McSema: Static Translation of X86 Instructions to LLVM

ARTEM DINABURG, ARTEM@TRAILOFBITS.COM

ANDREW RUEF, ANDREW@TRAILOFBITS.COM

“This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA).”
“The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government.”
Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

About Us

Artem

- Security Researcher
- blog.dinaburg.org

Andrew

- PhD Student, University of Maryland
- Trail of Bits
- www.cs.umd.edu/~awruef

“This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA).”
“The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government.”
Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

What is McSema?

Translate existing programs into a representation that can be easily manipulated and reasoned about.

The representation we chose is LLVM IR.

“This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA).”
“The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government.”
Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

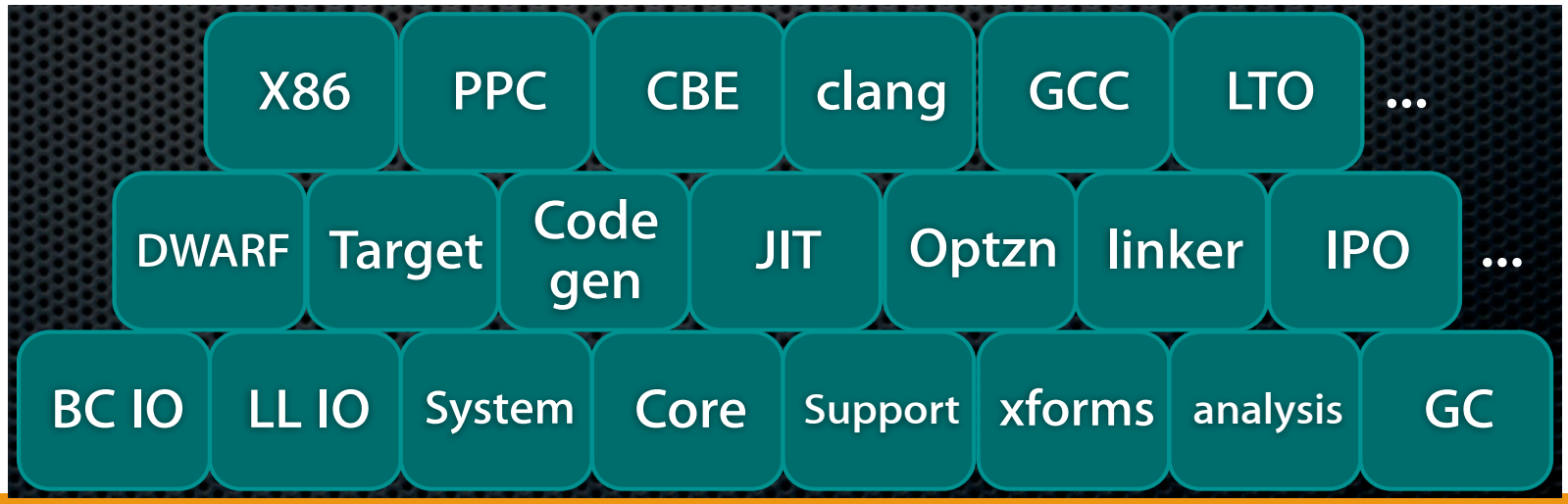
What is LLVM?

Modern Optimizing Compiler Infrastructure

- Infrastructure first, compiler second

Easy to learn and modify (for a compiler)

Very permissive licensing



"This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA)."
"The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government."
Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

What is LLVM IR?

Like a higher level assembly language

Typed, Static Single Assignment

Simplifies program analysis and transformation

```
define i32 @main(i32 %argc, i8** %argv) {
  %1 = alloca i32, align 4
  %2 = alloca i32, align 4
  %3 = alloca i8**, align 8
  store i32 0, i32* %1
  store i32 %argc, i32* %2, align 4
  store i8** %argv, i8*** %3, align 8
  %4 = call i32 (i8*, ...)* @printf(… <omitted>)
  ret i32 0 }
```

"This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA)."
"The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government."
Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Why translate x86 to LLVM IR?

