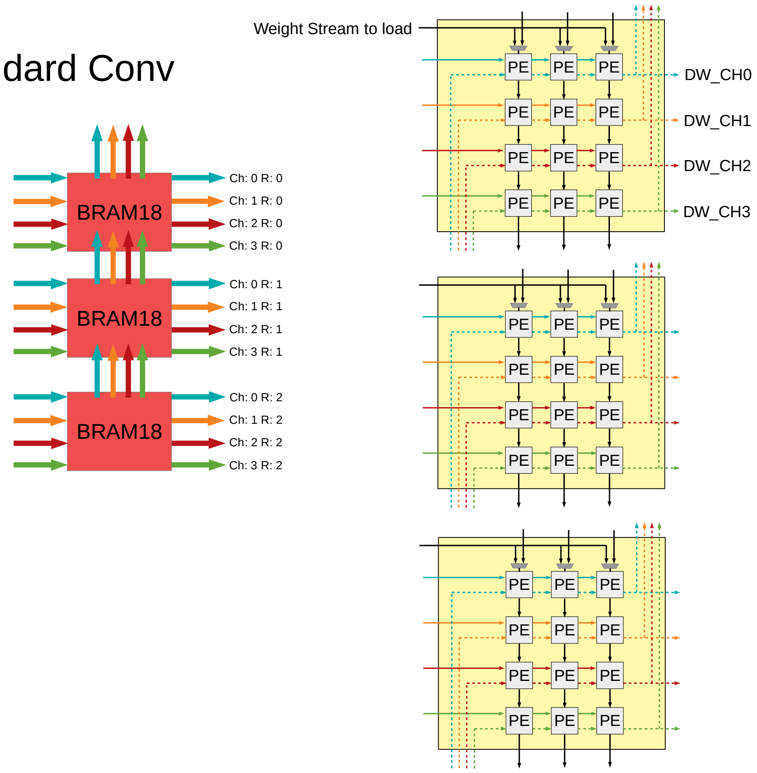
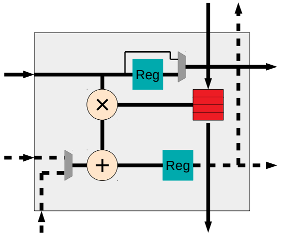
MLBlocks: Arming FPGA architectures with Dense & Low Precision units in classic column based manner

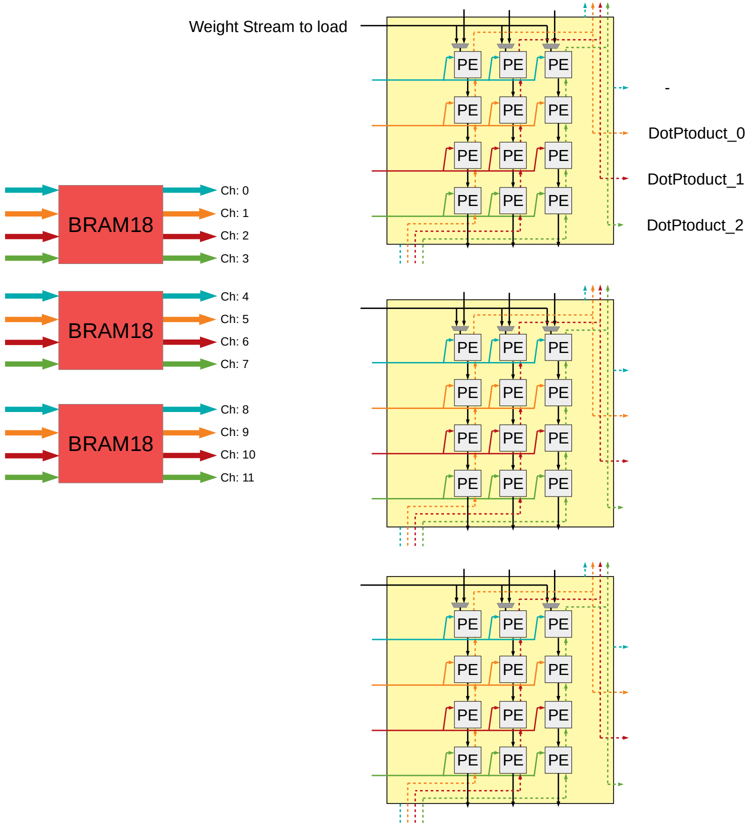
DWConv & Standard Conv



PE



PWConv & Matrix-Matrix Multiplication



Contributions:

1- Reconfigurable PE (**DSP size**, Systolic array, **Column based**)

6 times more 8x8 multiplier comparing to a DSP Block (two 8x8),

RS Data flow, High frequency, flexible data movement. Great for **SConv, DWConv, PWConv, Matrix-Matrix Multiplication**

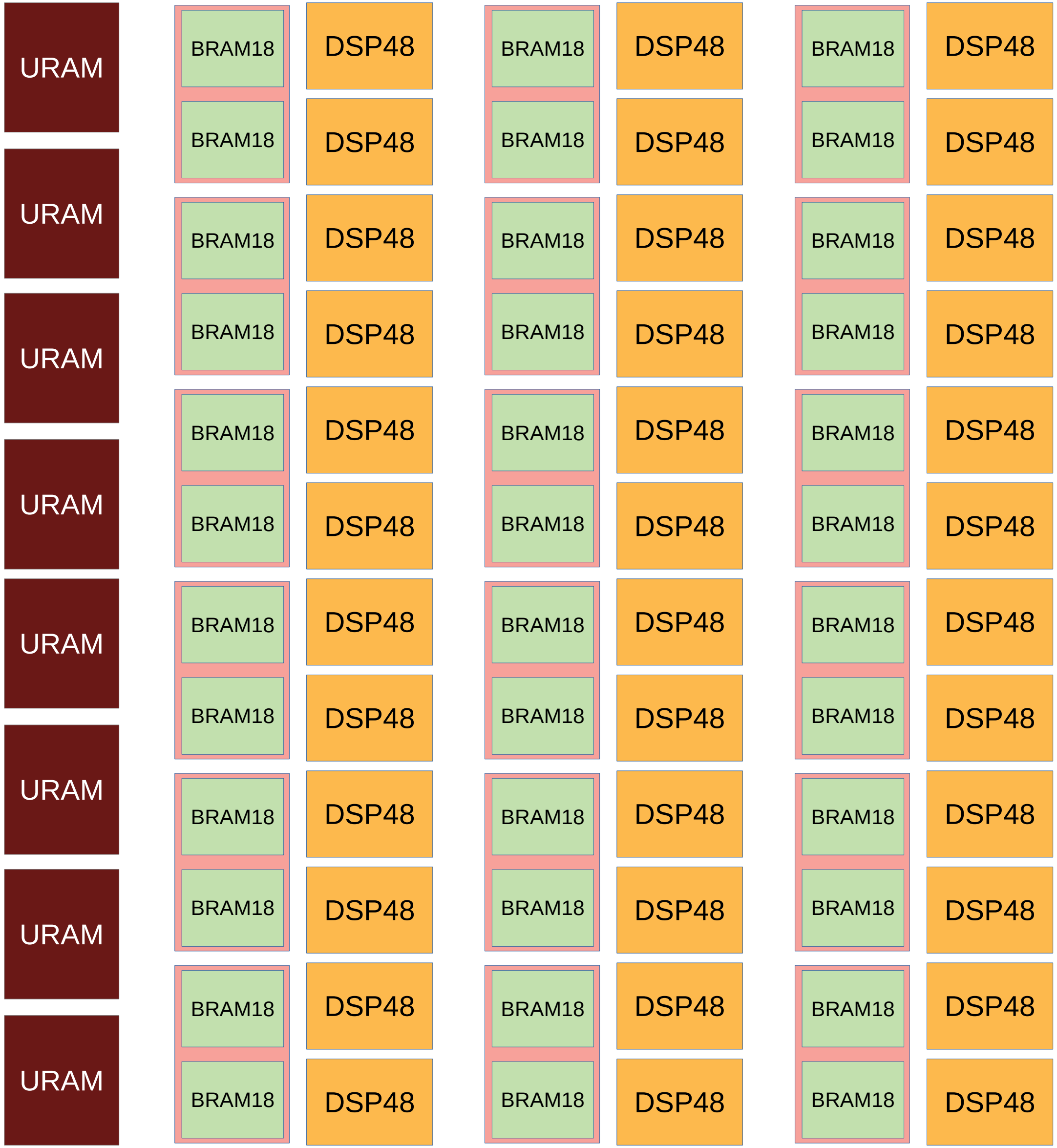
DPS – BRAM ratio 1/1 (same as Ultrascale+ arch), Low number of intermediary outputs in practice

