Undergraduate Project – CIM Accuracy Analysis and Multi-Exit Architecture

Speaker: 吳育丞 吳冠緯 陳宥辰

Mentor: Kevin Chris

Advisor: Prof. An-Yeu (Andy) Wu

Date:01/26/2022

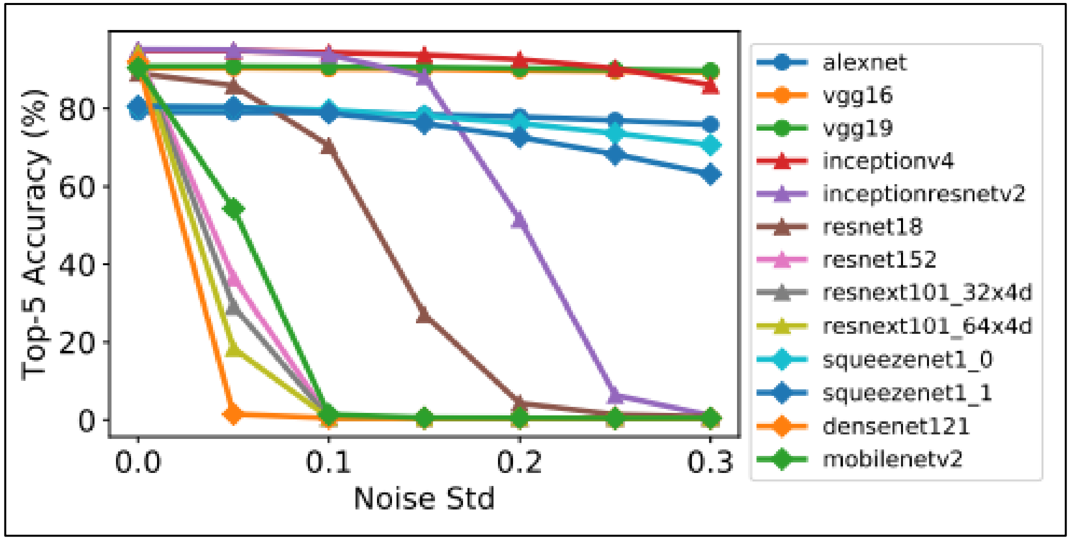
Outline

- Sensitivity Analysis of CIM Accuracy
  - Related Work: IEDM(2019)
  - Experimental Setting
  - Experimental Results Analysis
- Multi-Exit Architecture
- Summary & Future Work

Noise-Resilience of Different DNN Architectures (1/2)

- Compared to digital accelerators, **CIM accelerators are more sensitive to non-idealities** of the memory devices and the peripheral circuits.

- Different DNNs have different sensitivities to noise, but there is no clear relationship between the **sensitivity** and the **ideal accuracy**.