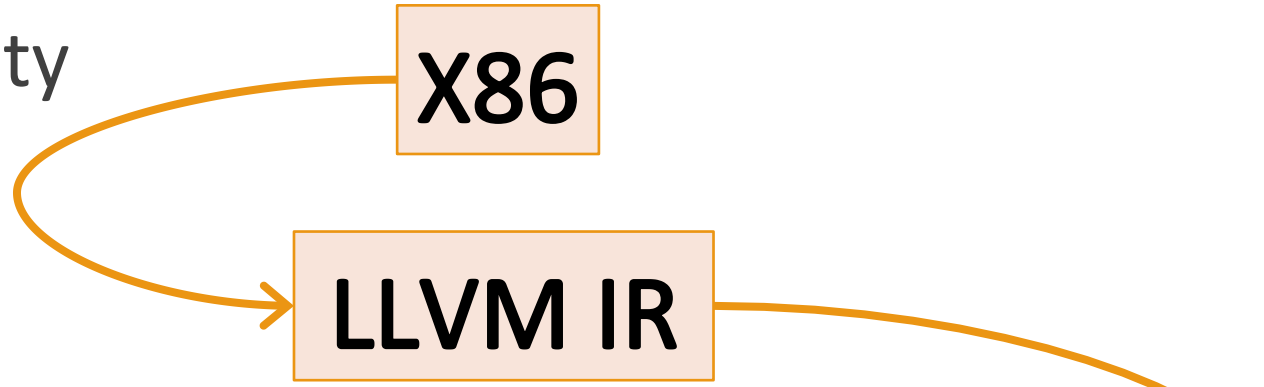
Use all existing LLVM tools

- Optimization
- Test Generation
- Model Checking

“This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA).”
“The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government.”
Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Why translate x86 to LLVM IR?

Portability

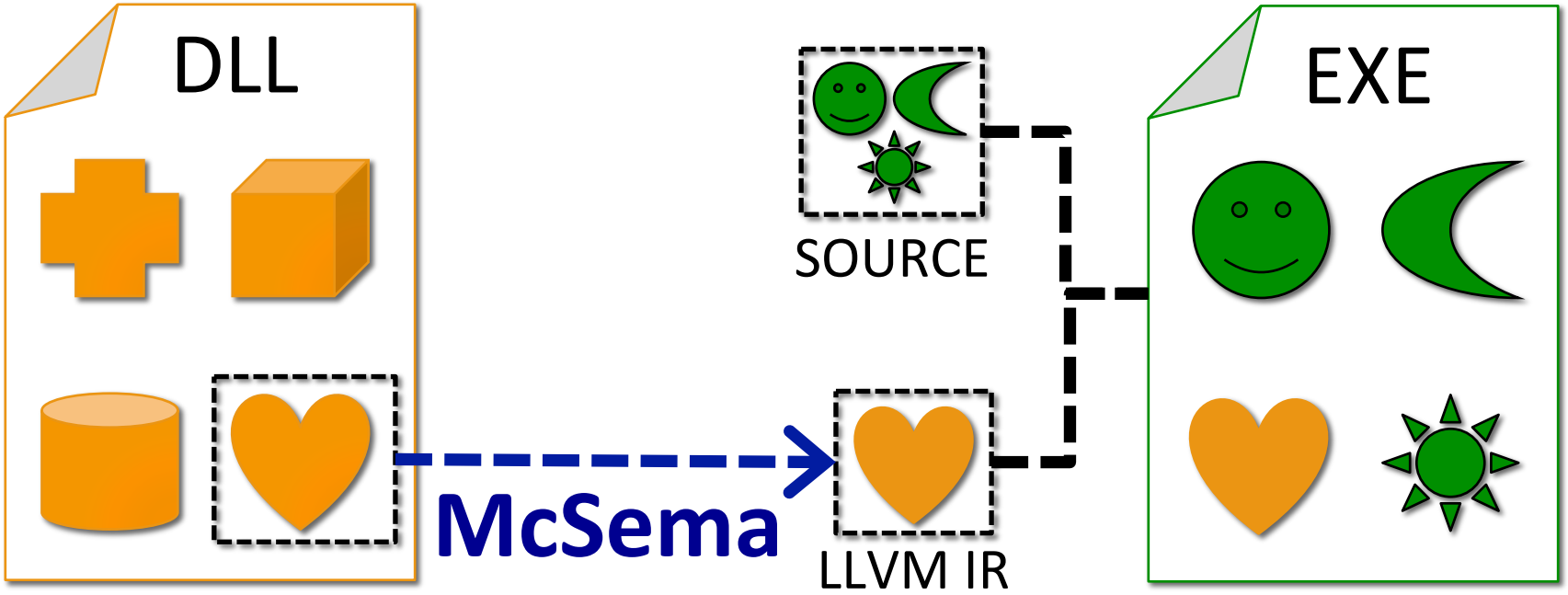


aarch64	arm	hexagon	mips	mips64
msp430	nvptx	nvptx64	ppc32	ppc64
r600	sparc	sparcv9	systemz	thumb
	x86	x86-64	xcore	

"This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA)."
"The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government."
Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Why translate x86 to LLVM IR?

