In [5]:

```
import brevitas.onnx as bo
from brevitas.quant_tensor import QuantTensor

ready_model_filename = "verification/model_fmnist_final.onnx"
```

In [6]:

```
from finn.util.visualization import showInNetron

showInNetron(ready_model_filename)
```

```
Stopping http://0.0.0.0:8081 (http://0.0.0.0:8081)
Serving 'verification/model_fmnist_final.onnx' at http://0.0.0.0:8081 (http://0.0.0.0:8081)
```

Out[6]:

In [2]:

```
import finn.builder.build_dataflow as build
import finn.builder.build_dataflow_config as build_cfg
import os
import shutil

model_file = "verification/model_fmnist_final.onnx"
#model_file = "model_v4noid_verified.onnx"

estimates_output_dir = "output_estimates_only"



#Delete previous run results if exist
if os.path.exists(estimates_output_dir):
    shutil.rmtree(estimates_output_dir)
    print(os.path.abspath(estimates_output_dir))
    print("Previous run results deleted!")


cfg_estimates = build.DataflowBuildConfig(
    output_dir          = estimates_output_dir,
    mvau_wwidth_max     = 80,
    target_fps          = 1000000,
    synth_clk_period_ns = 10.0,
    fpga_part               = "xczu7ev-ffvc1156-2-e",
    steps               = build_cfg.estimate_only_dataflow_steps,
    generate_outputs=[
        build_cfg.DataflowOutputType.ESTIMATE_REPORTS,
    ]
)
```

```
/workspace/finn/notebooks/mnist_ex/output_estimates_only
Previous run results deleted!
```

In [3]:

```
%%time
build.build_dataflow_cfg(model_file, cfg_estimates)
```

```
Building dataflow accelerator from verification/model_fmnist_final.onnx
Intermediate outputs will be generated in /home/mmirigaldi/finn_temp_mmirigaldi
Final outputs will be generated in output_estimates_only
Build log is at output_estimates_only/build_dataflow.log
Running step: step_qonnx_to_finn [1/8]
Running step: step_tidy_up [2/8]
Running step: step_streamline [3/8]
Running step: step_convert_to_hls [4/8]
Running step: step_create_dataflow_partition [5/8]
Running step: step_target_fps_parallelization [6/8]
Running step: step_apply_folding_config [7/8]
Running step: step_generate_estimate_reports [8/8]
Completed successfully
CPU times: user 959 ms, sys: 9.81 ms, total: 969 ms
Wall time: 966 ms
```

Out[3]:

```
0
```

In [4]:

```
! ls {estimates_output_dir}
```

```
auto_folding_config.json  intermediate_models  time_per_step.json
build_dataflow.log        report
```

In [5]:

```
! ls {estimates_output_dir}/report
```

```
estimate_layer_config_alternatives.json  estimate_network_performance.json
estimate_layer_cycles.json               op_and_param_counts.json
estimate_layer_resources.json
```

In [6]:

```
! cat {estimates_output_dir}/report/estimate_network_performance.json
```

```
{
  "critical_path_cycles": 96444,
  "max_cycles": 19760,
  "max_cycles_node_name": "ConvolutionInputGenerator_0",
  "estimated_throughput_fps": 5060.728744939272,
  "estimated_latency_ns": 964440.0
}
```

In [7]:

```
import json
def read_json_dict(filename):
    with open(filename, "r") as f:
        ret = json.load(f)
    return ret
```

In [8]:

```
read_json_dict(estimates_output_dir + "/report/estimate_layer_cycles.json")
```

Out[8]:

```
{'FMPadding_Batch_0': 1024,
 'ConvolutionInputGenerator_0': 19760,
 'StreamingFCLayer_Batch_0': 11760,
 'StreamingMaxPool_Batch_0': 1176,
 'FMPadding_Batch_1': 972,
 'ConvolutionInputGenerator_1': 14970,
 'StreamingFCLayer_Batch_1': 7840,
 'StreamingMaxPool_Batch_1': 294,
 'FMPadding_Batch_2': 968,
 'ConvolutionInputGenerator_2': 10240,
 'StreamingFCLayer_Batch_2': 19600,
 'StreamingFCLayer_Batch_3': 7840}
```

In [9]:

```
read_json_dict(estimates_output_dir + "/report/estimate_layer_resources.json")
```