Parameterized (for any budget limitation) – can integrate multi precision idea

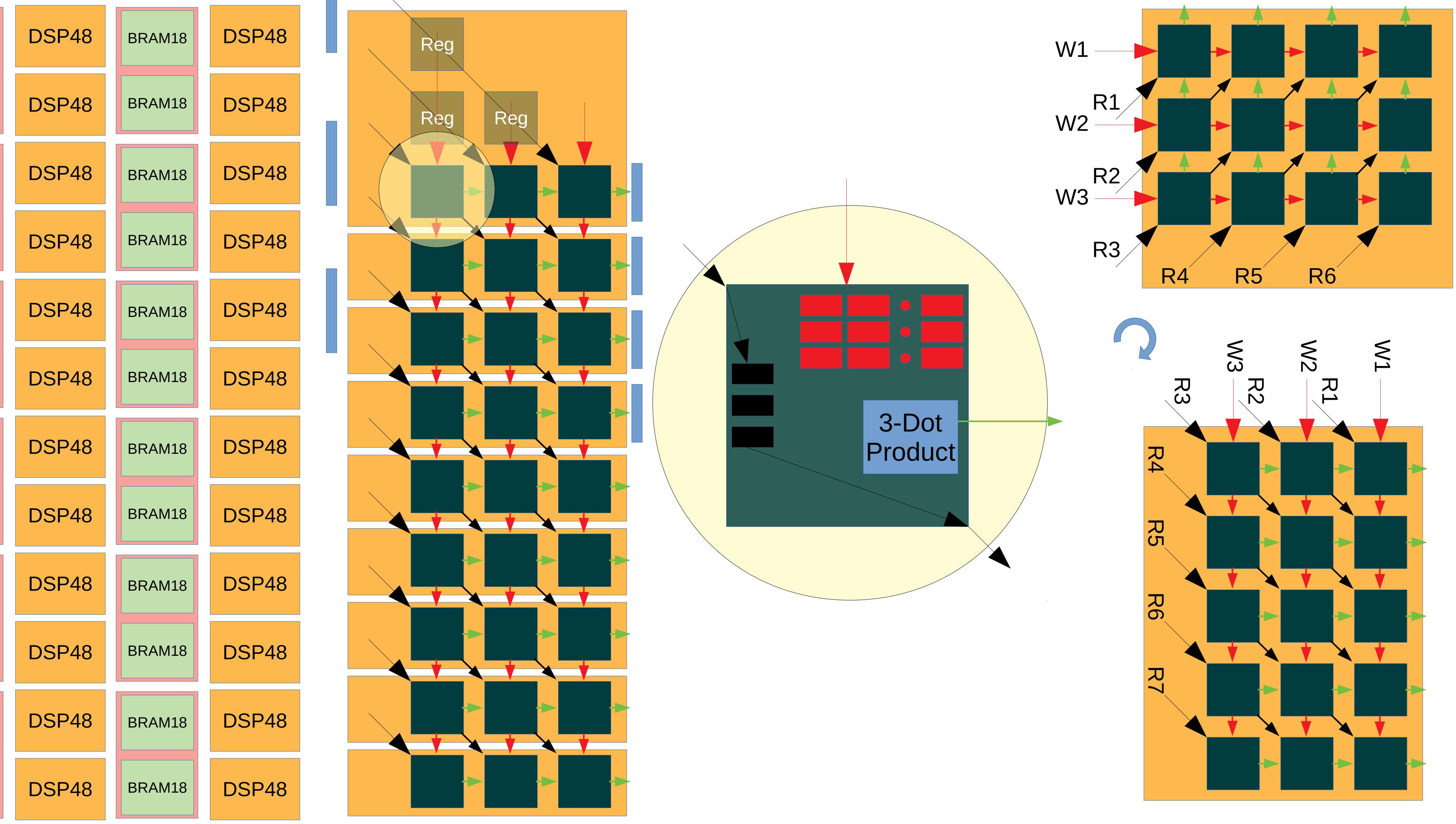
2- Compare with cascade paper (Prof. Nachiket)

