

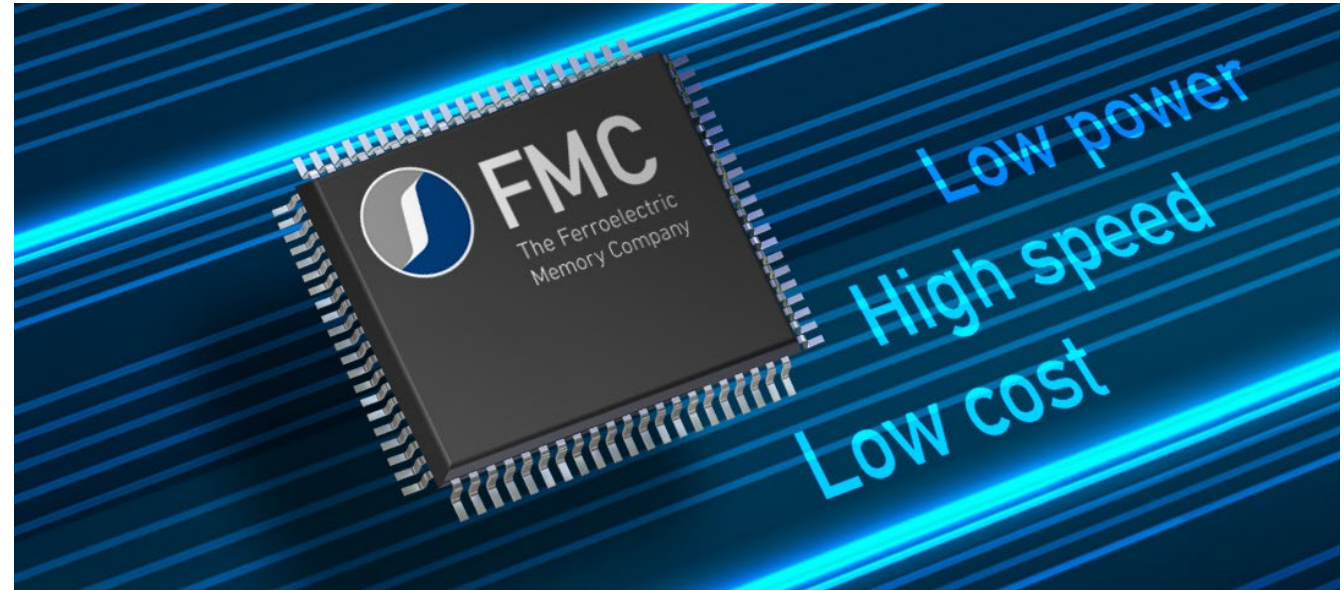


Flash Memory Summit

Near-memory computing – The hidden winner?

Marko Noack

Ferroelectric Memory GmbH



Outline



Flash Memory Summit

- Edge AI Computing
- CNNs
- Convolution
- Array architectures for IMC and NMC
- System architectures
- Results
- Fe-NVRAM

Current issues with AI computing

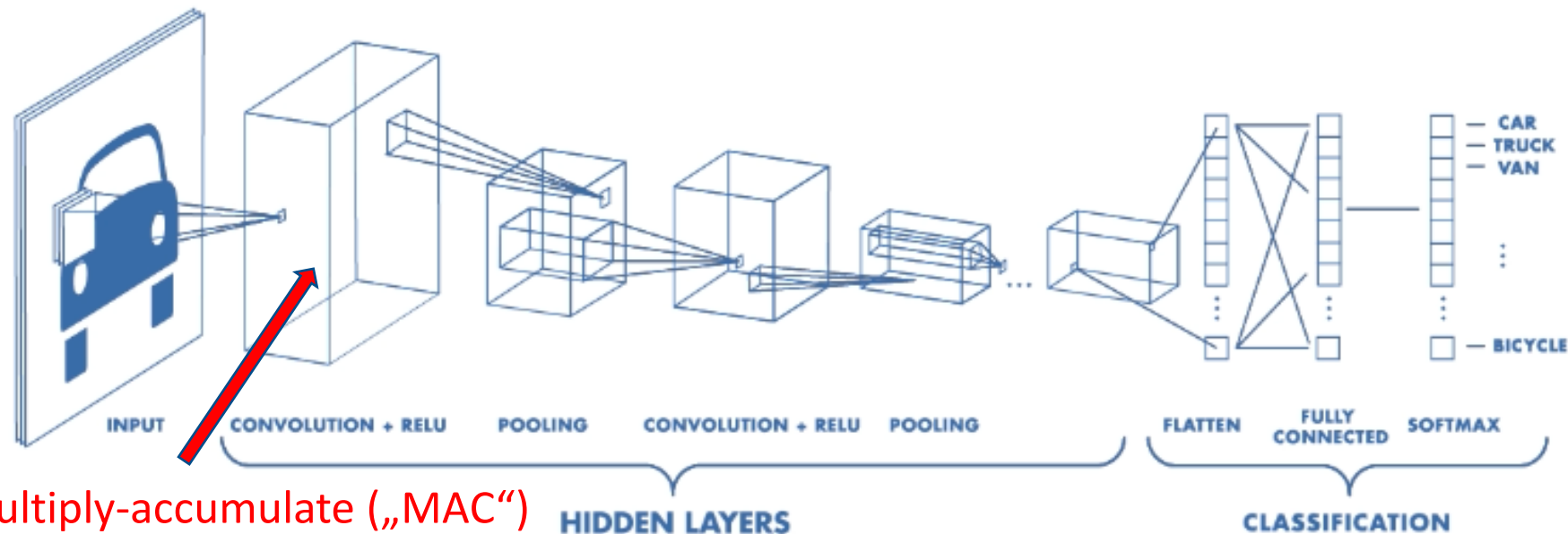
- Growing demand for real-time monitoring and control in industrial and automotive applications
- Network bandwidth limitations
- Large amount of data generated by sensors

Solution

- Moving resource intensive computation from cloud to edge devices
 - → reduced latency
 - → independent from network connection problems
 - → scalable
 - → cost reduction
 - → security enhancement

Common application of EDGE AI: Convolutional Neural Networks (CNNs)

- Speech and image (face) recognition and classification

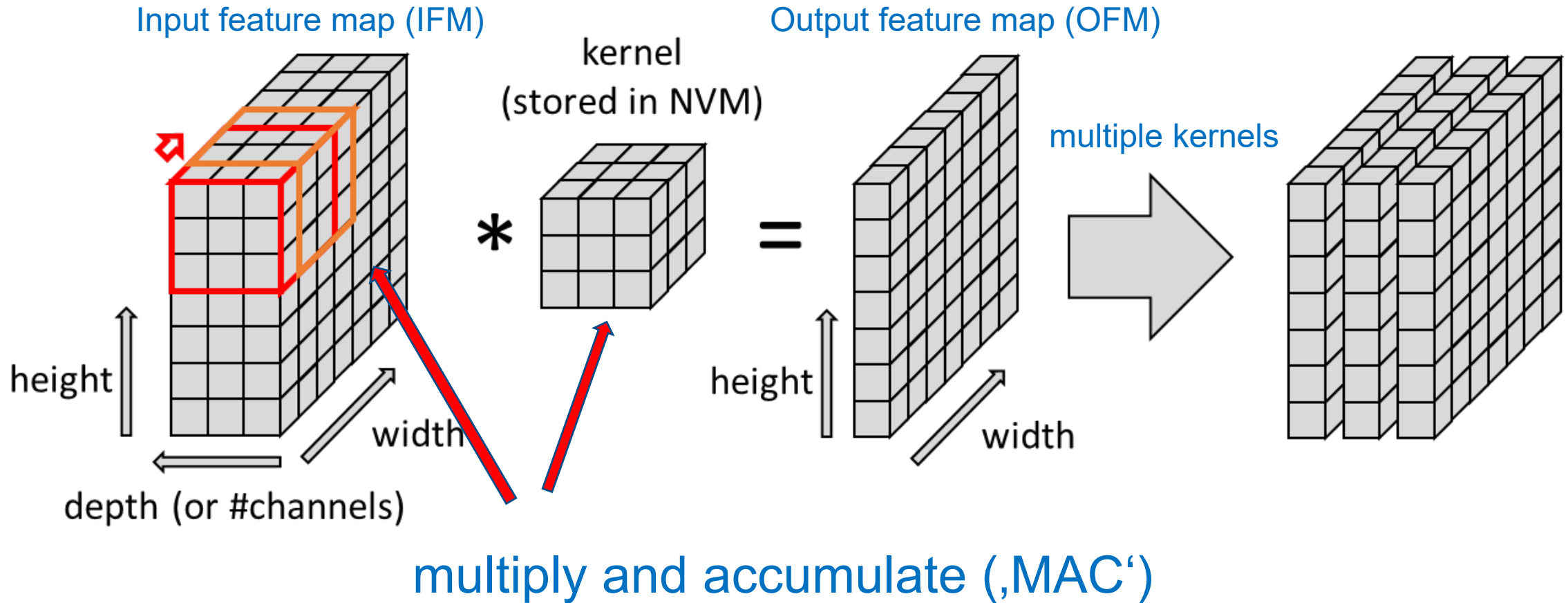


Many multiply-accumulate („MAC“) operations in convolution operation

Convolution operation



Flash Memory Summit



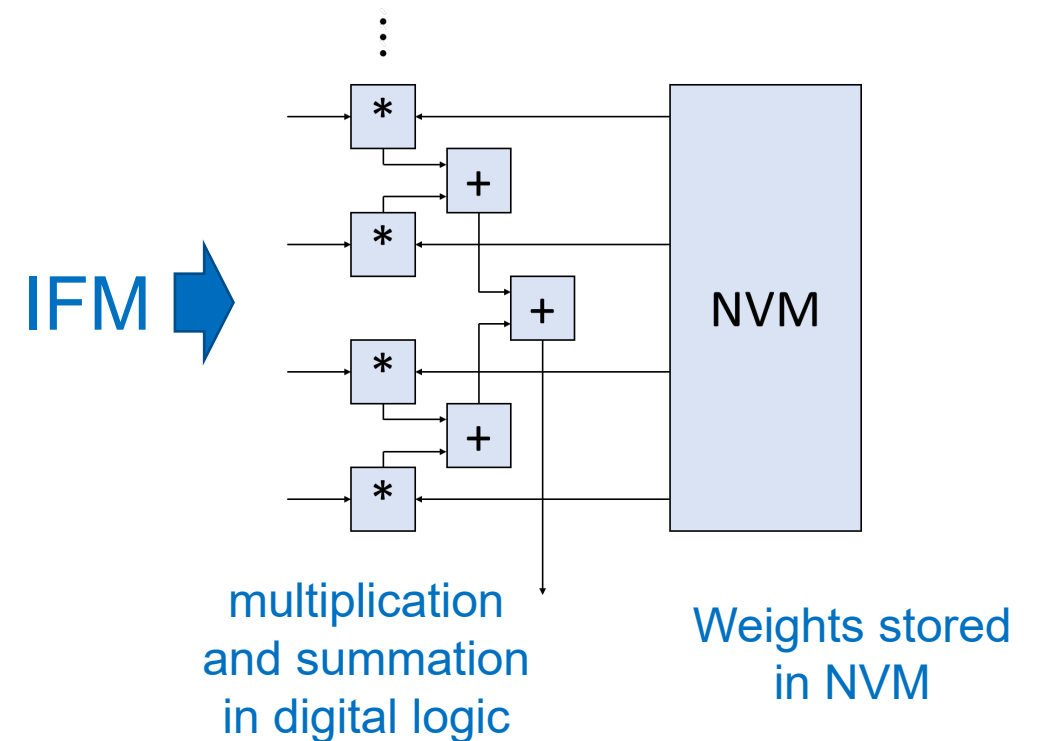
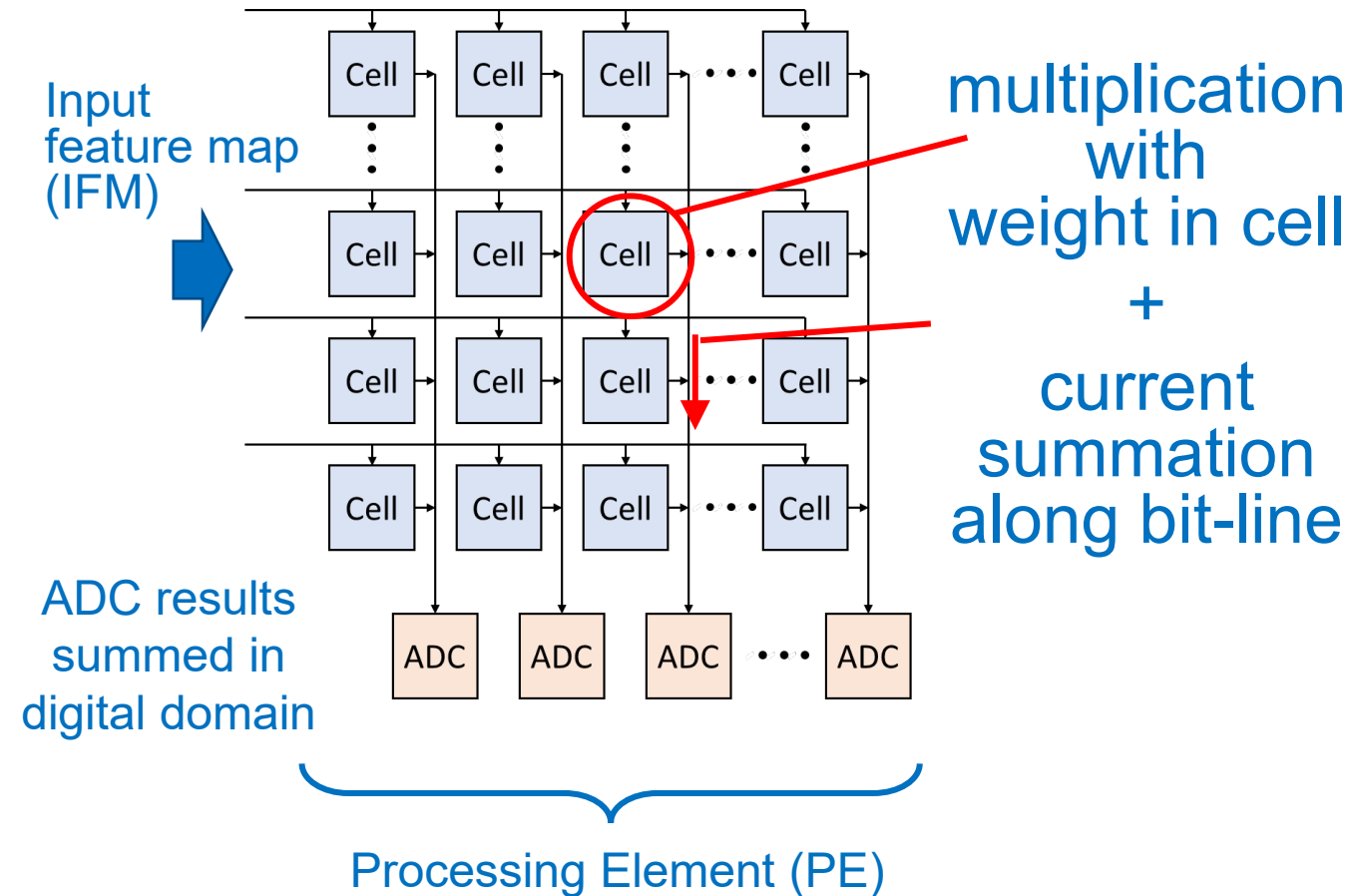
Comparison: Array Architecture



Flash Memory Summit

In-memory computing (IMC)

Near-memory computing (NMC)



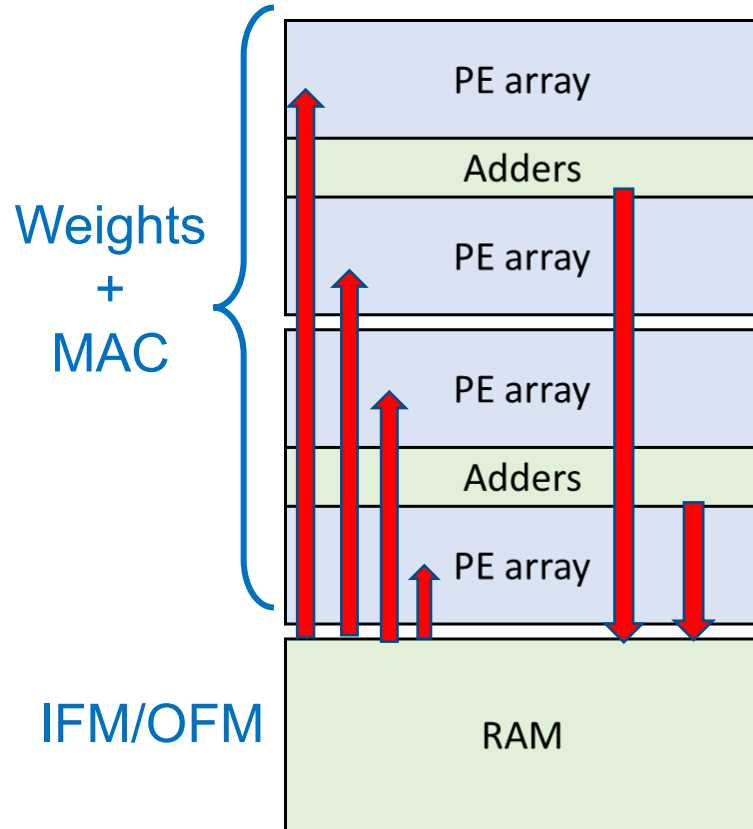
Comparison: System Architecture



Flash Memory Summit

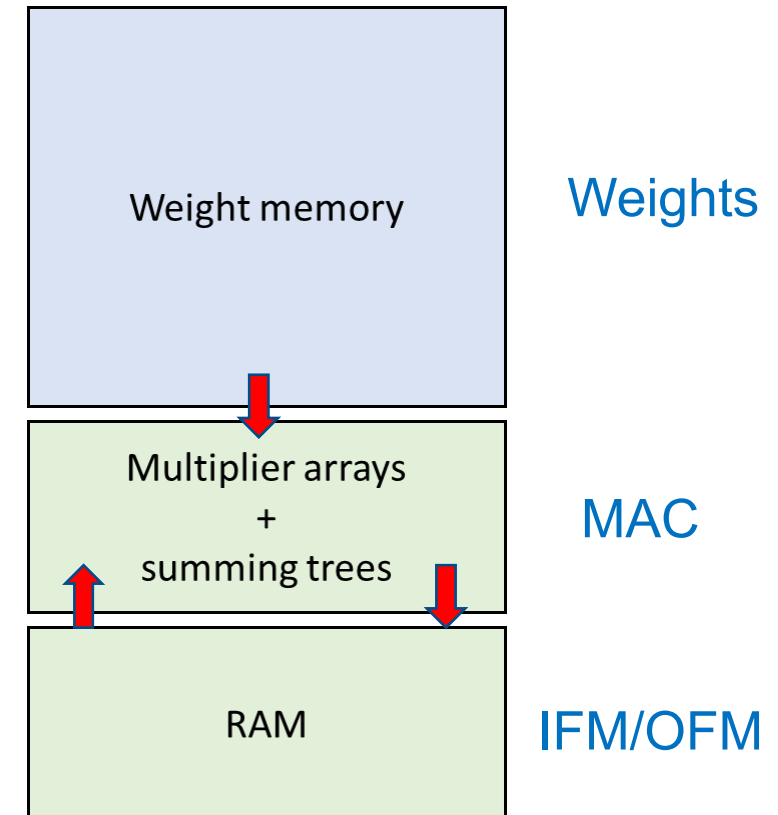
In-memory computing (IMC)

Near-memory computing (NMC)

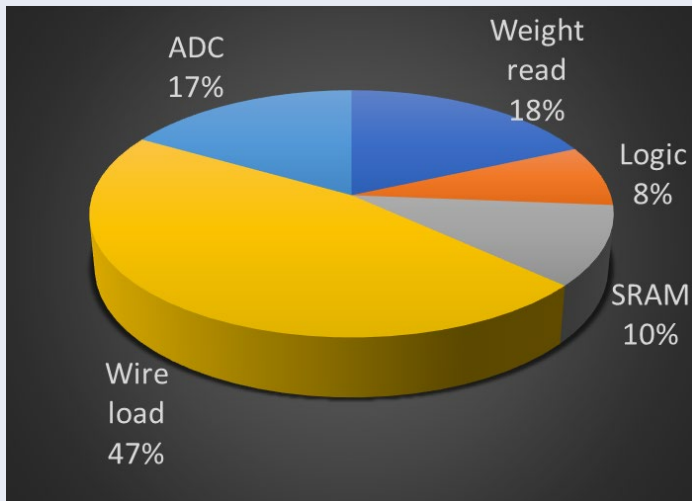
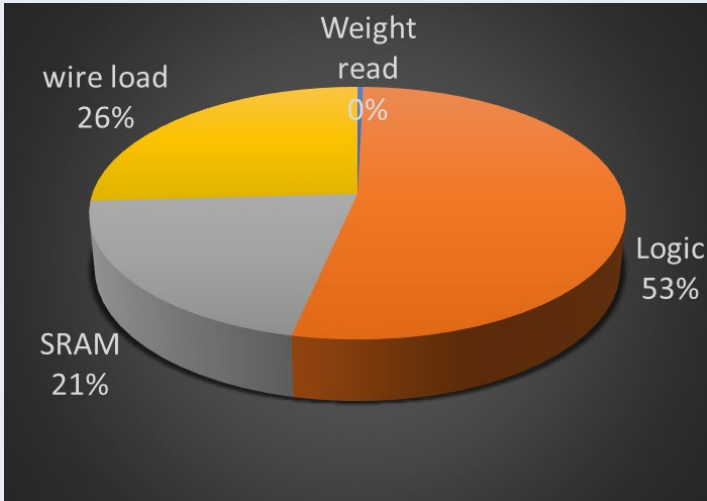


Comparable models for both approaches developed

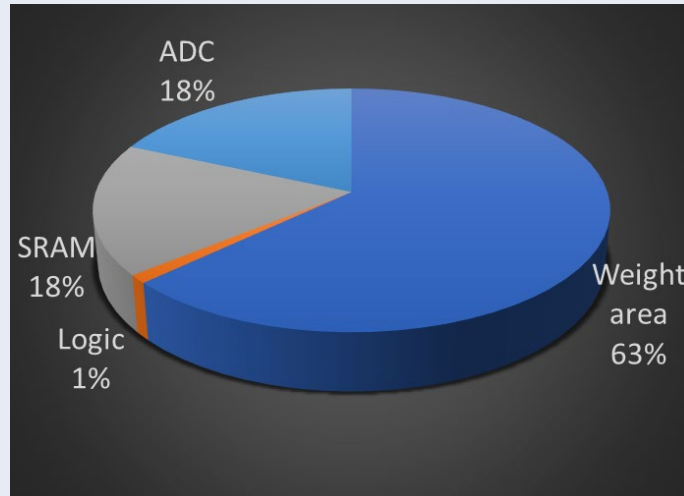
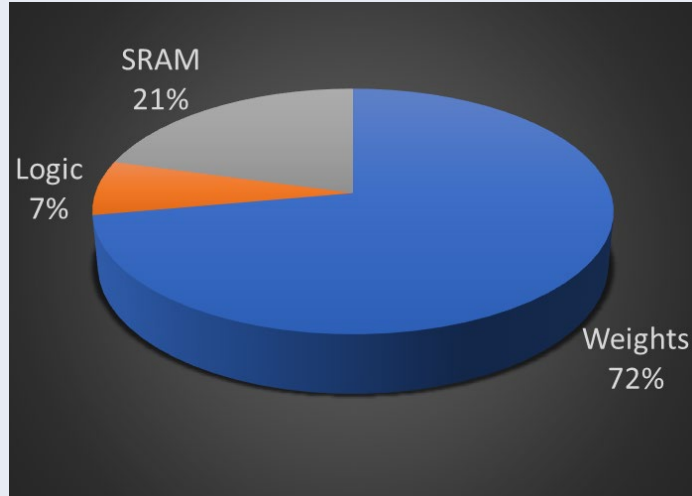
- FeFET as storage element
- Resnet50, 23M parameters stored as 4-bit weights
- Array and wire parasitics
- ADC, SRAM and logic power consumptions
- 28nm technology



