Algorithm-Hardware Co-design for Deformable Convolution

Qijing Huang*, Dequan Wang*, Yizhao Gao †, Yaohui Cai ‡,
Zhen Dong, Bichen Wu, Kurt Keutzer, John Wawrzynek

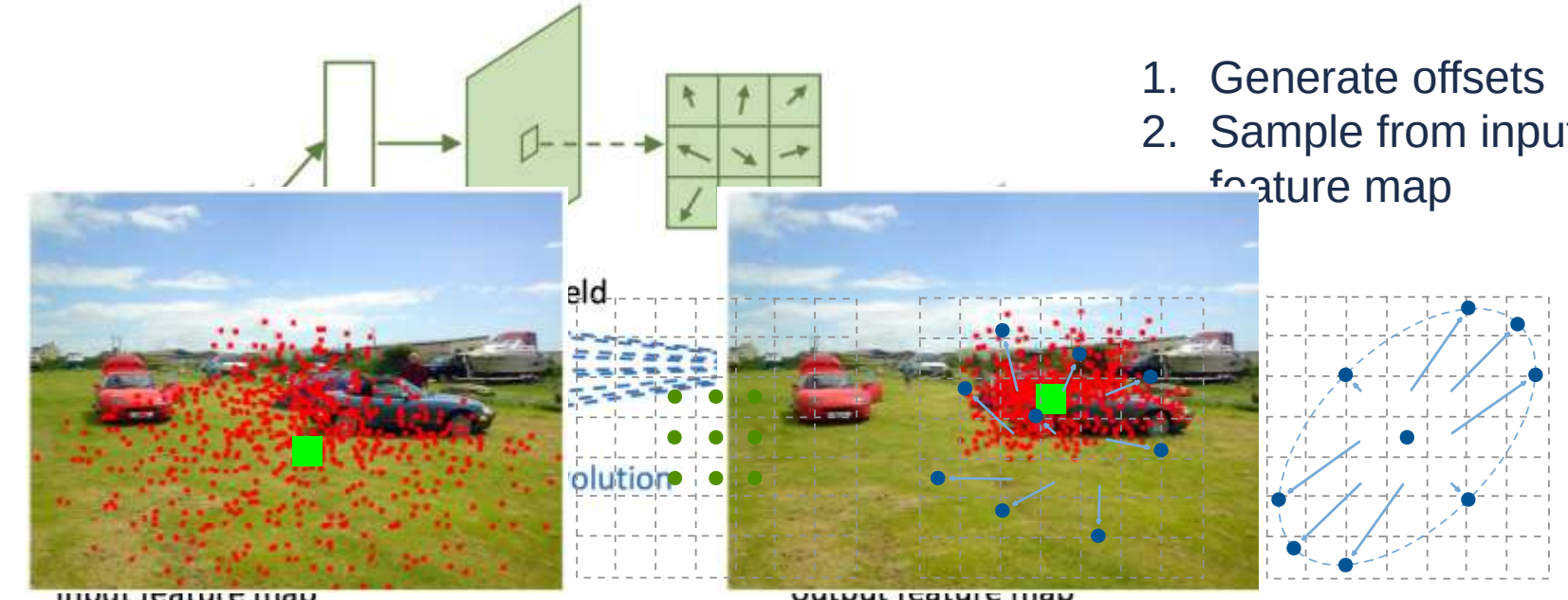
University of California, Berkeley
†University of Chinese Academy of Science
‡Peking University

Berkeley
UNIVERSITY OF CALIFORNIA

EMC2 Workshop @ NeurIPS 2019

Motivation

- **Deformable Convolution** is an input-adaptive dynamic operation that samples inputs from variable spatial locations
- Its sampling locations vary with:
  - Different in
  - Different ou
- It captures th
  - Scales
  - Aspect Rat
  - Rotation Ar
- Challenges:
  - Increasec
  - Irregular
    - Not frie



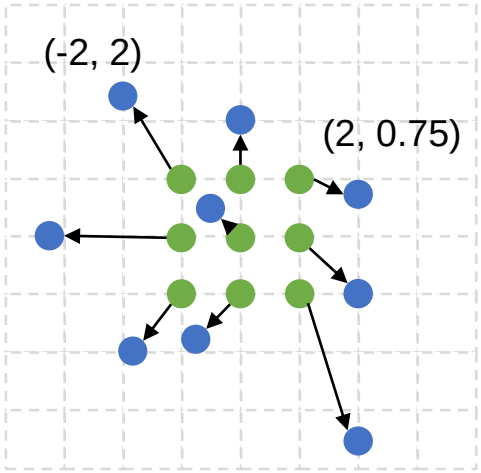
1. Generate offsets
2. Sample from input feature map