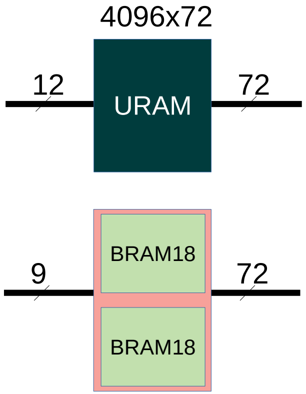
3- new suggestion to use each 18KBRAM as 36bit streamer using external controler circuit (delivering 662MHz) (in cascade paper: 18bit)

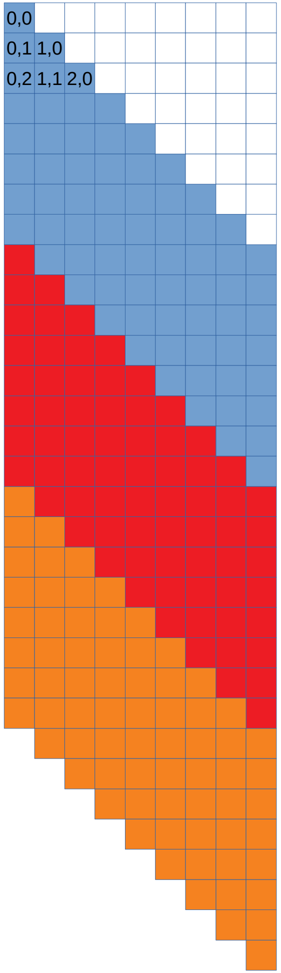
UltraScale+ architecture distribution:

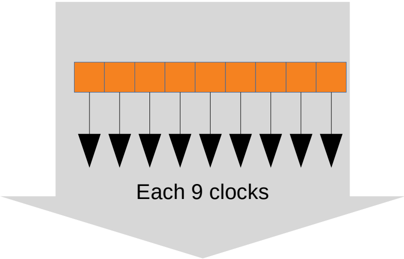

URAM
URAM
URAM
URAM
URAM
URAM
URAM
URAM
BRAM18 DSP48 BRAM18 DSP48 BRAM18 DSP48
BRAM18 DSP48 BRAM18 DSP48 BRAM18 DSP48
BRAM18 DSP48 BRAM18 DSP48 BRAM18 DSP48
BRAM18 DSP48 BRAM18 DSP48 BRAM18 DSP48
BRAM18 DSP48 BRAM18 DSP48 BRAM18 DSP48
BRAM18 DSP48 BRAM18 DSP48 BRAM18 DSP48
BRAM18 DSP48 BRAM18 DSP48 BRAM18 DSP48
BRAM18 DSP48 BRAM18 DSP48 BRAM18 DSP48
BRAM18 DSP48 BRAM18 DSP48 BRAM18 DSP48
BRAM18 DSP48 BRAM18 DSP48 BRAM18 DSP48
BRAM18 DSP48 BRAM18 DSP48 BRAM18 DSP48
BRAM18 DSP48 BRAM18 DSP48 BRAM18 DSP48

