Skymizer | ONNC Software Architecture Overview

Po-Yen Chen (poyenc@skymizer.com)

a

Open Neural Network Compiler (ONNC)

ONNC as a framework…

- "*Retargetable compilation framework designed specifically for proprietary deep learning accelerators.*"

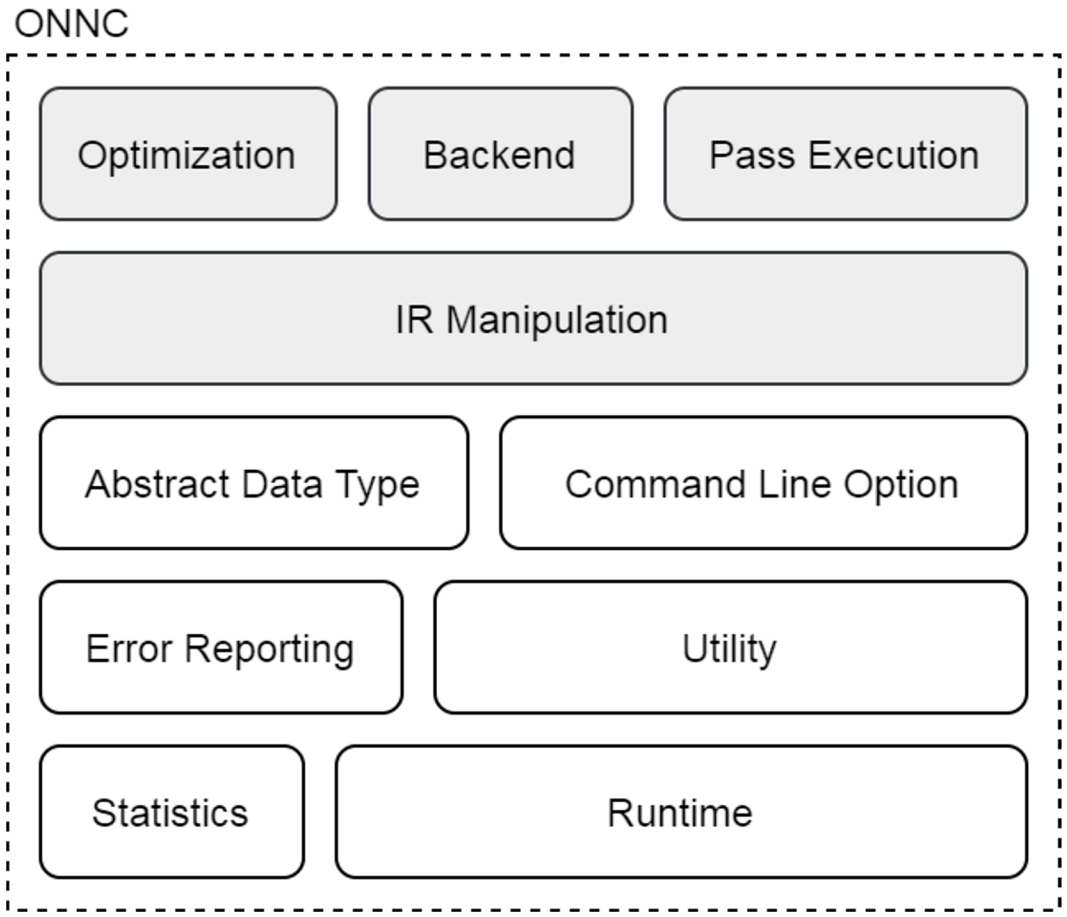
ONNC as a compiler…

- "*ONNC is the first open source compiler available for NVDLA-based hardware designs.*"