Out[9]:

```
{'FMPadding_Batch_0': {'BRAM_18K': 0,
  'BRAM_efficiency': 1,
  'LUT': 0,
  'URAM': 0,
  'URAM_efficiency': 1,
  'DSP': 0},
 'ConvolutionInputGenerator_0': {'BRAM_18K': 0,
  'BRAM_efficiency': 1,
  'LUT': 348,
  'URAM': 0,
  'URAM_efficiency': 1,
  'DSP': 0},
 'StreamingFCLayer_Batch_0': {'BRAM_18K': 2,
  'BRAM_efficiency': 0.016276041666666668,
  'LUT': 12702,
  'URAM': 0,
  'URAM_efficiency': 1,
  'DSP': 0},
 'StreamingMaxPool_Batch_0': {'BRAM_18K': 0,
  'BRAM_efficiency': 1,
  'LUT': 0,
  'URAM': 0,
  'URAM_efficiency': 1,
  'DSP': 0},
 'FMPadding_Batch_1': {'BRAM_18K': 0,
  'BRAM_efficiency': 1,
  'LUT': 0,
  'URAM': 0,
  'URAM_efficiency': 1,
  'DSP': 0},
 'ConvolutionInputGenerator_1': {'BRAM_18K': 0,
  'BRAM_efficiency': 1,
  'LUT': 348,
  'URAM': 0,
  'URAM_efficiency': 1,
  'DSP': 0},
 'StreamingFCLayer_Batch_1': {'BRAM_18K': 4,
  'BRAM_efficiency': 0.06510416666666667,
  'LUT': 15058,
  'URAM': 0,
  'URAM_efficiency': 1,
  'DSP': 0},
 'StreamingMaxPool_Batch_1': {'BRAM_18K': 0,
  'BRAM_efficiency': 1,
  'LUT': 0,
  'URAM': 0,
  'URAM_efficiency': 1,
  'DSP': 0},
 'FMPadding_Batch_2': {'BRAM_18K': 0,
  'BRAM_efficiency': 1,
  'LUT': 0,
  'URAM': 0,
  'URAM_efficiency': 1,
  'DSP': 0},
 'ConvolutionInputGenerator_2': {'BRAM_18K': 0,
  'BRAM_efficiency': 1,
  'LUT': 396,
  'URAM': 0,
  'URAM_efficiency': 1,
  'DSP': 0},
 'StreamingFCLayer_Batch_2': {'BRAM_18K': 2,
  'BRAM_efficiency': 0.6944444444444444,
  'LUT': 14758,
  'URAM': 0,
  'URAM_efficiency': 1,
  'DSP': 0},
 'StreamingFCLayer_Batch_3': {'BRAM_18K': 4,
  'BRAM_efficiency': 0.8506944444444444,
  'LUT': 434,
  'URAM': 0,
  'URAM_efficiency': 1,
  'DSP': 0},
 'total': {'BRAM_18K': 12.0, 'LUT': 44044.0, 'URAM': 0.0, 'DSP': 0.0}}
```

In [10]:

```
import finn.builder.build_dataflow as build
import finn.builder.build_dataflow_config as build_cfg
import os
import shutil

#model_file = "model_v4noid_verified.onnx"
model_file = "verification/model_fmnist_final.onnx"

rtlsim_output_dir = "output_ipstitch_ooc_rtlsim"

#Delete previous run results if exist
if os.path.exists(rtlsim_output_dir):
    shutil.rmtree(rtlsim_output_dir)
    print("Previous run results deleted!")

cfg_stitched_ip = build.DataflowBuildConfig(
    output_dir          = rtlsim_output_dir,
    mvau_wwidth_max     = 80,
    target_fps          = 100000,
    synth_clk_period_ns = 10.0,
    fpga_part               = "xczu7ev-ffvc1156-2-e",
    generate_outputs=[
        build_cfg.DataflowOutputType.STITCHED_IP,
        build_cfg.DataflowOutputType.RTLSIM_PERFORMANCE,
        #build_cfg.DataflowOutputType.OOC_SYNTH,
    ]
)
```

```
Previous run results deleted!
```

