CS 758: Advanced Topics in Computer Architecture

Lecture #16: ML: Memory

Professor Matthew D. Sinclair

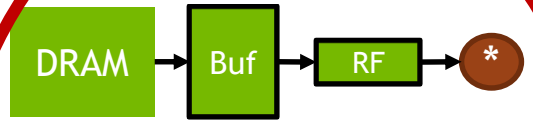
Some of these slides were developed by Tushar Krishna at Georgia Tech
Slides enhanced by Matt Sinclair

Hardware structures to exploit reuse

Goal of a good dataflow:

Algorithmic Data Reuse → Hardware Reuse

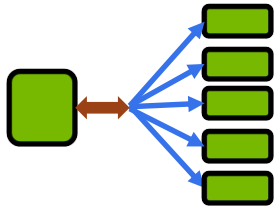
Temporal



Memory Hierarchy / Staging Buffers

E.g., custom memory hierarchies in accelerators.
- Open research topic

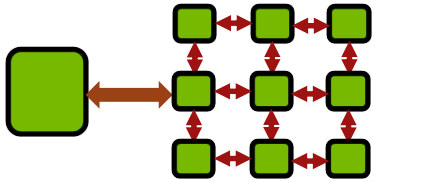
Spatial



Multicast Support

E.g., Hierarchical Bus in Eyeriss (ISCA 2016), Tree in MAERI (ASPLOS 2018)

Spatio-Temporal



Direct Neighbor-to-Neighbor Connections

E.g., TPU, local network in Eyeriss

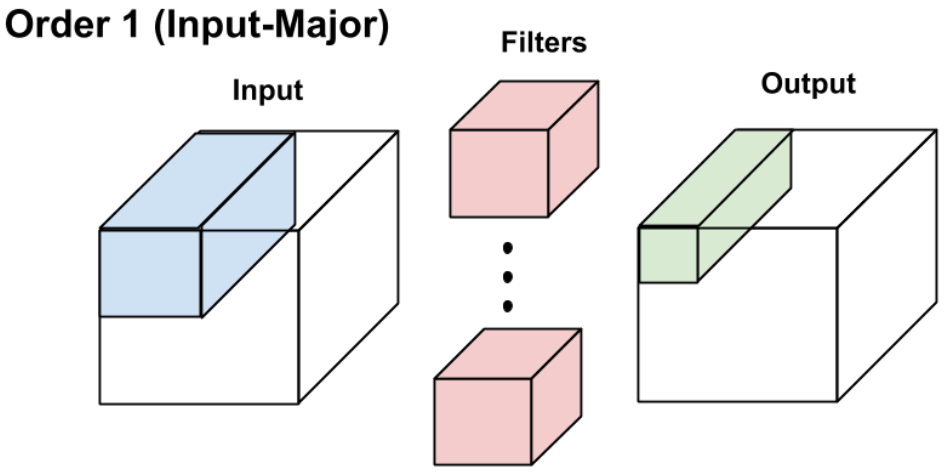
Why is it important?

Percentage of area devoted to on-chip buffers:

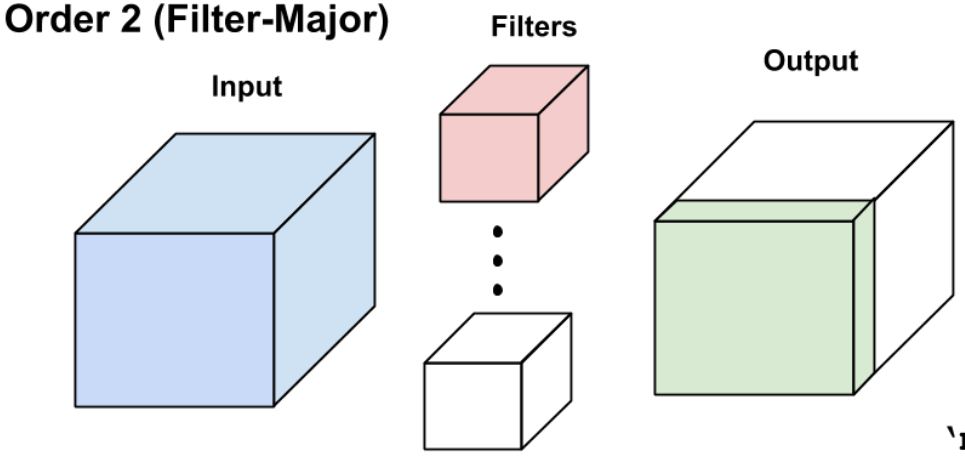
DaDianNao [5]:	48%	Eyeriss [6]:	40%-93%
EIE [18]:	93%	SCNN [35]:	57%
TPU [22]	35%	PuDianNao [27]	63%