skymizer

a

ONNC as a framework



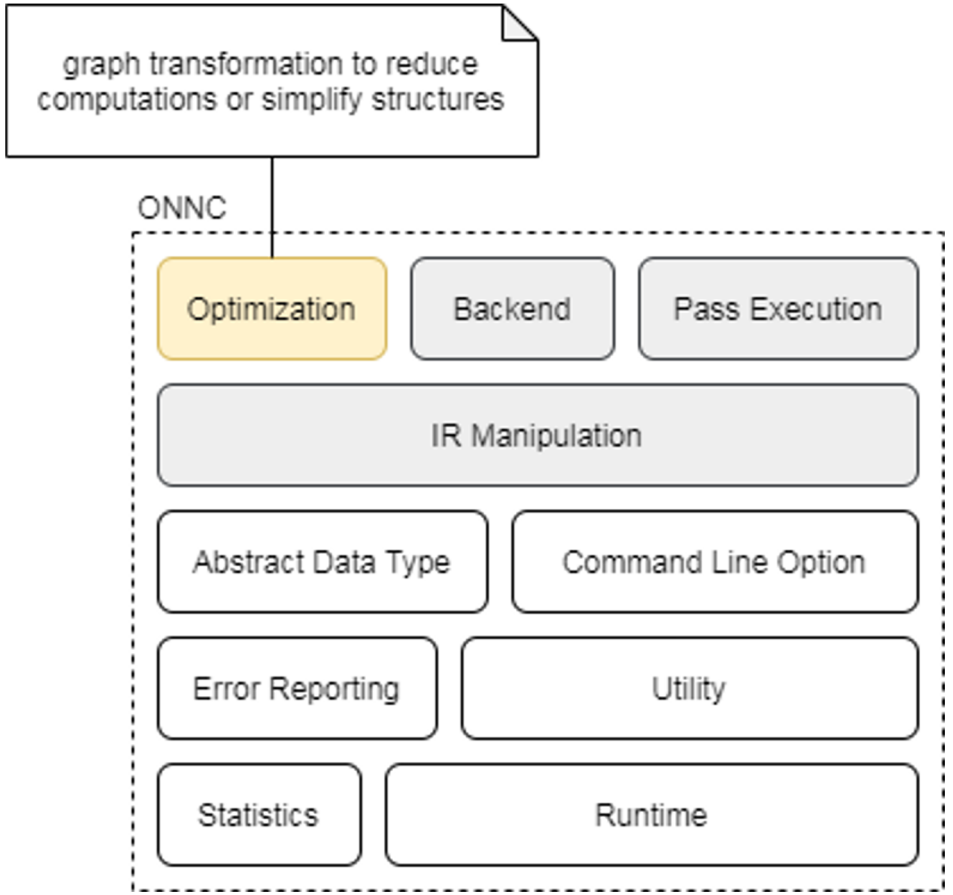
skymizer

a

ONNC as a framework - Optimization

Built-in optimizations

- Divide *GlobalAveragePool* into *AveragePool*
- Expand *BatchNormalization* into *Add* & *Mul*
- Replace *Gemm* by *Conv*
- Split *Conv* channel
- Propagate constant tensor
- Eliminate *Identity*
- Eliminate *Cast*



