XFDNN Compiler in a nutshell

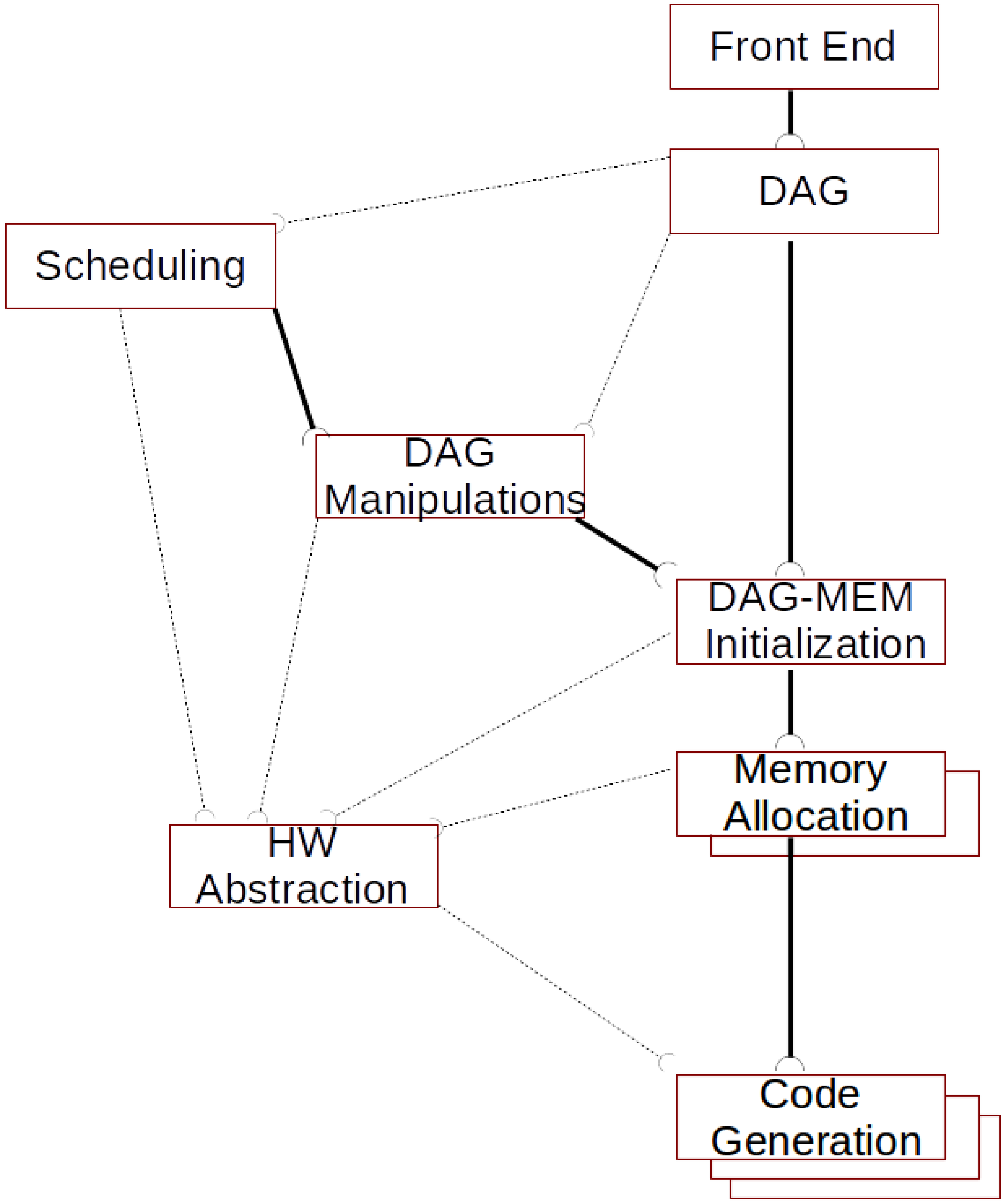
Name
Title
Date

XILINX.