Slide Courtesy: Michael Pellauer, NVIDIA

What dataflows did they consider? [Siu 2018]



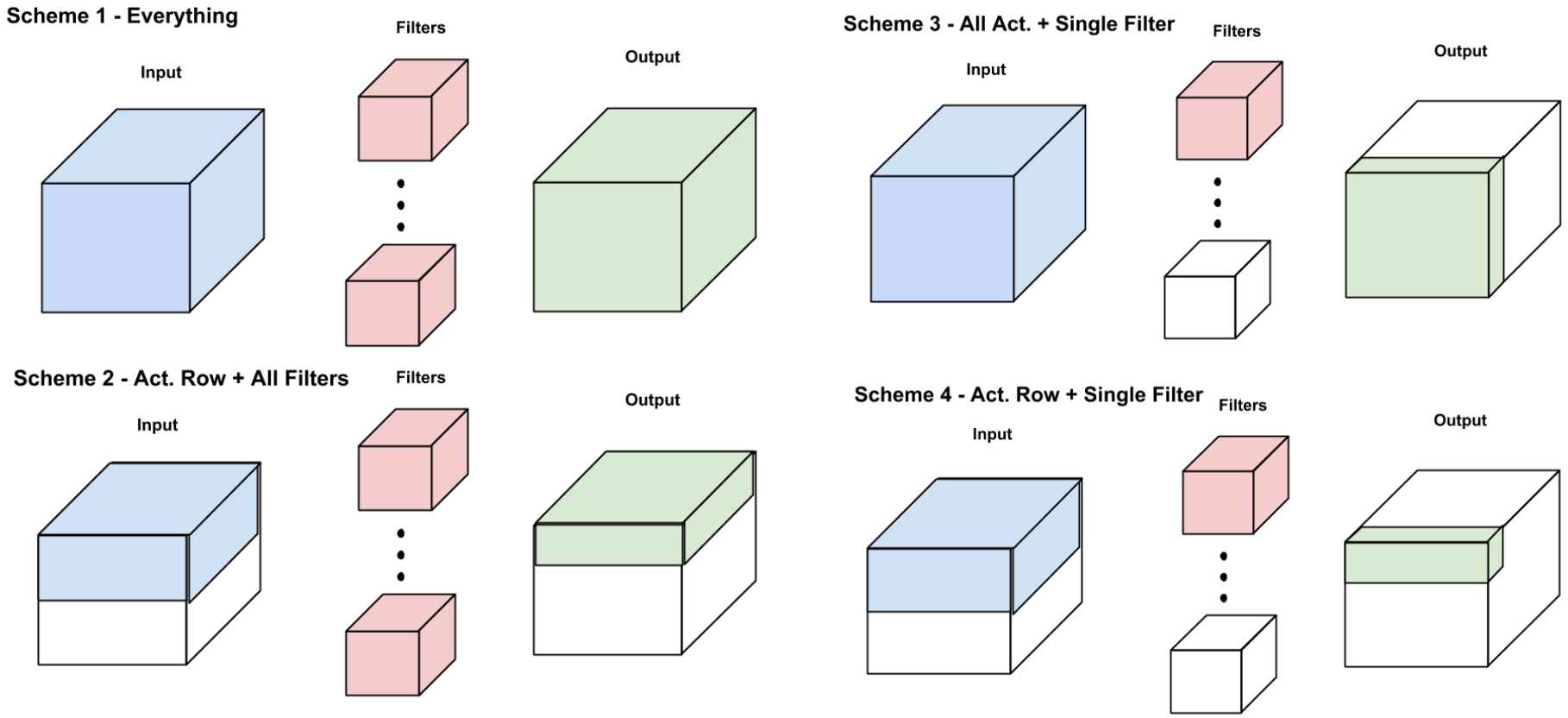

```
Order 1 (Input-Major)
For q = 1 .. Q
  For p = 1 .. P
    For k = 1 .. K
      For s = 1 .. S
        For r = 1 .. R
          For c = 1 .. C
            O[p][q][k] +=
              I[p*m+r-1][q*m+s-1][c]
              * F[k][r][s][c]
```




```
Order 2 (Filter-Major)
For k = 1 .. K
  For q = 1 .. Q
    For p = 1 .. P
      For s = 1 .. S
        For r = 1 .. R
          For c = 1 .. C
            O[p][q][k] +=
              I[p*m+r-1][q*m+s-1][c]
'm' is stride   * F[k][r][s][c]
```