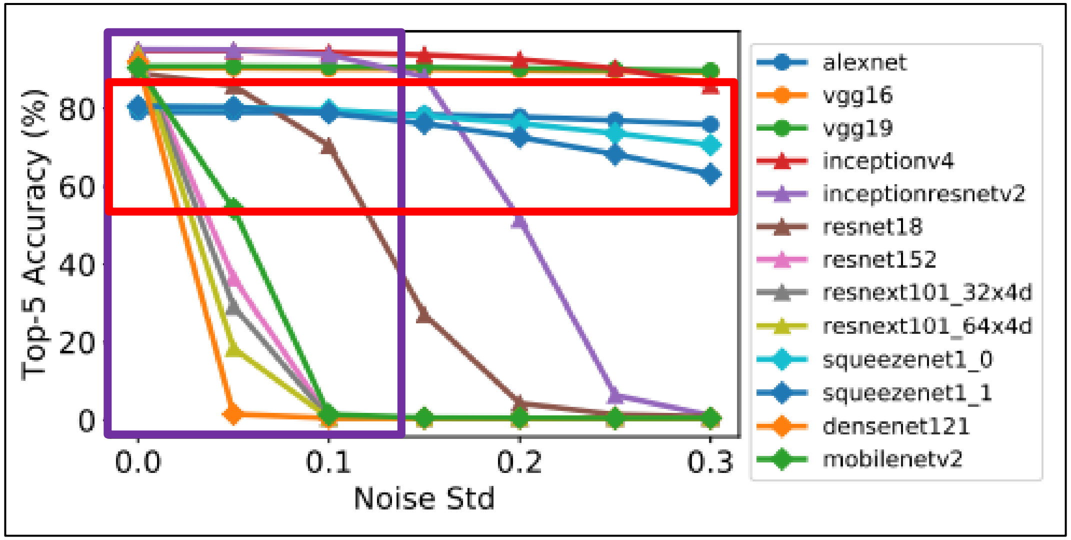


[1]

Noise-Resilience of Different DNN Architectures (2/2)

- The rank of DNNs in terms of **accuracy** may change from one noise standard deviation to another.

e.g. **Resnet** models have better accuracy than **Squeezenet** models without noise but **Squeezenet** models are more robust with bigger noise.



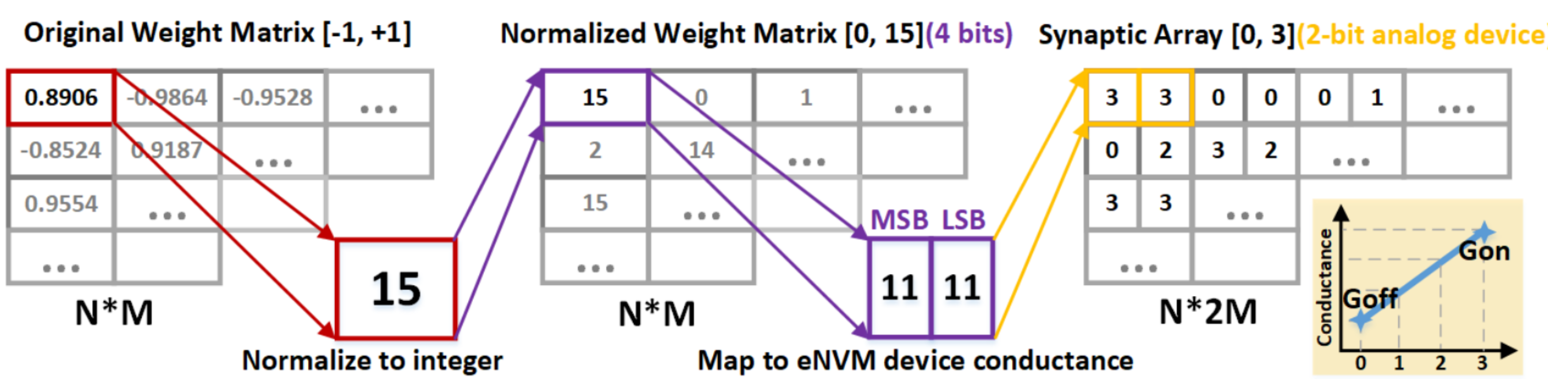
[1]

Outline

- Sensitivity Analysis of CIM Accuracy
  - Related Work: IEDM(2019)
  - Experimental Setting
  - Experimental Results Analysis
- Multi-Exit Architecture
- Summary & Future Work

Quantization of WAGE[2]
(Used in DNN+NeuroSim)

- **The range of the original weight is [-1,+1]**
- Mapping floating point weights to integers
  (Bit length of integers is decided by the synaptic weight precision)
- Due to limited resolution of memory cells, we may need to map single weight to multiple cells.
- In practice, we simulate with 8-bits weight precision and store each weight with two cells(4 bits per cell).

