

# **⇒GOing Down the Compilation Rabbit Hole**

**GOLAB 2023**

**Richard Rowland**

# whoami

## Richard Rowland

Place: Tokyo, Japan

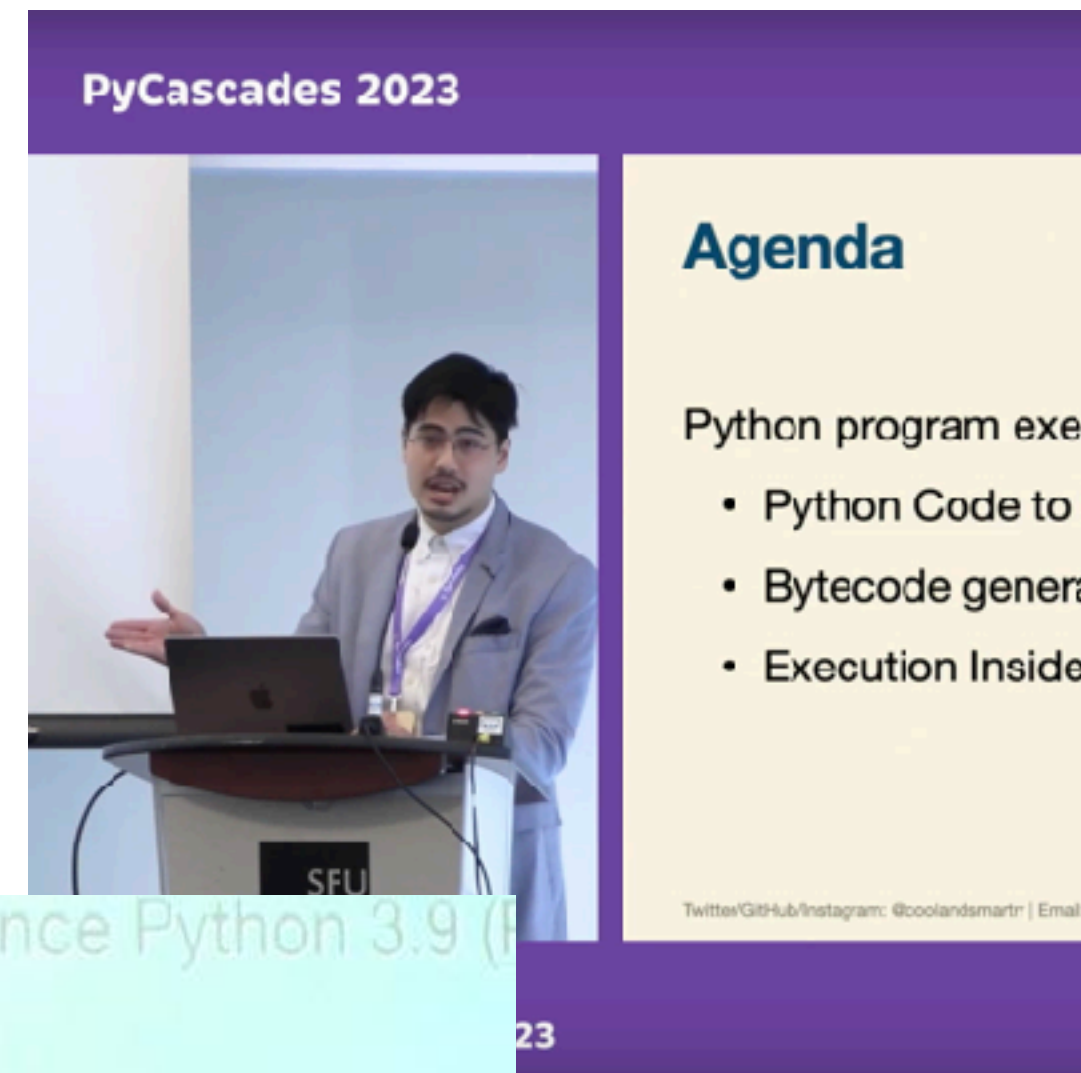
Job: Staff at OKCoinJapan, a global cryptocurrency exchange

Hobby: Photo Nerd! 📷  
(Follow my Insta / Share photo spots in Florence)

Projects: Created a method to semi-permanently store data in blockchain and P2P network (on GitHub)



# Previous Talks on Python Compiler



- Presented on Python Compiler at Python conferences
- Became curious about other languages's compilers



PyCascades 2023  
<https://www.youtube.com/watch?v=RcGshw0tzoc>

PyCon Ireland 2022  
<https://www.youtube.com/watch?v=u6XDRzLX6Eo>

# What is Compiling?

# Compiling

Turning source code into executable format



# Compilers

Programs used to compile programs



# Famous Compilers



gcc



llvm

# Compiling in Golang

```
> go build main.go
```



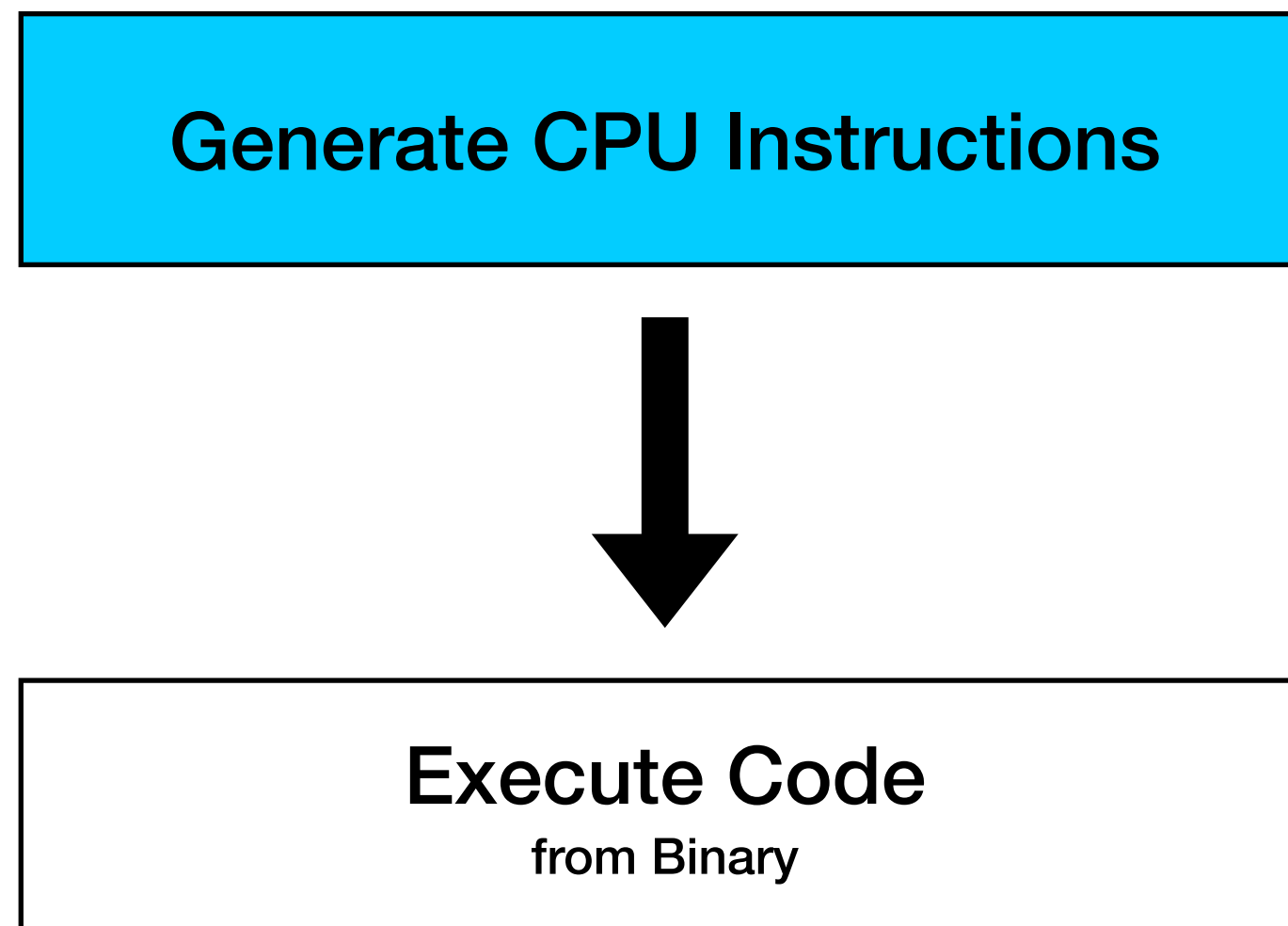
# About the Golang Compiler

- Golang project develops its own compiler (doesn't use GCC or LLVM)
- Source Code: <https://go.googlesource.com/go>
- Compiler code is stored under: `src/cmd/compile`

