What Is Interesting To Me Right Now …

Mark Horowitz

7/12/2018

Stanford | ENGINEERING
Electrical Engineering

APPLICATIONS TO HARDWARE

If You Are Building an Accelerator

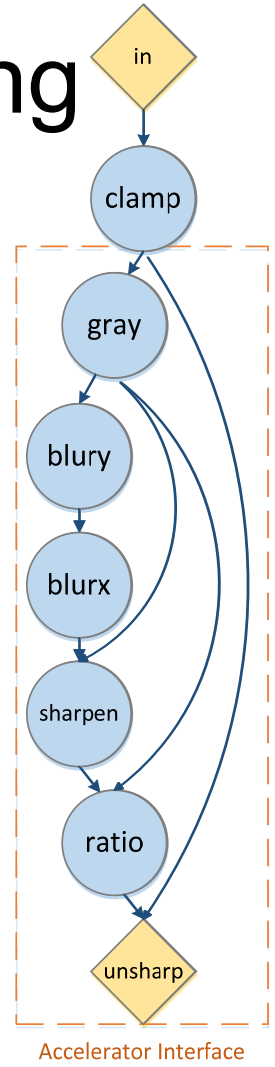
- The application better be parallel
- The application better have locality
- Performance/energy must be a critical issue

Remember It Must Be Cheap To Design

- Application designer accessible
- Little silicon expertise

Halide Example: Unsharp Masking

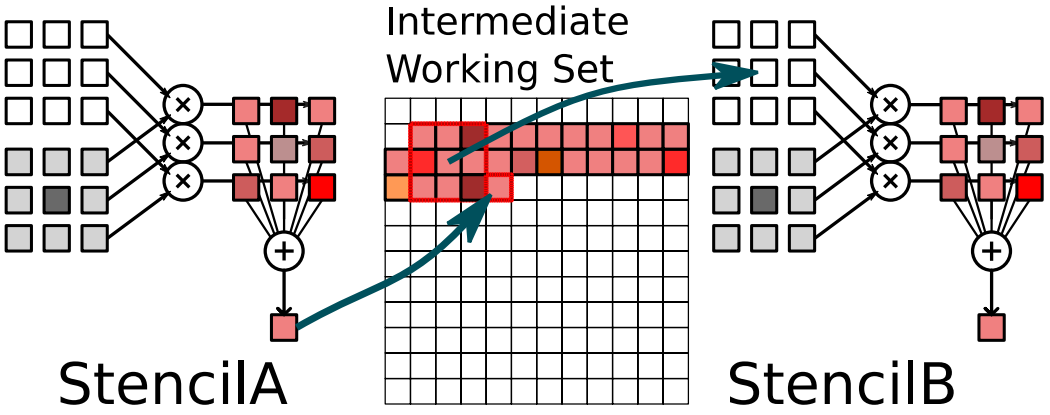
```
Func clamp, gray, blurx, blury, sharpen, ratio, unsharp;
Var x, y, c, xi, yi;

// The algorithm
clamp = BoundaryConditions::repeat_edge(in);
gray(x, y) = 0.3*clamp(0, x, y)+0.6*clamp(1, x, y)+0.1*clamp(2, x, y);
blury(x, y) = (gray(x, y-1) + gray(x, y) + gray(x, y+1)) / 3;
blurx(x, y) = (blury(x-1, y) + blury(x, y) + blury(x+1, y)) / 3;
sharpen(x, y) = 2 * gray(x, y) - blurx(x, y);
ratio(x, y) = sharpen(x, y) / gray(x, y);
unsharp(c, x, y) = ratio(x, y) * input(c, x, y);

// The schedule
unsharp.tile(x, y, xi, yi, 256, 256)
       .accelerate({clamp}, xi, x)
       .parallel(y).parallel(x);
in.fifo_depth(unsharp, 512);
gray.fifo_depth(ratio, 8);
// other schedules...
```