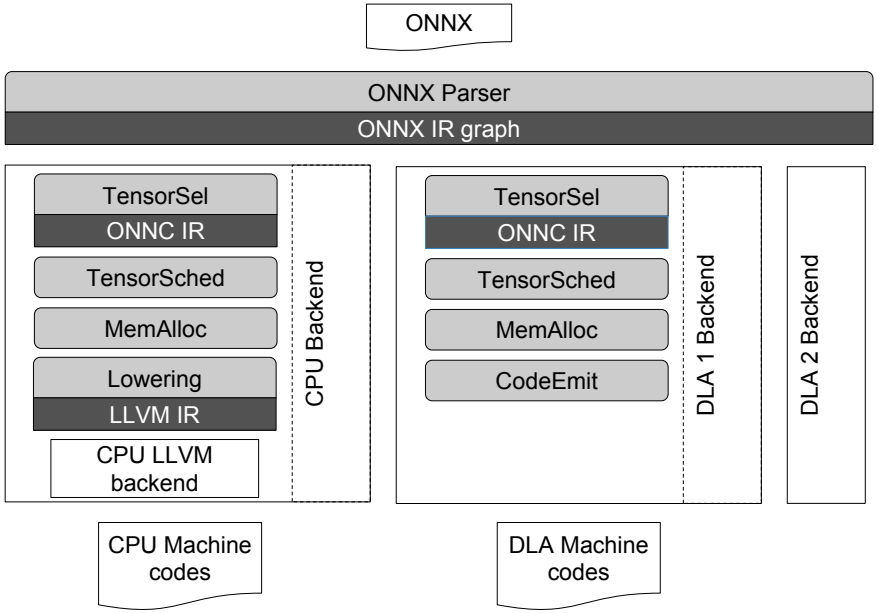
skymizer

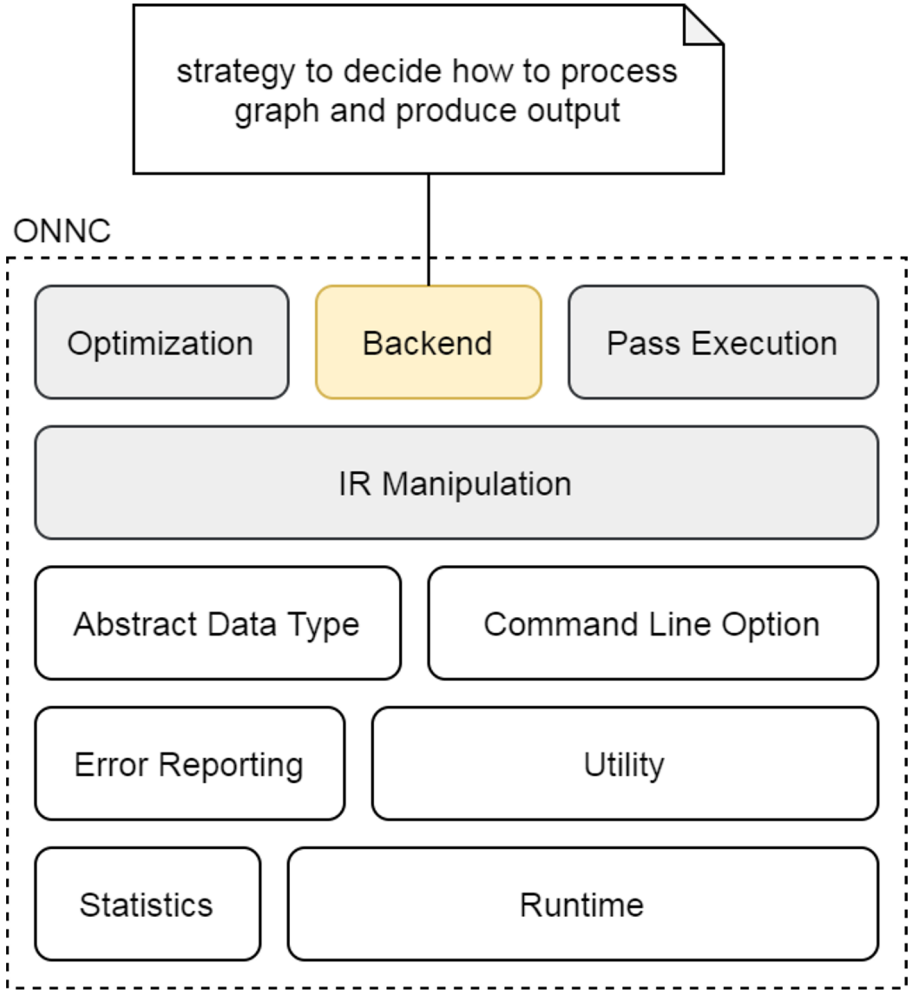
a

ONNC as a framework - Backend

Built-in Support

- x86 Backend: default backend (no output)
- *NVDLA* Backend: generate *NVDLA* loadable
- C Backend: create inference API in C