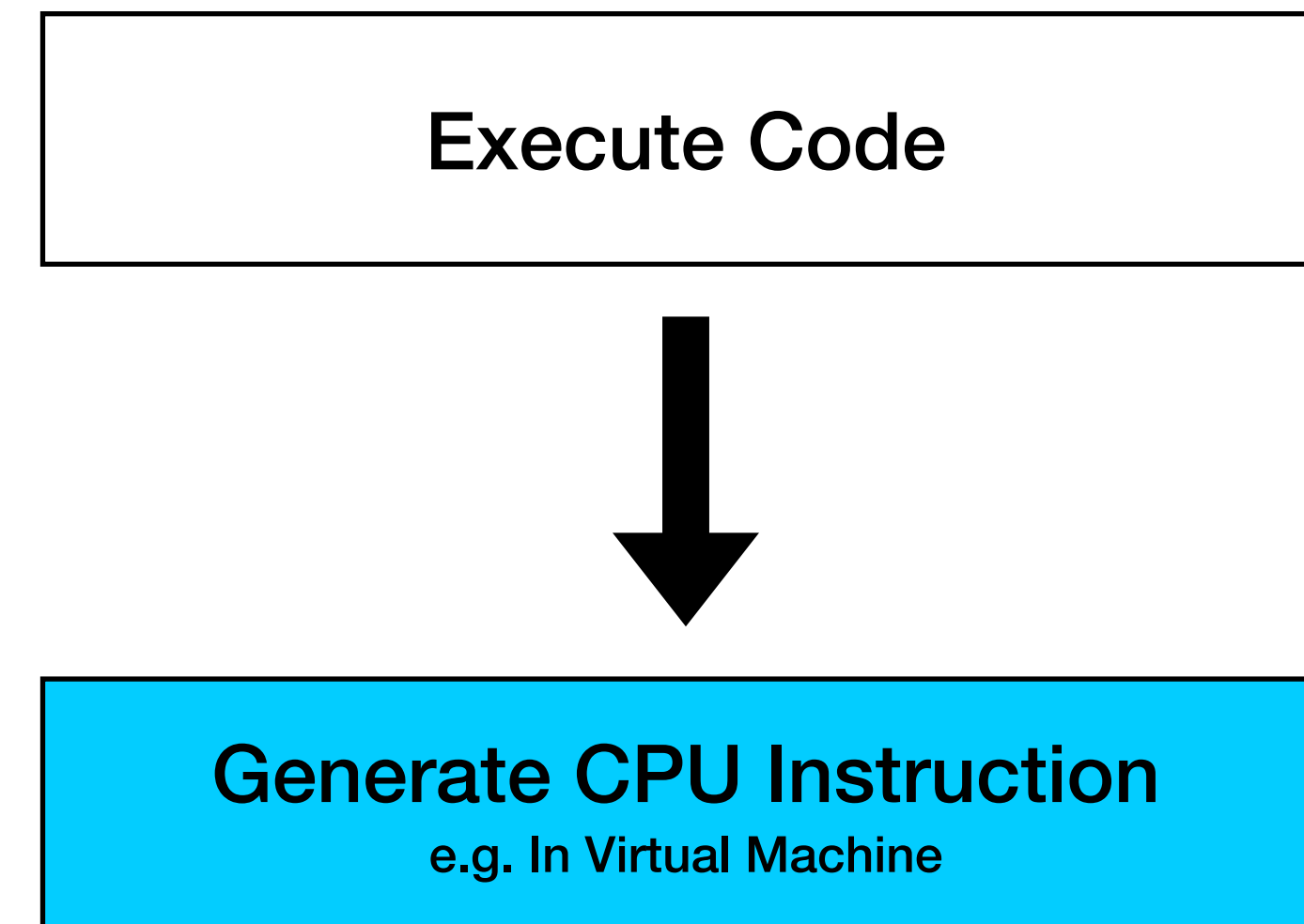
# Programming Language Execution Models

# Different models of program execution

## Compiled Languages



## Interpreted Languages



# Compiled vs Interpreted

## Examples of Programming Languages

### Compiled Languages

- **Golang**
- C, C++
- Rust
- Swift

### Interpreted Languages

- Javascript
- Python
- Ruby

# Compiled vs Interpreted

## Pros and Cons

	Compiled	Interpreted
Pros	<ul style="list-style-type: none"><li>• Fast execution</li></ul>	<ul style="list-style-type: none"><li>• No need to create binary for each platform (“Write Once, Run Everywhere”)</li></ul>
Cons	<ul style="list-style-type: none"><li>• Takes time to compile</li><li>• Need to compile for each platform</li></ul>	<ul style="list-style-type: none"><li>• Execution tends to be slower</li><li>• CPU/Memory Resource Overhead of Virtual Machine</li></ul>

# Comparing C and Golang Binary

# Binary Size - C vs Go

**C**

```
#include <stdio.h>

int main(){
    printf("Hello, World!\n");
    return 0;
}
```

Binary Size: 33 KB

**Go**

```
package main
import "fmt"

func main() {
    fmt.Println("Hello, World!")
}
```

Binary Size: 2 MB





# Syscalls - C vs Go

## C

```
execve("./main", ["/main"], 0xffffcd34e2c0 /* 4 vars */) = 0
set_tid_address(0xffff9cc74230) = 8
brk(NULL) = 0xaaaaad0305000
brk(0xaaaaad0307000) = 0xaaaaad0307000
mmap(0xaaaaad0305000, 4096, PROT_NONE, MAP_PRIVATE|MAP_FIXED|MAP_ANONYMOUS, -1, 0) = 0xaaaaad0305000
mprotect(0xaaaaac4cff000, 4096, PROT_READ) = 0
ioctl(1, TIOCGWINSZ, 0xfffff28fad48) = -1 ENOTTY (Not a tty)
writev(1, [{iov_base="Hello, World!", iov_len=13}, {iov_base="\n", iov_len=1}], 2) = 14
exit_group(0) = ?
+++ exited with 0 +++
```

