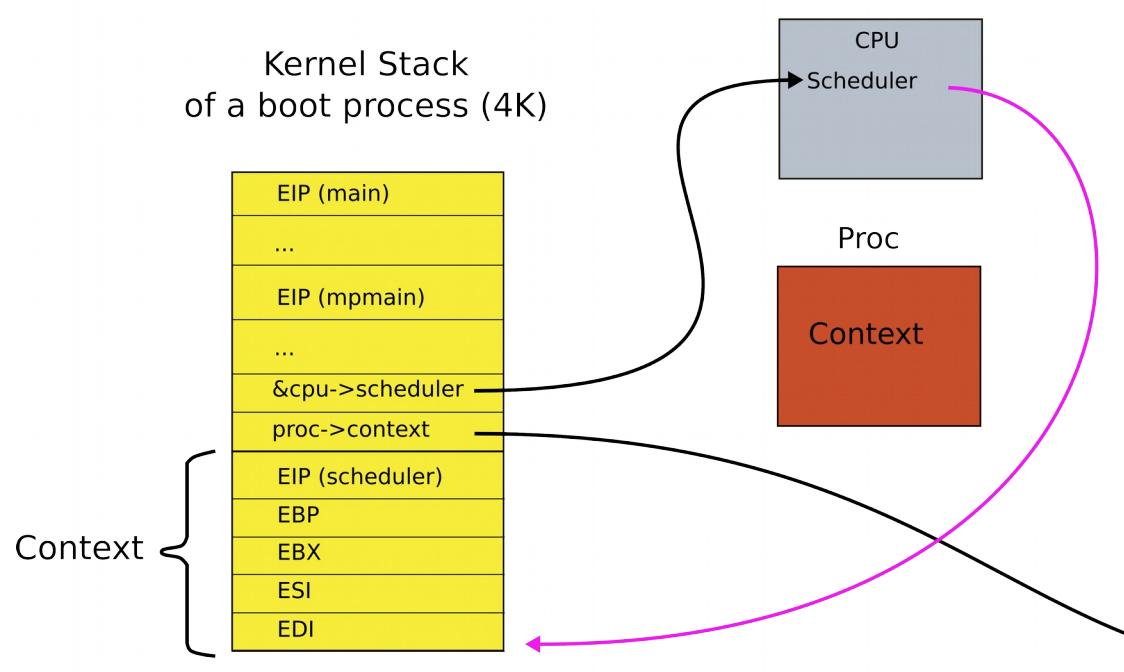
- Where does the `proc->context` point?

What does stack look like when scheduler() invokes swtch()?



- Where does the **proc->context** point?
 - Right the context (stack) of the next process to run