In [11]:

```
%%time
build.build_dataflow_cfg(model_file, cfg_stitched_ip)
```

```
Building dataflow accelerator from verification/model_fmnist_final.onnx
Intermediate outputs will be generated in /home/mmirigaldi/finn_temp_mmirigaldi
Final outputs will be generated in output_ipstitch_ooc_rtlsim
Build log is at output_ipstitch_ooc_rtlsim/build_dataflow.log
Running step: step_qonnx_to_finn [1/17]
Running step: step_tidy_up [2/17]
Running step: step_streamline [3/17]
Running step: step_convert_to_hls [4/17]
Running step: step_create_dataflow_partition [5/17]
Running step: step_target_fps_parallelization [6/17]
Running step: step_apply_folding_config [7/17]
Running step: step_generate_estimate_reports [8/17]
Running step: step_hls_codegen [9/17]
Running step: step_hls_ipgen [10/17]
Running step: step_set_fifo_depths [11/17]
Running step: step_create_stitched_ip [12/17]
Running step: step_measure_rtlsim_performance [13/17]
Running step: step_out_of_context_synthesis [14/17]
Running step: step_synthesize_bitfile [15/17]
Running step: step_make_pynq_driver [16/17]
Running step: step_deployment_package [17/17]
Completed successfully
CPU times: user 39.6 s, sys: 825 ms, total: 40.5 s
Wall time: 6min 57s
```

Out[11]:

```
0
```

In [12]:

```
! ls {rtlsim_output_dir}/stitched_ip
```

```
all_verilog_srcs.txt                   finn_vivado_stitch_proj.xpr
finn_vivado_stitch_proj.cache          ip
finn_vivado_stitch_proj.hbs            make_project.sh
finn_vivado_stitch_proj.hw             make_project.tcl
finn_vivado_stitch_proj.ip_user_files  vivado.jou
finn_vivado_stitch_proj.srcs           vivado.log
```

In [13]:

```
! ls {rtlsim_output_dir}/report
```

```
estimate_layer_resources_hls.json  rtlsim_performance.json
```

In [14]:

```
#! cat {rtlsim_output_dir}/report/ooc_synth_and_timing.json
```

In [15]:

```
! cat {rtlsim_output_dir}/report/rtlsim_performance.json
```

```
{
  "cycles": 46680,
  "runtime[ms]": 0.4668,
  "throughput[images/s]": 2142.2450728363324,
  "DRAM_in_bandwidth[Mb/s]": 1.6795201371036845,
  "DRAM_out_bandwidth[Mb/s]": 0.0856898029134533,
  "fclk[mhz]": 100.0,
  "N": 1,
  "latency_cycles": 46680
}
```

In [16]:

```
! cat {rtlsim_output_dir}/final_hw_config.json
```

```
{
  "Defaults": {},
  "StreamingFIFO_0": {
    "ram_style": "auto",
    "depth": 32,
    "impl_style": "rtl"
  },
  "FMPadding_Batch_0": {
    "SIMD": 1
  },
  "StreamingFIFO_1": {
    "ram_style": "auto",
    "depth": 256,
    "impl_style": "rtl"
  },
  "ConvolutionInputGenerator_0": {
    "SIMD": 1,
    "ram_style": "distributed"
  },
  "StreamingDataWidthConverter_Batch_0": {
```

In [17]:

```
import finn.builder.build_dataflow as build
import finn.builder.build_dataflow_config as build_cfg
import os
import shutil

#model_file = "model_v4noid_verified.onnx"
model_file = "verification/model_fmnist_final.onnx"

final_output_dir = "output_final"

#Delete previous run results if exist
if os.path.exists(final_output_dir):
    shutil.rmtree(final_output_dir)
    print("Previous run results deleted!")

cfg = build.DataflowBuildConfig(
    output_dir          = final_output_dir,
    mvau_wwidth_max     = 80,
    target_fps          = 1000000,
    synth_clk_period_ns = 10.0,
    board               = "ZCU104",
    fpga_part               = "xczu7ev-ffvc1156-2-e",
    shell_flow_type     = build_cfg.ShellFlowType.VIVADO_ZYNQ,
    generate_outputs=[
        build_cfg.DataflowOutputType.BITFILE,
        build_cfg.DataflowOutputType.PYNQ_DRIVER,
        build_cfg.DataflowOutputType.DEPLOYMENT_PACKAGE,
    ]
)
```

```
Previous run results deleted!
```