Syscalls: 9 times

## Go

```
execve("./main", ["/main"], 0xffffcd5741c0 /* 8 vars */) = 0
sched_getaffinity(0, 8192, [0, 1]) = 8
openat(AT_FDCWD, "/sys/kernel/mm/transparent_hugepage/hpage_pmd_size", O_RDONLY) = 3
read(3, "2097152\n", 20) = 8
close(3) = 0
mmap(NULL, 262144, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) = 0xffff953ed000
mmap(NULL, 131072, PROT_NONE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) = 0xffff953cd000
mmap(NULL, 1048576, PROT_NONE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) = 0xffff952cd000
mmap(NULL, 8388608, PROT_NONE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) = 0xffff94acd000
mmap(NULL, 67108864, PROT_NONE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) = 0xffff90acd000
mmap(NULL, 536870912, PROT_NONE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) = 0xffff70acd000
mmap(NULL, 8388608, PROT_NONE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) = 0xffff702cd000
mmap(0x40000000000, 67108864, PROT_NONE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) = 0x40000000000
mmap(NULL, 33554432, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) = 0xffff6e2cd000
mmap(NULL, 1133584, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) = 0xffff6e1b8000
mmap(0x40000000000, 4194304, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_FIXED|MAP_ANONYMOUS, -1, 0) = 0x40000000000
mmap(0xffff953cd000, 131072, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_FIXED|MAP_ANONYMOUS, -1, 0) = 0xffff953cd000
mmap(0xffff952cd000, 4096, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_FIXED|MAP_ANONYMOUS, -1, 0) = 0xffff952cd000
mmap(0xffff94acf000, 4096, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_FIXED|MAP_ANONYMOUS, -1, 0) = 0xffff94acf000
mmap(0xffff90add000, 4096, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_FIXED|MAP_ANONYMOUS, -1, 0) = 0xffff90add000
mmap(0xffff70b4d000, 4096, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_FIXED|MAP_ANONYMOUS, -1, 0) = 0xffff70b4d000
mmap(0xffff702cd000, 12288, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_FIXED|MAP_ANONYMOUS, -1, 0) = 0xffff702cd000
mmap(NULL, 1048576, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) = 0xffff6e0b8000
mmap(NULL, 65536, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) = 0xffff6e0a8000
mmap(NULL, 65536, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) = 0xffff6e098000
rt_sigprocmask(SIG_SETMASK, NULL, [], 8) = 0
sigaltstack(NULL, {ss_sp=NULL, ss_flags=SS_DISABLE, ss_size=0}) = 0
sigaltstack({ss_sp=0x4000004000, ss_flags=0, ss_size=32768}, NULL) = 0
rt_sigprocmask(SIG_SETMASK, [], NULL, 8) = 0
gettid() = 453
rt_sigaction(SIGHUP, NULL, {sa_handler=SIG_DFL, sa_mask=[], sa_flags=0}, 8) = 0
rt_sigaction(SIGHUP, {sa_handler=0x6c5f0, sa_mask=~[], sa_flags=SA_ONSTACK|SA_RESTRT|SA_SIGINFO}, NULL, 8) = 0
rt_sigaction(SIGINT, NULL, {sa_handler=SIG_DFL, sa_mask=[], sa_flags=0}, 8) = 0
rt_sigaction(SIGINT, {sa_handler=0x6c5f0, sa_mask=~[], sa_flags=SA_ONSTACK|SA_RESTRT|SA_SIGINFO}, NULL, 8) = 0
rt_sigaction(SIGQUIT, NULL, {sa_handler=SIG_DFL, sa_mask=[], sa_flags=0}, 8) = 0
...
...
...
```

Syscalls: 209 times

# Why does Go make many syscalls?

- Overhead of Golang runtime
- Go functions enabled by runtime
  - Garbage Collection
  - Goroutine scheduler
  - Channel operations

# Golang Compiler(s)

# What is Golang?

A programming language specification

Golang program execution environment

# What is Golang?

A programming language specification

**Programmer's  
Concern**

---

Golang program execution environment

**Golang's Concern**

# Golang Specification

<https://go.dev/ref/spec>

# Golang Specification

## Rules to Follow

Notation

Source code representation

Constants

Variables

Properties of types and values

Blocks

Declarations and scope

Expressions

Statements

Built-in functions

Packages

Program initialization and execution

Errors

Run-time panics

System considerations

Type unification rules

**If you follow the specs,  
you can build your own compiler!**



**If you follow the specs,  
you can build your own compiler!**

**(You have to maintain it too.)**

# Different Golang Compilers



gc (go compiler)



gcc (gccgo)



llvm (llgo)

# Different Golang Compilers

## Caveats



gc (go compiler)

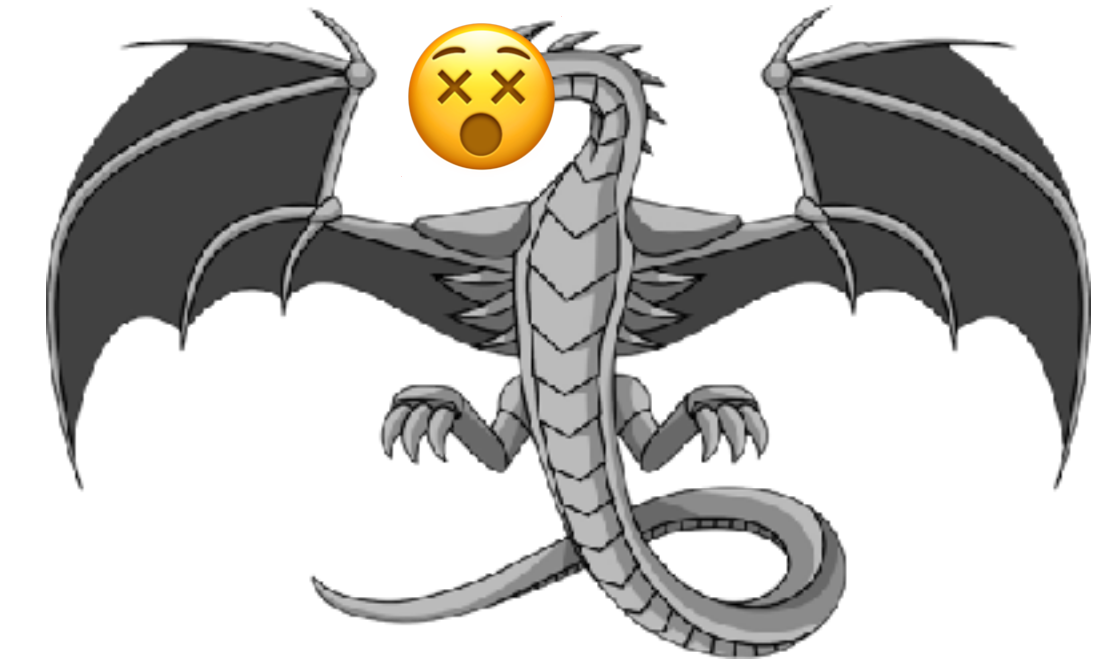
Latest Version



gcc (gccgo)

Up to Go ver. 1.18

No generics support [1]



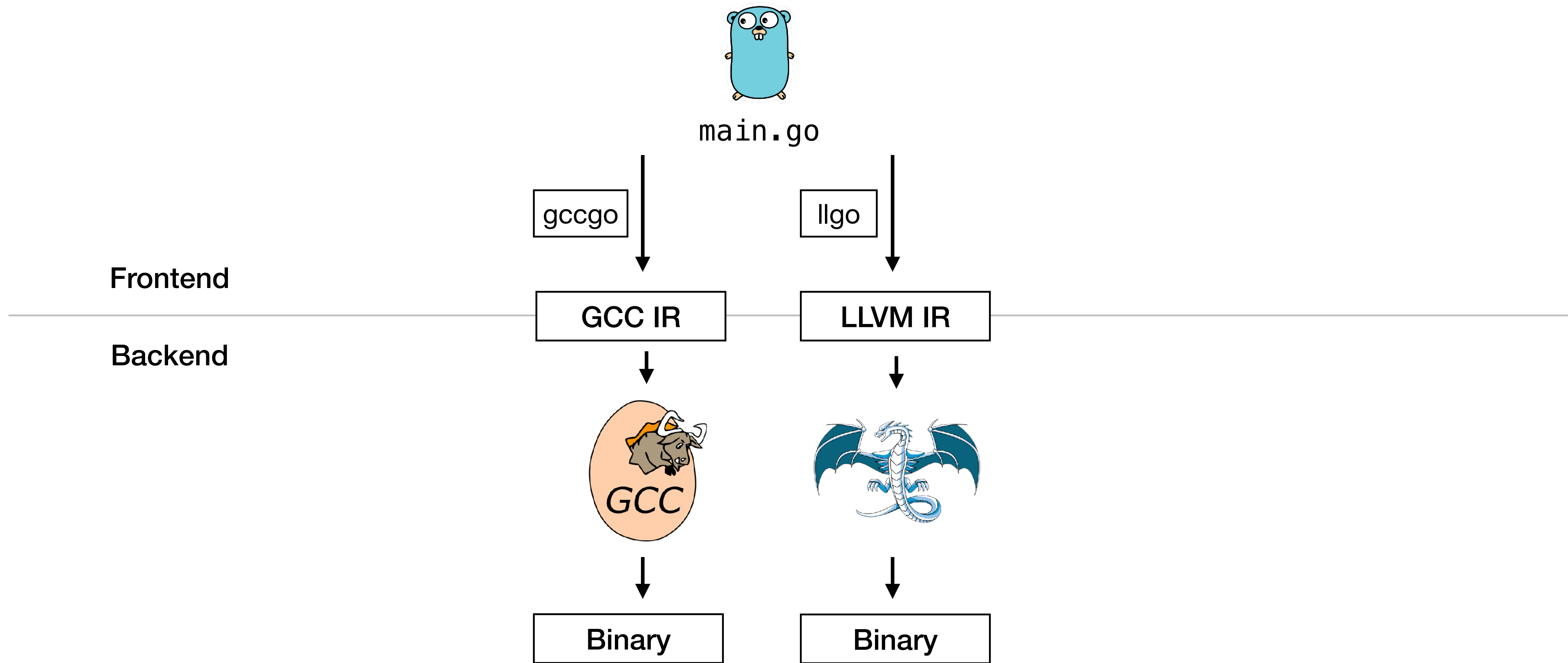
Ilvm (Ilgo)