- We save the context of the scheduler
- Restore the context of the next process

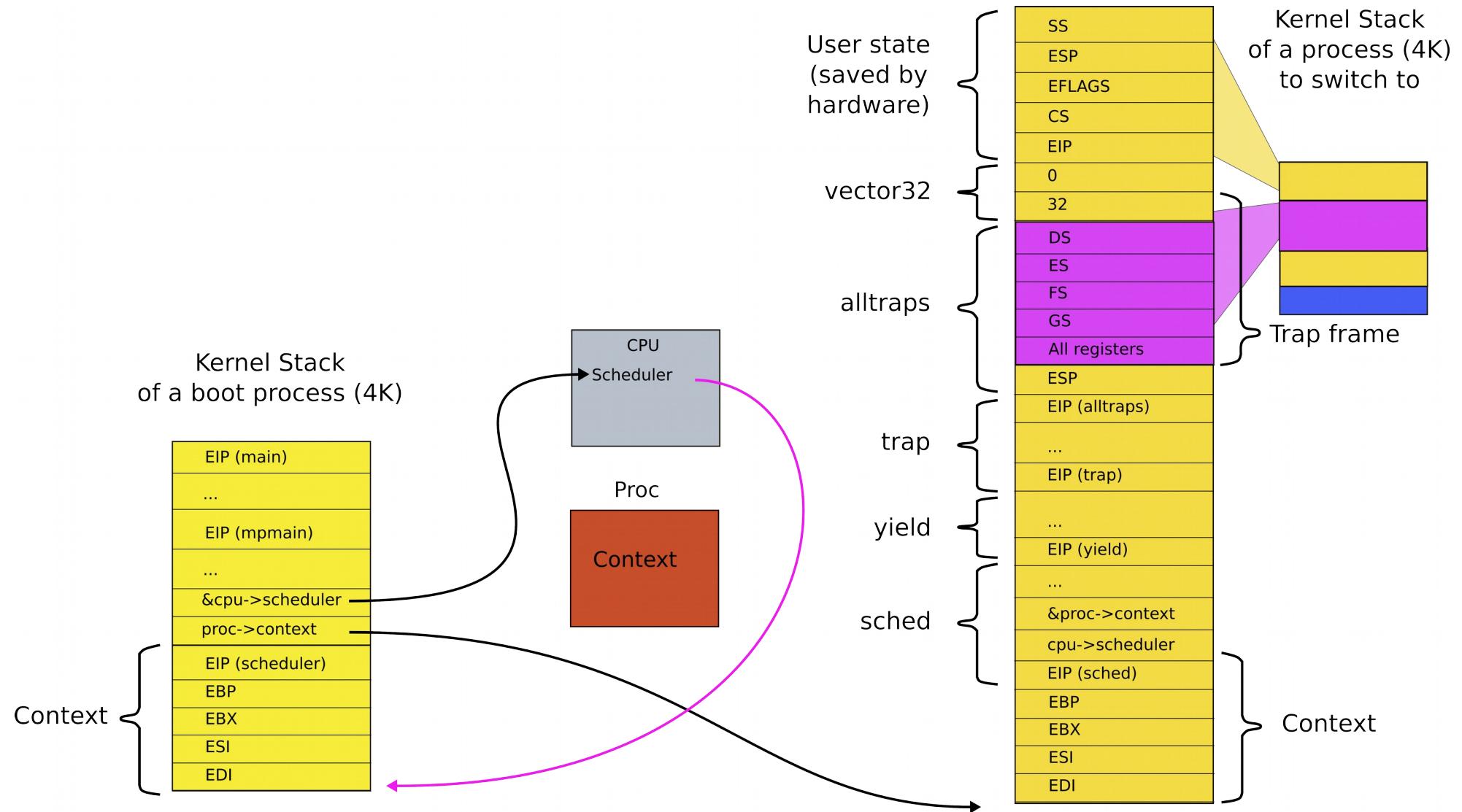


- Remember, from inside the scheduler we invoked `swtch()` as

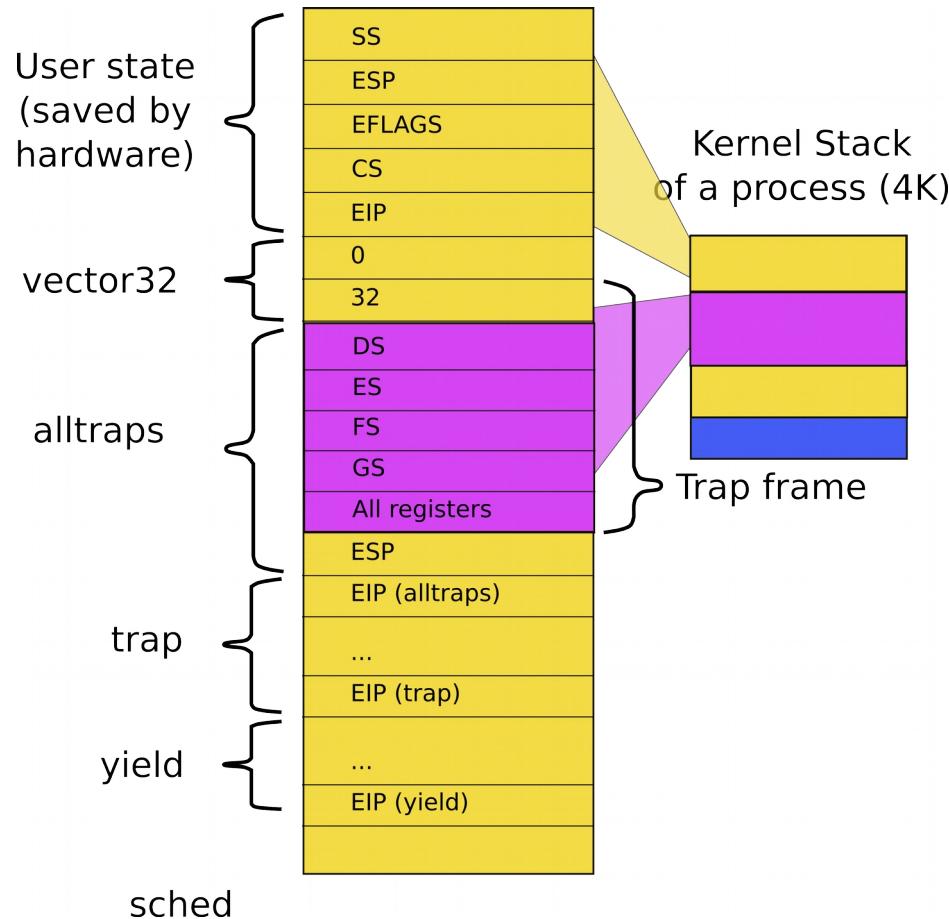
```
2478 swtch(&cpu->scheduler,  
           proc->context);
```