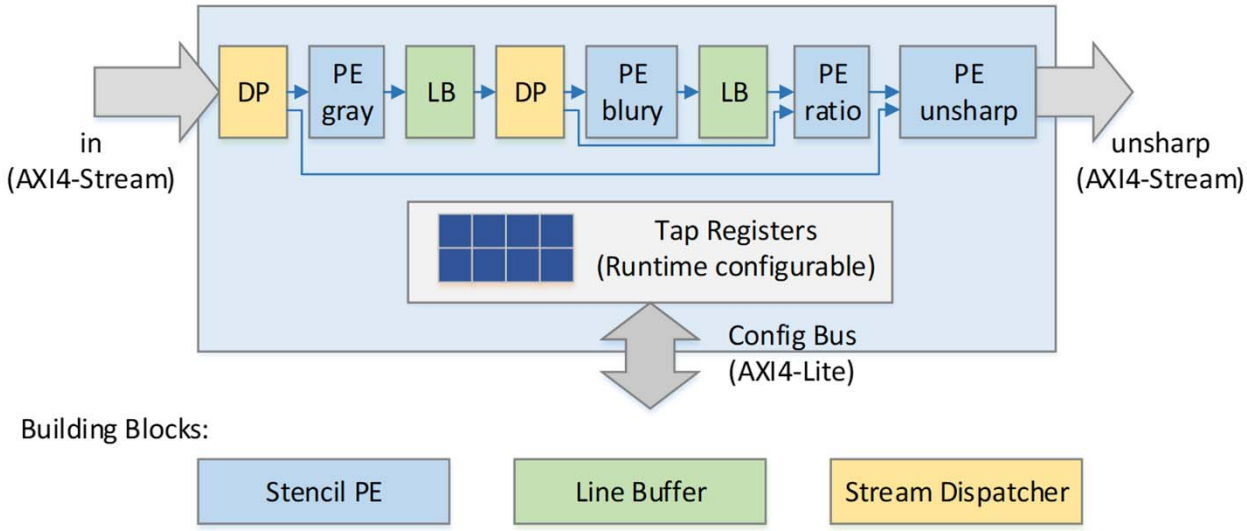
Architecture Template:

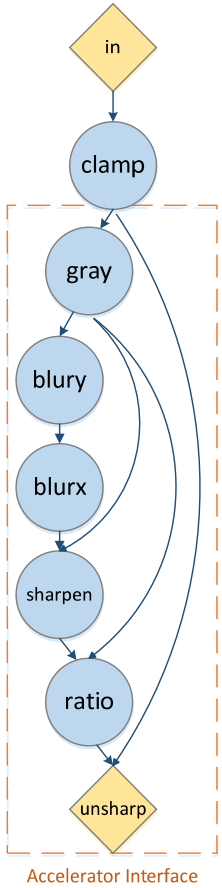
Stencil Functions and Line Buffers

- Stencil functions consume sliding windows of data
  - Huge locality
- To capture this locality need to buffer a few lines
  - Line buffer is the hardware buffer block.



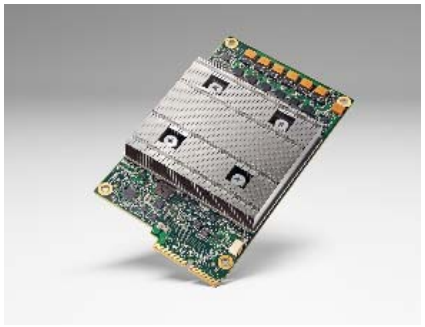
Dataflow IR Transformation



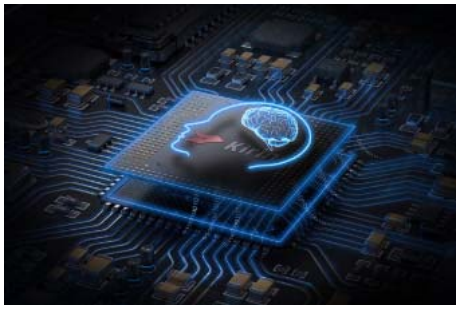


Pu, Jing, et al. "Programming heterogeneous systems from an image processing DSL." ACM Transactions on Architecture and Code Optimization (TACO) 14.3 (2017): 26