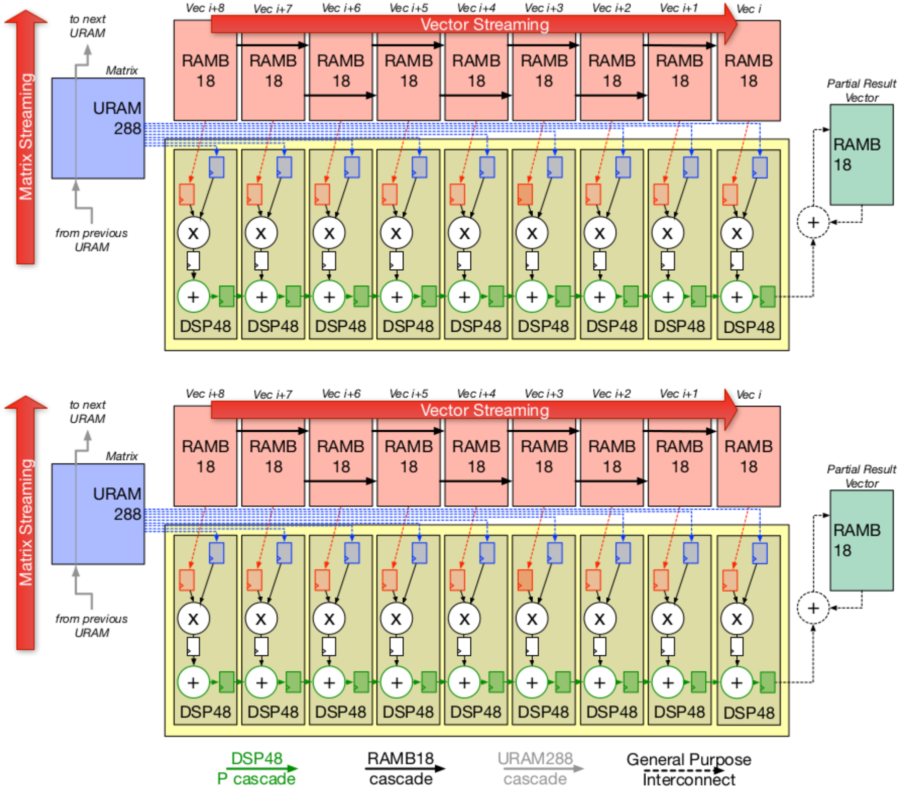
My amassing MLBlocks world

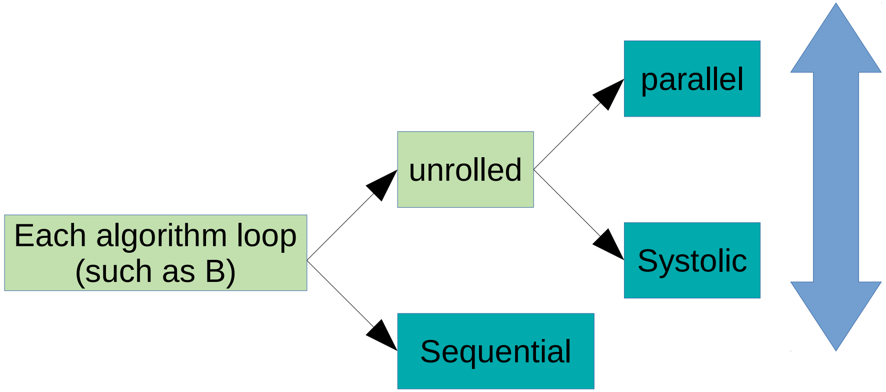

DSP48
BRAM18
DSP48
DSP48
BRAM18
DSP48
DSP48
BRAM18
DSP48
DSP48
BRAM18
DSP48
DSP48
BRAM18
DSP48
DSP48
BRAM18
DSP48
DSP48
BRAM18
DSP48
DSP48
BRAM18
DSP48
DSP48
BRAM18
DSP48
DSP48
BRAM18
DSP48
DSP48
BRAM18
DSP48
DSP48
BRAM18
DSP48
Reg
Reg
Reg
3-Dot Product
W1
R1
W2
R2
W3
R3
R4
R5
R6
W3
W2
W1
R3
R2
R1
R4
R5
R6
R7


4096x72
12
URAM
72
9
BRAM18
72
BRAM18


0,0
0,1 1,0
0,2 1,1 2,0


Each 9 clocks


Matrix Streaming
to next URAM
Matrix
URAM 288
from previous URAM
Vec i+8
Vec i+7
Vec i+6
Vec i+5
Vec i+4
Vec i+3
Vec i+2
Vec i+1
Vec i
Vector Streaming
RAMB 18
X
+
DSP48
Partial Result Vector
RAMB 18
DSP48 P cascade
RAMB18 cascade
URAM288 cascade
General Purpose Interconnect



Dis Parallel:

1- more fan in and outs (since we are talking about small Pes it is fine)

Dis Systolic:

1- tougher scheduling, rythmic scheduling

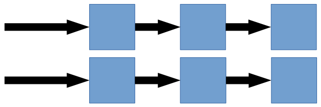
2- prevent circuit fusions (less optimization)