Basic steps and basic ideas

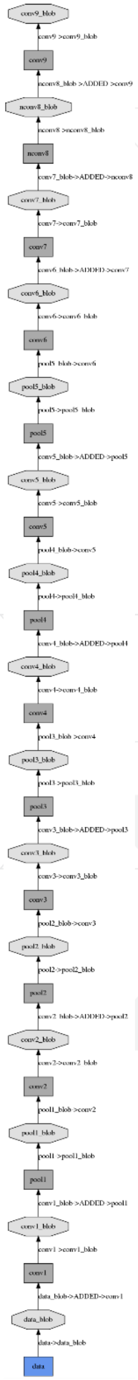
- Front end+Hardware specification
- Ruled based graph manipulations
- Scheduling exploration
- Graph Memory initialization
- AM only Memory Allocation
- DDR Memory
- Code Generation

Compiler


Front End
DAG
Scheduling
DAG Manipulations
DAG-MEM Initialization
Memory Allocation
HW Abstraction
Code Generation

Front end

- **The native format of the CNN dictates the CNN Structure**
  - We use mostly Caffe, TensorFlow
  - We MXNet, Keras, and a custom format for the math engine
  - Some Optimizations applied differently as func front end.
  - A graph of layers and volumes/tensors



HW abstraction

- **Hardware abstraction:**
  - DDR + Slices + Instruction encoding
    - Volume allocation abstraction, operations abstractions.
  - Slice = DSP + AM
  - Compute engine
    - DSP 28-56V2,
    - DSP 96 V3,
    - DSP 1-MathEngine and 32-64-Darius
  - Active memory
    - Every PCE has a custom AM (abstraction/custom Volume allocation)
  - Auxiliary small block (V3) + AM
    - For first layer