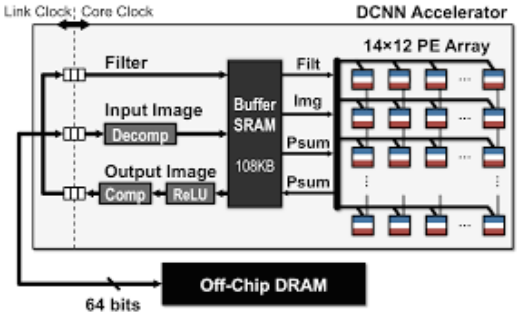
DNN Accelerators



Google TPU



Huawei Kirin NPU



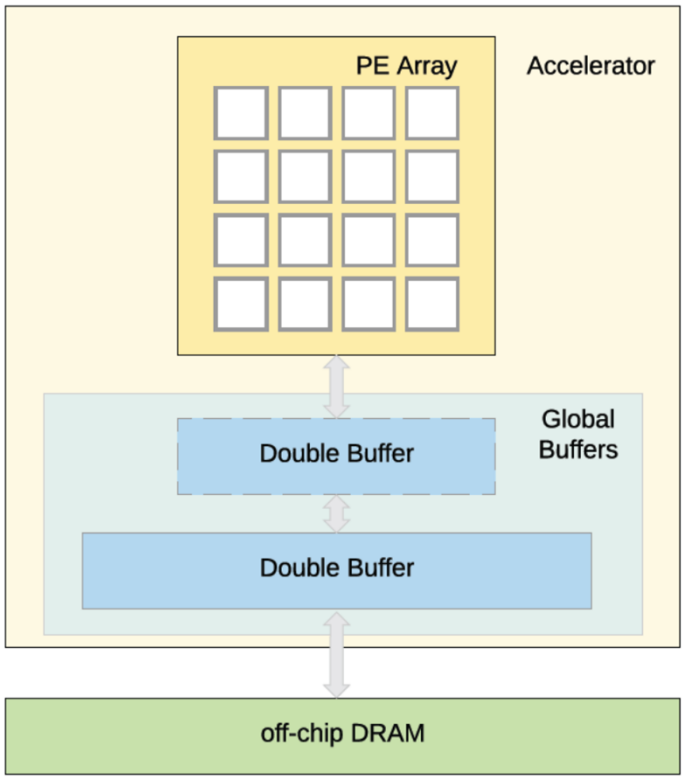
Eyeriss

Can represent them as different schedules to a Halide program.

C. Zheng, et al (FPGA)

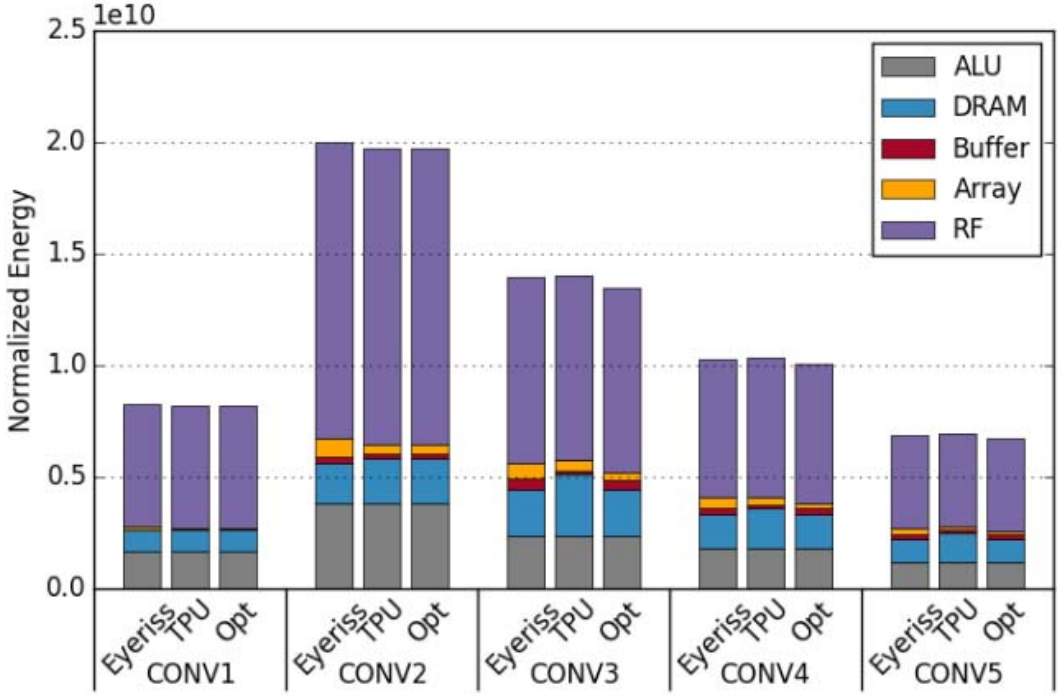
[1] https://cloudplatform.googleblog.com/2016/05/Google-supercharges-machine-learning-tasks-with-custom-chip.html
[2] https://www.electronicsweekly.com/news/ifa-2017-huawei-reveals-low-power-kirin-970-mobile-ai-chipset-2017-09/
[3] http://eyeriss.mit.edu/