Dropped from repo in 2020

Demonstrates difficulty of maintaining  
Mentioned for illustrative purposes only

[1] <https://go.dev/doc/install/gccgo>

# Using Different Compiler Backends



# Why Use Other Compiler Backends?

## When to Use

- Using Optimizations not implemented in gc (standard impl.)
- Targeting platforms not supported by gc
- Dynamic Linking of Libraries
- Making sure Implementation doesn't become Specification

# How Golang gets compiled

# Different Golang Compilers

We'll talk about the official compiler



gc (go compiler)

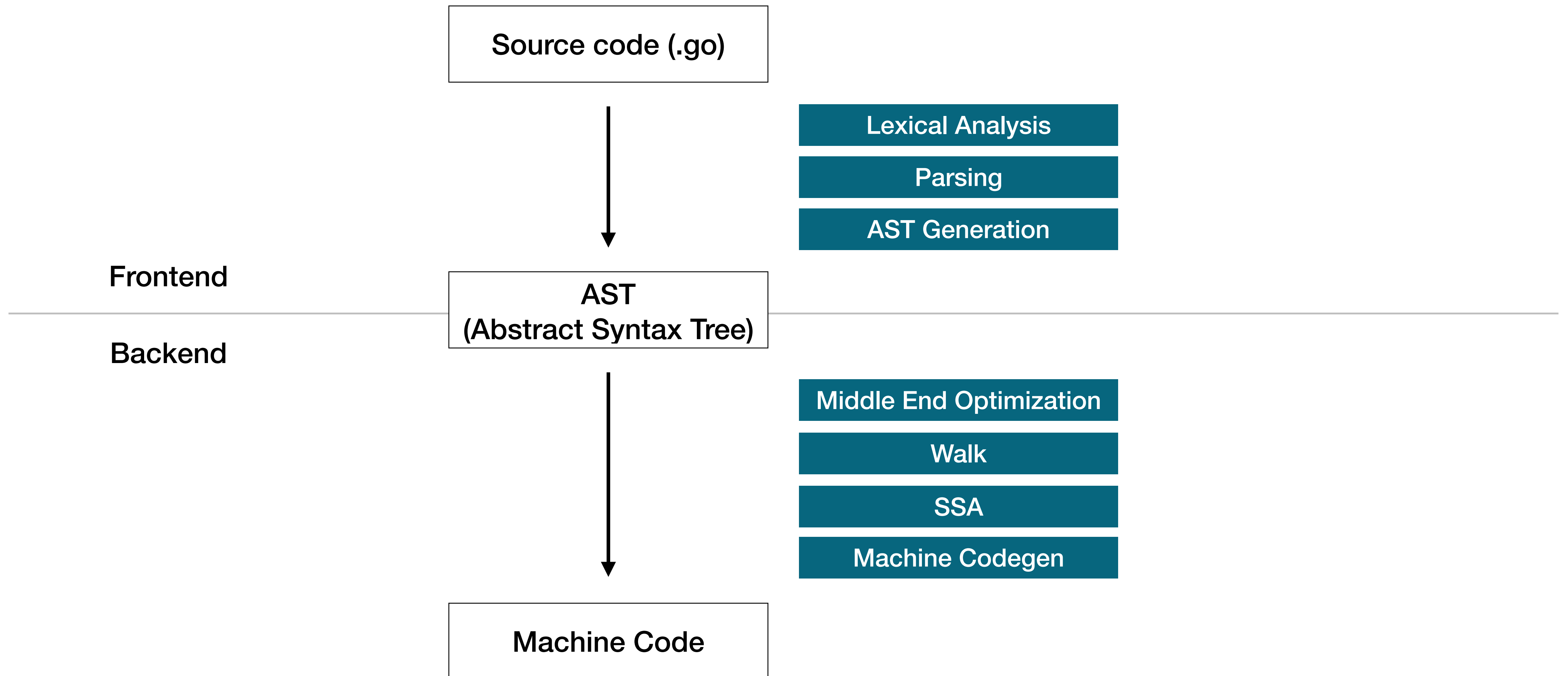


~~gcc (gccgo)~~



~~llvm (llgo)~~

# Golang Compiler Overview





# Tokenization (Lexical Analysis)

- Groups characters into tokens (unit of meaning)
- Takes place in:
  - `src/cmd/compile/internal/syntax/tokens.go` (List of tokens)
  - `src/cmd/compile/internal/syntax/tokenizer.go` (Turns code in tokens)
- List of Tokens

\_EOF  
\_Name  
\_Literal  
\_Operator  
\_AssignOp  
\_IncOp  
\_Assign  
\_Define  
\_Arrow  
\_Star  
\_Lparen

\_Lbrack  
\_Lbrace  
\_Rparen  
\_Rbrack  
\_Rbrace  
\_Comma  
\_Semi  
\_Colon  
\_Dot  
\_DotDotDot  
\_Break

\_Case  
\_Chan  
\_Const  
\_Continue  
\_Default  
\_Defer  
\_Else  
\_Fallthroug  
\_For  
\_Func

\_Go  
\_Goto  
\_If  
\_Import  
\_Interface  
\_Map  
\_Package  
\_Range  
\_Return  
\_Select  
\_Struct

\_Switch  
\_Type  
\_Var  
tokenCount  
Break  
Continue  
Fallthrough  
Goto  
Go  
Defer  
Def

Not  
Recv  
Tilde  
OrOr  
OrOr  
AndAnd  
EqL  
Neq  
Lss  
Leq  
Gtr  
Geq

Add  
Sub  
Or  
Xor  
Mul  
Div  
Rem  
And  
AndNot  
Shl  
Shr

Lexical Analysis

Parsing

AST Generation

Middle End Optimization

Walk

SSA

Machine Codegen

main.go

```
package main
import "fmt"

func main() {
    fmt.Println("Hello, World!")
}
```



main.go tokenized

```
Token: PACKAGE, Literal: package
Token: IDENT, Literal: main
Token: IMPORT, Literal: import
Token: STRING, Literal: "fmt"
Token: FUNC, Literal: func
Token: IDENT, Literal: main
Token: LPAREN, Literal: (
Token: RPAREN, Literal: )
Token: LBRACE, Literal: {
Token: IDENT, Literal: fmt
Token: PERIOD, Literal: .
Token: IDENT, Literal: Println
Token: LPAREN, Literal: (
Token: STRING, Literal: "Hello, World!"
Token: RPAREN, Literal: )
Token: RBRACE, Literal: }
Token: EOF, Literal:
```

Tokenized using "scanner" pkg

# Parsing (AST Generation)

Lexical Analysis

Parsing

AST Generation

Middle End Optimization

Walk

SSA

Machine Codegen

- Takes place in:
  - `src/cmd/compile/internal/syntax/parser.go` (Parsing)
  - `src/cmd/compile/internal/syntax/nodes.go` (Noder)
- Takes tokens and builds AST (Abstract Syntax Tree)