$B = B_{Seq} \times B_{par} \times B_{Sys}$

# of Physical MAC: $\times B_{par} \times B_{Sys}$

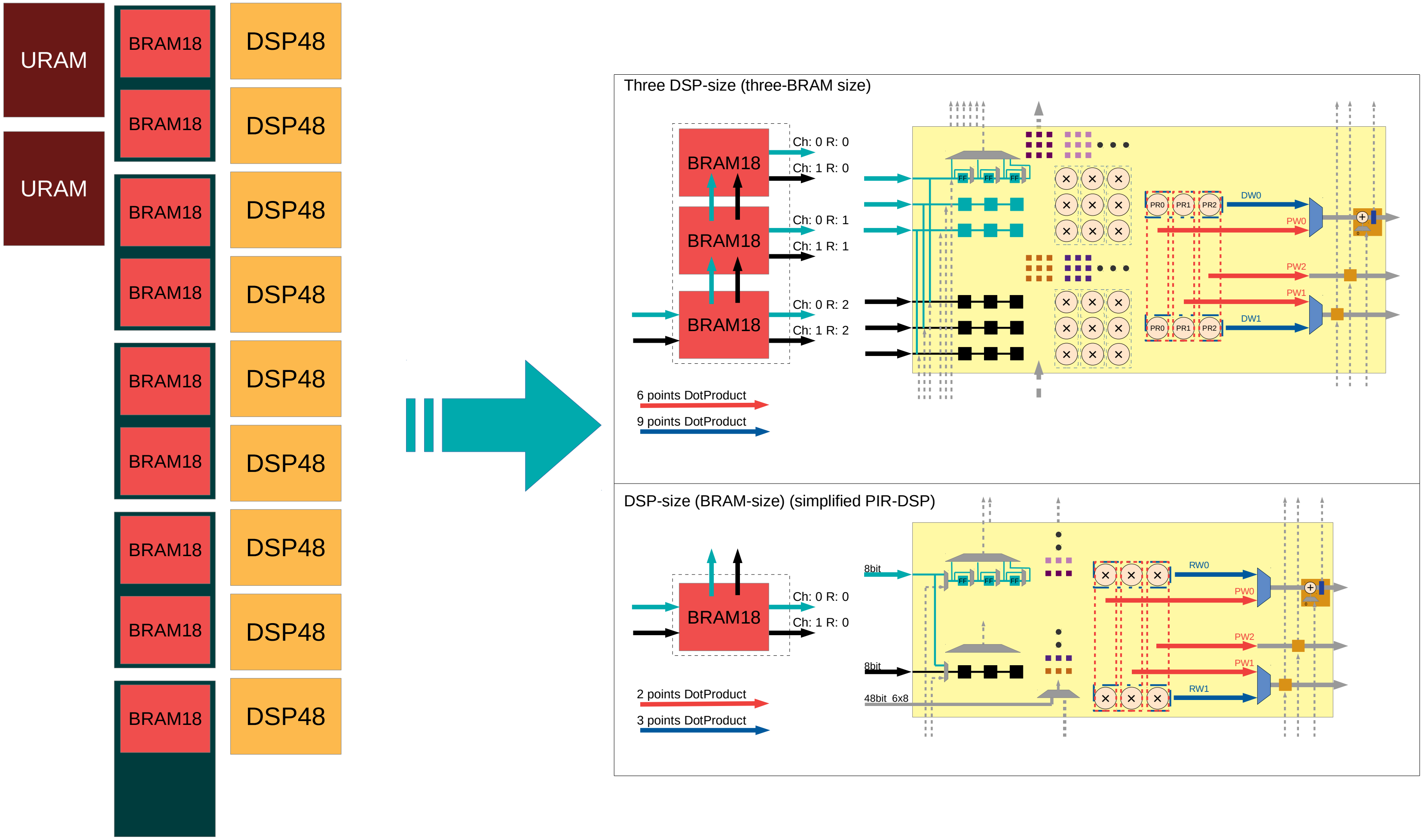
# of Input: $\times B_{par}$ (without internal serial to parallel)