Data Movement Order

Heuristics for memory BW reduction



Data Mapping ("Stationary") Behavior at Shared (L2) Buffer

What about at L1?

Results

TABLE III: AM and WM sizes for Scheme 1. Double buffering is assumed for activations.

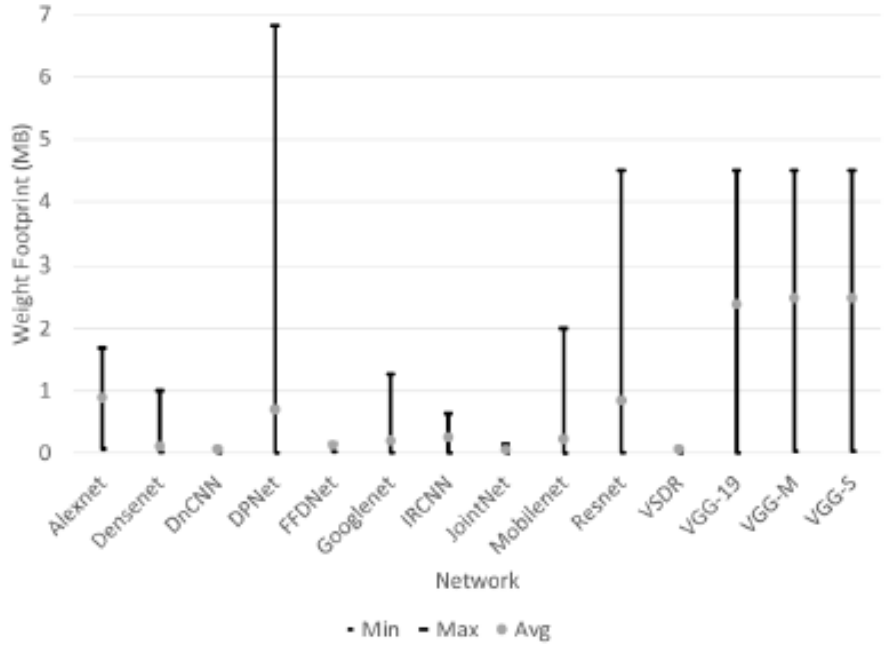
	WM Size (MB)		AM Size (MB)
	Conv	FC	Double
AlexNet	4.45	111.81	0.85
GoogleNet	11.38	1.95	0.77
VGG-M	12.45	163.91	0.82
VGG-S	12.45	183.81	0.56
VGG-19	38.18	235.81	12.25
MobileNet	6.08	1.95	2.30
DenseNet-121	13.10	1.95	2.38
DPNet-92	66.48	5.13	3.27
ResNet-50	44.74	3.91	2.30
DnCNN	1.27	-	506.25
FFDNet	1.31	-	189.84
IRCNN	1.80	-	506.25
JointNet	1.07	-	27.06
VDSR	1.27	-	506.25