# Parsing (AST Generation)

## What is an AST?

- AST is representation of source file in tree form
- Nodes correspond to elements
  - expressions
  - declarations
  - statements

Lexical Analysis

Parsing

AST Generation

Middle End Optimization

Walk

SSA

Machine Codegen

# Parsing (AST Generation)

## Example of an AST

Lexical Analysis

Parsing

AST Generation

Middle End Optimization

Walk

SSA

Machine Codegen

```
x := 2 + 3 * 6
```

# Parsing (AST Generation)

## Example of an AST

Lexical Analysis

Parsing

AST Generation

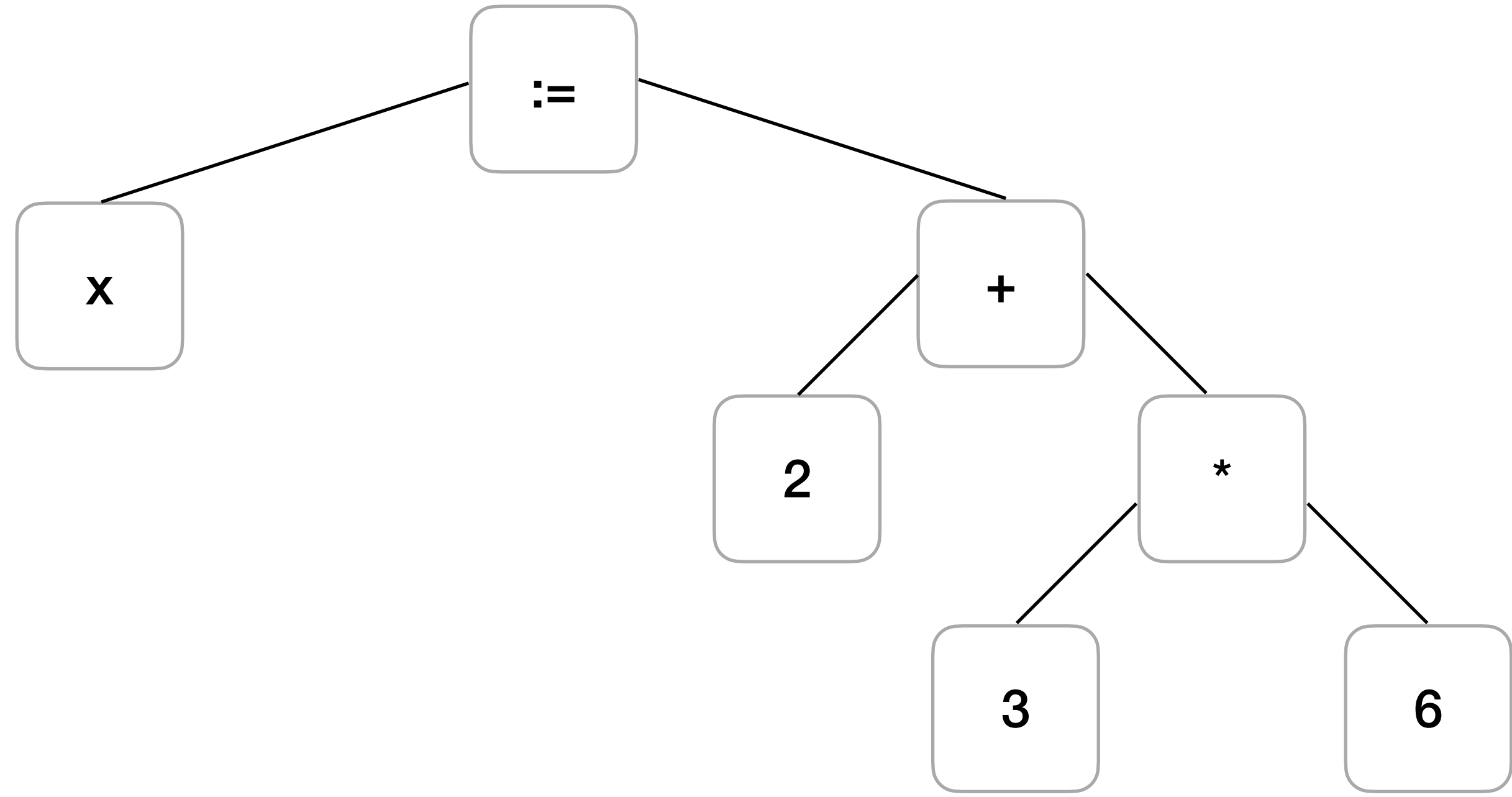
Middle End Optimization

Walk

SSA

Machine Codegen

```
x := 2 + 3 * 6
```



# Parsing (AST Generation)

## “Hello World” in AST

Lexical Analysis

Parsing

AST Generation

Middle End Optimization

Walk

SSA

Machine Codegen

main.go

```
fmt.Println("Hello, World!")
```



main.go AST

```
Body: *ast.BlockStmt {  
  Lbrace: 4:6  
  List: []ast.Stmt {  
    *ast.ExprStmt {  
      X: *ast.CallExpr {  
        Fun: *ast.SelectorExpr {  
          X: *ast.Ident {  
            NamePos: 5:5  
            Name: "fmt"  
          }  
          Sel: *ast.Ident {  
            NamePos: 5:9  
            Name: "Println"  
          }  
        }  
        Lparen: 5:15  
        Args: []ast.Expr {  
          *ast.BasicLit {  
            ValuePos: 5:16  
            Kind: STRING  
            Value: "\"Hello, World!\""  
          }  
        }  
        Ellipsis: 0:0  
        Rparen: 5:30  
      }  
    }  
  }  
  Rbrace: 6:1  
}
```

# Middle End Optimization

Lexical Analysis

Parsing

AST Generation

Middle End Optimization

Walk

SSA

Machine Codegen

- Perform optimization “passes”
- Takes place in:
  - `src/cmd/compile/internal/deadcode` (Dead code elimination)
  - `src/cmd/compile/internal/inline` (Function call inlining)
  - `src/cmd/compile/internal/devirtualize` (Devirtualize known interface method call)
  - `src/cmd/compile/internal/escape` (Escape analysis)



# Middle End Optimization

## Examples of Optimizations

1. Inlining
2. Escape Analysis

Lexical Analysis

Parsing

AST Generation

**Middle End Optimization**

Walk

SSA

Machine Codegen

# Middle End Optimization

## 1. Inlining - Highlighting Optimizations

- Embed called functions into caller functions
- Saves overhead of starting new stack with new function call
- Budget of 80 nodes (If called function exceeds 80 nodes, not inlined)

Lexical Analysis

Parsing

AST Generation

Middle End Optimization

Walk

SSA

Machine Codegen

# Middle End Optimization

## 2. Escape Analysis - Highlighting Optimizations

- “Escape” variables from stack territory to heap territory
  - Some variables “outlive” current stack frame
- Makes variables accessible even after function disappear

# Walk

Lexical Analysis

Parsing

AST Generation

Middle End Optimization

Walk

SSA

Machine Codegen

- Takes place in: `cmd/compile/internal/walk`
- Decompose complex statements
- Desugar high-level Go construct
  - Examples
    - switch statement → binary search or jump tables
    - map and channel → runtime calls

# SSA Conversion

Lexical Analysis

Parsing

AST Generation

Middle End Optimization

Walk

**SSA**

Machine Codegen

- SSA (Single-Static Assignment)
- Takes place in:
  - `src/cmd/compile/internal/ssagen` (Convert IR to SSA)
  - `src/cmd/compile/internal/ssa` (SSA passes and rules)

# SSA Conversion

## What is SSA (Single Static Assignment) ?

- Lower-level IR
- All Variables changed into Immutable Variables (hence “Single Static”)
- Easier to optimize code in SSA form
- Introduced in Go 1.7 (amd64) and 1.8 (all other archs)
- Technique also used by GCC and LLVM