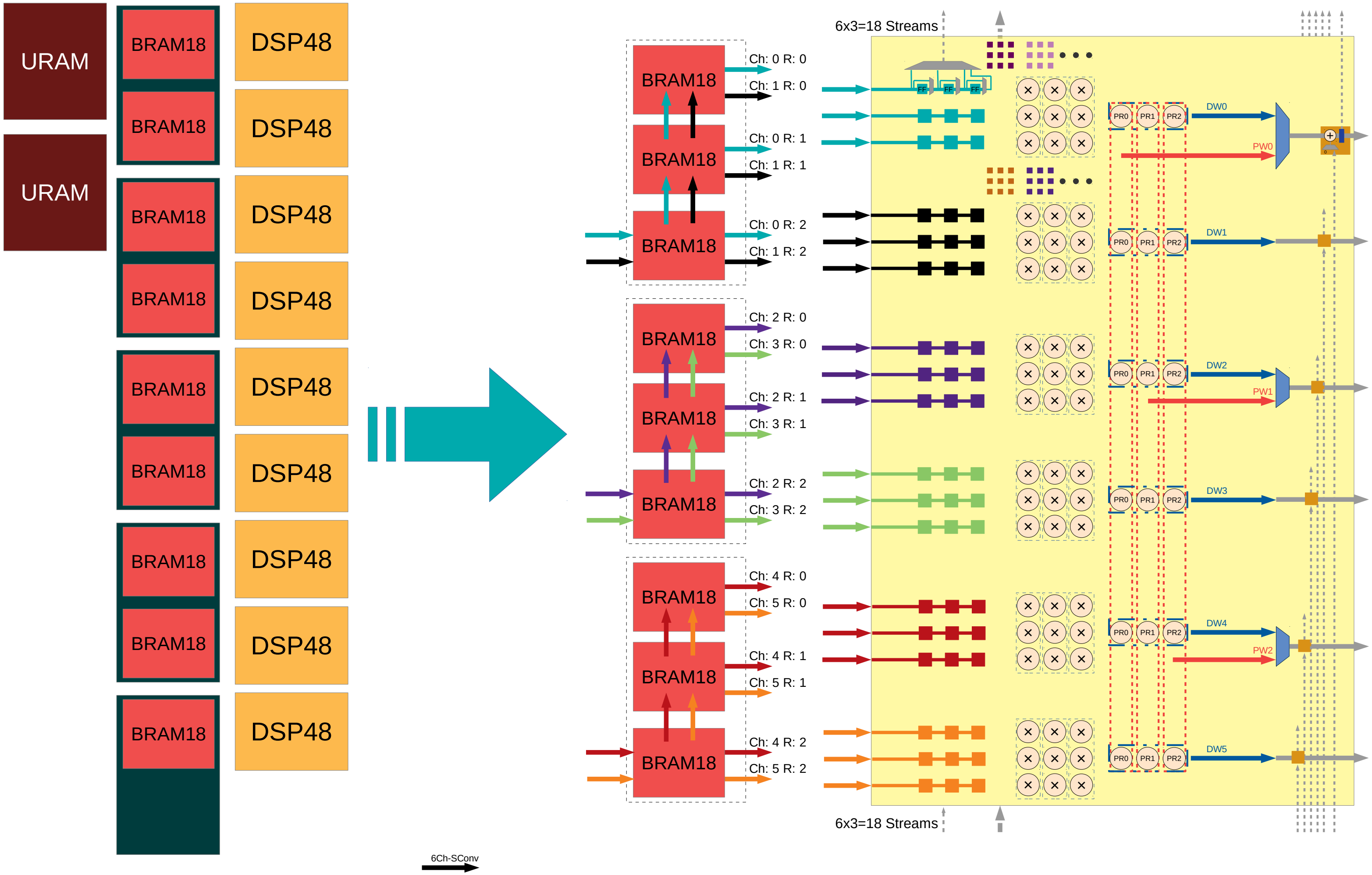
# of Output: $\times B_{par}$



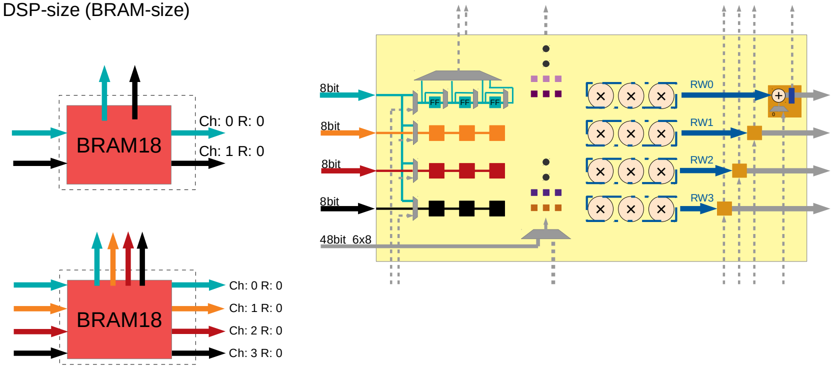
Params = {right side indexes}

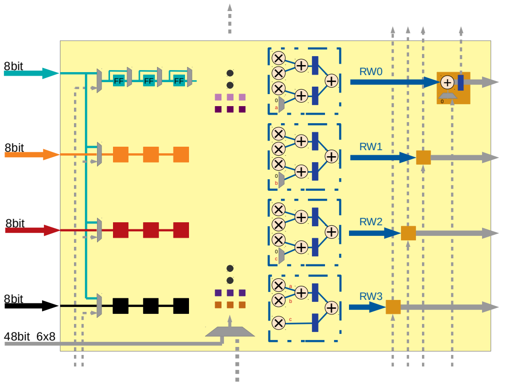
```
for $param^i_{sch\_0}$:  $0 \rightarrow sch^i\_0$
    for $param^i_{sch\_1}$:  $0 \rightarrow sch^i\_1$
        for $param^i_{sch\_2}$:  $0 \rightarrow sch^i\_2$

            for $param^i_{seq}$:  $0 \rightarrow comp\_seq^i$
                for $param^i_{un}$:  $0 \rightarrow comp\_un^i$
```


URAM
URAM
BRAM18
DSP48
Three DSP-size (three-BRAM size)
BRAM18
BRAM18
BRAM18
Ch: 0 R: 0
Ch: 1 R: 0
Ch: 0 R: 1
Ch: 1 R: 1
Ch: 0 R: 2
Ch: 1 R: 2
FF
PR0
PR1
PR2
DW0
PW0
PW2
PW1
DW1
6 points DotProduct
9 points DotProduct
