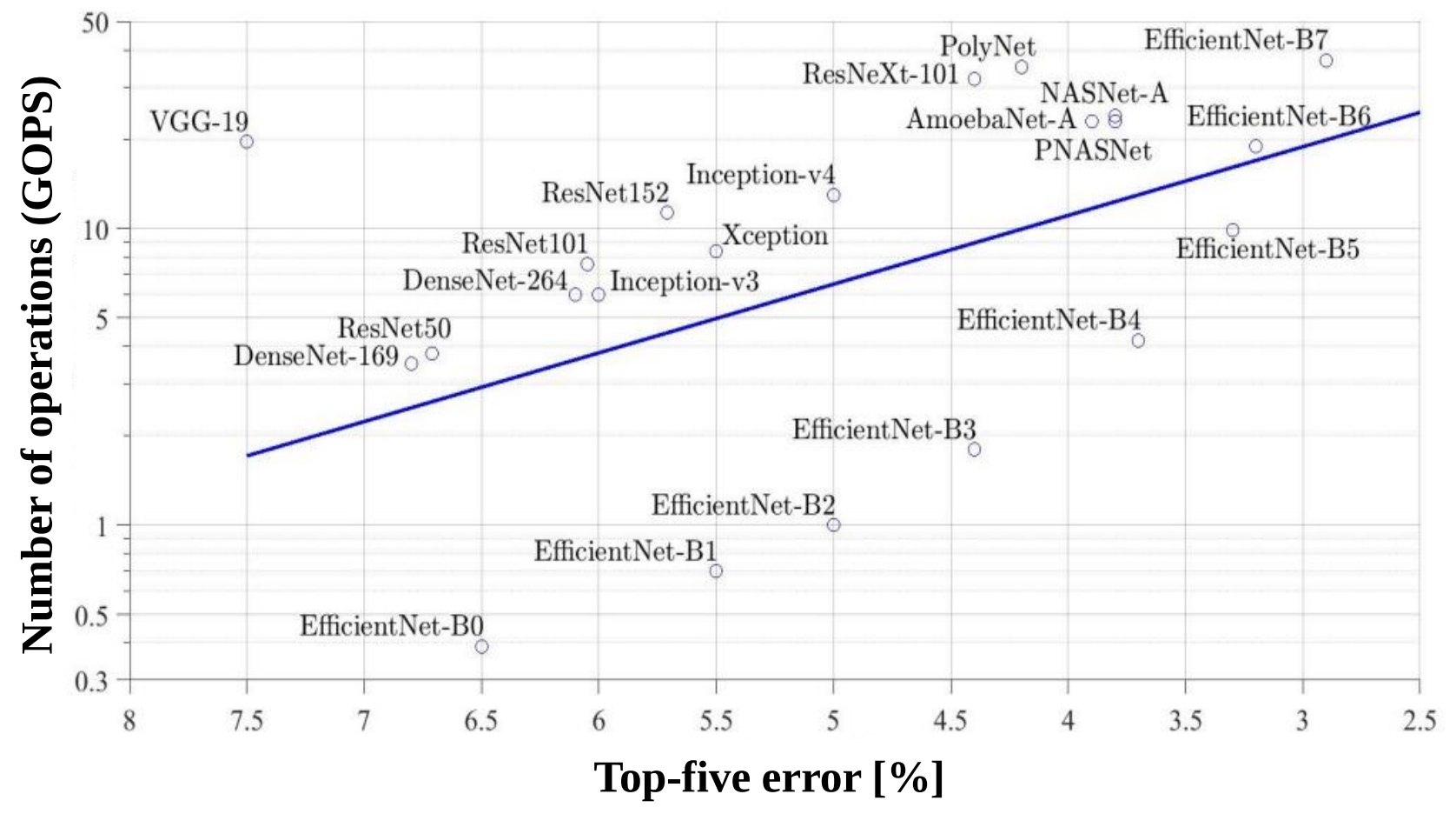
Photonic Reconfigurable Accelerators for Efficient Inference of CNNs with Mixed-Sized Tensors

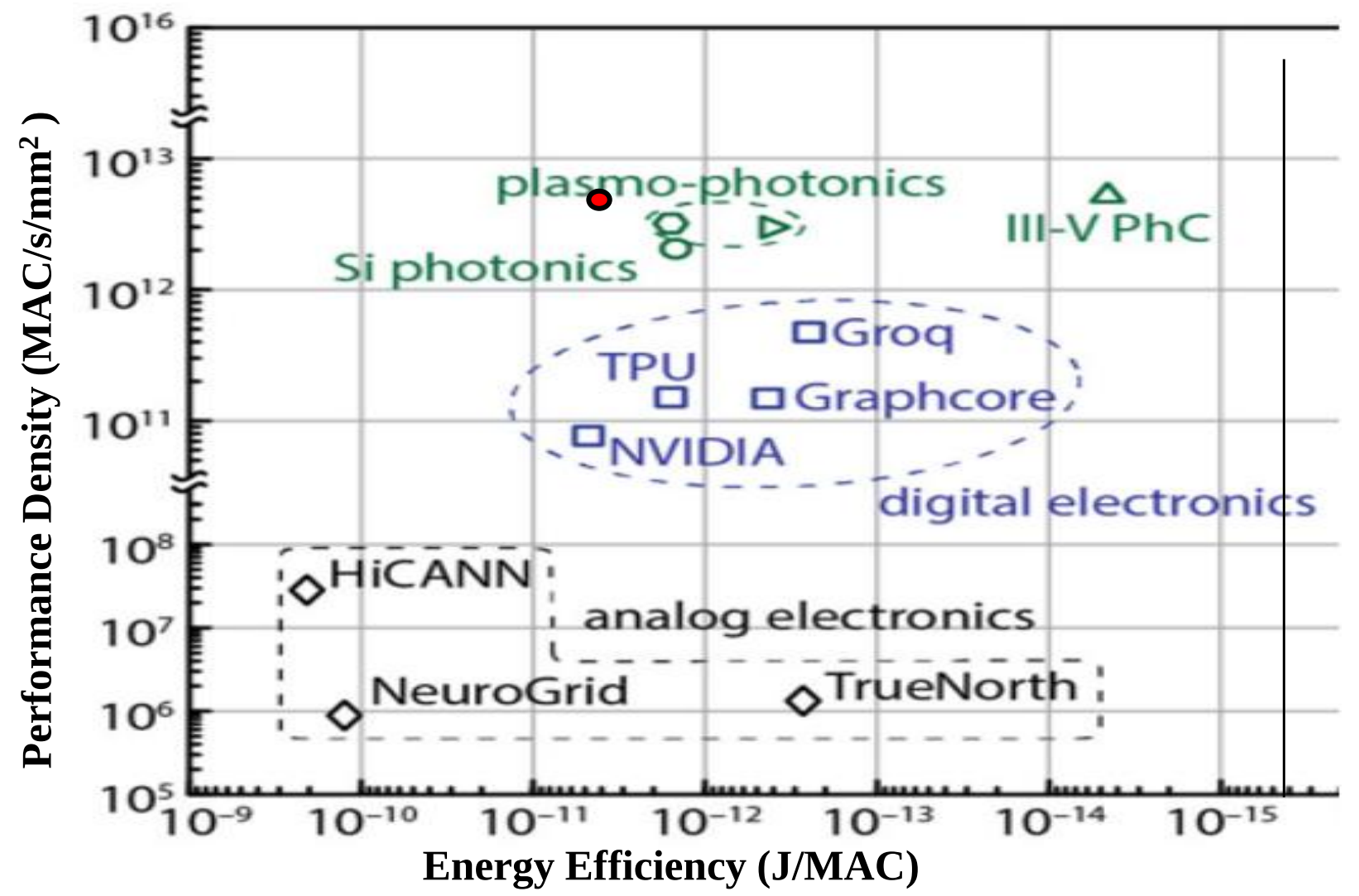
Sairam Sri Vatsavai, Ishan G Thakkar
Electrical and Computer Engineering, University of Kentucky, Lexington, KY 40508
ssr226@uky.edu, igthakkar@uky.edu