Architecture Template for DNNs



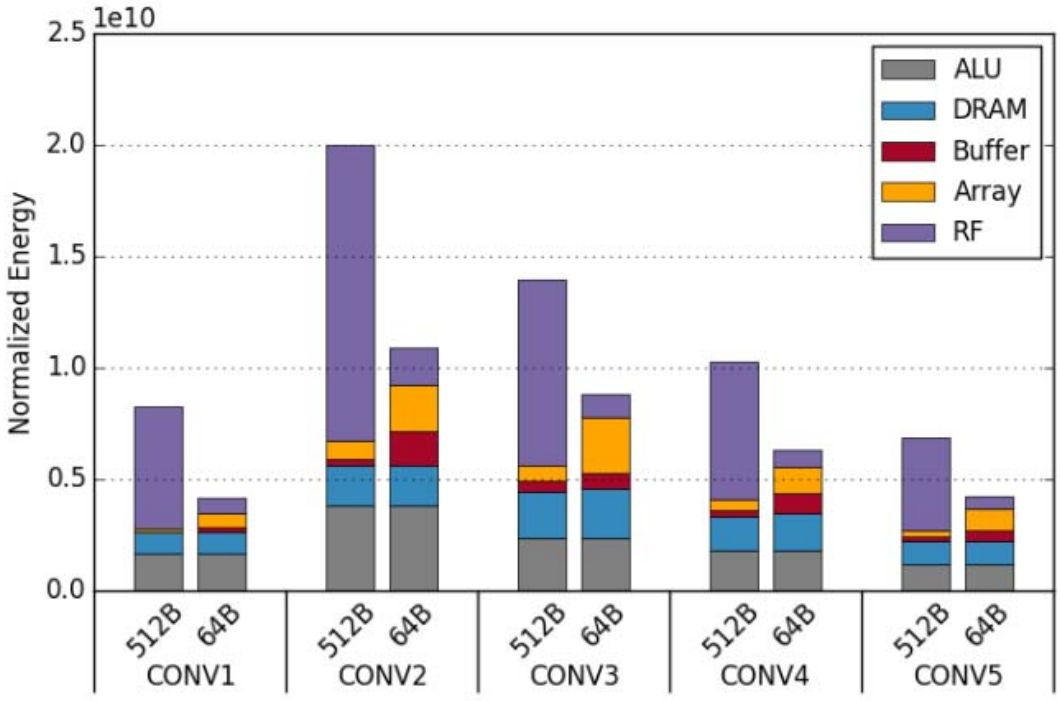
Dataflow Impact

- Most dataflows achieve close-to-optimal energy efficiency



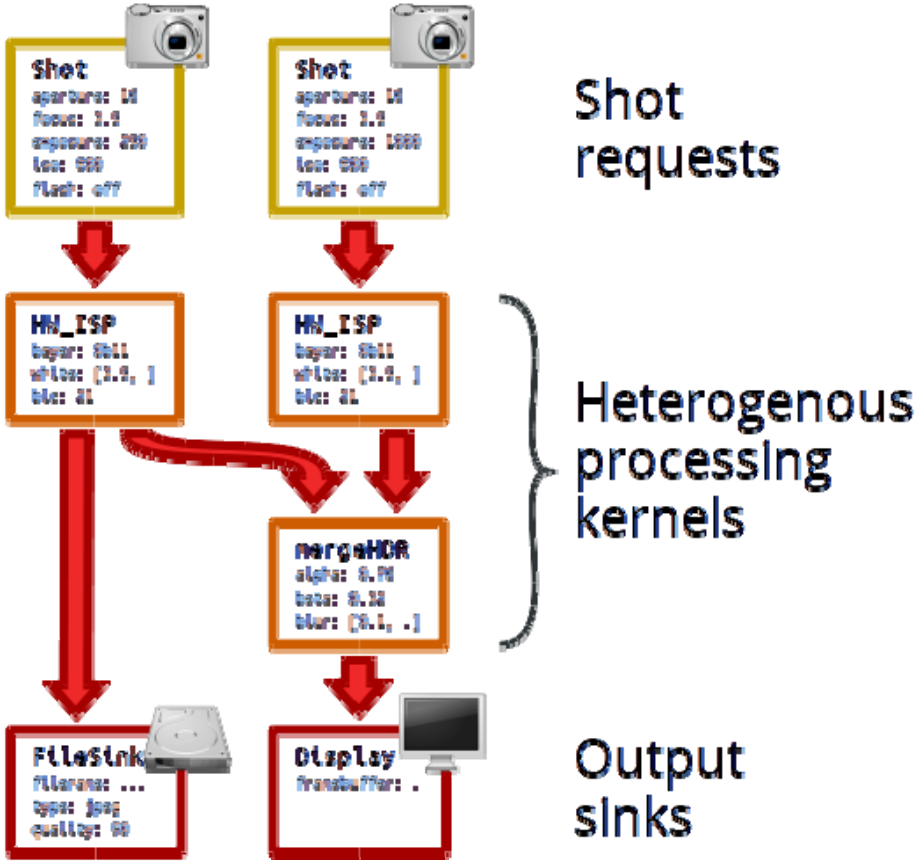
Energy Breakdown Comparison

- Use a smaller register file size can greatly improve overall energy efficiency, by reducing register file energy