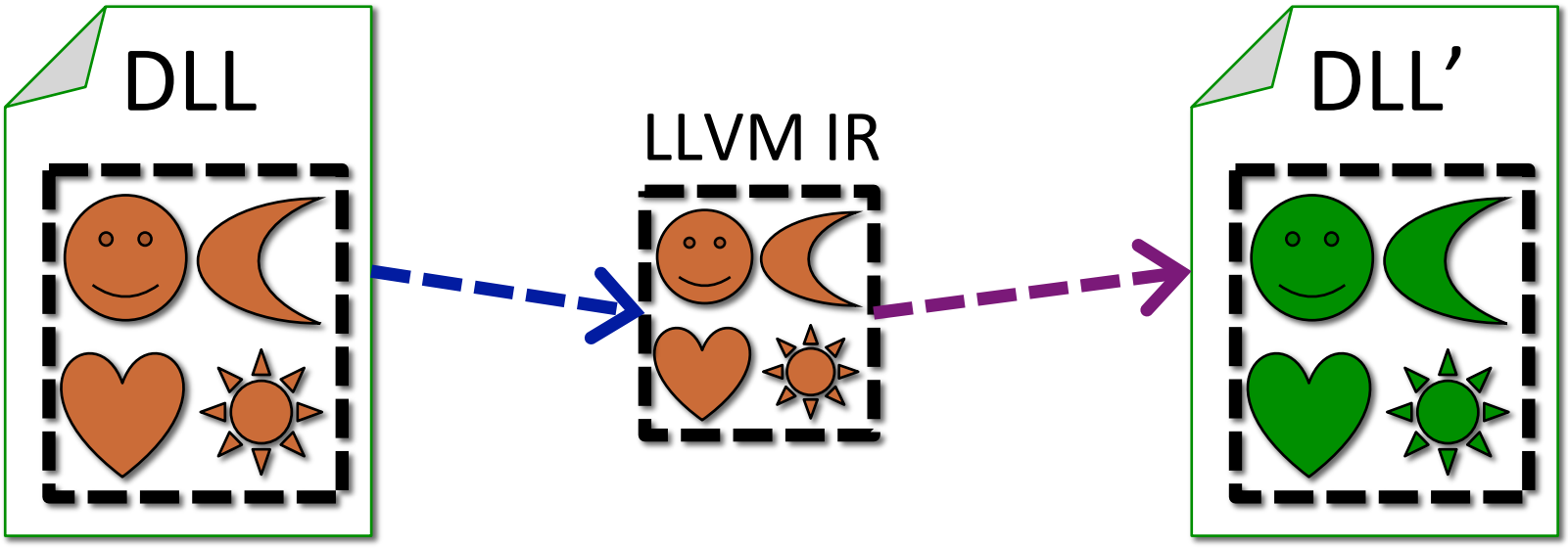
Foreign Code Integration and Re-Use



"This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA)."
"The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government."
Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Why translate x86 to LLVM IR?

Add obfuscation and/or security to existing code.



"This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA)."
"The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government."
Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Demo 1

“This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA).”
“The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government.”
Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Prior Work

- Dagger
- Second Write
- Fracture
  - Draper Lab
- BAP
  - CMU

“This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA).”
“The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government.”
Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Why McSema

Open Source

Documentation and Unit Tests

FPU and SSE Support (incomplete)

Modular architecture

- Separate control flow recovery from translation
- Designed to translate code from arbitrary sources
- Control flow graphs specified as Google protocol buffers

"This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA)."
"The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government."
Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Open Source

McSema is DARPA funded.

It is open sourced.