Results

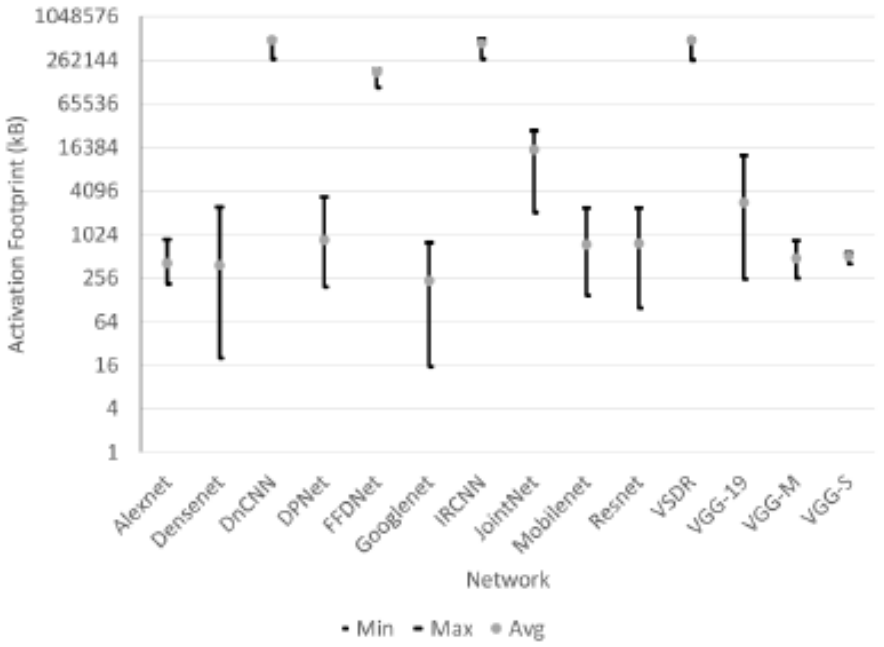
TABLE VI: Total memory requirements for Schemes 2–4. Double buffering is assumed, wherever possible. All values are in MB. The grayed out configurations violate our single, per value off-chip access invariant.

	Scheme 2		Scheme 3				Scheme 4			
	Conv	FC	Working Set Size				Working Set Size			
Network			Full Layer	64 Filters	16 Filters	1 Filter	Full Layer	64 Filters	16 Filters	1 Filter
AlexNet	4.507	111.867	3.8	1.41	0.99	0.8588	3.007	0.617	0.197	0.0658
GoogleNet	11.421	1.991	2.91	1.19	0.88	0.7766	2.181	0.461	0.151	0.0476
VGG-M	12.549	164.009	9.82	1.95	1.1	0.838	9.099	1.229	0.379	0.117
VGG-S	12.55	183.91	9.56	1.69	0.84	0.578	9.1	1.23	0.38	0.118
VGG-19	38.34	235.97	21.25	13.38	12.53	12.268	9.16	1.29	0.44	0.178
MobileNet	6.162	2.032	4.32	2.55	2.363	2.3039	2.102	0.332	0.145	0.0859
DenseNet-121	13.183	2.033	3.51	2.63	2.45	2.3844	1.213	0.333	0.153	0.0874
DPNet-92	66.61	5.26	12.36	3.9	3.43	3.2798	9.22	0.76	0.29	0.1398
ResNet-50	44.795	3.965	8.8	3.43	2.58	2.318	6.555	1.185	0.335	0.073
DnCNN	2.67	-	506.39	506.32	506.285	506.2522	1.54	1.47	1.435	1.4022
FFDNet	2.41	-	190.16	190	189.893	189.8433	1.42	1.26	1.153	1.1033
IRCNN	6.02	-	507.27	506.88	506.57	506.268	5.24	4.85	4.54	4.238
JointNet	1.26	-	27.27	27.2	27.095	27.0622	0.4	0.33	0.225	0.1922
VDSR	2.67	-	506.39	506.32	506.285	506.2522	1.54	1.47	1.435	1.4022