In [18]:

```
%%time
build.build_dataflow_cfg(model_file, cfg)
```

```
Building dataflow accelerator from verification/model_fmnist_final.onnx
Intermediate outputs will be generated in /home/mmirigaldi/finn_temp_mmirigaldi
Final outputs will be generated in output_final
Build log is at output_final/build_dataflow.log
Running step: step_qonnx_to_finn [1/17]
Running step: step_tidy_up [2/17]
Running step: step_streamline [3/17]
Running step: step_convert_to_hls [4/17]
Running step: step_create_dataflow_partition [5/17]
Running step: step_target_fps_parallelization [6/17]
Running step: step_apply_folding_config [7/17]
Running step: step_generate_estimate_reports [8/17]
Running step: step_hls_codegen [9/17]
Running step: step_hls_ipgen [10/17]
Running step: step_set_fifo_depths [11/17]
Running step: step_create_stitched_ip [12/17]
Running step: step_measure_rtlsim_performance [13/17]
Running step: step_out_of_context_synthesis [14/17]
Running step: step_synthesize_bitfile [15/17]
Running step: step_make_pynq_driver [16/17]
Running step: step_deployment_package [17/17]
Completed successfully
CPU times: user 38.3 s, sys: 675 ms, total: 38.9 s
Wall time: 38min 56s
```

Out[18]:

```
0
```

In [19]:

```
! ls {final_output_dir}/bitfile
```

```
finn-accel.bit  finn-accel.hwh
```

In [20]:

```
! ls {final_output_dir}/driver
```

```
driver.py  driver_base.py  finn  runtime_weights  validate.py
```

In [21]:

```
! ls {final_output_dir}/report
```

```
estimate_layer_resources_hls.json  post_synth_resources.xml
post_route_timing.rpt
```

In [22]:

```
! ls {final_output_dir}/deploy
```

```
bitfile  driver
```

In [23]:

```
! cp -r data2 {final_output_dir}/deploy/driver
```

In [24]:

```
! ls {final_output_dir}/deploy/driver
```

```
data2  driver.py  driver_base.py  finn  runtime_weights  validate.py
```

In [1]:

```
from shutil import make_archive
make_archive('deploy-final-on-pynq', 'zip', "output_final/deploy")
```

Out[1]:

```
'/workspace/finn/notebooks/mnist_ex/deploy-final-on-pynq.zip'
```

In [29]:

```
! zip -r mnist_ex.zip mnist_ex
```

```
  adding: mnist_ex/ (stored 0%)
  adding: mnist_ex/state_dict_self-trained.pth (deflated 28%)
  adding: mnist_ex/model_fmnist_notebook.onnx (deflated 72%)
  adding: mnist_ex/output_final/ (stored 0%)
  adding: mnist_ex/output_final/report/ (stored 0%)
  adding: mnist_ex/output_final/report/estimate_layer_resources_hls.json (deflated 82%)
  adding: mnist_ex/output_final/report/post_synth_resources.xml (deflated 96%)
  adding: mnist_ex/output_final/report/post_route_timing.rpt (deflated 95%)
  adding: mnist_ex/output_final/final_hw_config.json (deflated 85%)
  adding: mnist_ex/output_final/auto_folding_config.json (deflated 79%)
  adding: mnist_ex/output_final/deploy/ (stored 0%)
  adding: mnist_ex/output_final/deploy/driver/ (stored 0%)
  adding: mnist_ex/output_final/deploy/driver/data2/ (stored 0%)
  adding: mnist_ex/output_final/deploy/driver/data2/FashionMNIST/ (stored 0%)
  adding: mnist_ex/output_final/deploy/driver/data2/FashionMNIST/.ipynb_checkpoints/ (stored 0%)
  adding: mnist_ex/output_final/deploy/driver/data2/FashionMNIST/raw/ (stored 0%)
  adding: mnist_ex/output_final/deploy/driver/data2/FashionMNIST/raw/t10k-images-idx3-ubyte (deflated
44%)
  adding: mnist_ex/output_final/deploy/driver/data2/FashionMNIST/raw/FashionMNIST/ (stored 0%)
  adding: mnist_ex/output_final/deploy/driver/data2/FashionMNIST/raw/FashionMNIST/processed/ (stored
```

In [ ]: