遇见 OSPP : vision to reality

卫林泉 | weilinquan@buaa.edu.cn
北京航空航天大学 | Buddy Compiler开源社区贡献者

Buddy Compiler

Buddy Compiler 简介和愿景

Buddy Compiler 是领域特定的编译器框架，

致力于打造基于 MLIR 和 RISC-V 的软硬件协同设计生态。

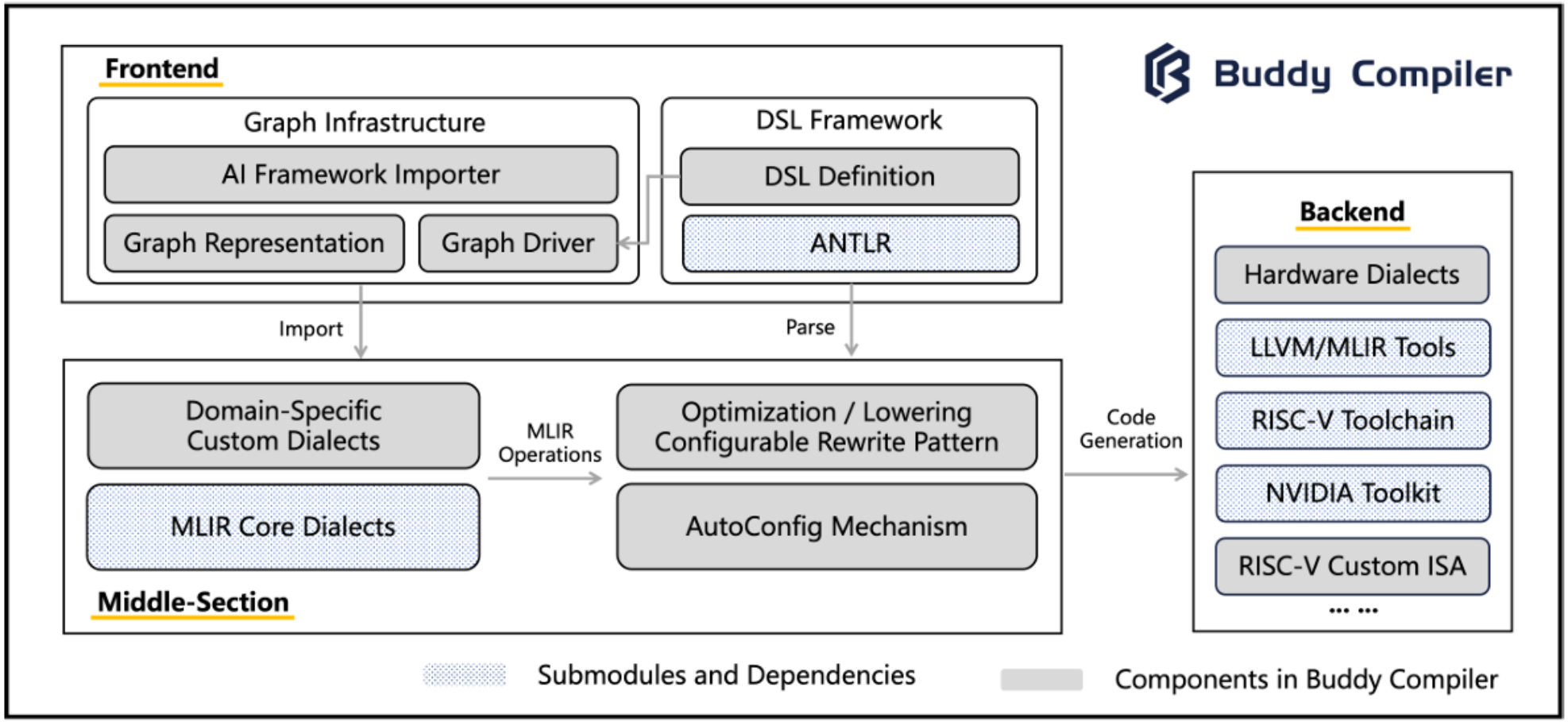
我们的目标是实现从 DSL 到 DSA 的编译流程和协同设计。

我们的愿景是让领域特定的协同设计不再困难。

Buddy Compiler 图层概述

Buddy Compiler

“Buddy System” for Domain-Specific Compilers | MLIR-Based Compilation Framework for Deep Learning Co-Design



Homepage: https://buddy-compiler.github.io/
GitHub: https://github.com/buddy-compiler

> AI 大模型的爆发为软硬件设计带来了新的抓手和机会，
> 我们的愿景是让领域特定的协同设计不再困难，
> 一条AI模型到硬件的标准端到端通路不可或缺。

Buddy Compiler AI推理--OSPP开源项目

Buddy Compiler

Buddy Compiler Frontend Ecosystem

PyTorch

PyTorch 2.x TorchDynamo

Graph Capture → FX Graph → Aten IR

```
class GraphModule(torch.nn.Module):
 def forward(self, ...):
   ... = torch.ops.aten.arange.start(...)
   ... = torch.ops.aten.unsqueeze.default(...)
   ... = torch.ops.aten.view.default(...)
   ... ...
   return [... ...]
```

Buddy Compiler As TorchDynamo Custom Compiler

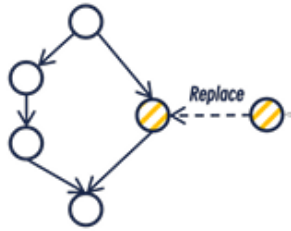
Buddy Compiler

Graph Infrastructure

Dynamo Importer (Aten IR)

↓

Graph Representation (Buddy Graph Ops Registry)




```
def kernel(... ...) {
  vector a = ... ...
  vector b = ... ...
  vector c = ... ...
  c = a x b + c
  out[i] = c
}
```

DSL Framework

Standalone DSL (DSL Grammar)

↓

ANTLR (Parser Rules / Visitor)

MLIR (Linalg Dialect / TOSA Dialect / Math Dialect / Vector Dialect ... ...)

