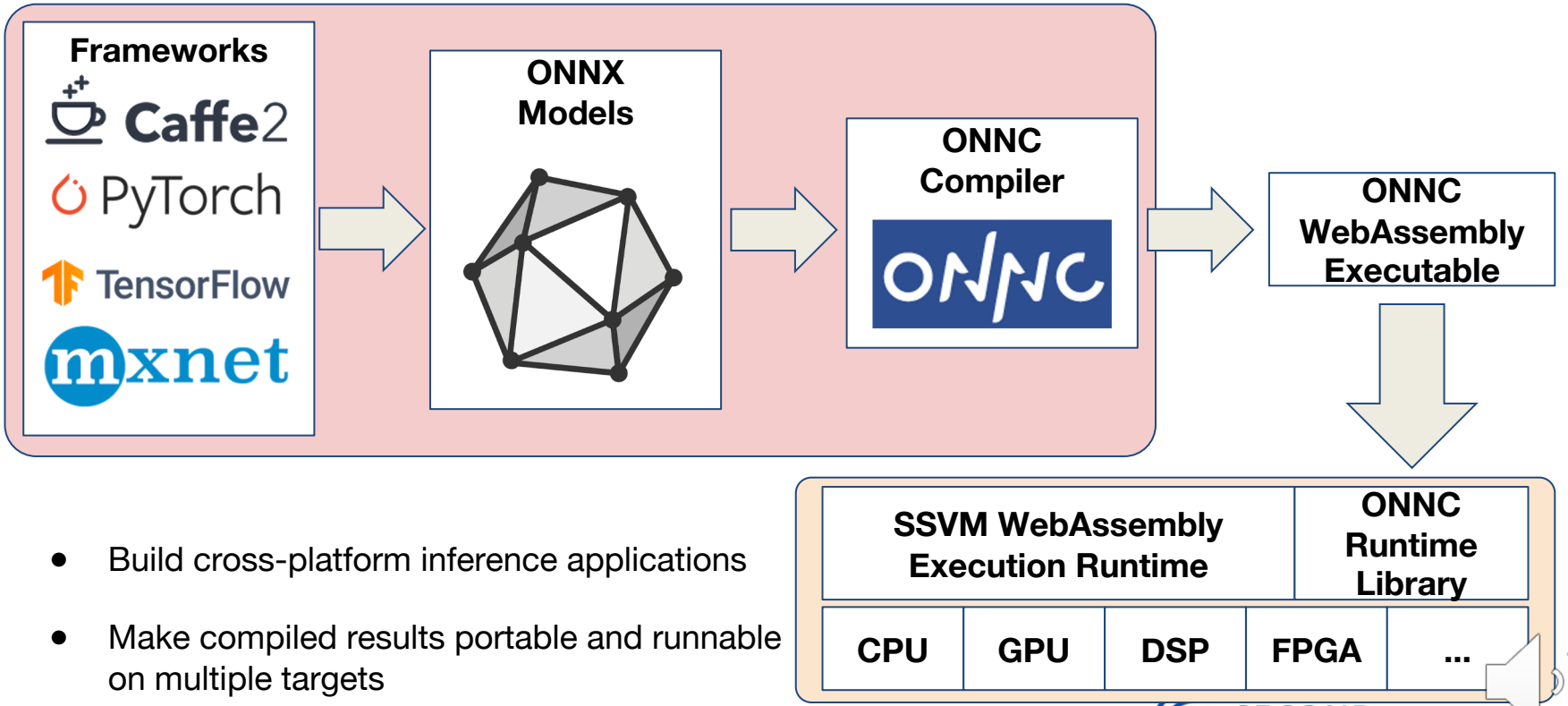
ONNC-WASM Project

A Joint Project with Skymizer

Hung-Ying, Tai
hydai@secondstate.io

Project Goal



- Build cross-platform inference applications
- Make compiled results portable and runnable on multiple targets