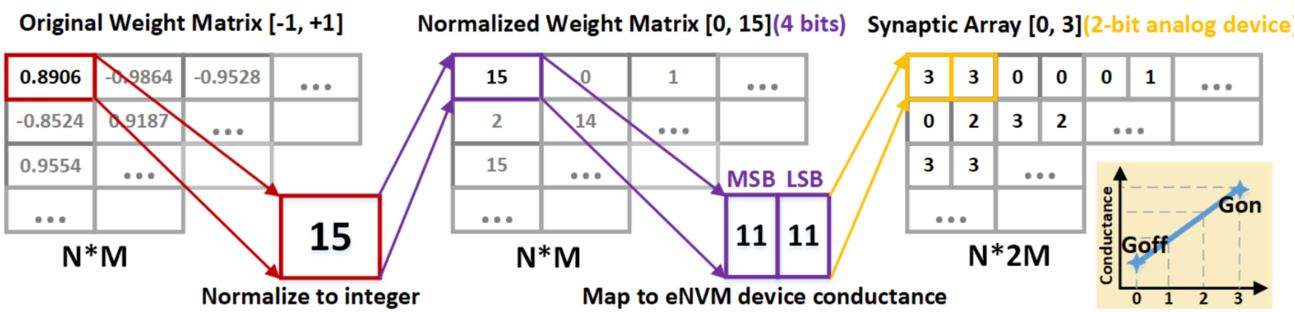


Modified Quantization[3]

(The training process is more stable)

- Difference: The range of the original weight is not [-1, 1]

Clamp the original weight to [-1, 1]
→ original weight / max of |weight|

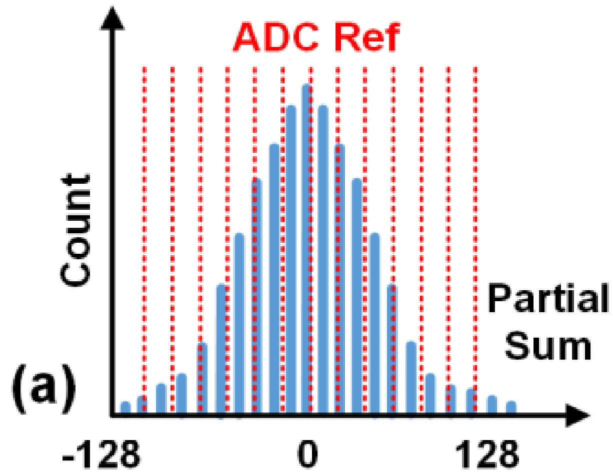


Same as WAGE

Rescaling → Output * max of original |weight|

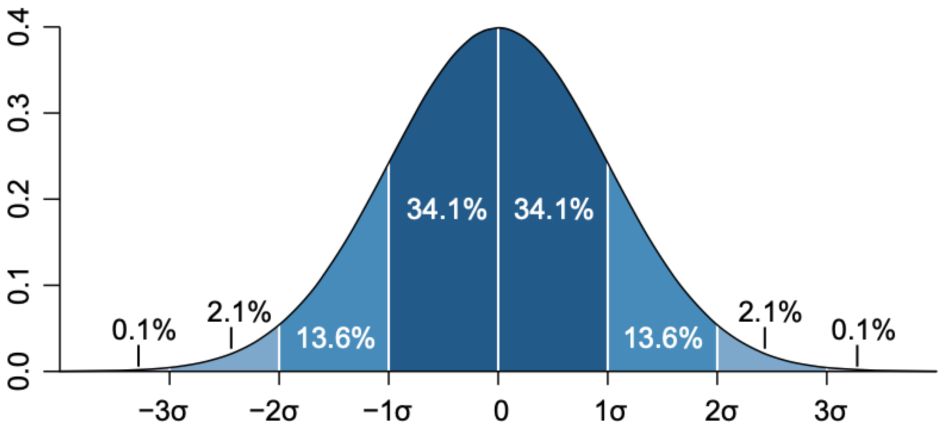
Simulation of Hardware Non-Ideal Effects (1/2)

- **ADC quantization:**
  - Simulate ADC quantization effect by quantizing partial sums of each synaptic array **linearly**.
  - We set **ADC resolution = 10** (**Control Variable**) in inference simulation to avoid its effect overshadowing the effect of conductance variation.