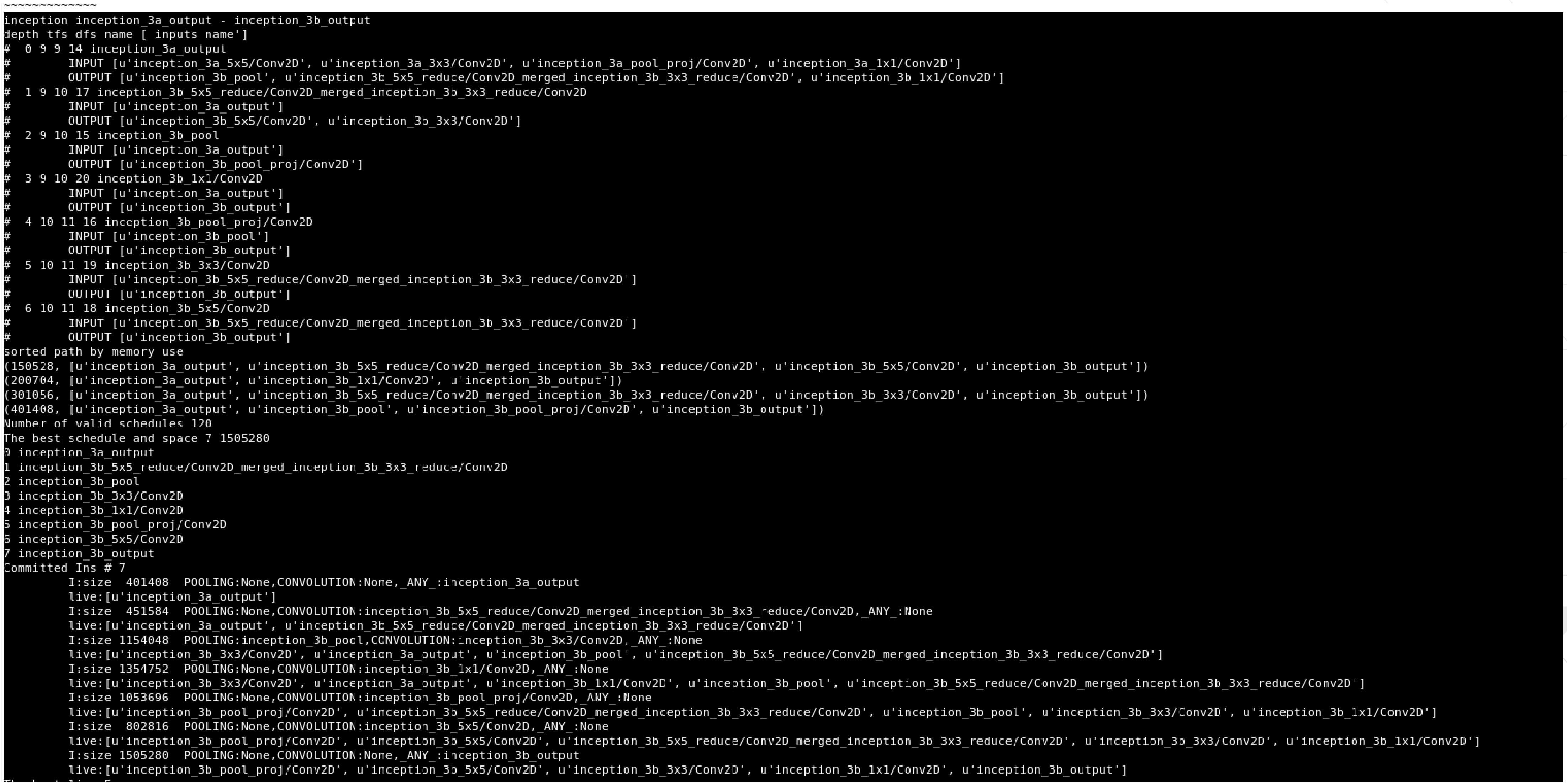
Rule-based graph manipulations

- **BatchNorm+Conv + BatchNorm + Scale + Relu**
  - Convolution
- **BN+Scale + Relu**
  - BN
- **BatchNorm into Scale**
- **conv_1x1_s2 = conv_1x1 + maxpool**
- **Conv+pool = pipelined**
- **Basically:**
  - If possible we represent CNN optimizations as graph manipulations
  - We apply as soon as possible.

Scheduling

- **Caffe native networks come with a default schedule**
  - The prototext describes the execution
- **TensorFlow the schedule by ops**
  - It is not assured to be topological
- **The compiler starts taking control and create multiple schedules**
- **Identifies inceptions**
  - Concat → Concat
  - ElementWise Addition → ElementWise
  - Small subgraphs
- **Creates all schedules**
  - Filter only valid,
  - Find the best time, best space, the minimum alive volumes
- **Stitch the subgraph schedules to create a set of schedules.**