SCAN ME

Introduction



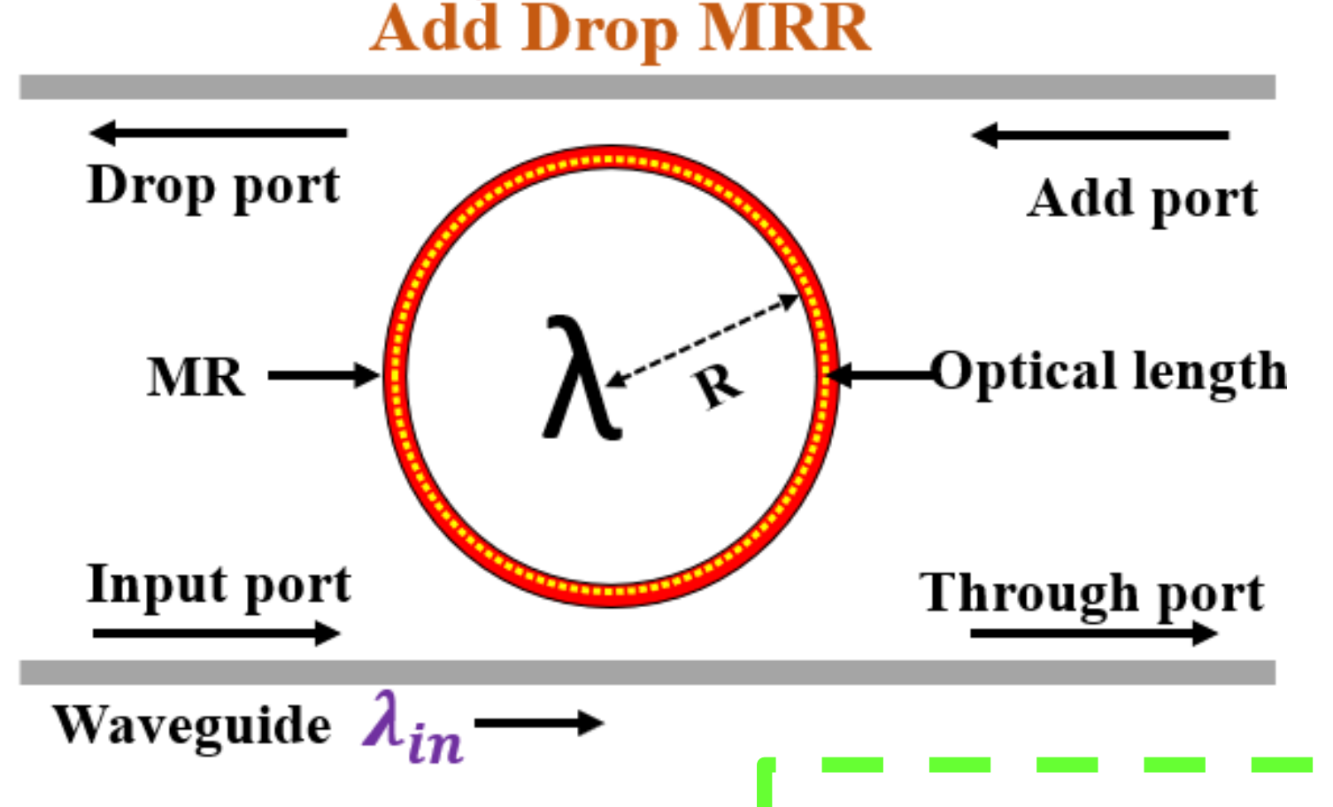
Source : Lukas Baischer Axriv 2021



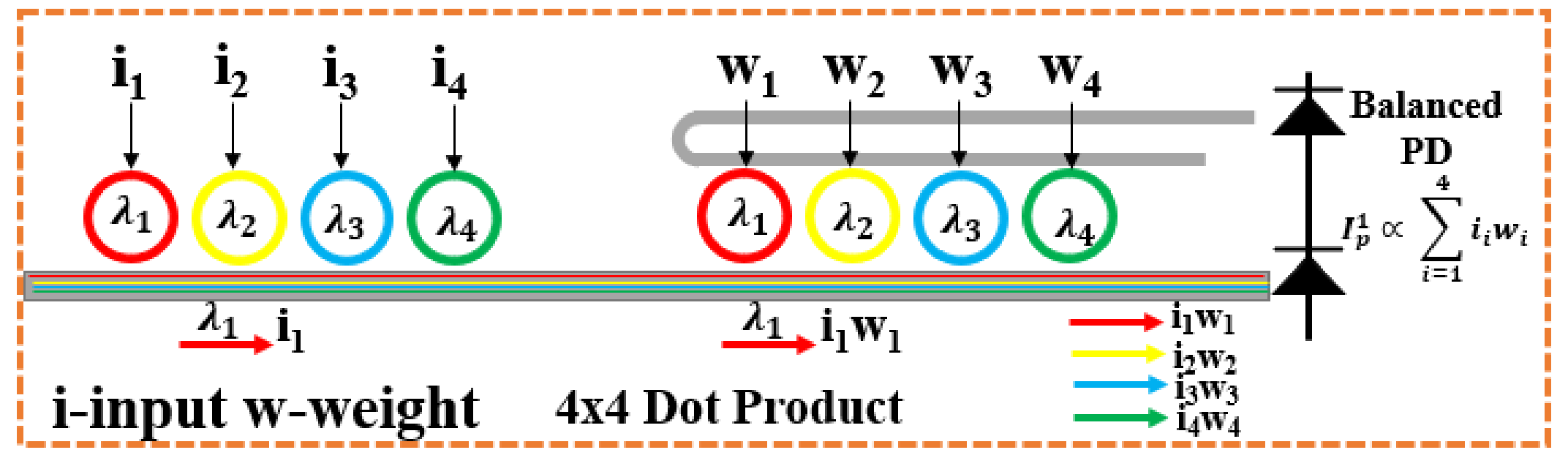
Source: Angelina R. Totovi´c, Optics Express 2021

Microring Resonator (MRR)

Add Drop MRR



Dot product operation with MRRs



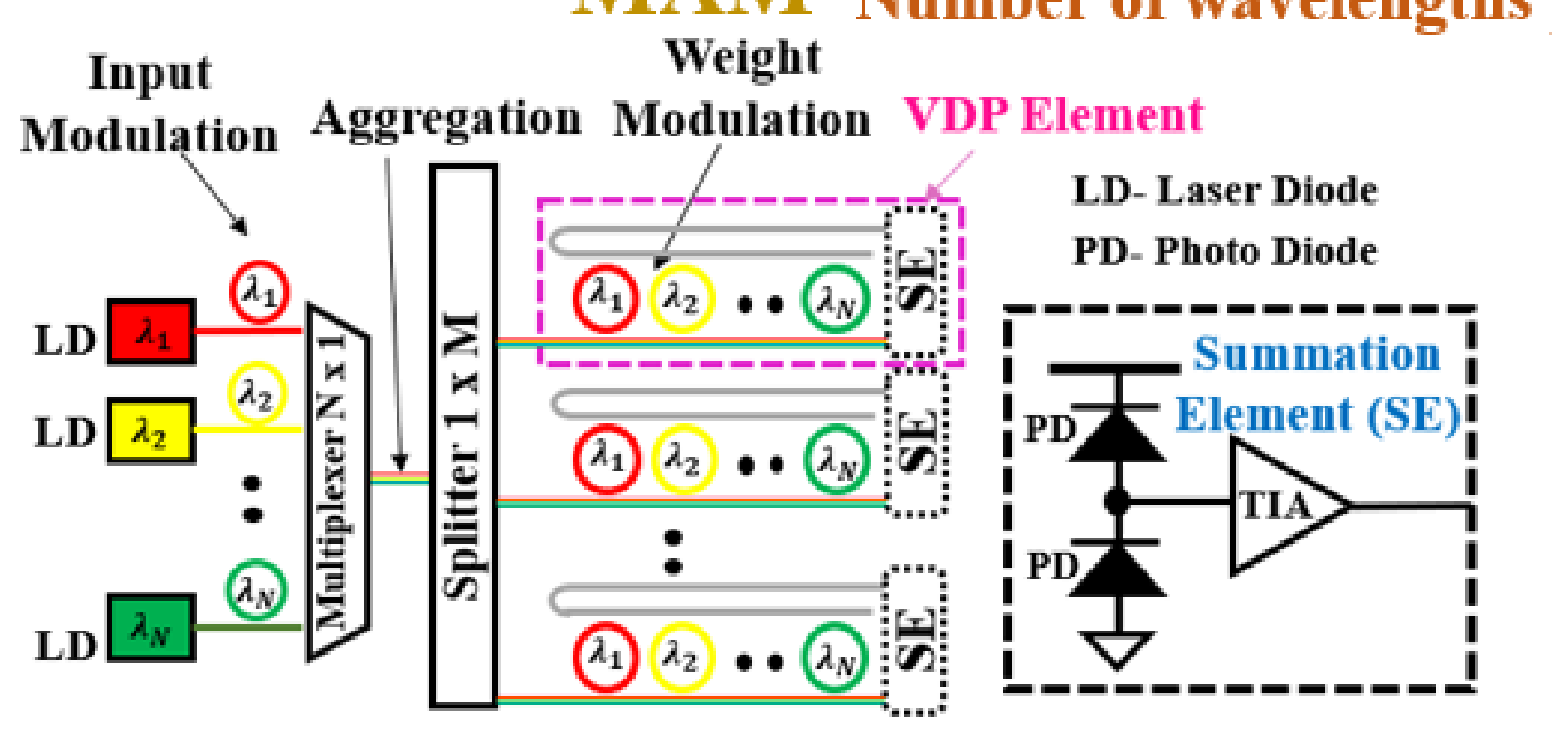
MRRs can perform VDP operation

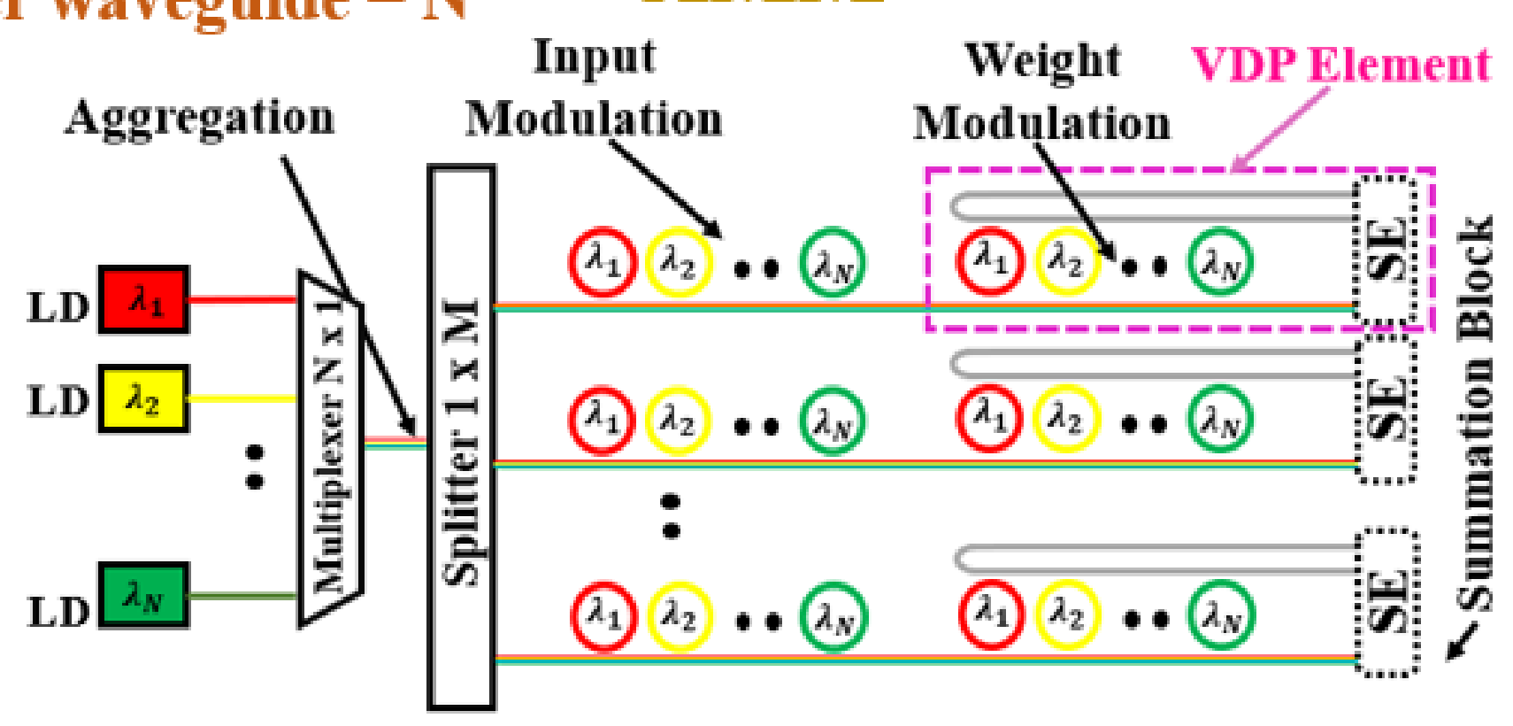
MRR-based CNN Accelerators

Classification of MRR-based Accelerators

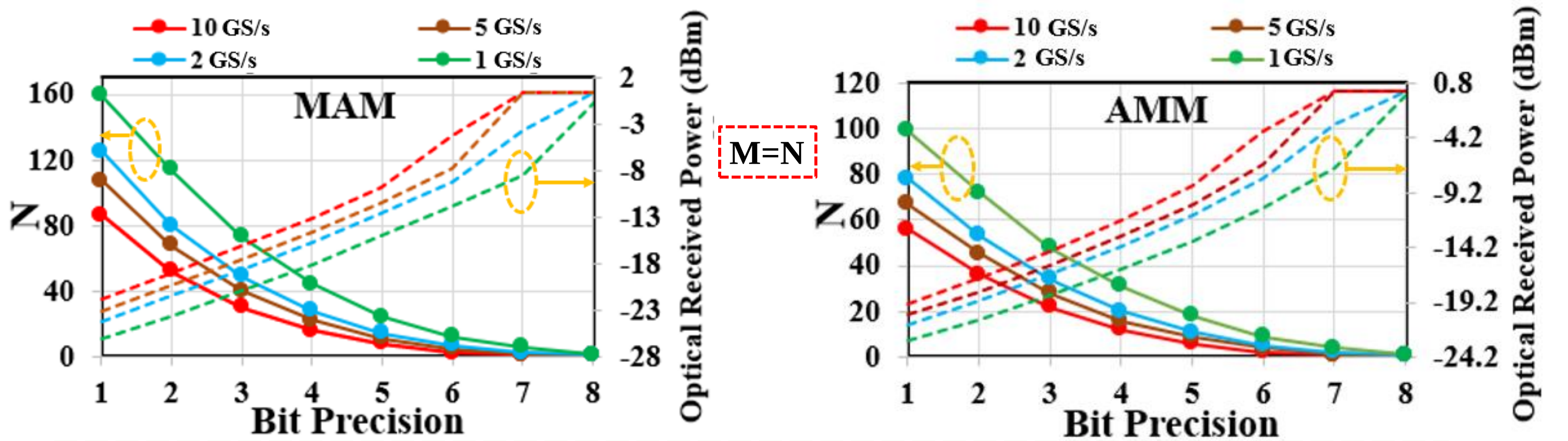
Number of VDPEs = M

MAM Number of wavelengths per waveguide = N AMM





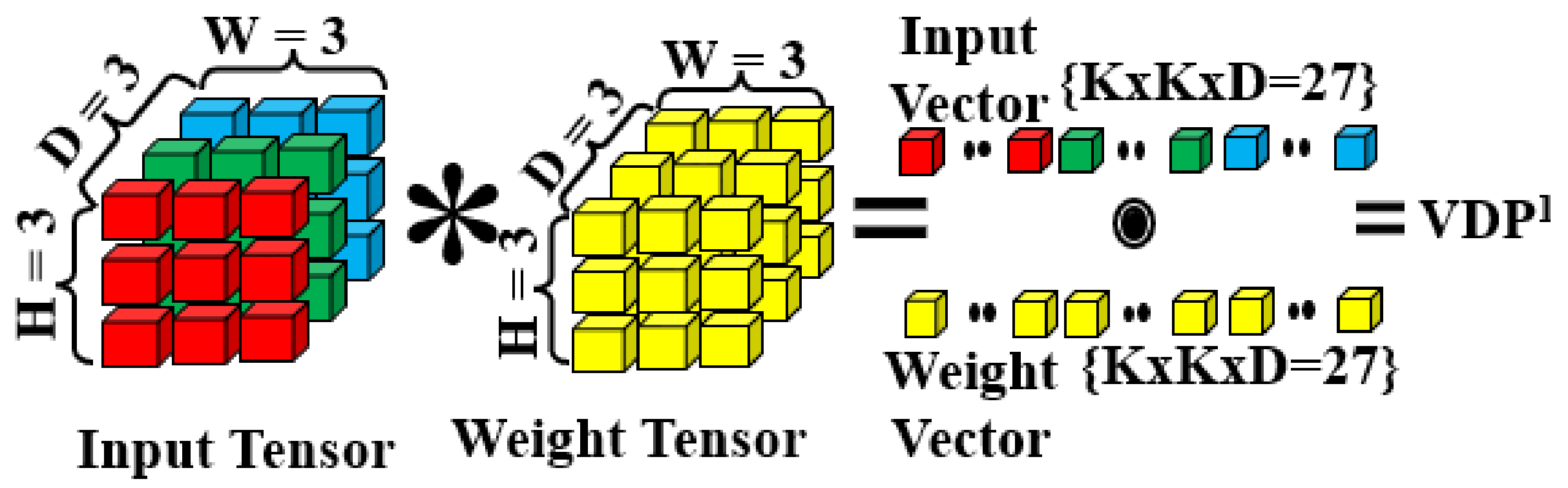
Scalability of MAM and AMM Organizations



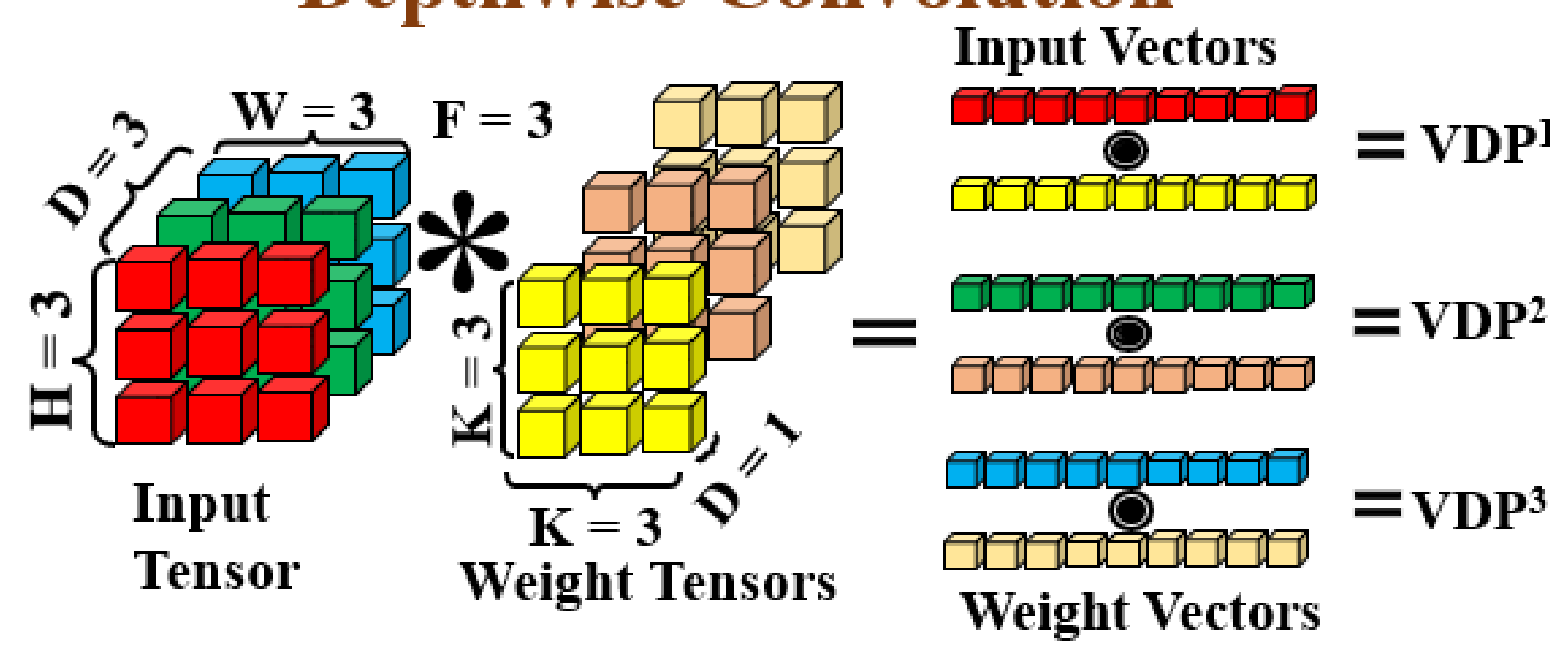
N decreases with datarate for given bit precision

Acceleration of Convolution Operations

Standard Convolution



Depthwise Convolution



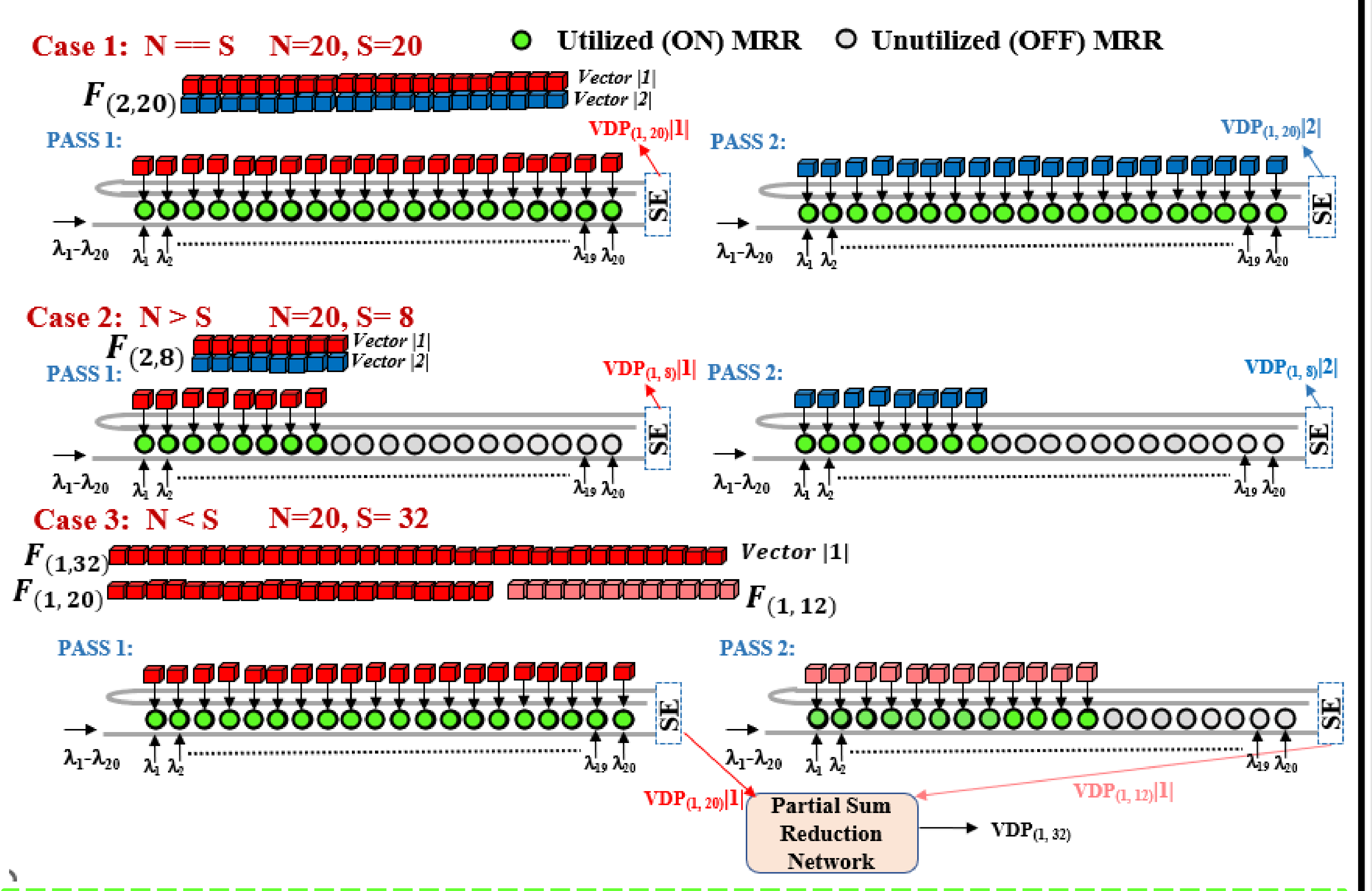
F = No of Kernels; S = Size of Vector

Model	Convolution	Tensor Shape (K, K, D)	F	S
EfficientNet_B7	DC	(3, 3, 1)	25024	9
	DC	(5, 5, 1)	45216	25
	PC	(1, 1, 8)	288	8
	PC	(1, 1, 12)	2016	12
	PC	(1, 1, 16)	64	16
	PC	(1, 1, 20)	3360	20
	PC	(1, 1, 32)	312	32
	PC	(1, 1, 40)	9600	40
	PC	(1, 1, 48)	2016	48
	PC	(1, 1, 56)	13440	56
	PC	(1, 1, 64)	48	64
	PC	(1, 1, 80)	3360	80
	PC	(1, 1, 96)	29952	96
	PC	(1, 1, 160)	21120	160
	PC	(1, 1, 192)	56	192
	PC	(1, 1, 224)	13440	224
	PC	(1, 1, 288)	452	288
	PC	(1, 1, 384)	29952	384
	PC	(1, 1, 480)	780	480
	PC	(1, 1, 640)	14080	640
	PC	(1, 1, 960)	2064	960
	PC	(1, 1, 1344)	2960	1344
	PC	(1, 1, 2304)	6496	2304
	PC	(1, 1, 3840)	2400	3840
	SC	(3, 3, 3)	64	27
	FC	(2560, 1, 1)	1	2560

Vector Size Requirement of CNNs varies widely

Need for Reconfigurability

Mapping of Convolution Weight Matrix



Fixed-size VDPE leads to underutilization or partial sum latency

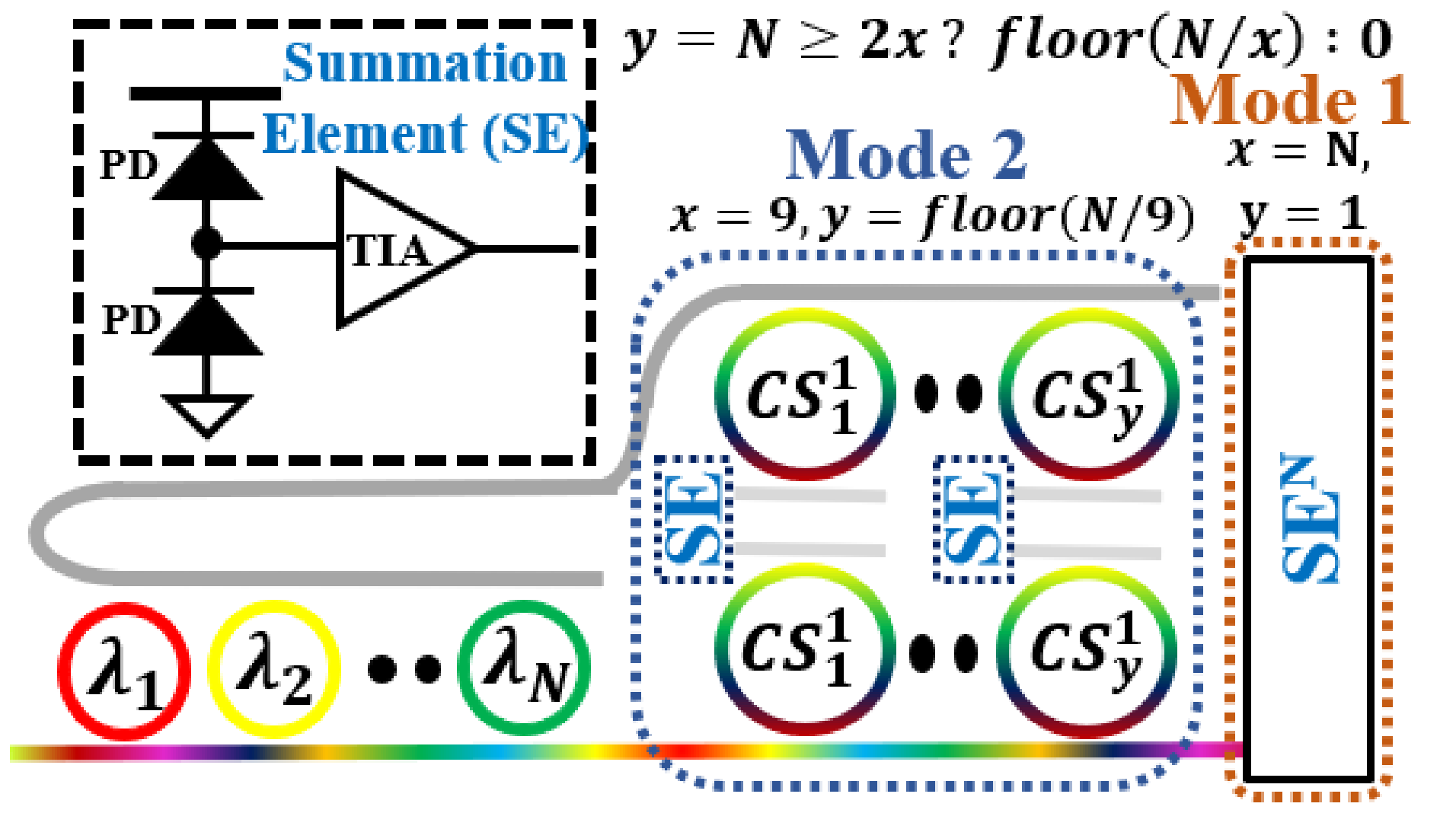
Proposed Reconfigurable Architecture and Mapping

Reconfigurable VDPE

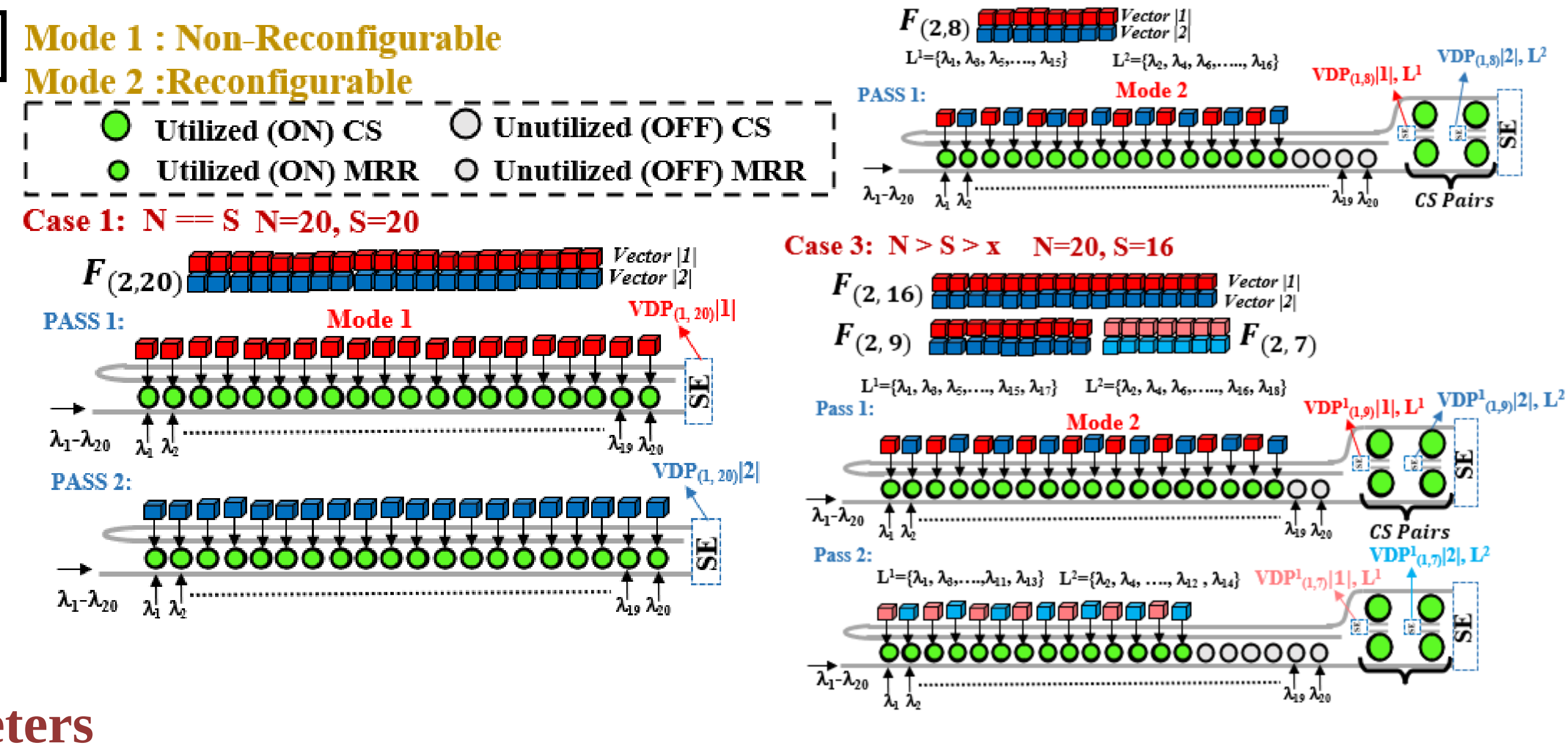
$CS = Comb\ Switch\ \ y = \#\ of\ CS\ of\ one\ type\ x = sizeof(L)$

$L = \{\lambda_i \in \{\lambda_1, \lambda_2, \ldots \lambda_N\};\ \lambda_i = \lambda_{1+iy};\ i \in z;\ i \geq 0\}$

$y = N \geq 2x\ ?\ floor(N/x) : 0$



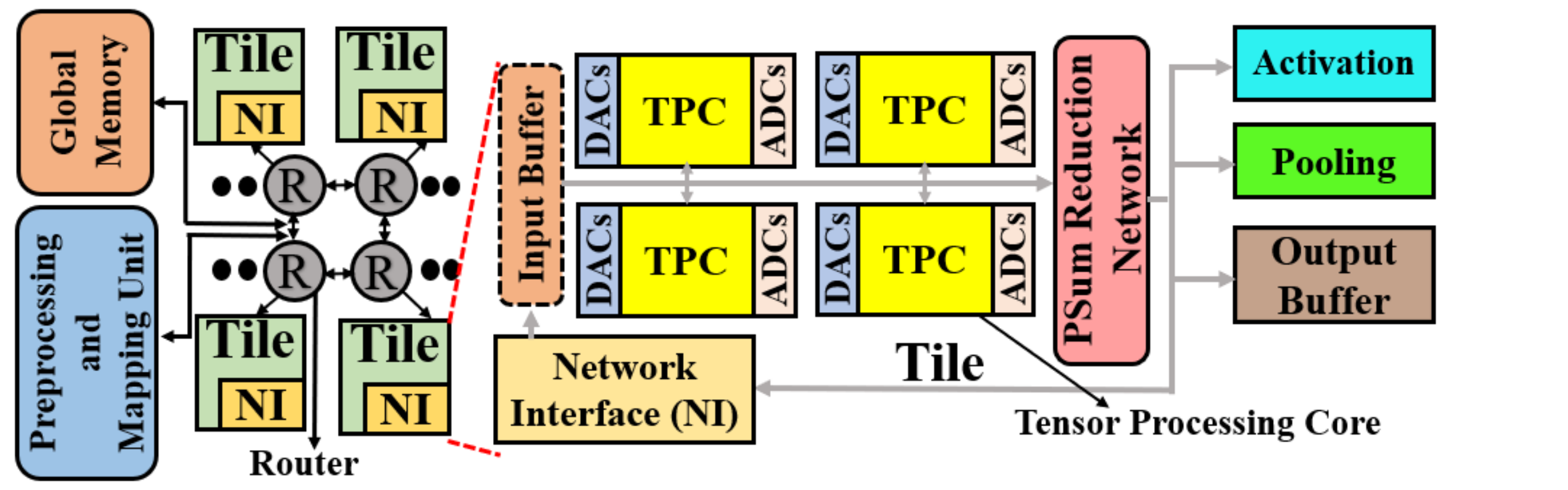
Reconfigurable VDPE Operation



Comb Switches Design Parameters (Ansys Lumerical Simulations)

Data Rate (DR) (GS/s)	1	3	5
RAMM TPC			
N	31	20	16
CS_{FSR}	4.83nm	5 nm	NA
Radius	18.17 μm	17.5 μm	NA
No of CS Pairs	3	2	0
Insertion Loss (dB)	0.029	0.028	0
RMAM TPC			
N	43	28	22
CS_{FSR}	4.65 nm	5.35nm	4.54 nm
Radius	18.98 μm	16.2 μm	19.49 μm
No of CS Pairs	4	3	2
Insertion Loss (dB)	0.029	0.026	0.031

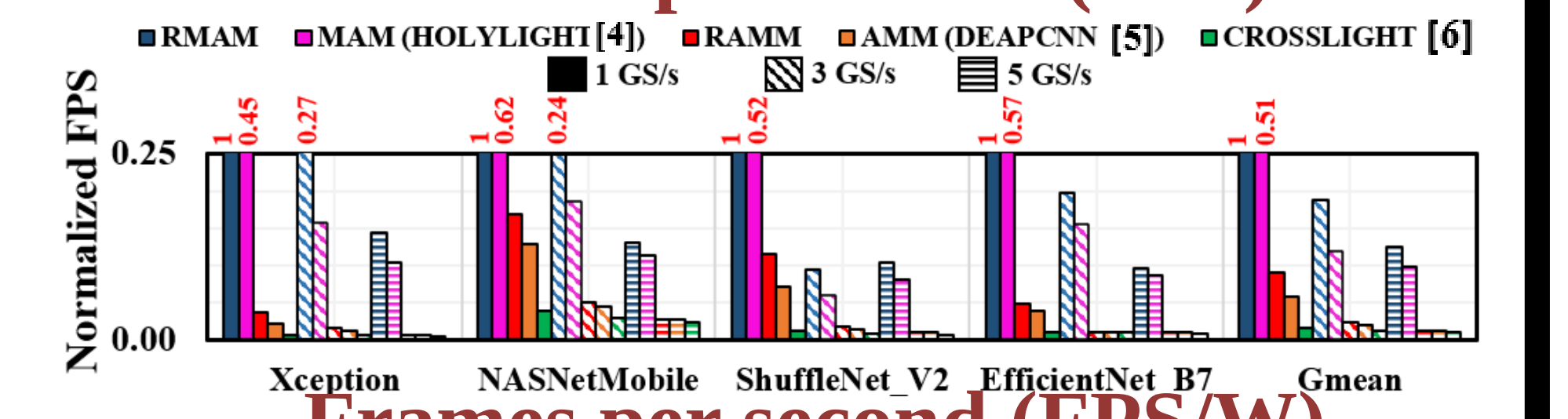
System Level Implementation



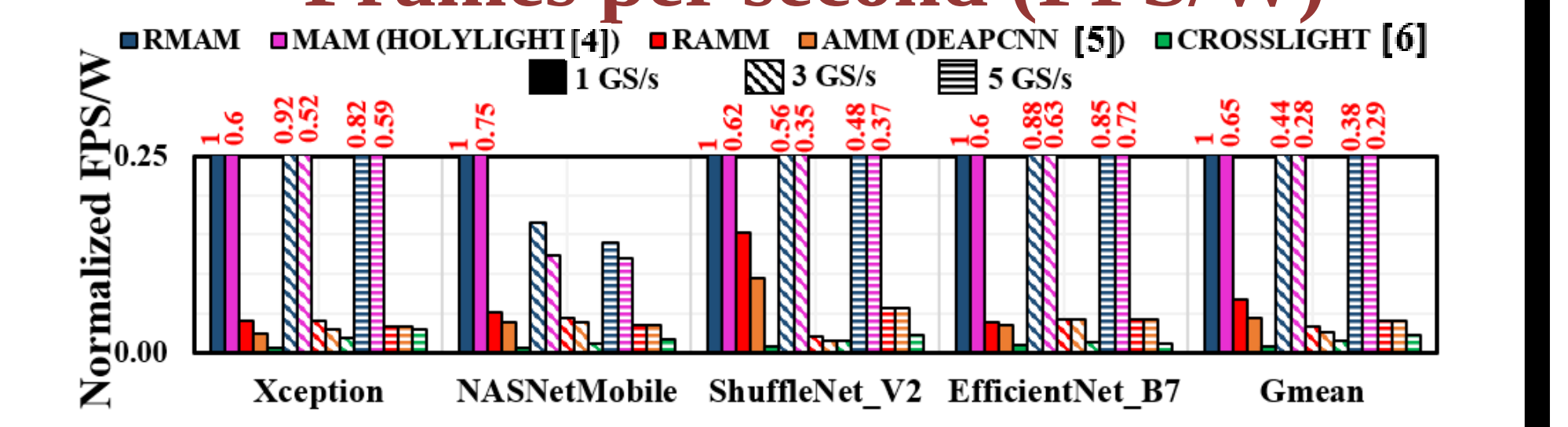
Reconfigurable VDPEs improve MRR utilization and throughput of MAM and AMM organizations

Evaluation

Frames per second (FPS)



Frames per second (FPS/W)



- We compare our RAMM and RMAM accelerator architectures with the baseline AMM (DEAPCNN [5]), MAM (HOLYLIGHT [4]) and the latest variant of AMM design (CROSSLIGHT [6])
- We evaluate accelerators at 4-bit precision and across different DRs such as 1 GS/s, 3 GS/s, and 5 GS/s.
- Results are normalized to RMAM at 1 GS/s.
- Our area proportionate outlook, provides improvements on gmean over the considered CNNs) up to 1.8× in frames-per-second (FPS), and up to 1.5× in FPS/W.

Conclusions

- We presented our novel reconfigurable VDPE design to introduce flexibility in Photonic MRR-based CNN accelerators.
- Our reconfigurable VDPE employs set of comb switches to enable dynamic maximization of the size compatibility between VDPEs and the CNN tensors that are processed using the VDPEs.
- Our evaluation of reconfigurable VDPE equipped -AMM (RAMM) and –MAM (RMAM) on modern CNNs with mixed-sized tensors show substantial improvements in Frames-Per-Second (FPS) and FPS/W (energy efficiency), compared to the photonic MRR-based accelerators from prior work.

References

[1] Angelina R. Totovi´c et al., Compute with Light”, in Optics Express 2021.
[2] A. N. Tait et al., “Microring Weight Banks”, in JSTQE 2016.
[3] L. Yang et al., “On-chip optical matrix-vector multiplier”, in OPIP 2013.
[4] W. Liu et al., “Holylight a nanophotonic accelerator for deep learning in data centers”, in DATE 2019.
[5] V. Bangari et al., “Digital Electronics and Analog Photonics for Convolutional Neural Networks (DEAP-CNNs)”, in JSTQE, 2019.
[6] F.Sunn et al., “CrossLight: A Cross-Layer Optimized Silicon Photonic Neural Network Accelerator”, in DAC, 2021.
[7] A. Biberman et al., “Silicon microring resonator-based broadband comb switch for wavelength-parallel message routing”, in LEOS 2007.
[8] M. Tan et al., “Efficientnet: Rethinking model scaling for convolutional neural networks”, CoRR, 2019.
[9] K. He et al., “Deep residual learning for image recognition”, 2015.
[10] F. Chollet, “Xception: Deep learning with depthwise separable convolutions”, CoRR, 2016.
[11] A. G. Howard et al., “Mobilenets: Efficient convolutional neural networks for mobile vision applications”, CoRR, 2017.
[12] B. Zoph et al., “Learning transferable architectures for scalable image recognition”, CoRR, 2017.
[13] X. Zhang et al., “Shufflenet: An extremely efficient convolutional neural network for mobile devices”, CoRR, 2017.
[14] M. A. Al-Qadasi ., "Scaling up silicon photonic-based accelerators: Challenges and opportunities", APL Photonics, 2022.