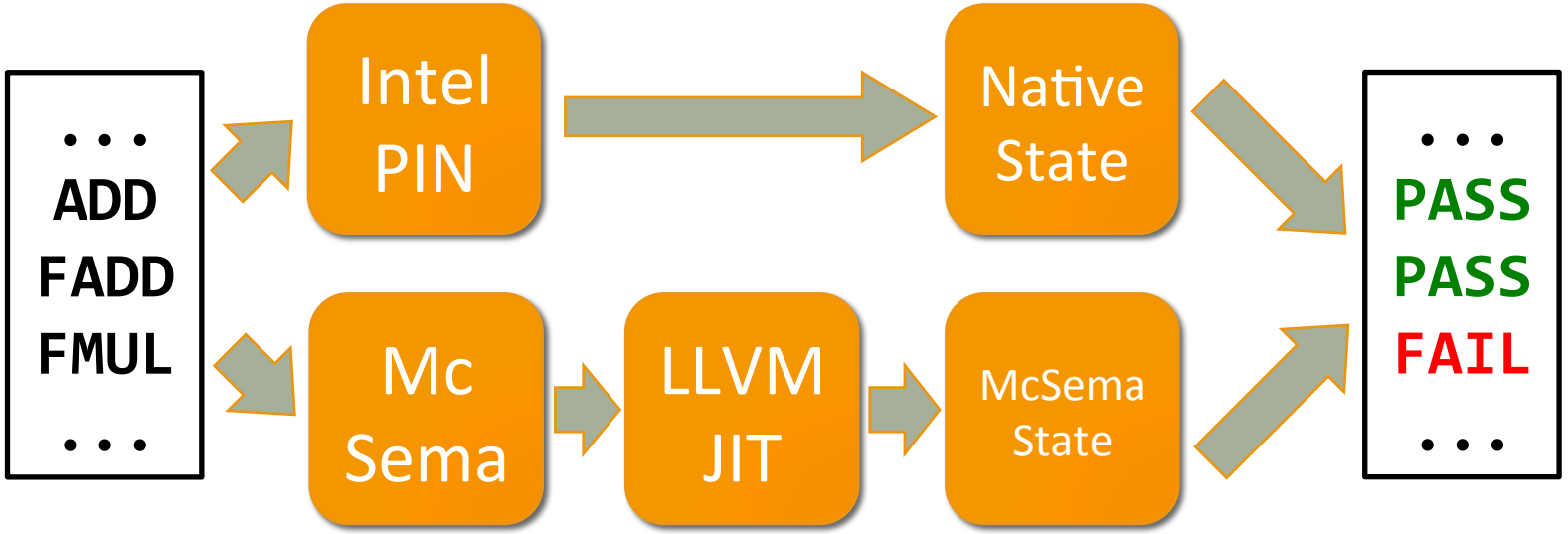
Permissively licensed.

“This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA).”
“The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government.”
Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Unit Tests

Google test powered unit test for instruction semantics

Compares McSema CPU context to native CPU state



"This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA)."
"The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government."
Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

FPU And SSE Support

Nearly Complete FPU Support

- Many instructions
- Some core issues remain:
  - Precision Control
  - Rounding Control

SSE Support is architecturally implemented

- Register state is complete
- Needs more instructions

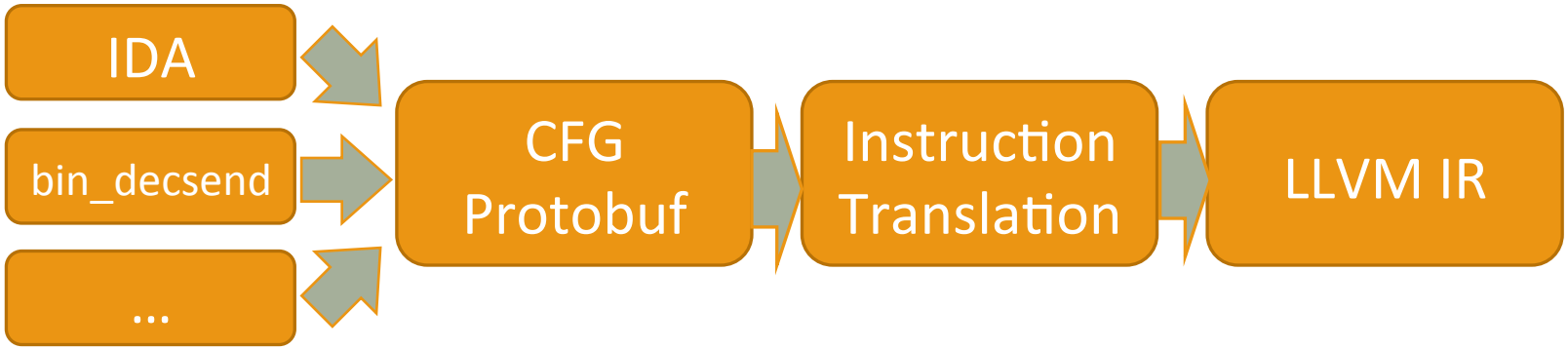
"This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA)."
"The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government."
Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