Per layer results



(a) Weights

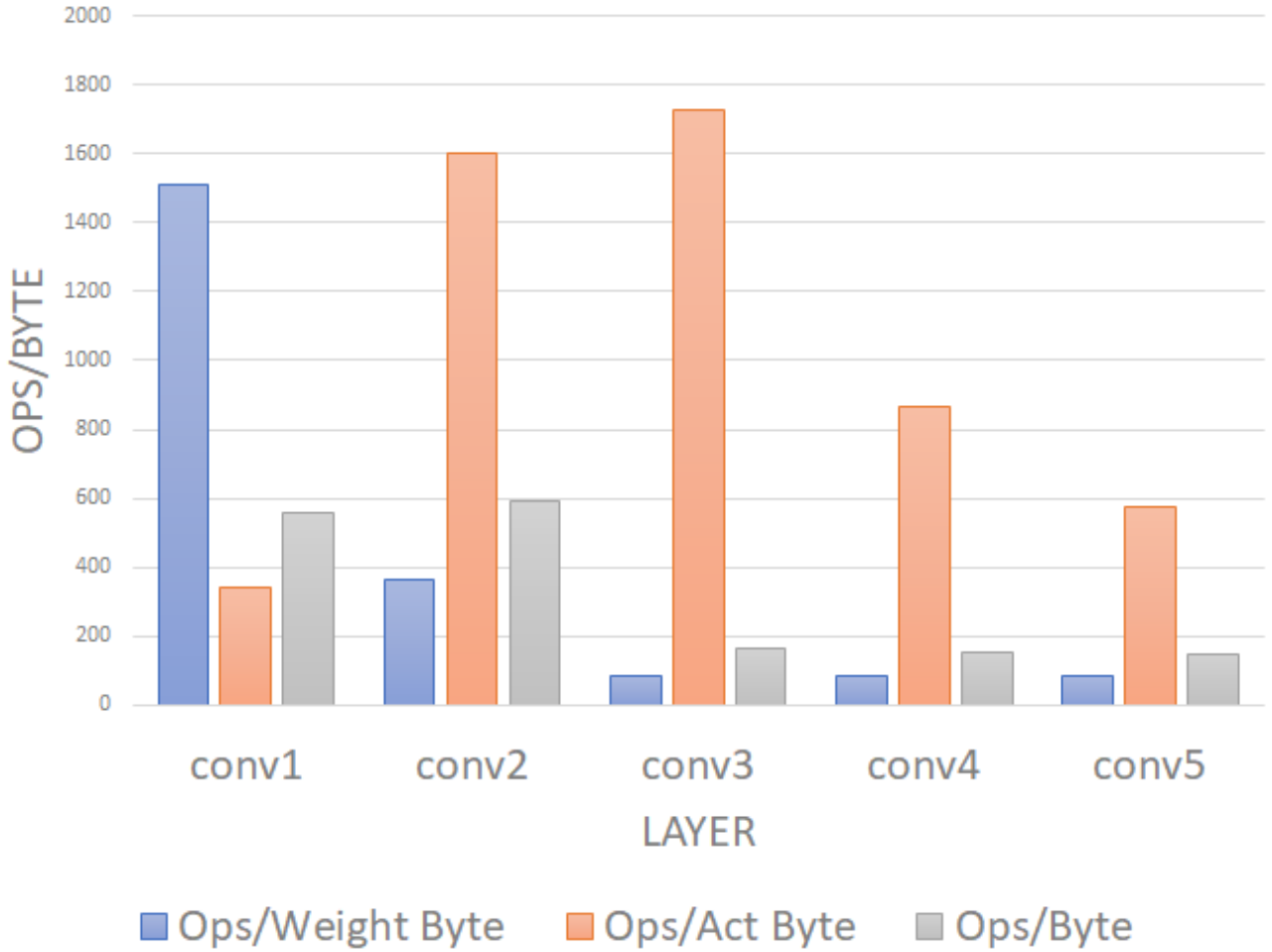


(b) Activations ($\log_2$ scale)