Lexical Analysis

Parsing

AST Generation

Middle End Optimization

Walk

SSA

Machine Codegen

# SSA Conversion

## Example of SSA (Single Static Assignment) Conversion

Lexical Analysis

Parsing

AST Generation

Middle End Optimization

Walk

SSA

Machine Codegen

```
//before
```

```
a := 1
```

```
a = 2
```

```
b = a + 1
```



rename vars

```
//after
```

```
a1 := 1
```

```
a2 := 2
```

```
b2 = a2 + 1
```

a is assigned value *twice*

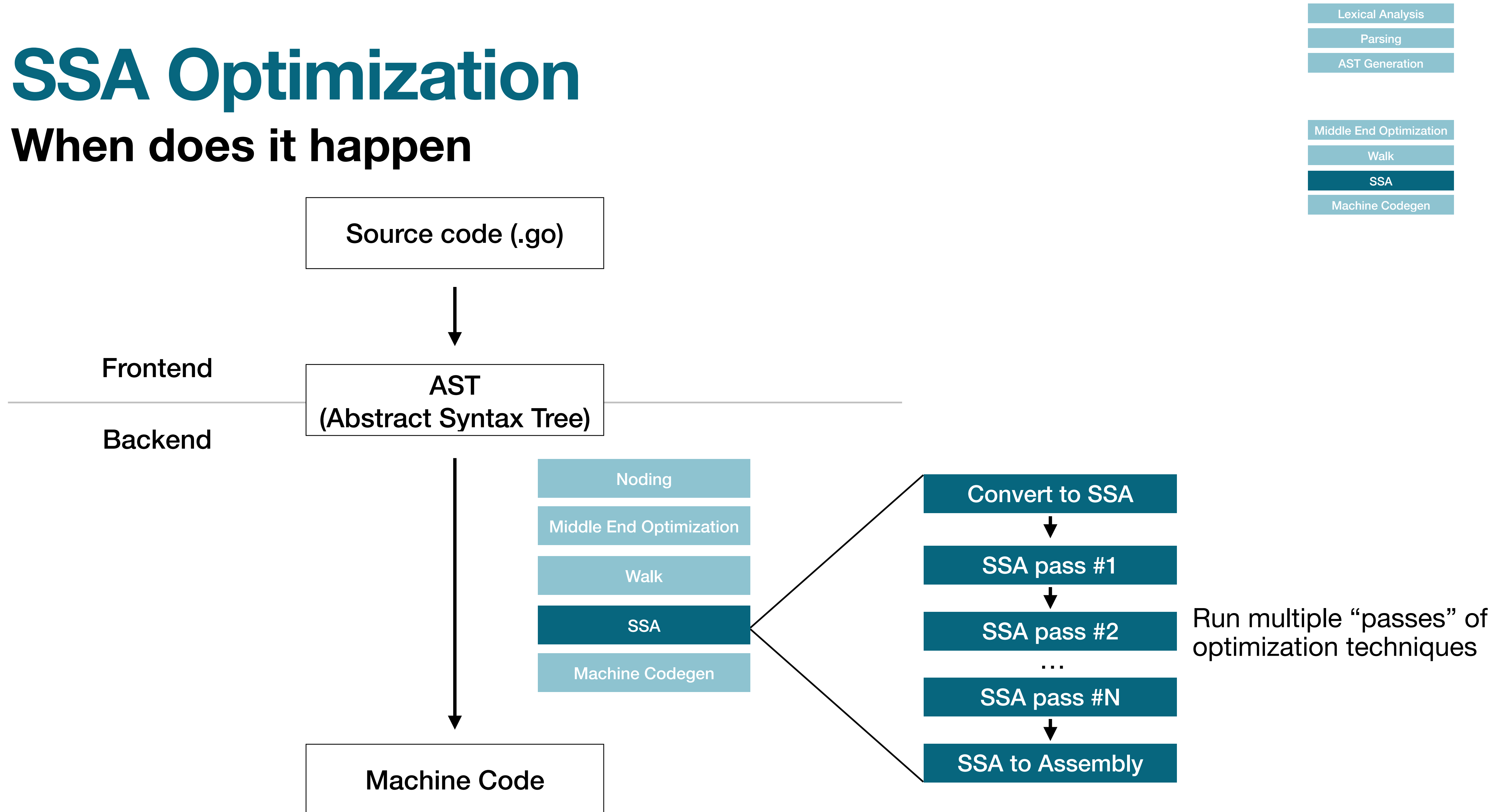
a1 is assigned value *once*

a2 is assigned value *once*

No variable gets value assigned > 1 times

# SSA Optimization

## When does it happen





# SSA Optimization

## Kinds of Optimization Passes

Lexical Analysis

Parsing

AST Generation

Middle End Optimization

Walk

SSA

Machine Codegen

PhiElim  
CopyElim  
ShortCircuit  
Decompose  
CSE  
PhiOpt

NilCheckElim  
Prove  
BCE  
Loop  
Fuse  
DSE

WriteBarrier  
insertLoopRes  
chedChecks  
Critical  
LikelyAdjust  
Layout

FlagAlloc  
RegAlloc  
LoopRotate

# SSA Optimization

## Kinds of Optimization Passes

Lexical Analysis

Parsing

AST Generation

Middle End Optimization

Walk

SSA

Machine Codegen

PhiElim  
CopyElim  
ShortCircuit  
Decompose  
CSE  
PhiOpt

NilCheckElim  
Prove  
BCE  
Loop  
Fuse  
DSE

WriteBarrier  
insertLoopRes  
chedChecks  
Critical  
LikelyAdjust  
Layout

FlagAlloc  
RegAlloc  
LoopRotate

# SSA Optimization

## Ex. Common Subexpression Elimination

(Reduce redundant expressions)

Without SSA

```
y = x + 2
```

```
...
```

```
z = x + 2
```

Need to traverse code between  
`y = ...` and `z = ...`

With SSA

```
y = x + 2
```

```
...
```

```
z = y
```

No need to traverse code since  
variable is assigned value once

# SSA Conversion

## How to inspect Golang SSA

- GOSSAFUNC
  - Build flag to output SSA compilation in HTML
  - Shows code through each step of compilation
  - invoke by `GOSSAFUNC=main go build main.go`

Lexical Analysis

Parsing

AST Generation

Middle End Optimization

Walk

SSA

Machine Codegen

# SSA Conversion

## See your intermediate code in GOSSAFUNC

Lexical Analysis

Parsing

AST Generation

Middle End Optimization

Walk

SSA

Machine Codegen

The screenshot displays the GOSSAFUNC SSA conversion tool interface. It is divided into several vertical panes:

- main**: Contains the source code for the `start` function, including variable declarations and function calls like `Printf`, `scanf`, and `strcpy`.
- AST**: Shows the abstract syntax tree representation of the source code.
- opt**: Displays the code after various optimization passes, including `early deadcode`, `early phiins`, and `early phiins`.
- dse**: Shows the code after Dead Store Elimination (DSE) and other dead code removal passes.
- addressing modes**: Shows the code after addressing mode optimization.
- late lower**: Shows the code after late lowering of instructions.
- lowered deadcode for cse**: Shows the code after lowered dead code removal for Constant Side Effects (CSE).
- lowered cse**: Shows the code after lowered CSE.
- elim unread autos**: Shows the code after eliminating unread automatic variables.
- tighten tuple selectors**: Shows the code after tightening tuple selectors.
- lowered deadcode**: Shows the code after another round of lowered dead code removal.
- checkLower**: Shows the code after the `checkLower` pass.
- late phielim**: Shows the code after late phi elimination.
- late copyelim**: Shows the code after late copy elimination.
- tighten**: Shows the code after the `tighten` pass.
- late deadcode**: Shows the code after late dead code removal.
- critical**: Shows the code after the `critical` pass.
- phi tighten**: Shows the code after phi tightening.
- likelyadjust**: Shows the code after likely adjust.
- layout**: Shows the code after the `layout` pass.
- schedule**: Shows the code after the `schedule` pass.
- late nilcheck**: Shows the code after late nil checking.
- flagalloc**: Shows the code after flag allocation.
- regalloc**: Shows the code after register allocation.
- genssa**: Shows the final SSA code, including the `main` function and the `start` function, with SSA variables and phi nodes.

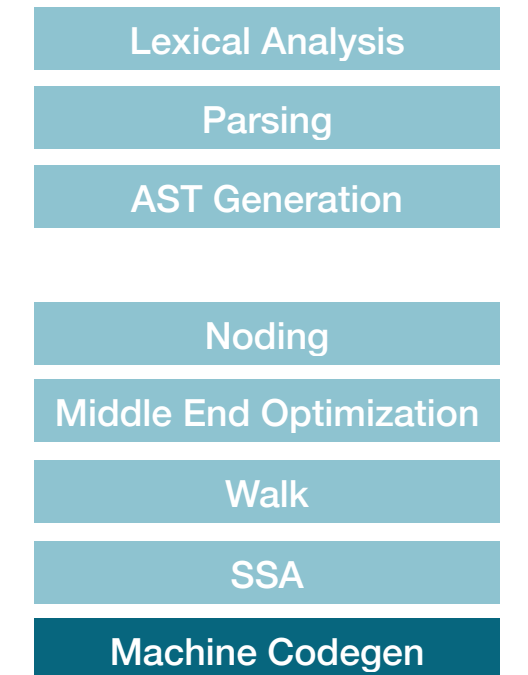
# Machine Code Generation

Lexical Analysis
Parsing
AST Generation
Noding
Middle End Optimization
Walk
SSA
Machine Codegen

- Turn the SSA representation into CPU Instruction
- Takes place in:
  - `src/cmd/compile/internal/ssa` (SSA lowering + arch-specific passes)
  - `src/cmd/compile/internal/obj` (Machine code generation)

# Machine Code Generation

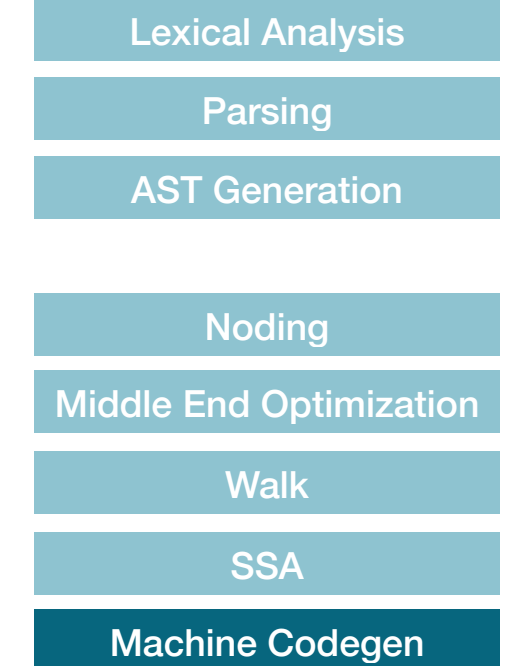
## Lowering Pass



- “lowers” the abstraction of SSA code from machine-independent into machine-specific
- ex. ARM64: memory operands are possible, so maybe use load-store ops

# Machine Code Generation

## Examples of Final Code Passes



- Dead code elimination
- Move value closer to point of usage
- Remove local variables that are never read
- Register allocation
- Stack frame layout: assign stack offsets to local variables
- Pointer liveness analysis
  - compute which on-stack pointers are alive at each safepoint in Garbage Collector



# Machine Code Generation

## Object File Writing

- Transform SSA into obj.Prog instructions
- Write out final object file
  - contains reflect data, export data, and debugging info

Lexical Analysis

Parsing

AST Generation

Noding

Middle End Optimization

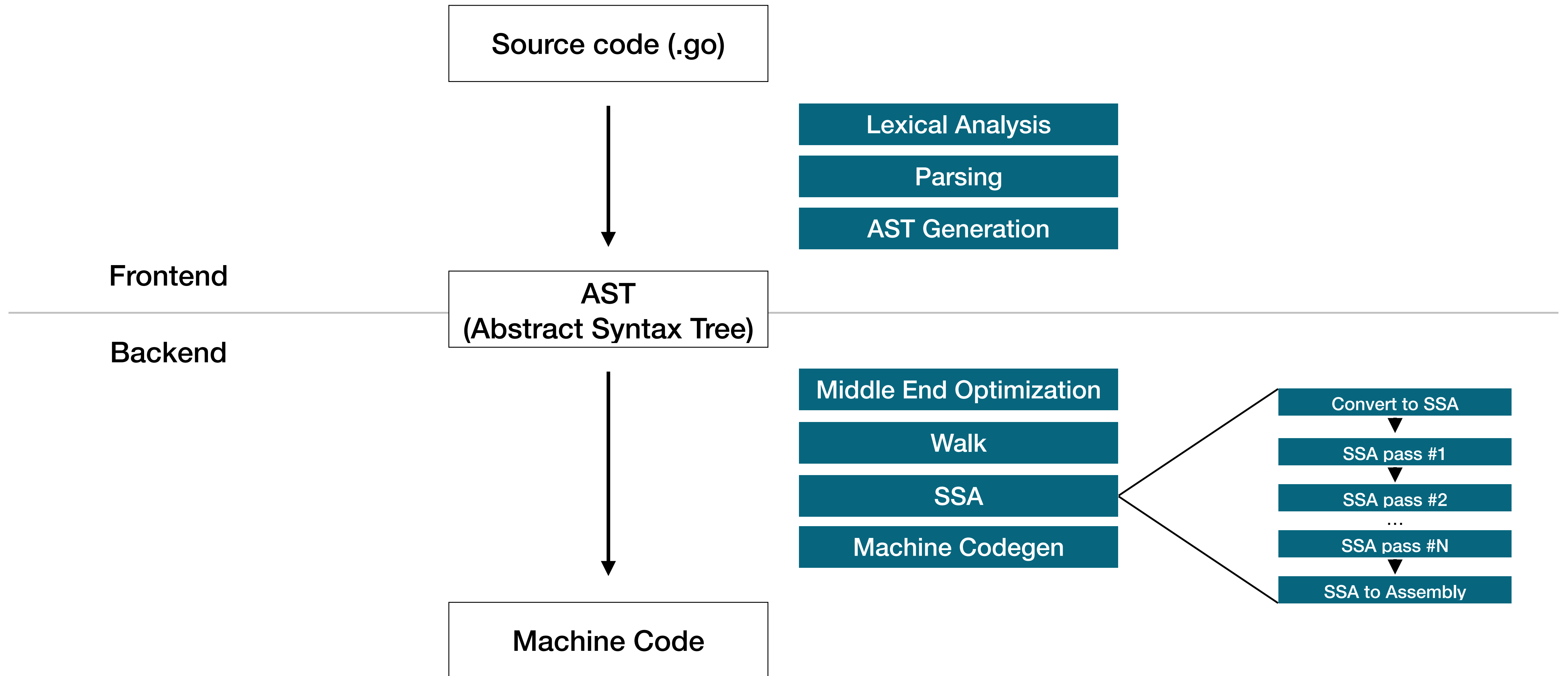
Walk

SSA

Machine Codegen

# Conclusion

# What We Covered



# Resources

## Compiler Internals

Introduction to the Go compiler (<https://go.dev/src/cmd/compile/README>)

A Journey With GO (<https://medium.com/a-journey-with-go>)

Eli Bendersky (<https://eli.thegreenplace.net/tag/go>)

Hacking Go Compiler Internals (<https://speakerdeck.com/moriyoshi/hacking-go-compiler-internals-2nd-season>)

## SSA

Introduction to the Go compiler's SSA backend (<https://go.dev/src/cmd/compile/internal/ssa/README>)