SECOND STATE skymizer

Why WebAssembly(Wasm)

Portability

Platform Independent

Efficiency

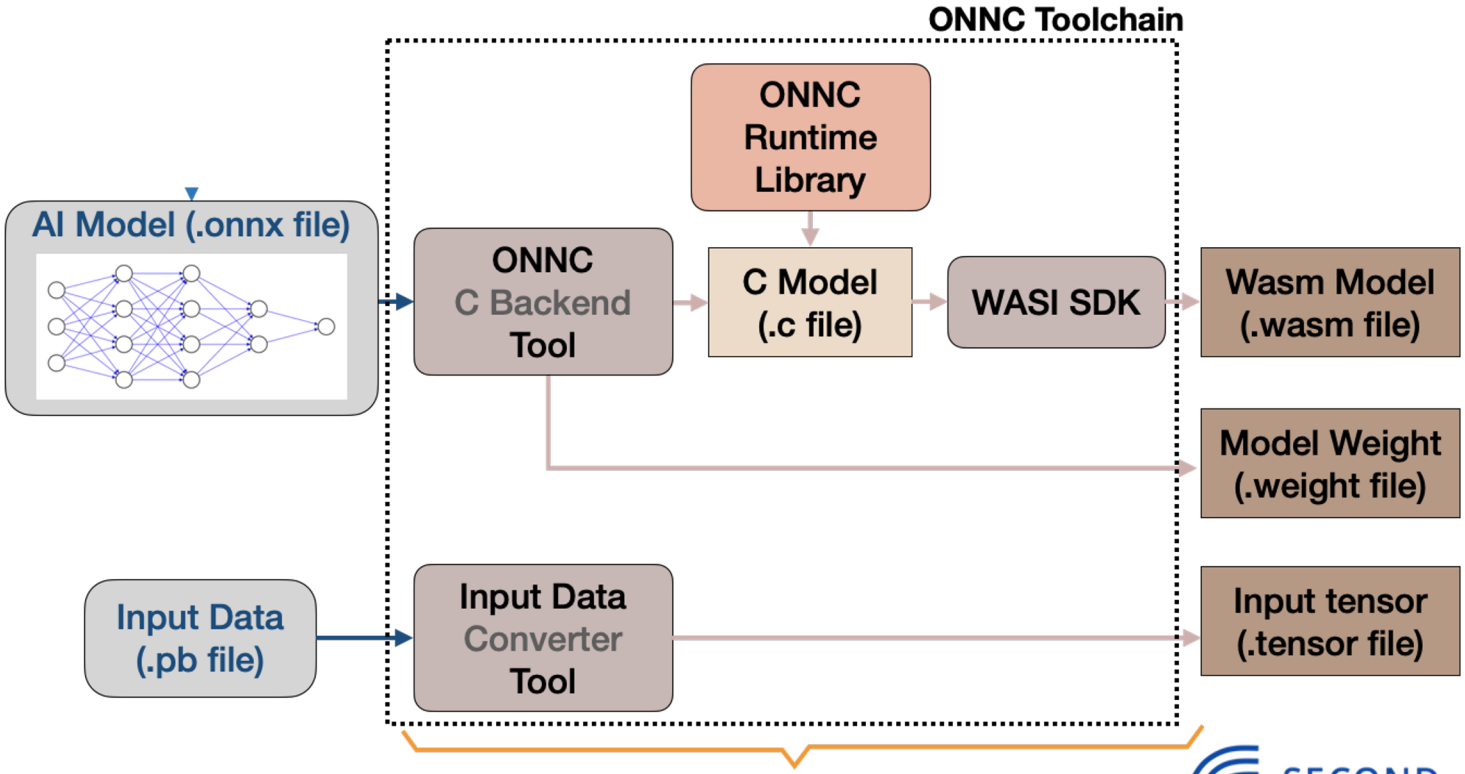
Low level as assembly

Safety

Sandboxed Environment