Buddy Compiler

```
root@9d5b995149a1:~/buddy-mlir/build/bin# numactl --cpunodebind=1
./buddy-llama-run
LLaMA 2 Inference Powered by Buddy Compiler

Please send a message:
>>> How old are you?

[Log] Vocab file: "/root/buddy-mlir/examples/BuddyLlama/vocab.txt"
[Log] Tokenize time: 56.2298ms
[Log] Loading params...
[Log] Params file: "/root/buddy-mlir/examples/BuddyLlama/arg0.data
"
[Log] Params load time: 52.2917s

[Iteration 0] Token: <0x0A> | Time: 20.443s
[Iteration 1] Token: <0x0A> | Time: 18.4413s
[Iteration 2] Token: I | Time: 15.7251s
[Iteration 3] Token: _am | Time: 14.6714s
[Iteration 4] Token: _ | Time: 15.5225s
[Iteration 5] Token: 2 | Time: 14.4288s
[Iteration 6] Token: 5 | Time: 13.83s
[Iteration 7] Token: _years | Time: 14.0442s
[Iteration 8] Token: _old | Time: 13.8333s
[Iteration 9] Token: . | Time: 13.8439s
[Iteration 10] Token: </s> | Time: 13.9806s

[Input] How old are you?
[Output] <0x0A><0x0A>I am 25 years old.
```

“软硬件协同设计本质上是“人”的协同，
开源社区工作本质上也是“人”的连结！”

Buddy Compiler @ LFX Mentorship

Buddy Compiler

THE LINUX FOUNDATION

RISC-V | Vector Programming | Instruction Set Architecture

RISC-V Mentorship: MLIR Convolution Vectorization

Mentees

Prathamesh Tagore

- DIP Dielect Corr2D Operation
- Vectorization Pass
- Performance Evaluation

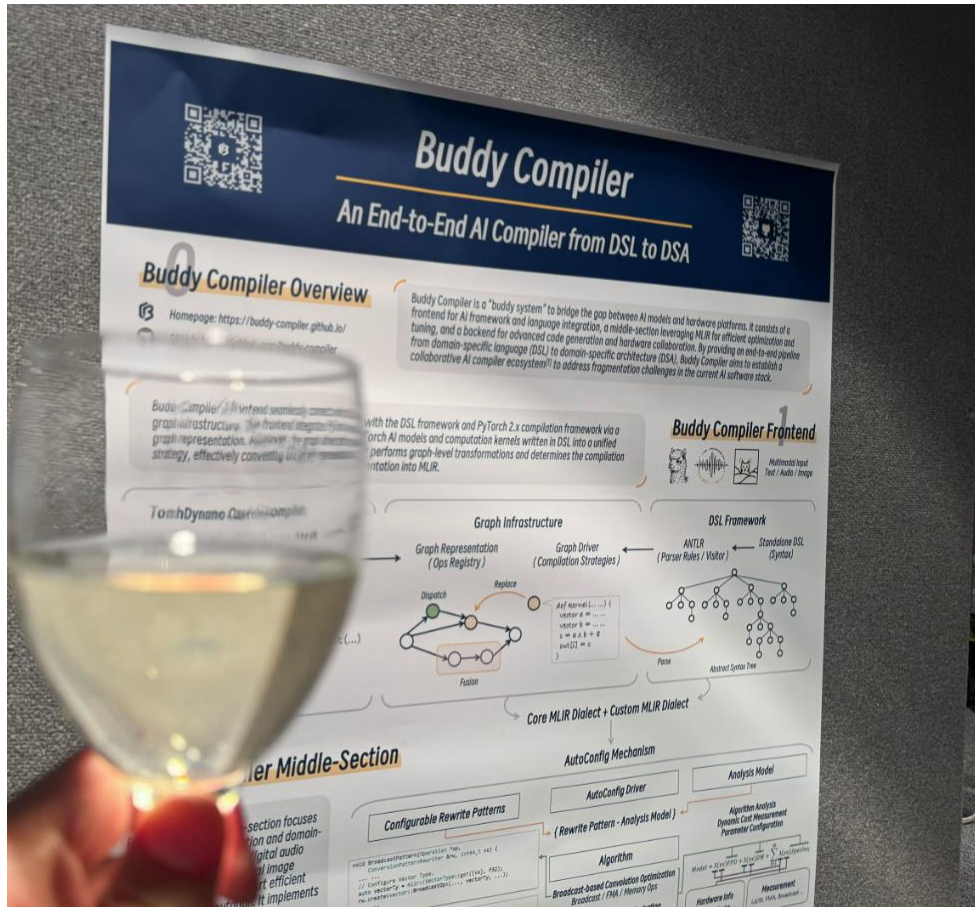
Joe Wu

- Convolution Operation Optimization

Ahmat Hamdan

- MemRef Descriptor Implementation and Test
- Pooling Operation Optimization

Mentors



https://www.youtube.com/watch?v=1pe3G3UQRkQ&t=142s

“

Buddy Compiler 社区始于兴趣，基于创新！

我们要遇见更多人，碰撞更多想法，付出更多努力，

一起做出有意思、有意义的事情！

”

Buddy Compiler

Thanks

weilinquan@buaa.edu.cn