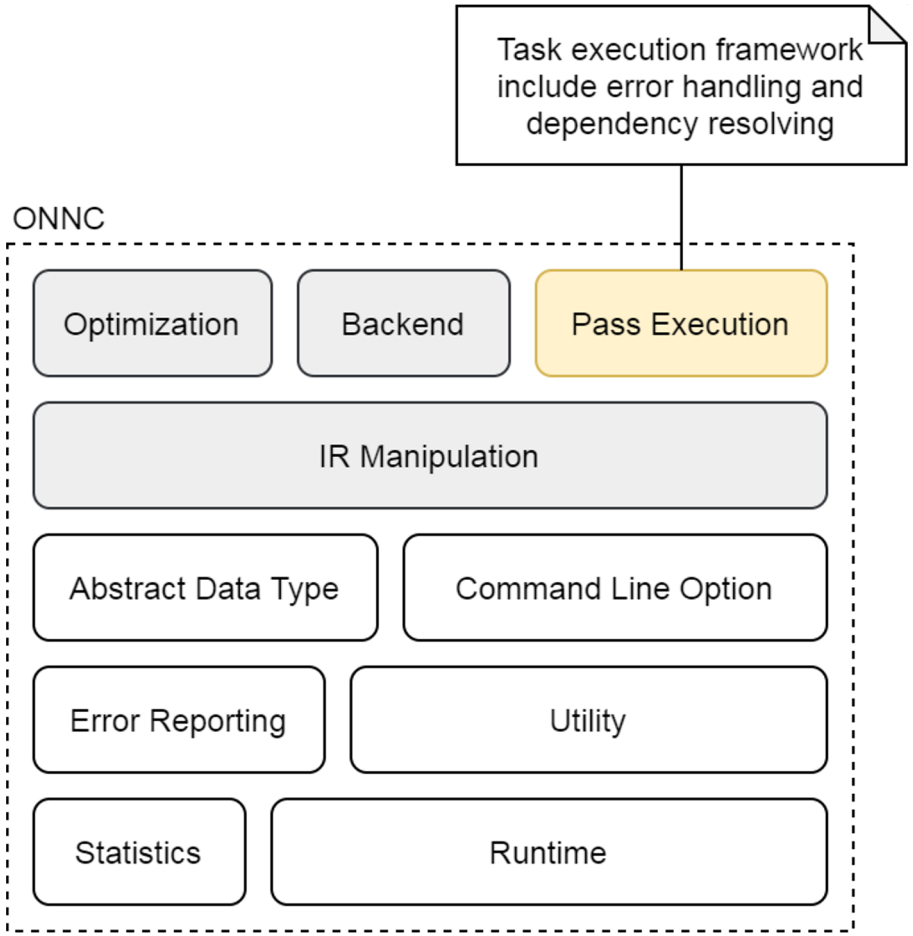
skymizer

a

ONNC as a framework – Pass Execution

Types involved in task execution

- *Pass*: task itself
- *PassManager*: manage pass instances and run/stop passes
- *AnalysisUsage*: define dependency between passes