DSP-size (BRAM-size) (simplified PIR-DSP)
BRAM18
Ch: 0 R: 0
Ch: 1 R: 0
8bit
8bit
48bit 6x8
RW0
PW0
PW2
PW1
RW1
2 points DotProduct
3 points DotProduct

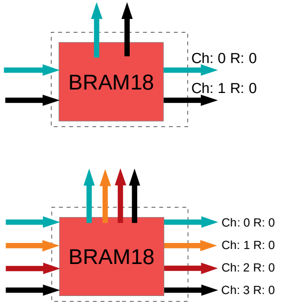

URAM
URAM
BRAM18
DSP48
BRAM18
Ch: 0 R: 0
Ch: 1 R: 0
Ch: 0 R: 1
Ch: 1 R: 1
Ch: 0 R: 2
Ch: 1 R: 2
Ch: 2 R: 0
Ch: 3 R: 0
Ch: 2 R: 1
Ch: 3 R: 1
Ch: 2 R: 2
Ch: 3 R: 2
Ch: 4 R: 0
Ch: 5 R: 0
Ch: 4 R: 1
Ch: 5 R: 1
Ch: 4 R: 2
Ch: 5 R: 2
6x3=18 Streams
FF
PR0
PR1
PR2
DW0
DW1
DW2
DW3
DW4
DW5
PW0
PW1
PW2
6x3=18 Streams
6Ch-SConv

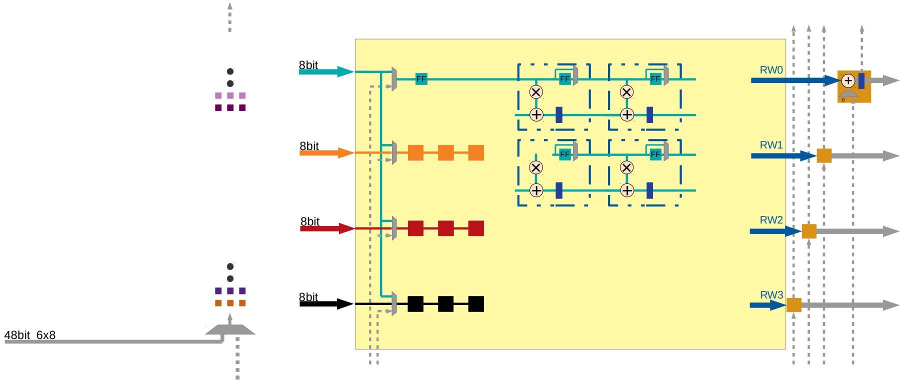
DSP-size (BRAM-size)





DSP-size (BRAM-size)


BRAM18
Ch: 0 R: 0
Ch: 1 R: 0
BRAM18
Ch: 0 R: 0
Ch: 1 R: 0
Ch: 2 R: 0
Ch: 3 R: 0


8bit
8bit
8bit
8bit
FF
FF
FF
FF
FF
RW0
RW1
RW2
RW3
48bit 6x8