Inception Example of scheduling



```
~~~~~~~~~~~~~~
inception inception_3a_output - inception_3b_output
depth tfs dfs name [ inputs name']
#  0 9 9 14 inception_3a_output
#         INPUT [u'inception_3a_5x5/Conv2D', u'inception_3a_3x3/Conv2D', u'inception_3a_pool_proj/Conv2D', u'inception_3a_1x1/Conv2D']
#         OUTPUT [u'inception_3b_pool', u'inception_3b_5x5_reduce/Conv2D_merged_inception_3b_3x3_reduce/Conv2D', u'inception_3b_1x1/Conv2D']
#  1 9 10 17 inception_3b_5x5_reduce/Conv2D_merged_inception_3b_3x3_reduce/Conv2D
#         INPUT [u'inception_3a_output']
#         OUTPUT [u'inception_3b_5x5/Conv2D', u'inception_3b_3x3/Conv2D']
#  2 9 10 15 inception_3b_pool
#         INPUT [u'inception_3a_output']
#         OUTPUT [u'inception_3b_pool_proj/Conv2D']
#  3 9 10 20 inception_3b_1x1/Conv2D
#         INPUT [u'inception_3a_output']
#         OUTPUT [u'inception_3b_output']
#  4 10 11 16 inception_3b_pool_proj/Conv2D
#         INPUT [u'inception_3b_pool']
#         OUTPUT [u'inception_3b_output']
#  5 10 11 19 inception_3b_3x3/Conv2D
#         INPUT [u'inception_3b_5x5_reduce/Conv2D_merged_inception_3b_3x3_reduce/Conv2D']
#         OUTPUT [u'inception_3b_output']
#  6 10 11 18 inception_3b_5x5/Conv2D
#         INPUT [u'inception_3b_5x5_reduce/Conv2D_merged_inception_3b_3x3_reduce/Conv2D']
#         OUTPUT [u'inception_3b_output']
sorted path by memory use
(150528, [u'inception_3a_output', u'inception_3b_5x5_reduce/Conv2D_merged_inception_3b_3x3_reduce/Conv2D', u'inception_3b_5x5/Conv2D', u'inception_3b_output'])
(200704, [u'inception_3a_output', u'inception_3b_1x1/Conv2D', u'inception_3b_output'])
(301056, [u'inception_3a_output', u'inception_3b_5x5_reduce/Conv2D_merged_inception_3b_3x3_reduce/Conv2D', u'inception_3b_3x3/Conv2D', u'inception_3b_output'])
(401408, [u'inception_3a_output', u'inception_3b_pool', u'inception_3b_pool_proj/Conv2D', u'inception_3b_output'])
Number of valid schedules 120
The best schedule and space 7 1505280
0 inception_3a_output
1 inception_3b_5x5_reduce/Conv2D_merged_inception_3b_3x3_reduce/Conv2D
2 inception_3b_pool
3 inception_3b_3x3/Conv2D
4 inception_3b_1x1/Conv2D
5 inception_3b_pool_proj/Conv2D
6 inception_3b_5x5/Conv2D
7 inception_3b_output
Committed Ins # 7
         I:size  401408  POOLING:None,CONVOLUTION:None,_ANY_:inception_3a_output
         live:[u'inception_3a_output']
         I:size  451584  POOLING:None,CONVOLUTION:inception_3b_5x5_reduce/Conv2D_merged_inception_3b_3x3_reduce/Conv2D,_ANY_:None
         live:[u'inception_3a_output', u'inception_3b_5x5_reduce/Conv2D_merged_inception_3b_3x3_reduce/Conv2D']
         I:size 1154048  POOLING:inception_3b_pool,CONVOLUTION:inception_3b_3x3/Conv2D,_ANY_:None
         live:[u'inception_3b_3x3/Conv2D', u'inception_3a_output', u'inception_3b_pool', u'inception_3b_5x5_reduce/Conv2D_merged_inception_3b_3x3_reduce/Conv2D']
         I:size 1354752  POOLING:None,CONVOLUTION:inception_3b_1x1/Conv2D,_ANY_:None
         live:[u'inception_3b_3x3/Conv2D', u'inception_3a_output', u'inception_3b_1x1/Conv2D', u'inception_3b_pool', u'inception_3b_5x5_reduce/Conv2D_merged_inception_3b_3x3_reduce/Conv2D']
         I:size 1053696  POOLING:None,CONVOLUTION:inception_3b_pool_proj/Conv2D,_ANY_:None
         live:[u'inception_3b_pool_proj/Conv2D', u'inception_3b_5x5_reduce/Conv2D_merged_inception_3b_3x3_reduce/Conv2D', u'inception_3b_pool', u'inception_3b_3x3/Conv2D', u'inception_3b_1x1/Conv2D']
         I:size  802816  POOLING:None,CONVOLUTION:inception_3b_5x5/Conv2D,_ANY_:None
         live:[u'inception_3b_pool_proj/Conv2D', u'inception_3b_5x5/Conv2D', u'inception_3b_5x5_reduce/Conv2D_merged_inception_3b_3x3_reduce/Conv2D', u'inception_3b_3x3/Conv2D', u'inception_3b_1x1/Conv2D']
         I:size 1505280  POOLING:None,CONVOLUTION:None,_ANY_:inception_3b_output
         live:[u'inception_3b_pool_proj/Conv2D', u'inception_3b_5x5/Conv2D', u'inception_3b_3x3/Conv2D', u'inception_3b_1x1/Conv2D', u'inception_3b_output']
```

Memory/Graph initialization

- **For every schedule**
- **Volume Space requirement (as if in AM)**
- **Data dependency as schedule iterations**
  - When a volume will be used in the schedule
- **Estimate space requirements per each step**
  - Minimum memory requirements
- **Initialize replication information for V3**
  - Volume replication
  - Input – layer – output
- **Compute naïve memory requirements**
- **REPLICATION optimizations here**

What is replication

- The main differences between V2 and V3
  - V2 aligned by width but
- A volume is aligned by channels V3
  - DSP = 96
- The volume will have padded space Module 96
  - 96*int(Channels // 96) + **channels%96**
- Padding in V3 can be REPLICATION or ZEROs
  - Zeros no op
- Replication can improve throughput if Conv is not 1x1
  - Replication 1, 2 … several times
  - Memory constraints

Memory Allocation AM

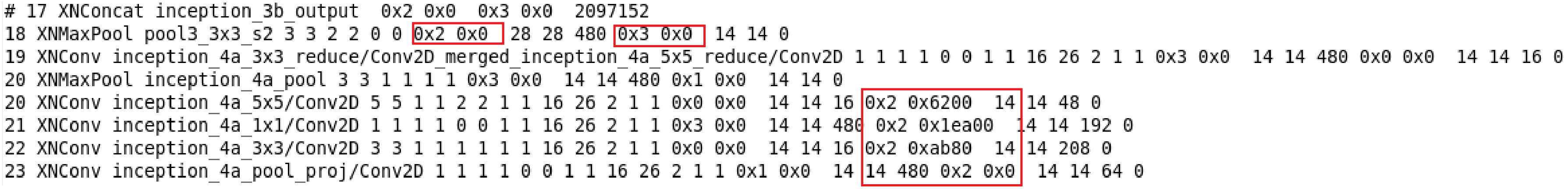
- **We have a set of heuristics**
  - Top, bottom, bysize
- **Following the schedule we allocate and deallocate space**
- **Every AM has different policy:**
  - Math engine: uses buffers
    - One volume into a buffer (unless concatenation)
    - No alignments requirements
  - V2 needs volumes aligned by width
  - V3 need volumes aligned by channels, So Darius
- **DDR has its own volume space policies**

Memory Allocation DDR

- **No heuristic worked (fine we have one last resort)**
- **We introduce DDR and a two level memory**
- **A layer is computed once**
  - Its output volume can be in different memories and locations as the computation goes along
  - Its size in DDR is different from the size in AM
- **We have HW gather—scatter memory operations**
  - We have an offline memory management
- **HW allows to execute operation directly from DDR**
  - V2: Compiler create code to compute all tiles
    - Gather, Tiled operation, Scatter
  - V3: Compiler estimate the maximum Tile
    - HW schedule the operation, double buffering and hiding data movements.

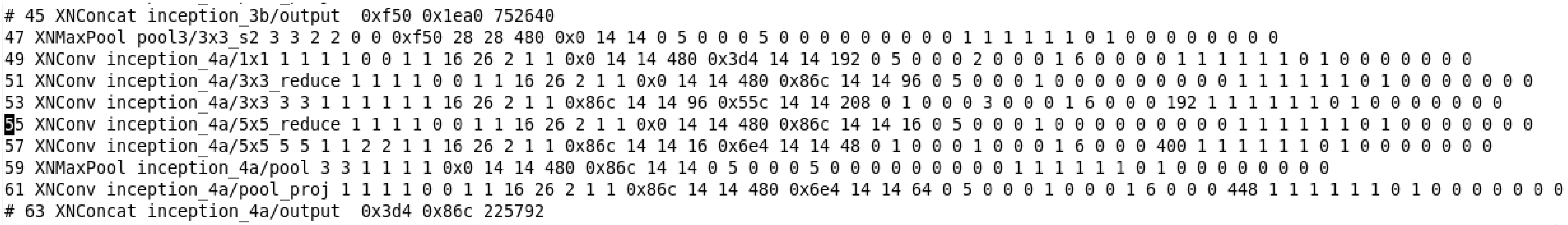
Code Generation

Math: addresses are buffer + offset and schedule optimized conv//pool



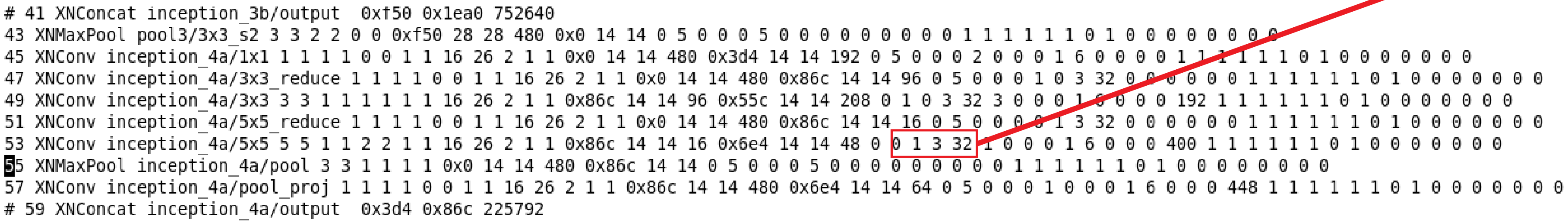
```
# 17 XNConcat inception_3b_output  0x2 0x0  0x3 0x0  2097152
18 XNMaxPool pool3_3x3_s2 3 3 2 2 0 0 0x2 0x0  28 28 480 0x3 0x0  14 14 0
19 XNConv inception_4a_3x3_reduce/Conv2D_merged_inception_4a_5x5_reduce/Conv2D 1 1 1 1 0 0 1 1 16 26 2 1 1 0x3 0x0  14 14 480 0x0 0x0  14 14 16 0
20 XNMaxPool inception_4a_pool 3 3 1 1 1 1 0x3 0x0  14 14 480 0x1 0x0  14 14 0
20 XNConv inception_4a_5x5/Conv2D 5 5 1 1 2 2 1 1 16 26 2 1 1 0x0 0x0  14 14 16 0x2 0x6200  14 14 48 0
21 XNConv inception_4a_1x1/Conv2D 1 1 1 1 0 0 1 1 16 26 2 1 1 0x3 0x0  14 14 480 0x2 0x1ea00  14 14 192 0
22 XNConv inception_4a_3x3/Conv2D 3 3 1 1 1 1 1 1 16 26 2 1 1 0x0 0x0  14 14 16 0x2 0xab80  14 14 208 0
23 XNConv inception_4a_pool_proj/Conv2D 1 1 1 1 0 0 1 1 16 26 2 1 1 0x1 0x0  14 14 480 0x2 0x0  14 14 64 0
```

V3: with replication “off”



```
# 45 XNConcat inception_3b/output  0xf50 0x1ea0 752640
47 XNMaxPool pool3/3x3_s2 3 3 2 2 0 0 0xf50 28 28 480 0x0 14 14 0 5 0 0 0 5 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 1 0 0 0 0 0 0 0 0
49 XNConv inception_4a/1x1 1 1 1 1 0 0 1 1 16 26 2 1 1 0x0 14 14 480 0x3d4 14 14 192 0 5 0 0 0 2 0 0 0 1 6 0 0 0 0 1 1 1 1 1 1 0 1 0 0 0 0 0 0 0
51 XNConv inception_4a/3x3_reduce 1 1 1 1 0 0 1 1 16 26 2 1 1 0x0 14 14 480 0x86c 14 14 96 0 5 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 1 0 0 0 0 0 0 0
53 XNConv inception_4a/3x3 3 3 1 1 1 1 1 1 16 26 2 1 1 0x86c 14 14 96 0x55c 14 14 208 0 1 0 0 0 3 0 0 0 1 6 0 0 0 192 1 1 1 1 1 1 0 1 0 0 0 0 0 0 0
55 XNConv inception_4a/5x5_reduce 1 1 1 1 0 0 1 1 16 26 2 1 1 0x0 14 14 480 0x86c 14 14 16 0 5 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 1 0 0 0 0 0 0 0
57 XNConv inception_4a/5x5 5 5 1 1 2 2 1 1 16 26 2 1 1 0x86c 14 14 16 0x6e4 14 14 48 0 1 0 0 0 1 0 0 0 1 6 0 0 0 400 1 1 1 1 1 1 0 1 0 0 0 0 0 0 0
59 XNMaxPool inception_4a/pool 3 3 1 1 1 1 0x0 14 14 480 0x86c 14 14 0 5 0 0 0 5 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 1 0 0 0 0 0 0 0 0
61 XNConv inception_4a/pool_proj 1 1 1 1 0 0 1 1 16 26 2 1 1 0x86c 14 14 480 0x6e4 14 14 64 0 5 0 0 0 1 0 0 0 1 6 0 0 0 448 1 1 1 1 1 1 0 1 0 0 0 0 0 0 0
# 63 XNConcat inception_4a/output  0x3d4 0x86c 225792
```

V3: with replication “on”



```
# 41 XNConcat inception_3b/output  0xf50 0x1ea0 752640
43 XNMaxPool pool3/3x3_s2 3 3 2 2 0 0 0xf50 28 28 480 0x0 14 14 0 5 0 0 0 5 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 1 0 0 0 0 0 0 0 0
45 XNConv inception_4a/1x1 1 1 1 1 0 0 1 1 16 26 2 1 1 0x0 14 14 480 0x3d4 14 14 192 0 5 0 0 0 2 0 0 0 1 6 0 0 0 0 1 1 1 1 1 1 0 1 0 0 0 0 0 0 0
47 XNConv inception_4a/3x3_reduce 1 1 1 1 0 0 1 1 16 26 2 1 1 0x0 14 14 480 0x86c 14 14 96 0 5 0 0 0 1 0 3 32 0 0 0 0 0 0 1 1 1 1 1 1 0 1 0 0 0 0 0 0 0
49 XNConv inception_4a/3x3 3 3 1 1 1 1 1 1 16 26 2 1 1 0x86c 14 14 96 0x55c 14 14 208 0 1 0 3 32 3 0 0 0 1 6 0 0 0 192 1 1 1 1 1 1 0 1 0 0 0 0 0 0 0
51 XNConv inception_4a/5x5_reduce 1 1 1 1 0 0 1 1 16 26 2 1 1 0x0 14 14 480 0x86c 14 14 16 0 5 0 0 0 0 1 3 32 0 0 0 0 0 0 1 1 1 1 1 1 0 1 0 0 0 0 0 0 0
53 XNConv inception_4a/5x5 5 5 1 1 2 2 1 1 16 26 2 1 1 0x86c 14 14 16 0x6e4 14 14 48 0 0 1 3 32 1 0 0 0 1 6 0 0 0 400 1 1 1 1 1 1 0 1 0 0 0 0 0 0 0
55 XNMaxPool inception_4a/pool 3 3 1 1 1 1 0x0 14 14 480 0x86c 14 14 0 5 0 0 0 5 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 1 0 0 0 0 0 0 0 0
57 XNConv inception_4a/pool_proj 1 1 1 1 0 0 1 1 16 26 2 1 1 0x86c 14 14 480 0x6e4 14 14 64 0 5 0 0 0 1 0 0 0 1 6 0 0 0 448 1 1 1 1 1 1 0 1 0 0 0 0 0 0 0
# 59 XNConcat inception_4a/output  0x3d4 0x86c 225792
```

16 channels :
1 replication session
3 replications of size
32 channels (min)

> Thank you