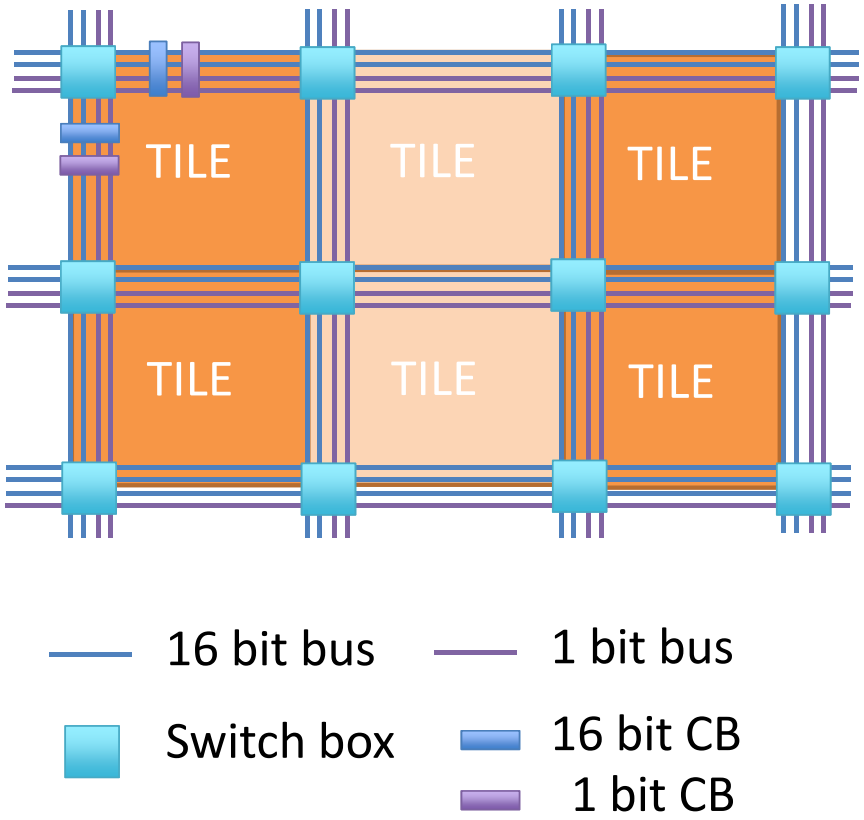


What Are Other Applications Classes?