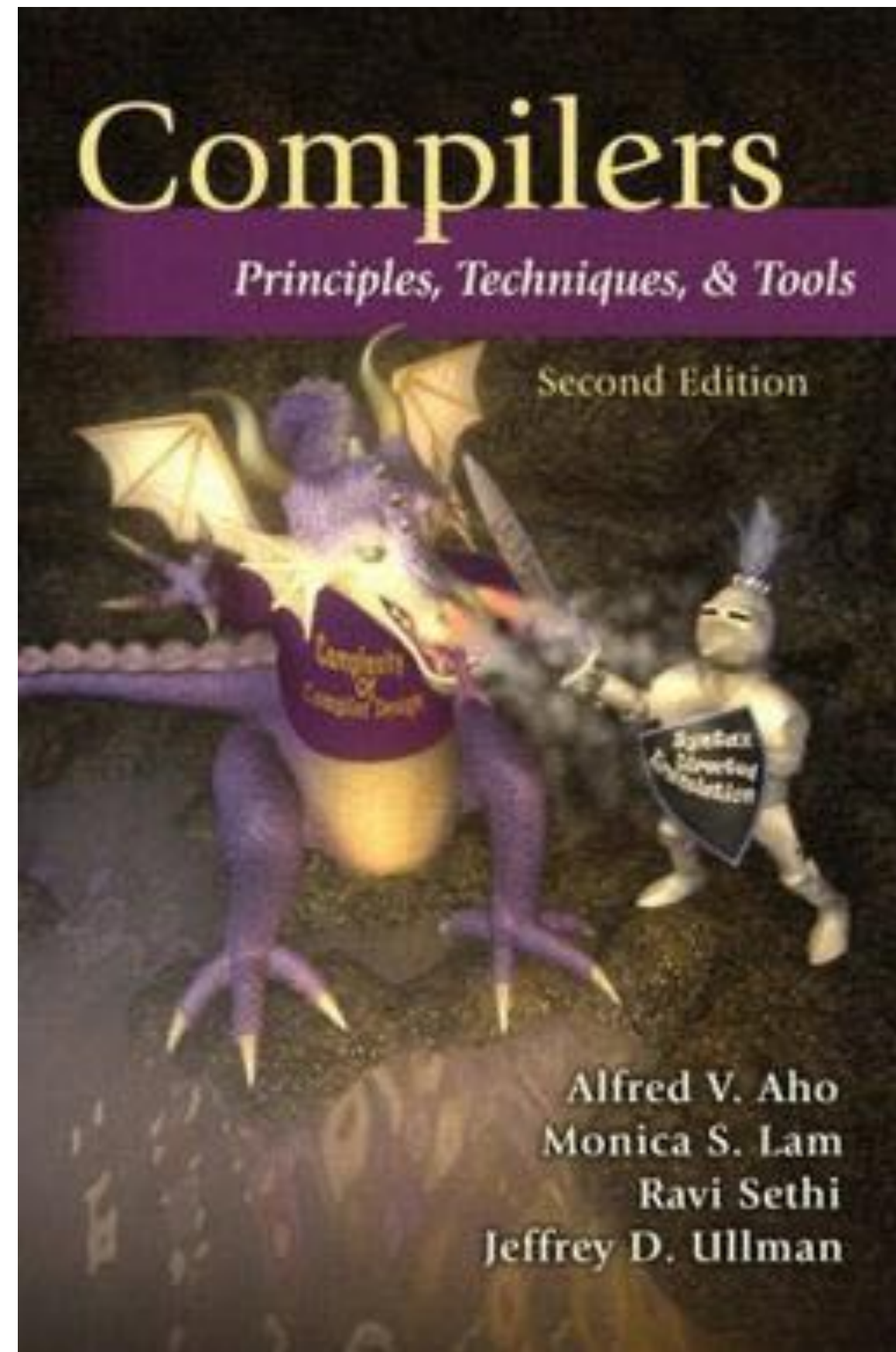
Keith Randall - Generating Better Machine Code with SSA (<https://youtu.be/uTMvKVma5ms>)

Inlining optimisations in Go (<https://dave.cheney.net/2020/04/25/inlining-optimisations-in-go>)

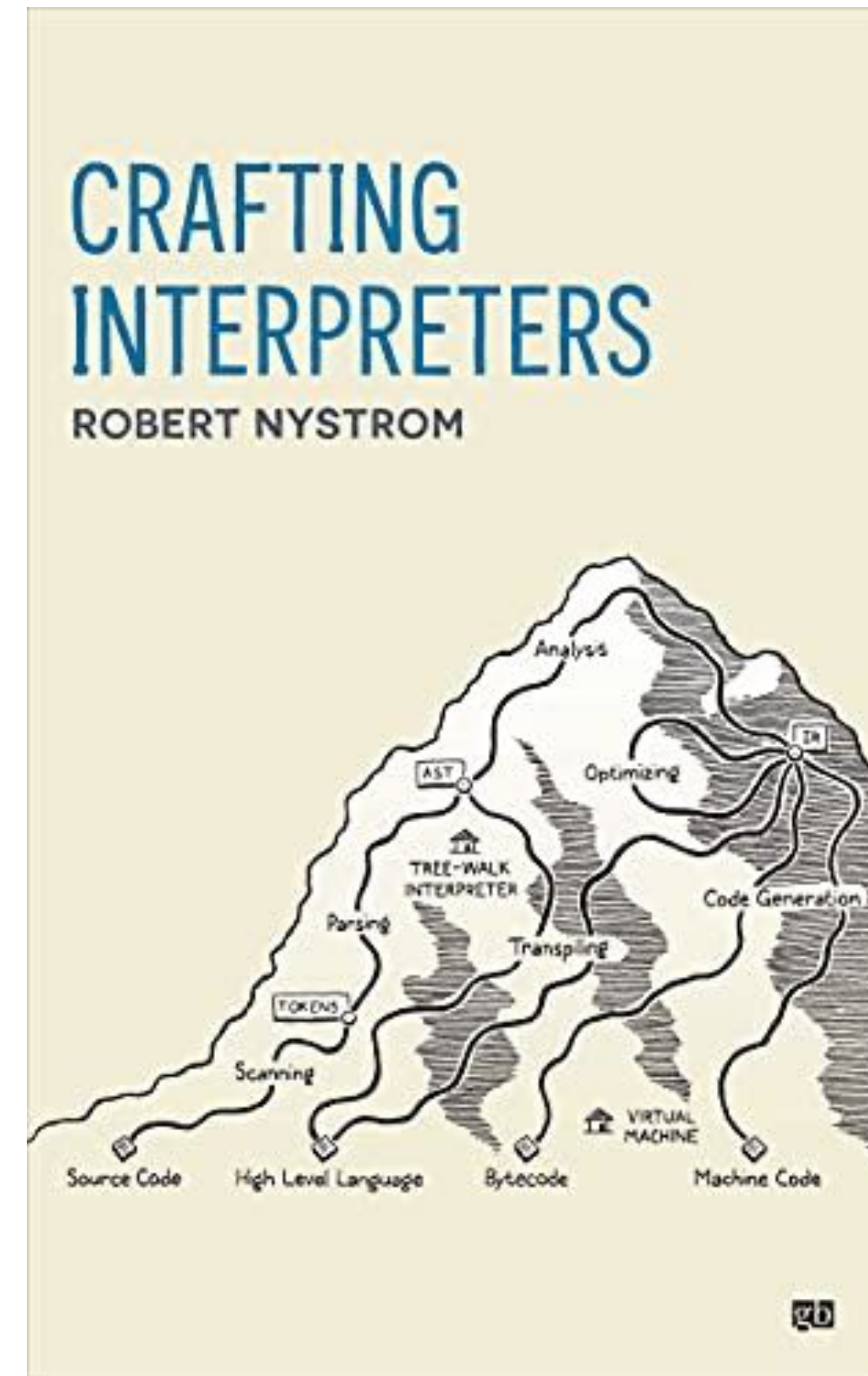
## GCCGO

Gccgo in GCC 4.7.1 (<https://go.dev/blog/gccgo-in-gcc-471>)

# Learn More on Compilers



Aho, A., Sethi, R. and Ullman, J. (2006).  
Compilers: Principles, Techniques, and Tools



Nystrom, R. (2021). Crafting interpreters

# exit()

➡ Thank you for listening!