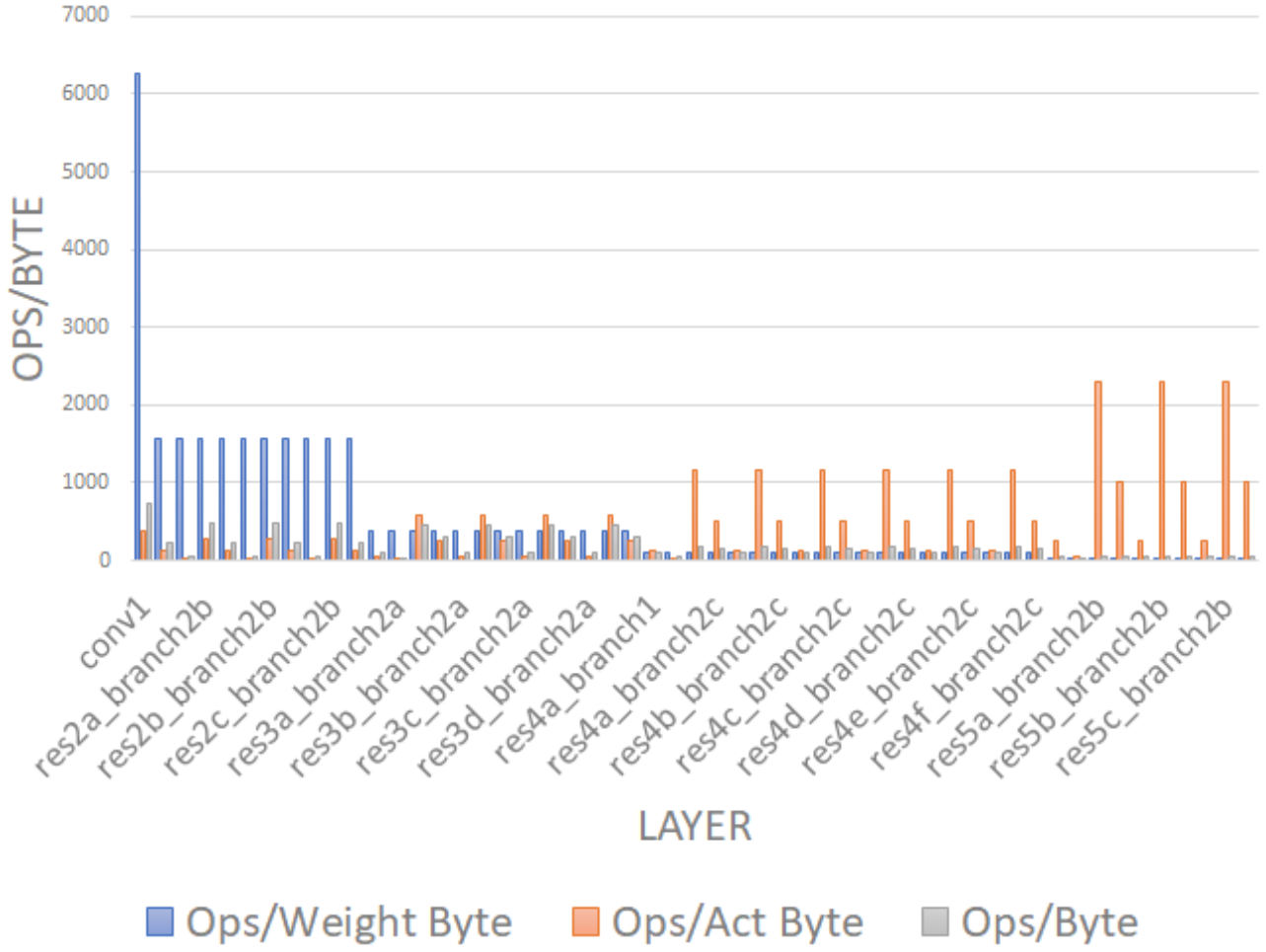
BW Requirements

Network	Scheme 1	Scheme 2			Scheme 3			Scheme 4		
	Mem	Avg BW	Peak BW	Mem	Avg BW	Peak BW	Mem	Avg BW	Peak BW	Mem
AlexNet	5.30	1.22	3.00	4.48	7.18	12.12	2.54	35.07	50.11	0.03
GoogleNet	12.15	7.75	128.00	11.40	9.91	41.78	2.03	78.70	356.57	0.02
VGG-M	13.27	1.16	7.49	12.50	11.71	12.12	5.32	59.72	126.30	0.06
VGG-S	13.01	0.70	15.96	12.50	7.03	7.09	5.06	42.08	63.83	0.06
VGG-19	50.43	1.42	7.11	38.27	2.73	41.80	16.75	24.74	168.71	0.09
MobileNet	8.37	18.55	912.29	6.12	11.49	41.89	4.30	148.01	971.76	0.04
DenseNet	15.48	10.88	16.00	13.15	4.60	32.00	3.38	70.56	272.00	0.04
DPNet	69.75	6.08	146.52	66.54	9.67	32.00	10.09	142.02	483.13	0.07
ResNet	47.03	5.38	64.00	44.76	12.46	41.80	6.80	112.02	601.14	0.04
DnCNN	507.52	3.74	75.85	1.98	0.001	0.001	506.32	4.10	76.21	0.70
FFDNet	191.15	2.62	18.96	1.84	0.004	0.004	190.00	3.33	19.67	0.53
IRCNN	508.05	0.84	75.85	3.91	0.001	0.001	506.88	1.00	76.21	2.12
JointNet	28.13	4.45	682.68	1.17	0.035	0.04	27.20	7.31	687.23	0.09
VDSR	507.52	3.75	227.56	1.97	0.001	0.001	506.32	4.10	227.56	0.70