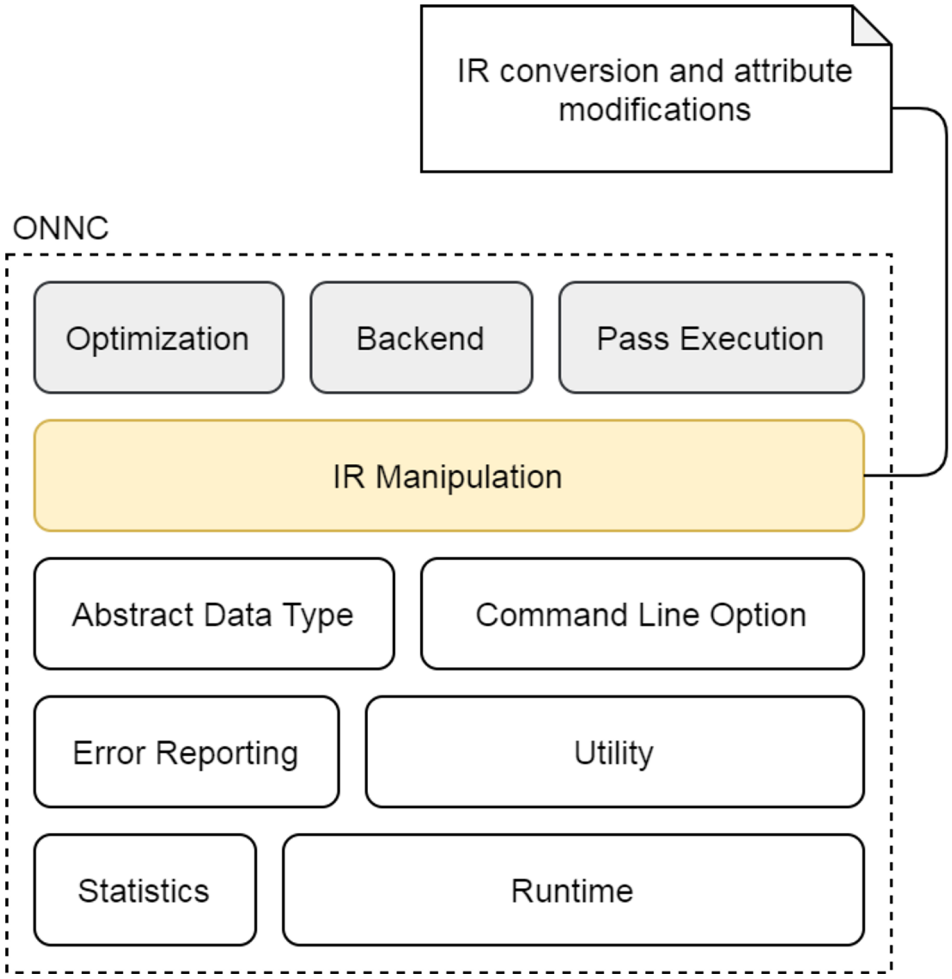
skymizer

a

ONNC as a framework – IR Manipulation

IR manipulation infrastructures

- *Lower*: convert ONNX IR to ONNC IR
- *Module*: the abstraction of model
- *ComputeGraph*: IR graph in model
- *ComputeOperator*: base type for every IR
- *Tensor*: base type for IR input/output storage
- *IRBuilder*: helper type for module creation

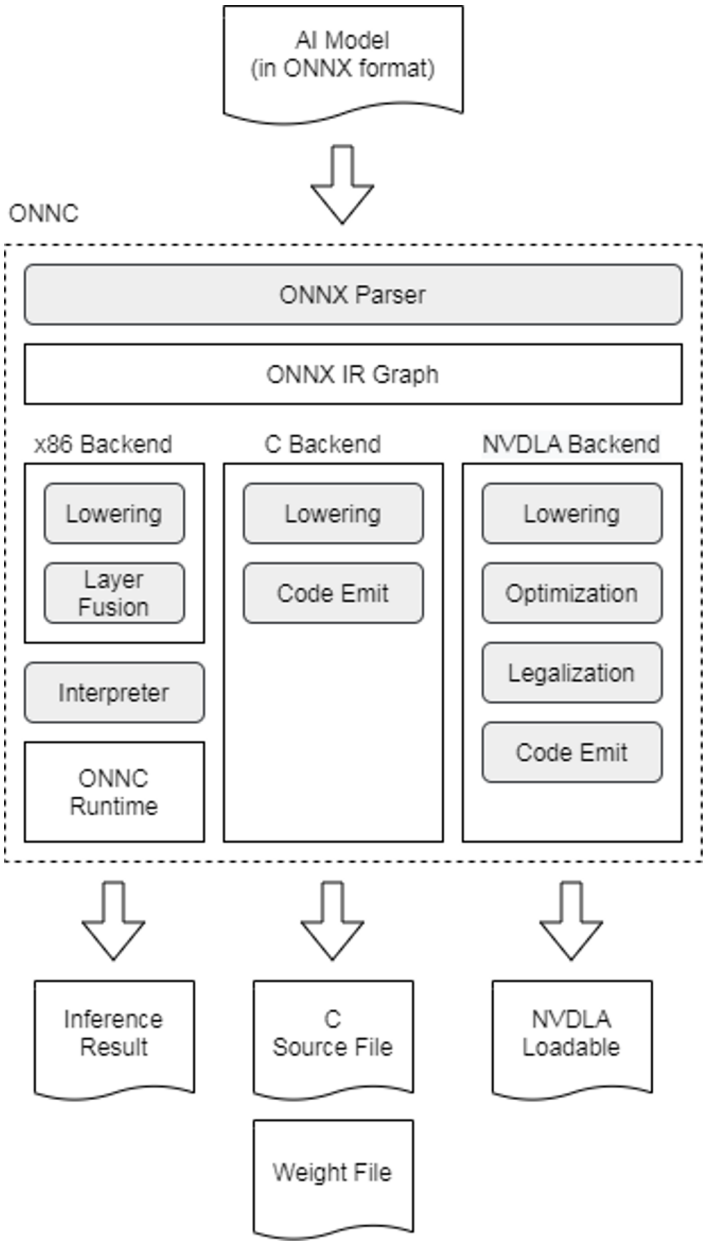


skymizer

a

ONNC as a compiler

- ONNC accepts ONNX model as input
- All built-in backends convert ONNX IR to ONNC IR before any actions
- Backend may do *target-dependent* transformations
- Backend can work alone or be used as frontend in any scenarios