McSema Architecture

Separate control flow recovery from translation

Designed to translate code from arbitrary sources

Control flow graphs specified as Google protocol buffers



"This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA)."

Control Flow Recovery

1) Start at the entry point

2) BFS through all discovered basic blocks

3) ???

4) Recover CFG

What could go wrong???

“This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA).”
“The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government.”
Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

CFG Recovery Challenges

Indirect Calls

- `JMP EAX`

Jump Tables

- `JMP [EAX*4+OFFSET]`

Mixed Code and Data

- `0x40040: RET`
- `0x40041: db 'H','e','l','l','o',' ','W','o','r','l','d','\0'`
- `0x40056: PUSH EBP`

Constant, Data, or Code?

- `0x40000: MOV EAX, 0x40040`

"This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA)."
"The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government."
Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

CFG Recovery Solutions

Relocation Entries

- Reliably identify pointers
- Required for ASLR on Windows

API Domain Knowledge

- Argument types to help solve code/data question
- Need to know about APIs later anyway

"This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA)."
"The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government."
Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

CFG Recovery Solutions

Let IDA do it!

- McSema comes with an IDAPython script to dump the CFG from IDA

Why IDA

- Countless man-hours spent on CFG recovery
- The CFG will be at least as good as what you see in IDA

"This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA)."
"The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government."
Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

CFG Recovery Solutions

In the future

- CFG recovery via symbolic execution
- Static call resolution drastically improves binary size
- Even external code vs. translated code would be a big improvement

“This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA).”
“The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government.”
Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Instruction Translation