Comparison: Power efficiency

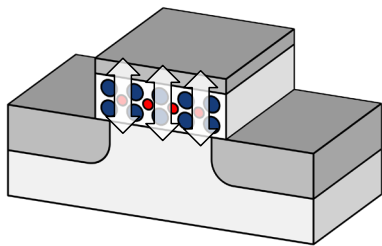
	IMC	NMC																						
TOPs/W	5.2	10.4																						
Energy breakdown	 <p>A 3D pie chart showing the energy breakdown for IMC. The largest slice is Wire load at 47%, followed by Weight read at 18%, ADC at 17%, SRAM at 10%, and Logic at 8%.</p> <table><thead><tr><th>Component</th><th>Percentage</th></tr></thead><tbody><tr><td>Wire load</td><td>47%</td></tr><tr><td>Weight read</td><td>18%</td></tr><tr><td>ADC</td><td>17%</td></tr><tr><td>SRAM</td><td>10%</td></tr><tr><td>Logic</td><td>8%</td></tr></tbody></table>	Component	Percentage	Wire load	47%	Weight read	18%	ADC	17%	SRAM	10%	Logic	8%	 <p>A 3D pie chart showing the energy breakdown for NMC. The largest slice is Logic at 53%, followed by wire load at 26%, SRAM at 21%, and Weight read at 0%.</p> <table><thead><tr><th>Component</th><th>Percentage</th></tr></thead><tbody><tr><td>Logic</td><td>53%</td></tr><tr><td>wire load</td><td>26%</td></tr><tr><td>SRAM</td><td>21%</td></tr><tr><td>Weight read</td><td>0%</td></tr></tbody></table>	Component	Percentage	Logic	53%	wire load	26%	SRAM	21%	Weight read	0%
Component	Percentage																							
Wire load	47%																							
Weight read	18%																							
ADC	17%																							
SRAM	10%																							
Logic	8%																							
Component	Percentage																							
Logic	53%																							
wire load	26%																							
SRAM	21%																							
Weight read	0%																							
	<ul style="list-style-type: none">• Wire load dominates → IFM data must be routed through PE arrays• ADC is crucial component and needs careful low-power design	<ul style="list-style-type: none">• Logic power consumption dominates• Wire load includes parasitic caps of routing channels• Weight reading power consumption small since weights are reused for convolution																						

Comparison: Area / Speed

	IMC	NMC																		
Estimated chip area	9.7 mm ²	8.2 mm ²																		
Area breakdown	 <table><tr><th>Component</th><th>Percentage</th></tr><tr><td>Weight area</td><td>63%</td></tr><tr><td>ADC</td><td>18%</td></tr><tr><td>SRAM</td><td>18%</td></tr><tr><td>Logic</td><td>1%</td></tr></table>	Component	Percentage	Weight area	63%	ADC	18%	SRAM	18%	Logic	1%	 <table><tr><th>Component</th><th>Percentage</th></tr><tr><td>Weights</td><td>72%</td></tr><tr><td>SRAM</td><td>21%</td></tr><tr><td>Logic</td><td>7%</td></tr></table>	Component	Percentage	Weights	72%	SRAM	21%	Logic	7%
Component	Percentage																			
Weight area	63%																			
ADC	18%																			
SRAM	18%																			
Logic	1%																			
Component	Percentage																			
Weights	72%																			
SRAM	21%																			
Logic	7%																			
TOPs	1.64	1.64																		
Clock frequency	400 Mhz	400 Mhz																		
	<ul style="list-style-type: none">Speed determined by ADC conversion rate	<ul style="list-style-type: none">Speed determined by data path (multipliers + summing tree)																		

Using Fe-NVRAM as NVM

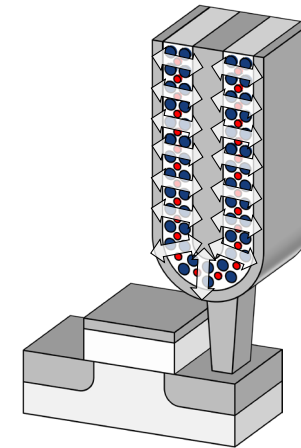
- NVM takes large portion of area in both cases
- FMC's new Fe-NVRAM cell creates a low power, low cost and high performance solution for NMC
- Chip area of NMC approach shrinks from 8.2mm² to 4.9mm²



FeFET cell



shrink factor 4



Fe-NVRAM cell

Conclusion

- FMC's new FeFET technology offers solutions for both IMC and NMC
- IMC is very energy efficient on array level
- But on system level our model shows that energy efficiency, area and performance of IMC and NMC are comparable
- using new Fe-NVRAM cell can reduce area drastically for NMC approach