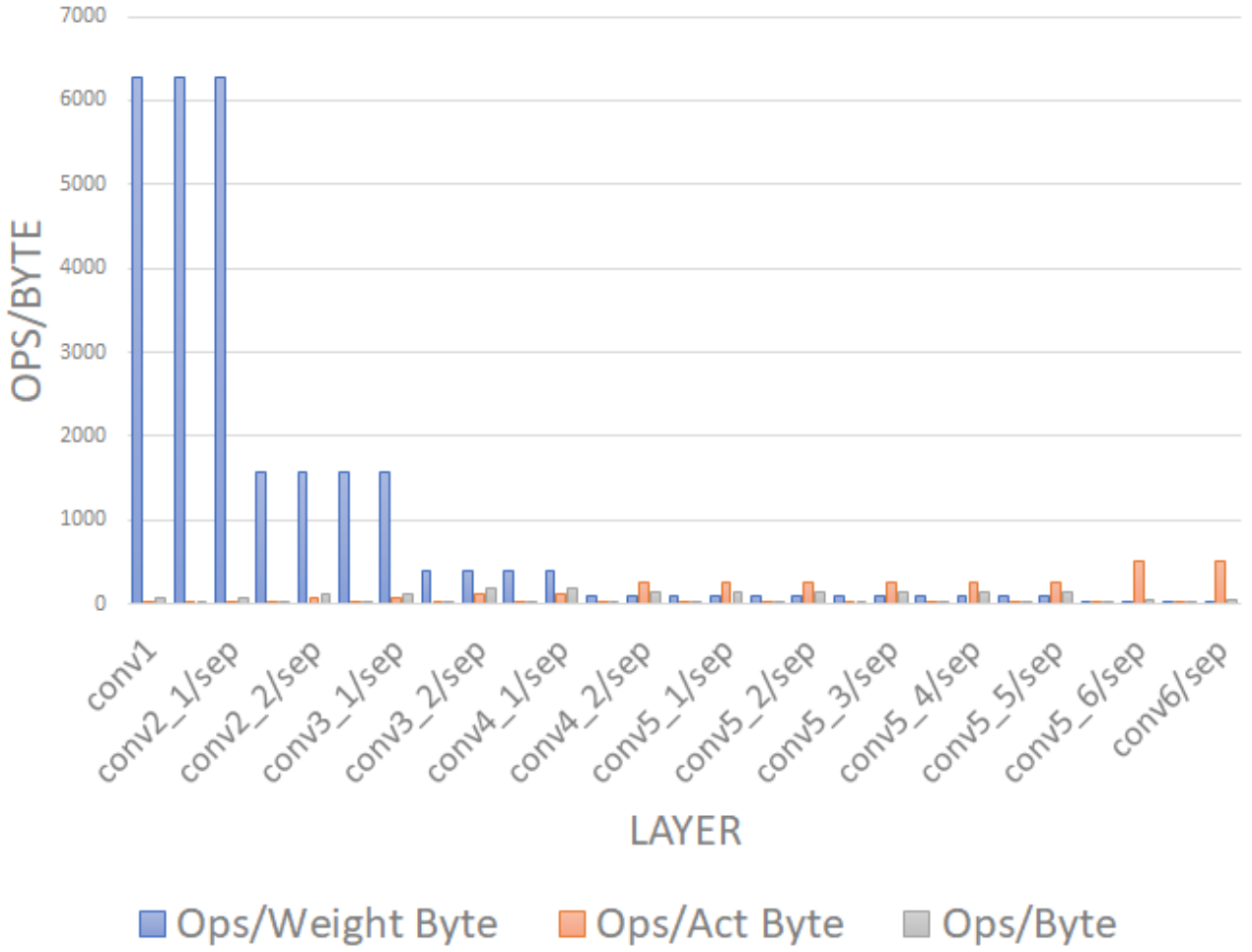
Reuse per layer – AlexNet (2012)