Modeling the CPU & Memory

- Instructions are operations on CPU and memory context

Modeling Functions

- Every function operates on a CPU Context

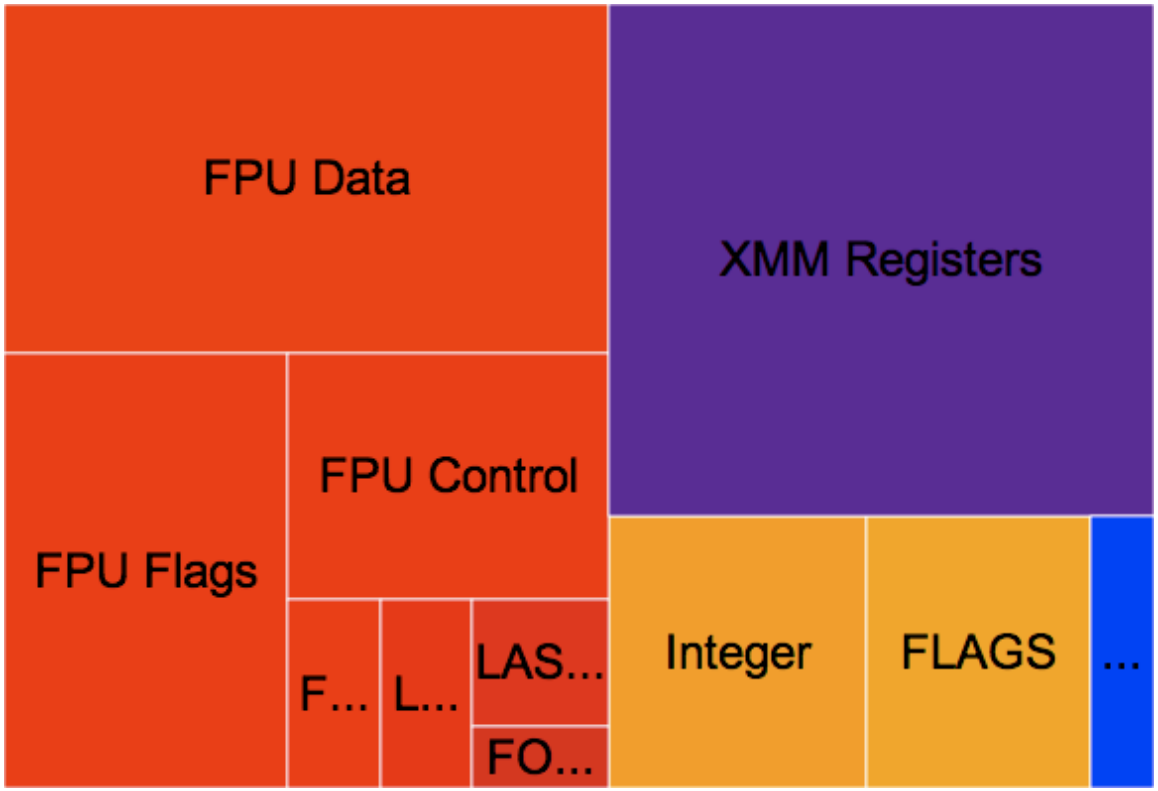
Edge Cases

- Callbacks
- Externals
- `CALL REG/MEM`

“This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA).”
“The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government.”
Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Instruction Translation: CPU

Model as operations on CPU context



"This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA)."
"The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government."
Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Instruction Translation: CPU

Explicit Flags Modification

```
// Compute AF.
WriteAFAddSub<width>(b, addRes, lhs, rhs);
// Compute SF.
WriteSF<width>(b, addRes);
// Compute ZF.
WriteZF<width>(b, addRes);
// Compute OF.
WriteOFAdd<width>(b, addRes, lhs, rhs);
// Compute PF.
WritePF<width>(b, addRes);
// Compute CF.
WriteCFAdd<width>(b, addRes, lhs);
```

Demo 2

“This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA).”

Instruction Translation: Memory Model

Manipulates actual memory

Stack pointer is set to a translator stack

Stack variable recovery would be ideal

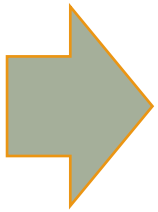
- Create LLVM IR `alloca` values for function stack locals
- Not always possible for sound variable recovery

“This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA).”
“The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government.”
Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Instruction Translation: Functions

Spill Context, Translate, Store Context

```
F(A,B):
   EAX = ESP[-4]
   EBX = ESP[-8]
   EAX += EBX
END
```



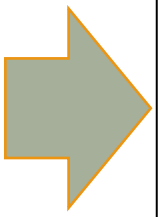
```
TRANSLATED_F(RegContext):
   VAR_EAX = RegContext.EAX
   VAR_EBX = RegContext.EBX
   VAR_ESP = RegContext.ESP
   VAR_EAX = VAR_ESP[-4]
   VAR_EBX = VAR_ESP[-8]
   VAR_EAX += VAR_EBX
   RegContext.EAX = VAR_EAX
   RegContent.EBX = VAR_EBX
   RegContent.ESP = VAR_ESP
END
```

"This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA)."
"The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government."
Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Instruction Translation: Lazy Translation

Let the optimizer make it better!

```
TRANSLATED_F(RegContext):
   VAR_EAX = RegContext.EAX
   VAR_EBX = RegContext.EBX
   VAR_ESP = RegContext.ESP
   VAR_EAX = VAR_ESP[-4]
   VAR_EBX = VAR_ESP[-8]
   VAR_EAX += VAR_EBX
   RegContext.EAX = VAR_EAX
   RegContent.EBX = VAR_EBX
   RegContent.ESP = VAR_ESP
END
```



```
OPTIMIZED_F(RegContext):
   VAR_ESP = RegContext.ESP
   VAR_EAX = VAR_ESP[-4]
   VAR_EBX = VAR_ESP[-8]
   RegContext.EAX =
      VAR_EAX + VAR_EBX
   RegContent.EBX = VAR_EBX
END
```

"This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA)."
"The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government."
Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Instruction Translation: Externals

