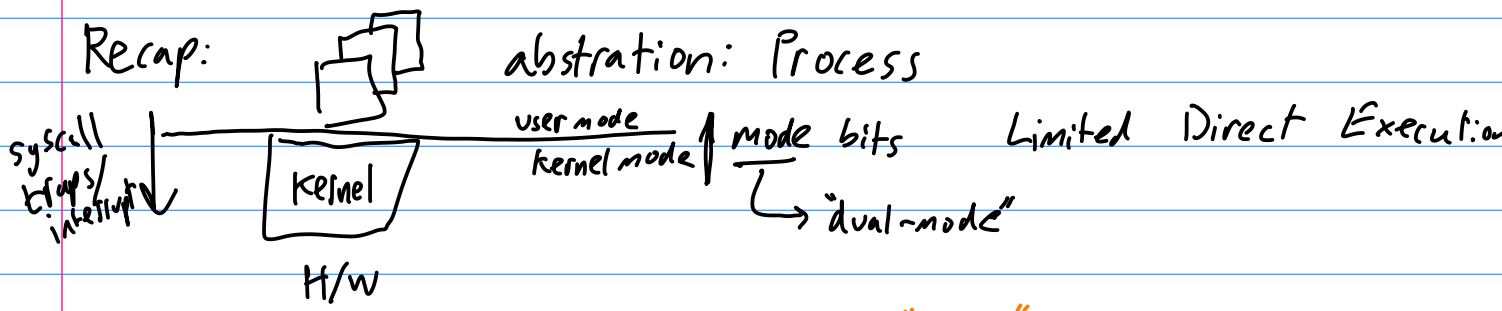
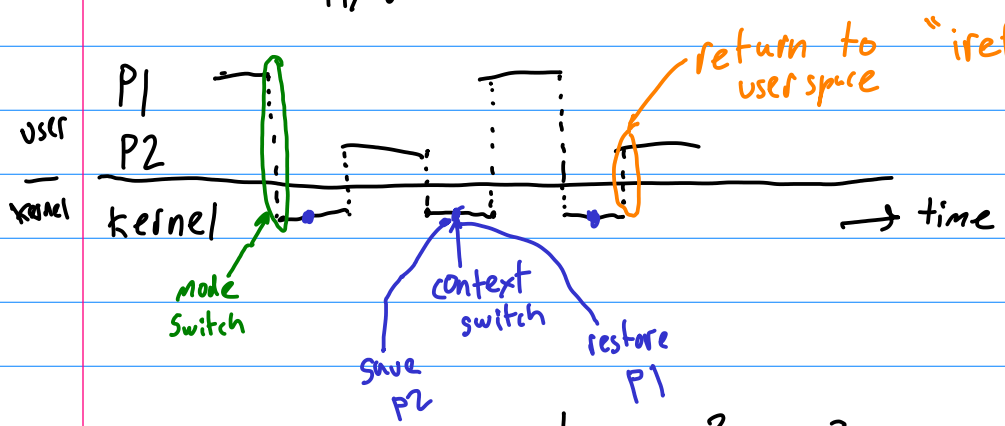
CS 3214 Lecture #3 "Unicode"

courses.cs.vt.edu/cs3214/williams/

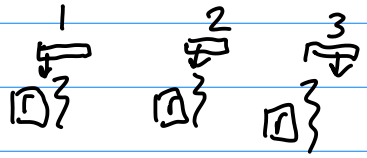
Office Hours: T/Th 1-2pm on Zoom

Recap: abstration: Process





user-threading stacks? reg? [address space is shared]



"scheduling"

1. cooperative scheduling: "yield"

preemption

Coming up: Process states: (runnable?) Management: (create / destroy) || next time

Detail about interacting w/ the system: ExO
character sets + unicode
"appreciation of details"

UNIX philosophy: small programs do one thing well
connect w/ universal text

cat foo | grep bar | wc

when do you need unicode/char sets?
- user input (web forms / web servers)
- text file processing (copying / conversion / validation / display)
- i18n internationalization (18)

kubernetes : k8s

byte - unit of digital info
octet : 8-bits
0 ... 255
0x00 ... 0xff

History: byte didn't have to be 8

C: uint_8 foo;
unsigned char foo;
CHAR_WIDTH
POSIX = 8

ints, shorts, longs
endianness

0x0102
1x256 + 2 = 258
1 + 256*2 = 513

Characters: abstract entities from some "alphabet"

~~XX~~ (3) 3 bits

variable-length encoding

00 01 10 1100 1101 1110

00 01 10 1100
A T F ♡

~~00 11 11~~

grapheme

characters "code points" U+0041 "A"

1, 114, 112 U+00C4 "Ä"

U+1F385  Santa

but... U+0308 umlaut over preceding char

U+0041 U+0308 "Ä"

how to encode
- in prog mem?
- on disk?

ASCII

Tradeoff: ease of processing vs. efficiency of storage

UTF-32 endianness BOM "byte order mark"
- 32 bits for each char S[10]
- santa 0x0001F385 4x overhead over ASCII

Python's Swift Linux wchar_t C11 char32_t

UTF-16
- 16 bits for each char (some have 4 bytes) 0xD83C 0xDF85
- Sometimes can index S[10]
- still wasteful, error-prone

Java Javascript

UTF-8
- 8 bits variable-length 1, 2, 3 or 4 bytes
- santa 0xF0 0x9F 0x8E 0x85

UTF-8 (cont'd) ← most common when unicode transmitted/ stored

- ASCII ← 1 byte 0x7A "z"
- space efficient
- indexing is hard/impossible s[0] ~~s[0]~~

Rust, Go

"crate"
"packages"

Given a seq. of bytes, can you tell the encoding?

- No
- can say something is invalid
- metadata is needed HTTP response:

  Content-Type: text/html; charset=UTF-8

- files? operator / program