Connect Hardware to User Process



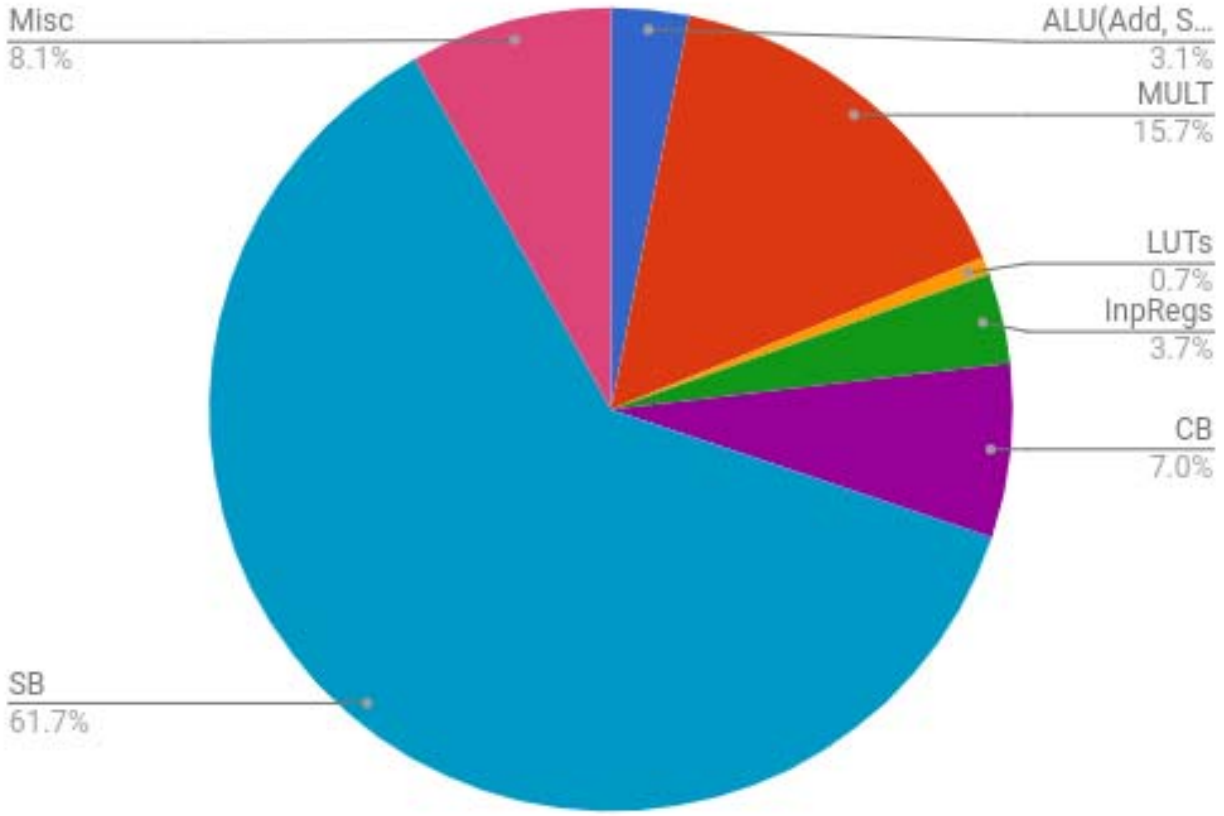
AHA CGRA Architecture



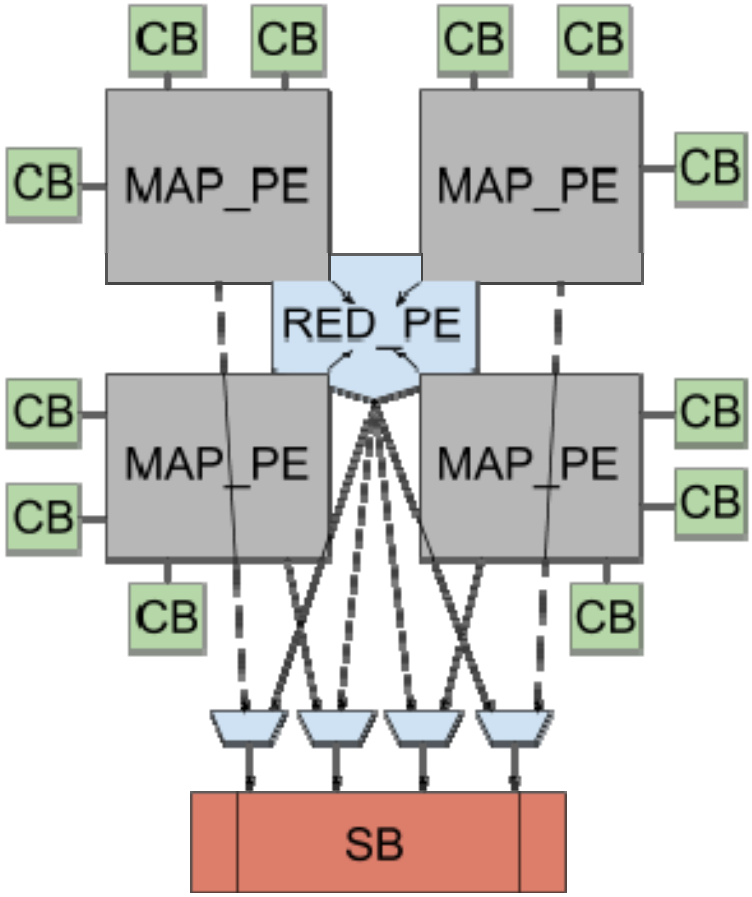
Design Space Exploration:

- Tile Design
- Programmable Routing
- Tile Clustering
- Memory Topology

Area Breakdown for Simple 2:1 PE



Clustering


CB
CB
CB
CB
CB
MAP_PE
MAP_PE
CB
RED_PE
CB
MAP_PE
MAP_PE
CB
CB
CB
CB
CB
SB