Parse Windows DLLs to extract API signatures

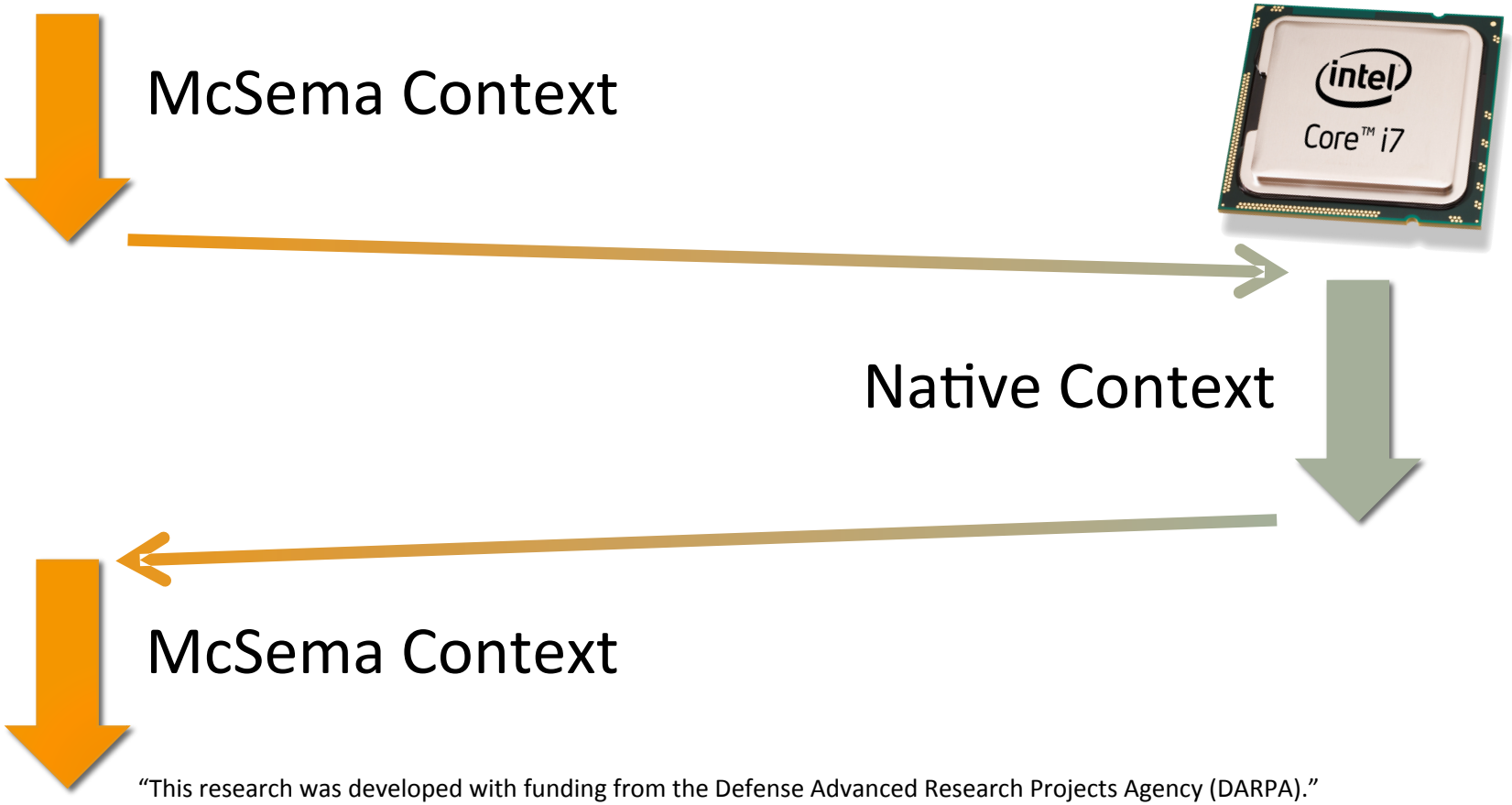
- Simple text-based format
- Easy to add custom mappings

Match import names

Emit as an `extern` function in LLVM IR

“This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA).”
“The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government.”
Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