skymizer

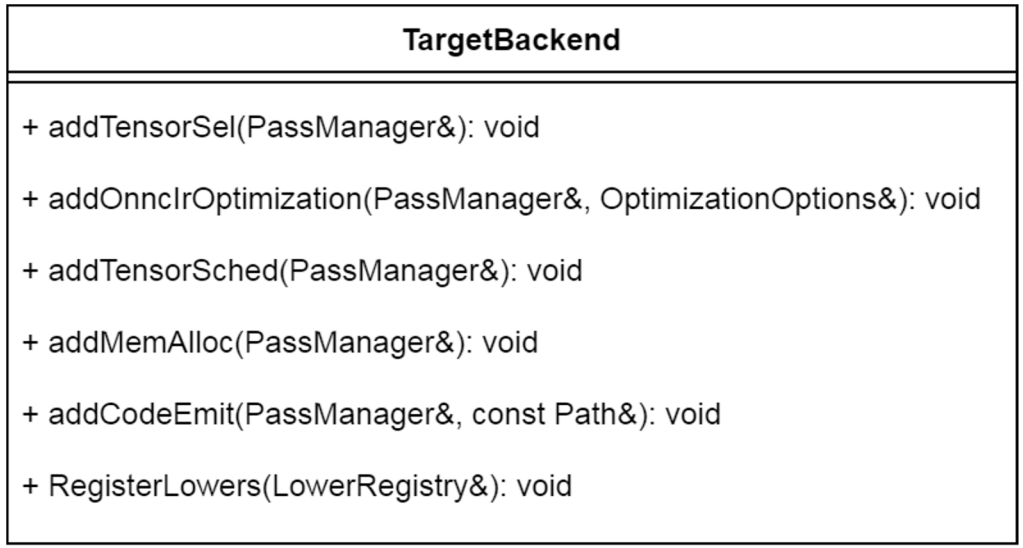
a

ONNC as a compiler – Implement a Backend (I)

TargetBackend is the base type for every backends

Backend lifetime events:

1. *addTensorSel*: convert ONNX IR to ONNC IR
2. *addOnncIrOptimization*: do any graph transformation
3. *addTensorSched*: schedule IR execution
4. *addMemAlloc*: analyze and allocate memory for storing result
5. *addCodeEmit*: generate output files



skymizer

a

ONNC as a compiler – Implement a Backend (II)

Steps to create your own backends…

1. Create a backend class inherited from *TargetBackend*
2. Override *RegsterLowers()* method - define valid ONNX operators
3. Override lifetime events if necessary
4. Register the backend class into ONNC
5. Rebuild ONNC

Detail covered in *Backend Developer Guide*,
or reference the example backend: *VanillaBackend*
or use prepared script *create-new-backend.sh*

```
# under onnc repository root, run script to create a new backend called Test (TestBackend)
$ ./script/create-new-backend.sh Test
# after build with new backend, use TestBackend to compiler model
$ onnc -mquadruple test <onnx model file>
```


skymizer

a

ONNC as a compiler – Implement a Backend (III)

Override *RegsterLowers()* method: define valid ONNX operators

lib/Target/Vanilla/VanillaBackend.cpp:115

```
115  void VanillaBackend::RegisterLowers(LowerRegistry& pRegistry) const
116  {
117    pRegistry.emplace<AddLower>();
118    pRegistry.emplace<AveragePoolLower>();
119    pRegistry.emplace<BatchNormalizationLower>();
120    pRegistry.emplace<ConcatLower>();
121    pRegistry.emplace<ConvLower>();
122    pRegistry.emplace<FlattenLower>();
123    pRegistry.emplace<GemmLower>();
```

skymizer

skymizer

Boost deep learning accelerator
with compiler technology

Skymizer Taiwan Inc.

CONTACT US

E-mail sales@skymizer.com **Tel** +886 2 8797 8337

HQ 12F-2, No.408, Ruiguang Rd., Neihu Dist., Taipei City 11492, Taiwan

BR Center of Innovative Incubator, National Tsing Hua University,
Hsinchu Taiwan

https://skymizer.com