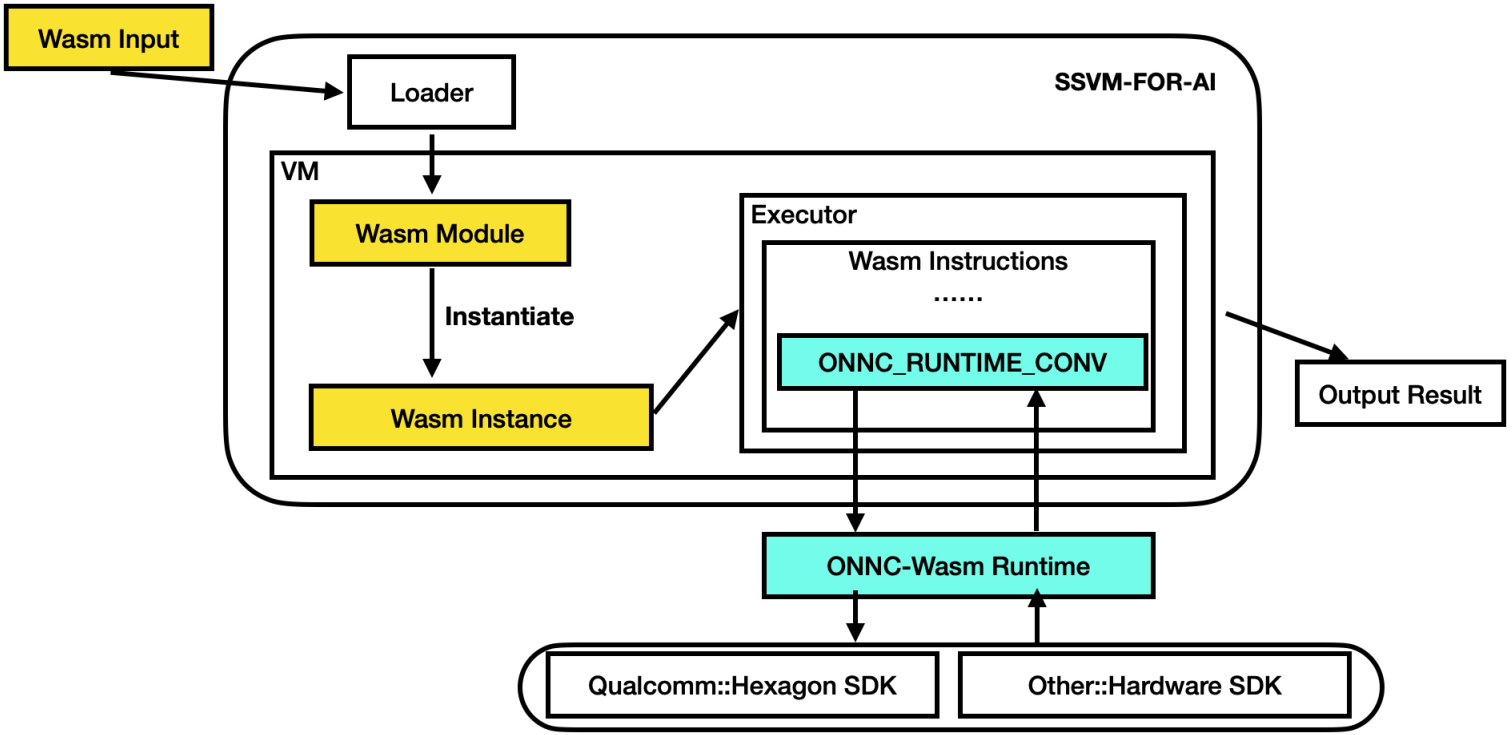
- Powered by W3C
- A **portable** binary compilation-target
- **Efficient**, **safe** and a **cross-platform** bytecode format
- Good for computation-intensive applications

ONNC-Wasm Backend Architecture



SECOND STATE skymizer

Second State Virtual Machine(SSVM) Architecture



SECOND STATE skymizer

SSVM Performance Benchmark

Benchmark name	native	ssvm	lucet(wasmtime)	WAVM	Nodejs v14.3.0
Runtime Initialize Time (nop 0)	**0.001**	0.004	**0.002**	0.054	0.055
Concat File (cat-sync 0)	**0.003**	**0.006**	0.577	0.032	0.066
nbody-c (n = 50000000)	**3.310**	3.716	4.616	3.750	**3.393**
nbody-cpp (n = 50000000)	**3.137**	3.773	4.699	**3.744**	3.967
fannkuch-redux-c 12	**23.699**	**27.711**	52.281	28.436	29.367
mandelbrot-c 15000	**9.400**	11.850	-	11.946	**10.485**
binary-trees-c 21	**15.691**	**12.917**	-	14.825	18.776

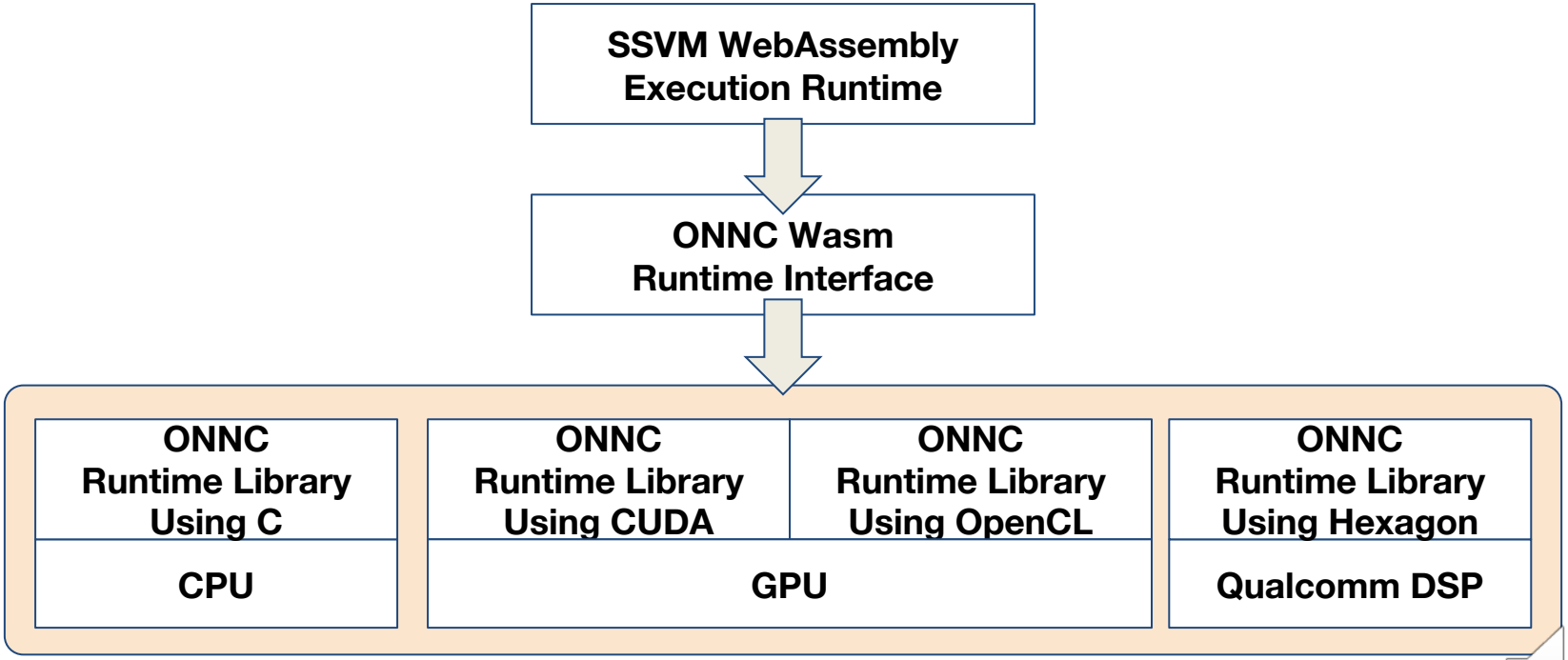
Runs on Intel(R) Xeon(R) CPU E5-2673 v4 @ 2.30GHz, Linux 5.4.0-1010-azure

Unit: Second, lower is better

SECOND STATE

skymizer

How to Support Multiple Hardware Backend



Labs

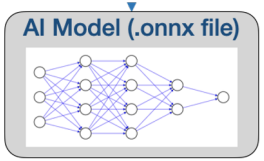
Object Detection with TinyYolo V1

Prepare Environment for ONNC-Wasm project

- Repo: https://github.com/onnc/onnc-wasm
- Clone onnc-wasm project
  - `git clone https://github.com/ONNC/onnc-wasm.git`
  - `cd onnc-wasm`
  - `git submodule add https://github.com/second-state/SSVM ssvm`
  - `cd ssvm && git checkout b5785ed6f8f24de2afa54ec6c28904e7607a0f12`
  - `cd ..`
- Download and Run docker image
  - `./scripts/start_docker_env.sh`

Convert AI Model to Wasm Model

ONNX Model



+

ONNC-Wasm Runtime Interface

```
void ONNC_RUNTIME_conv_float(
    void *onnc_runtime_context, const float *input_X, int32_t input_X_ndim,
    const int32_t *input_X_dims, const float *input_W, int32_t input_W_ndim,
    const int32_t *input_W_dims, const float *input_B, int32_t input_B_ndim,
    const int32_t *input_B_dims, float *output_Y, int32_t output_Y_ndim,
    const int32_t *output_Y_dims, const char *auto_pad, int32_t *dilations,
    int32_t number_of_dilations, int32_t group, int32_t *kernel_shape,
    int32_t number_of_kernel_shape, int32_t *pads, int32_t number_of_pads,
    int32_t *strides, int32_t number_of_strides);
```

Wasm Model

```
(import "onnc_runtime" "conv_float" (func (;0;) (type 0)))
(func (;27;) (type 9)
    (func (;26;) (type 9)
    (local i32 i32 i64)
    global.get 0
    i32.const 208
    i32.sub
    ...omitted...
  )
(memory (;0;) 2)
(global (;0;) (mut i32) (i32.const 86176))
(export "memory" (memory 0))
(export "main" (func 27))
```

Convert AI Model to Wasm Model

- Build ONNX model and ONNC runtime libraries
- `cd /home/onnc/workspace/models && \`
- `../scripts/build.sh ssvm /home/onnc/tiny_yolov1/model.onnx`

```
onnc@e3fcac2c35a3:/home/onnc/workspace$ cd /home/onnc/workspace/models && ../scripts/build.sh ssvm /home/onnc/tiny_yolov1/model.onnx
-- Configuring done
-- Generating done
-- Build files have been written to: /home/onnc/workspace/models/build-ssvm
[  1%] Built target detection
[ 15%] Built target onnc-wasm
[ 17%] Built target jpg2t
[ 97%] Built target onnc-runtime
[100%] Built target model.wasm
Install the project...
-- Install configuration: "Release"
-- Up-to-date: /home/onnc/workspace/models/out-ssvm/bin/model.wasm
-- Up-to-date: /home/onnc/workspace/models/out-ssvm/bin/model.weight
-- Up-to-date: /home/onnc/workspace/models/out-ssvm/bin/detection
-- Up-to-date: /home/onnc/workspace/models/out-ssvm/bin/jpg2t
-- Up-to-date: /home/onnc/workspace/models/out-ssvm/lib/libonnc-runtime.a
-- Up-to-date: /home/onnc/workspace/models/out-ssvm/lib/libonnc-wasm.so
```

Prepare Input Image

- Convert input JPEG file to Tensor file
- `./out-ssvm/bin/jpg2t /home/onnc/tiny_yolov1/test_data_set_0/input_0.jpg \`
- `input.tensor 448 448`
- Input Image:

Console:

```
onnc@e3fcac2c35a3:/home/onnc/workspace/models$ ./out-ssvm/bin/jpg2t \
> /home/onnc/tiny_yolov1/test_data_set_0/input_0.jpg \
> input.tensor 448 448
Crop image to (448 x 448)
```

SECOND STATE skymizer

Run Wasm Model with SSVM

Wasm models

```
(import "onnc_runtime" "conv_float" (func (;0;) (type 0)))
(func (;27;) (type 9)
    (func (;26;) (type 9)
    (local i32 i32 i64)
    global.get 0
    i32.const 208
    i32.sub
    ...omitted...
  )
(memory (;0;) 2)
(global (;0;) (mut i32) (i32.const 86176))
(export "memory" (memory 0))
(export "main" (func 27))
```

ONNC-Wasm Runtime Interface

```
void ONNC_RUNTIME_conv_float(
    void *onnc_runtime_context, const float *input_X, int32_t input_X_ndim,
    const int32_t *input_X_dims, const float *input_W, int32_t input_W_ndim,
    const int32_t *input_W_dims, const float *input_B, int32_t input_B_ndim,
    const int32_t *input_B_dims, float *output_Y, int32_t output_Y_ndim,
    const int32_t *output_Y_dims, const char *auto_pad, int32_t *dilations,
    int32_t number_of_dilations, int32_t group, int32_t *kernel_shape,
    int32_t number_of_kernel_shape, int32_t *pads, int32_t number_of_pads,
    int32_t *strides, int32_t number_of_strides);
```

ONNC-Wasm Runtime Using DSP

```
int onnc_runtime_conv_float(
    const float* tensor_x, int tensor_xLen, const int32_t* dim_x, int dim_xLen,
    ...,
    const int32_t* strides, int stridesLen
){
    ...
    int numWorkers = hvxInfo.numThreads;
    dspCV_synctoken_t token;
    dspCV_worker_pool_synctoken_init(&token, numWorkers);

    conv_2d_data_t data = {
        .token = &token,
```

Run Object Detection

- Run model with input Tensor file
- `../scripts/run.sh ssvm out-ssvm/ model input.tensor output.data`
- Console log:

```
onnc@e3fcac2c35a3:/home/onnc/workspace/models$ ../scripts/run.sh ssvm out-ssvm/ model input.tensor output.data
 Args : out-ssvm//bin/model.wasm
 Args : input.tensor
 Args : out-ssvm//bin/model.weight
 Info: Start running...
Output size: 5880
 --- Inference: SSVM cost 451 us, Host functions cost 5050421 us
 Info: Worker execution succeeded.
 ================  Statistics  ================
 Total execution time: 5383484 us
 Wasm instructions execution time: 288472 us
 Host functions execution time: 5095012 us
 Executed wasm instructions count: 19097
 Instructions per second: 66200
```

Retrieve Output Image from Output.data

- Detect result and combined it with input image file into output image file
- `./out-ssvm/bin/detection output.data \`
- `/home/onnc/tiny_yolov1/test_data_set_0/input_0.jpg output.jpg`
- Output.jpg: Console log:


```
onnc@e3fcac2c35a3:/home/onnc/workspace/models$ ./out-ssvm/bin/detection output.data \
> /home/onnc/tiny_yolov1/test_data_set_0/input_0.jpg output.jpg
cow: 54%, (0.536, 0.602)[0.522, 0.544]
```

SECOND STATE skymizer

Thank You!