- Hence, we save context of the scheduler into
`&cpu->scheduler`
- And restore
`proc->context`

Stacks and contexts inside the swtch()



Exiting back to user-level



- Stack of the process after context switch, i.e., inside `sched()`
 - Return as usual all the way to `alltrap()`

```
3004 alltraps:
```

```
...
```

```
3020 # Call trap(tf), where tf=%esp
```

```
3021 pushl %esp
```

```
3022 call trap
```

```
3023 addl $4, %esp
```

```
3024
```

```
3025 # Return falls through to trapret...
```

```
3026 .globl trapret
```

```
3027 trapret:
```

```
3028 popal
```

```
3029 popl %gs
```

```
3030 popl %fs
```

```
3031 popl %es
```

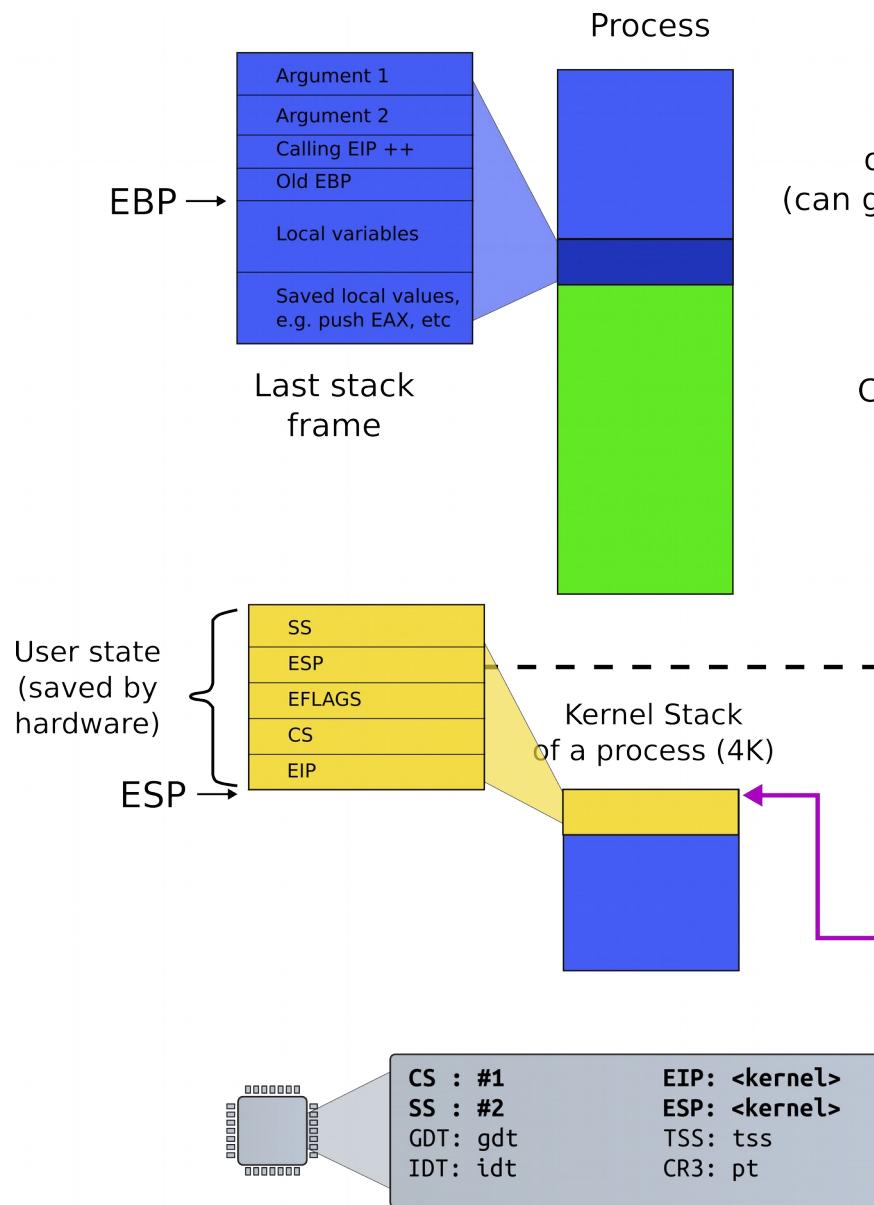
```
3032 popl %ds
```

```
3033 addl $0x8, %esp # trapno and errcode
```

```
3034 iret
```

alltraps(): exit into user-level

We're back to where we started, but in a new process



vector32

Summary

- We switch between processes now

Thank you