Sampling Locations (in red) for Different Output Pixels (in green)

Variable Receptive Fields

Algorithm-Hardware Codesign

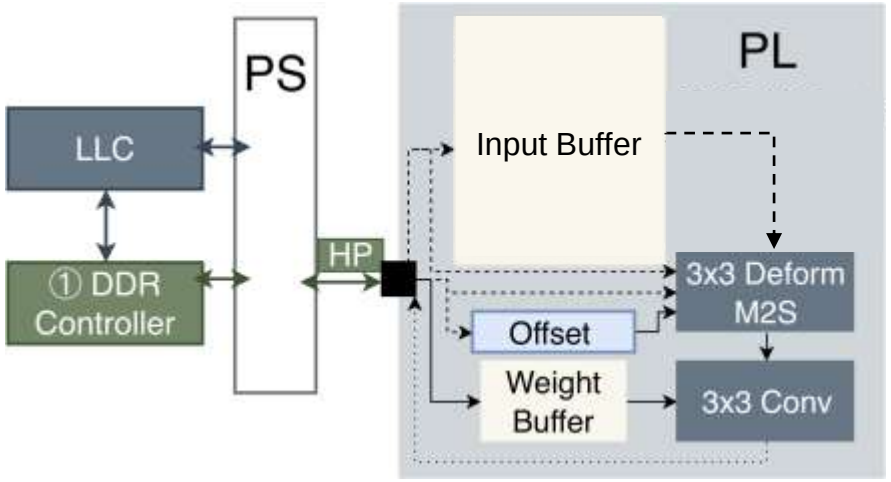
Algorithm Modification:



0. Original Deformable

Accuracy [1](mIoU ↑): **79.9**

Hardware Optimization:

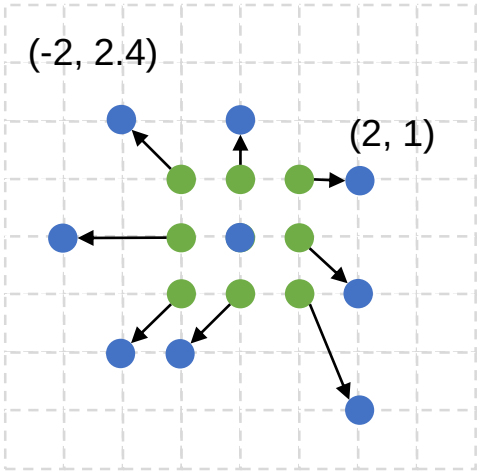


- **Preloads weights to on-chip buffer**
- **Loads input and offsets directly from DRAM**

[1] Accuracy for Semantic Segmentation on CityScapes

Algorithm-Hardware Codesign

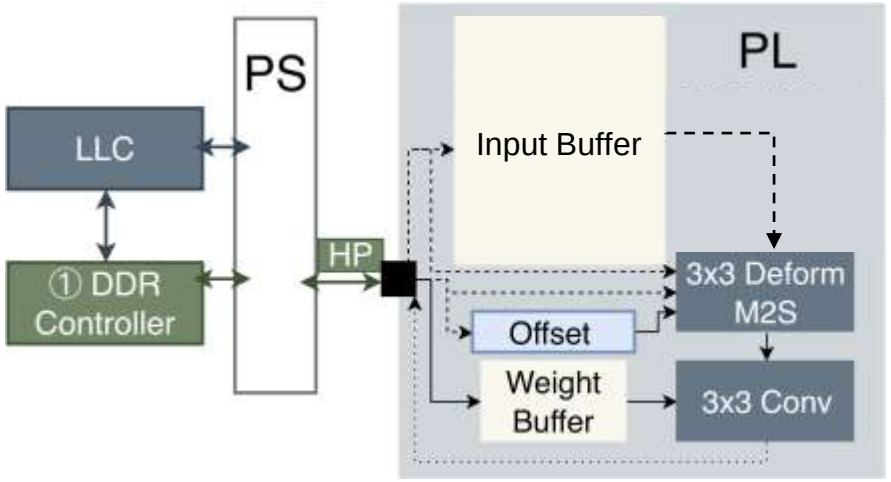
Algorithm Modification:



1. Rounded Offsets

Accuracy [1](mIoU ↑): **79.6** **↓ 0.3**

Hardware Optimization:

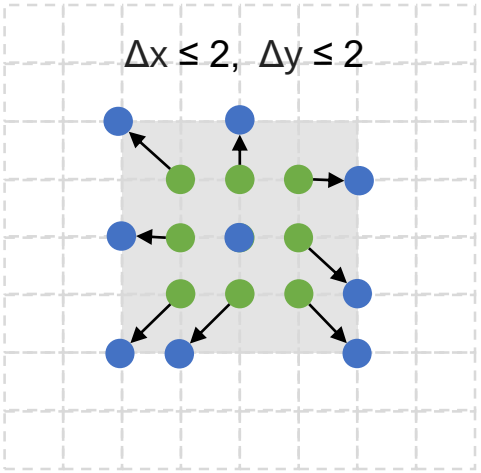


- **Reduces the computation for bilinear interpolation**

[1] Accuracy for Semantic Segmentation on CityScapes

Algorithm-Hardware Codesign

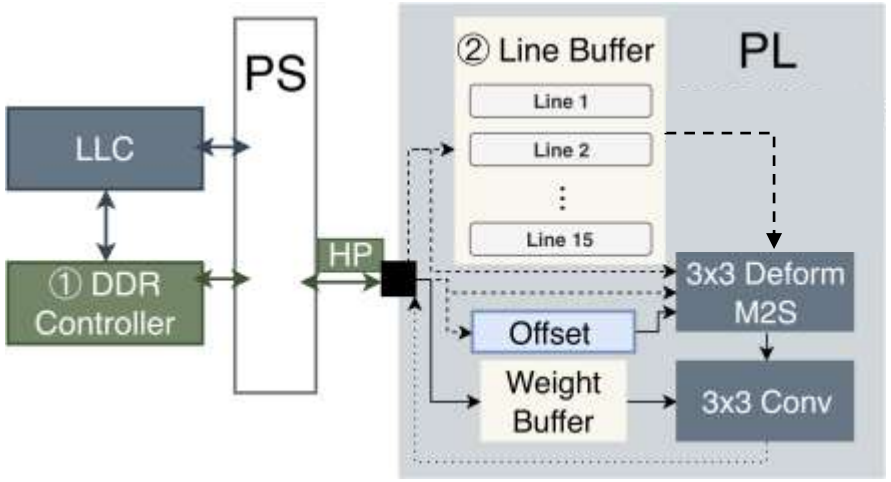
Algorithm Modification:



2. Bounded Range

Accuracy [1](mIoU ↑): **79.4** **↓ 0.2**

Hardware Optimization:



- **Buffers inputs in the on-chip line buffer to allow spatial reuse**

[1] Accuracy for Semantic Segmentation on CityScapes

Algorithm-Hardware Codesign

Hardware Performance

Algorithm Modification:

Hardware Optimization:

Operation	Original	Deformable	Bound (buffered)	Square (multi-ported)	Without LLC		With LLC	
					Latency (ms)	GOPs	Latency (ms)	GOPs
Full 3×3 Conv	✓				43.1	112.0	41.6	116.2
		✓			59.0	81.8	42.7	113.1
		✓	✓		43.4	111.5	41.8	115.5
		✓	✓	✓	43.4	111.5	41.8	115.6
Depthwise 3×3 Conv	✓				1.9	9.7	2.0	9.6
		✓			20.5	0.9	17.8	1.1
		✓	✓		3.0	6.2	3.4	5.5
		✓	✓	✓	2.1	9.2	2.3	8.2

- Our algorithm-hardware co-design methodology for the deformable convolution achieves a **1.36×** and **9.76×** speedup respectively for the *full* and *depthwise* deformable convolution on FPGA

5. Depthwise Convolution

3. Rectangular Shape

Reduced the total memory bandwidth

Feature Extractor	Operation	mIoU
DLA	DeformConv	70.9
ShuffleNetV2	DeformConv	70.1
ShuffleNetV2	DeformConv + Depthwise	68.0

Email: qijing.huang@berkeley.edu

[1] Accuracy for Semantic Segmentation on CityScapes