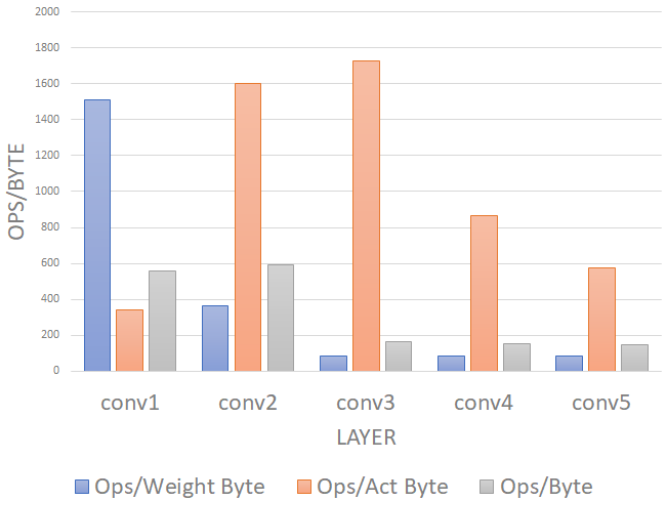
Reuse per layer: ResNet-50 (2015)



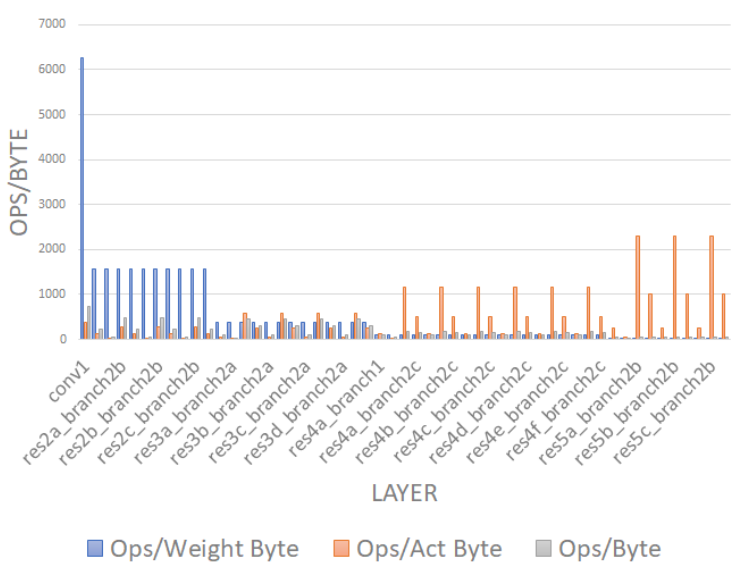
Reuse per layer: MobileNet (2018)



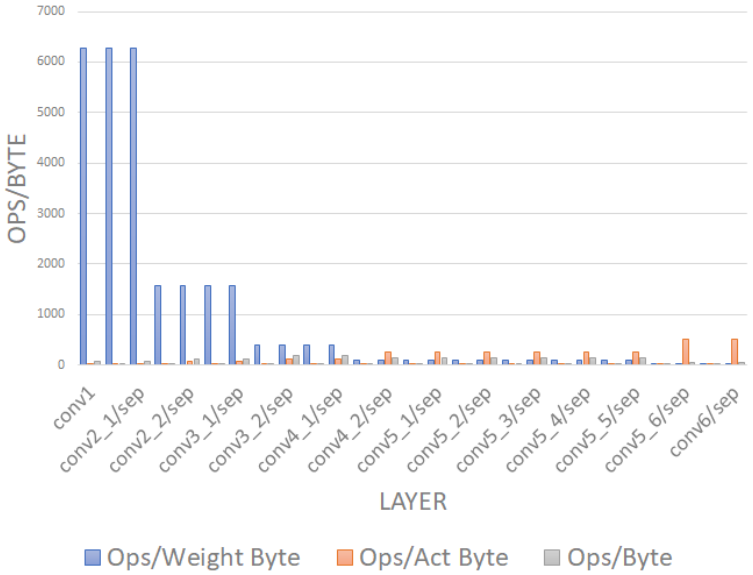
Reuse per layer



AlexNet
(2012)



ResNet-50
(2015)



MobileNet
(2018)