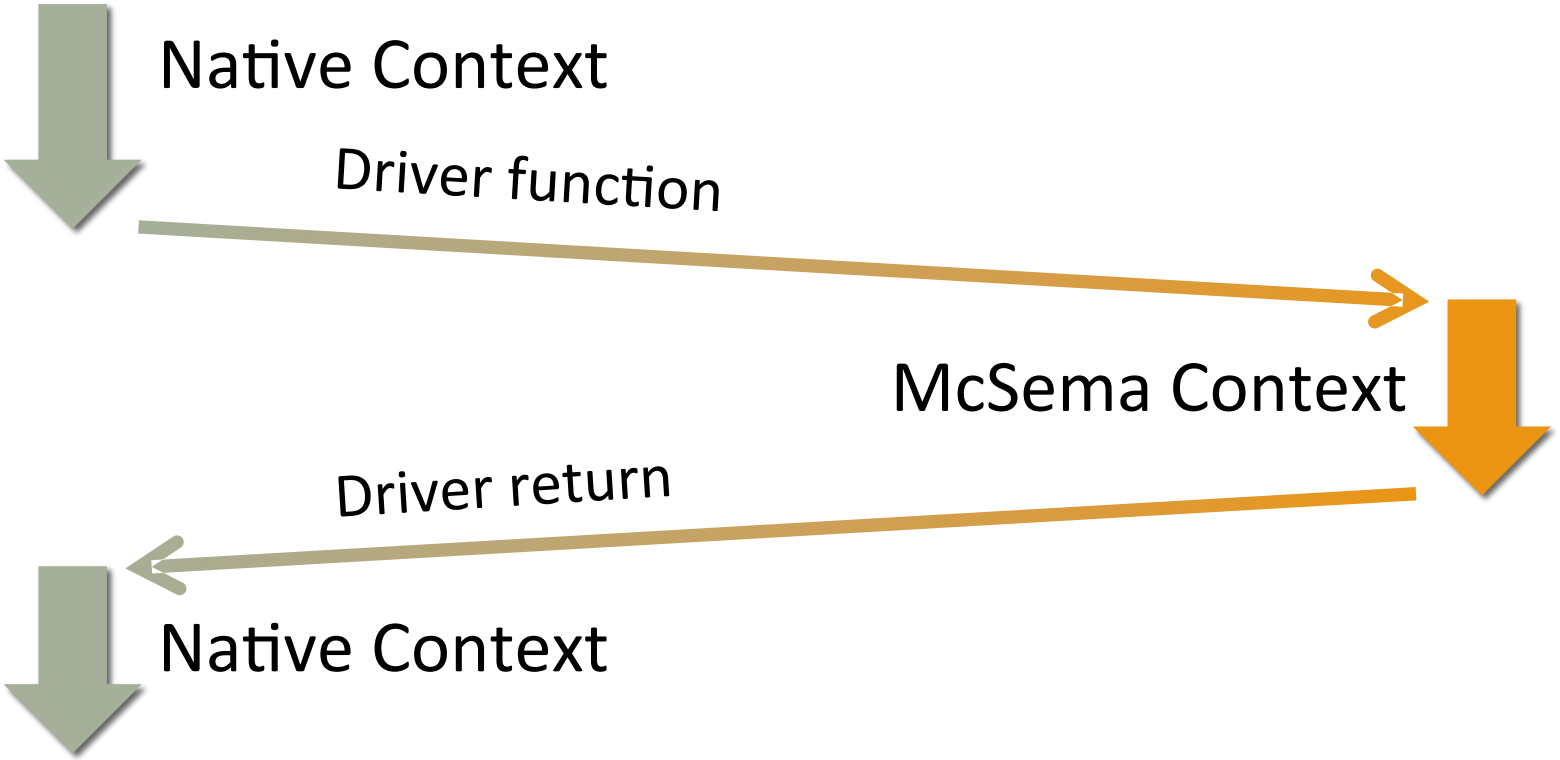
Instruction Translation: CALL REG/MEM



"This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA)."
"The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government."
Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Instruction Translation: Callbacks

Create ‘drivers’ that translate context



“This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA).”
“The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government.”
Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Instruction Translation: Context Switches

Native Context ⇔ McSema Context switch is complicated.

- Mostly because of stacks

Requires OS and Architecture specific voodoo

Currently implemented for x86 Windows

Fairly isolated to allow for portability

“This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA).”
“The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government.”
Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Development Progress: What Works

Integer instructions

Unit Tests

FPU registers

FPU instructions (some)

SSE registers

SSE instructions (very few)

Callbacks

External Calls

Jump Tables

Data References

"This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA)."
"The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government."
Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Demo 3

“This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA).”
“The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government.”
Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Development Progress: What Needs to be Done

FPU Instructions (some)

SSE Instructions (most)

Exceptions

Privileged instructions

Need more unit tests!

Better optimization

"This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA)."
"The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government."
Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Future Plans

More instructions support

Memory modeling

Optimization

Rigorous Testing

"This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA)."
"The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government."
Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Questions?

“This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA).”
“The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government.”
Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)