Simulation of Hardware Non-Ideal Effects (2/2)

- **Conductance variation:**
  - We simulate conductance variation effect with a normal distribution model centered at the ideal conductance of each cell.
  - **Inferencing with different standard deviations. (σ = 0~0.2)** (**Independent Variable**)



Resnet Architecture Used for Simulation

input

↓

Conv2d(3, 16, kernel=3)

↓

BasicBlock1(channel=16, kernel=3)*num

↓

BasicBlock2(channel=32, kernel=3)*num

↓

BasicBlock3(channel=64, kernel=3)*num

↓

ReLU

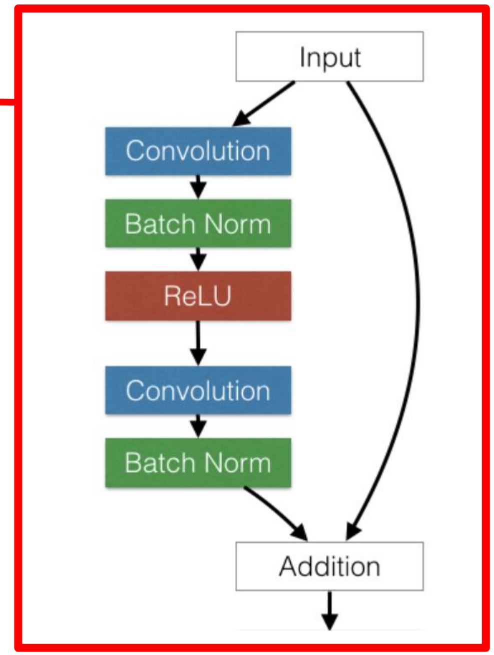
↓

⋮

FC

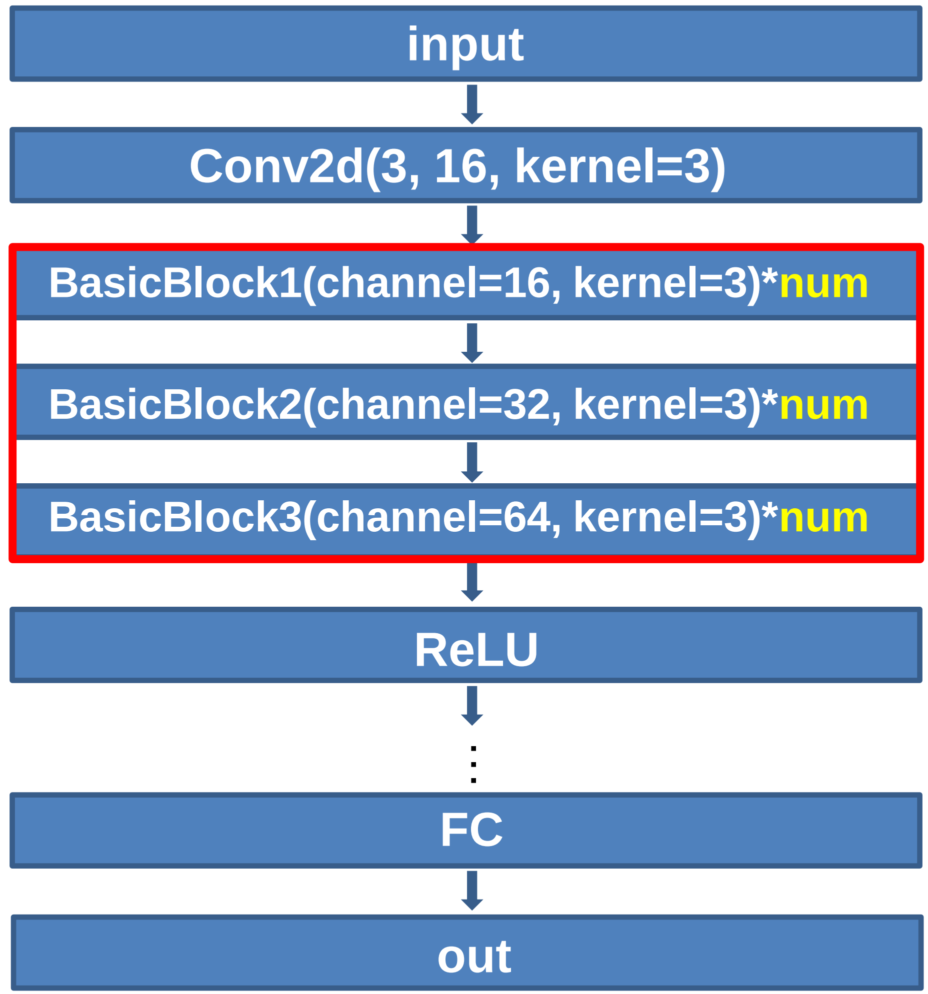
↓

out



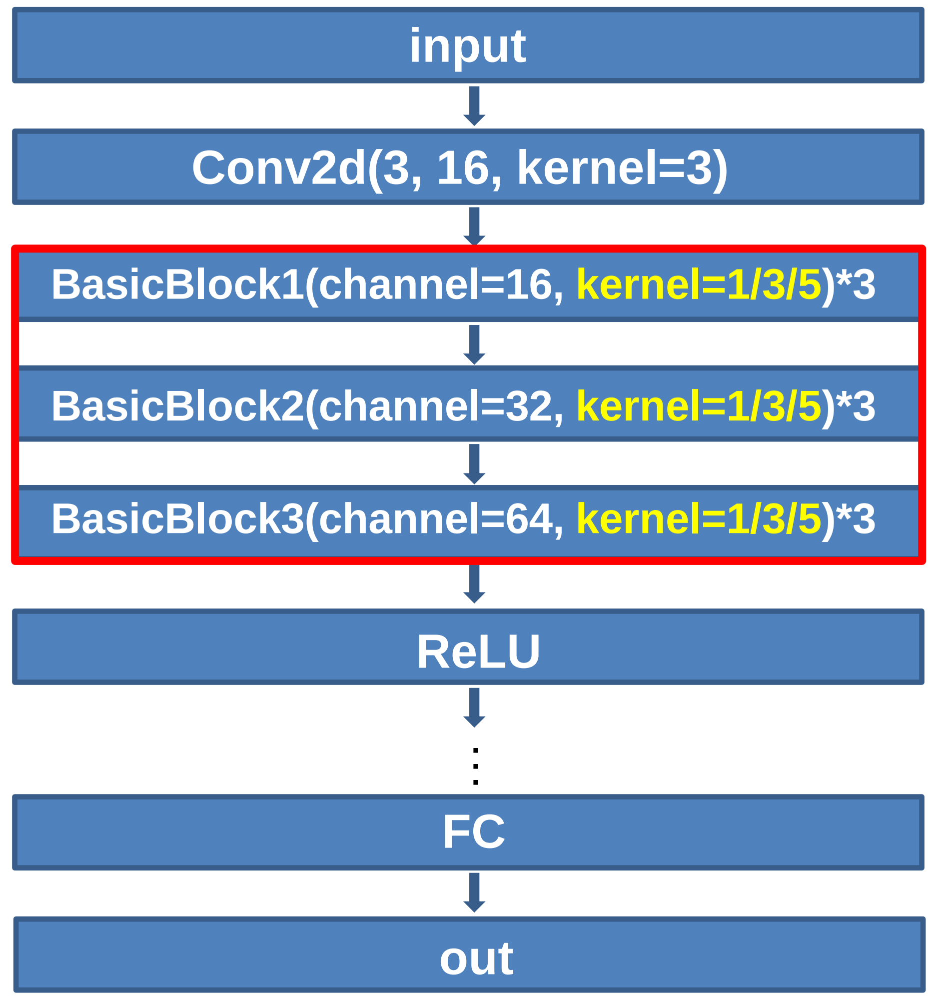
Independent Variables of Our Simulation

- Depth(Number of layers)
  - Resnet20(**num = 3**)
  - Resnet32(**num = 5**)
  - Resnet44(**num = 7**)
  - Resnet56(**num = 9**)



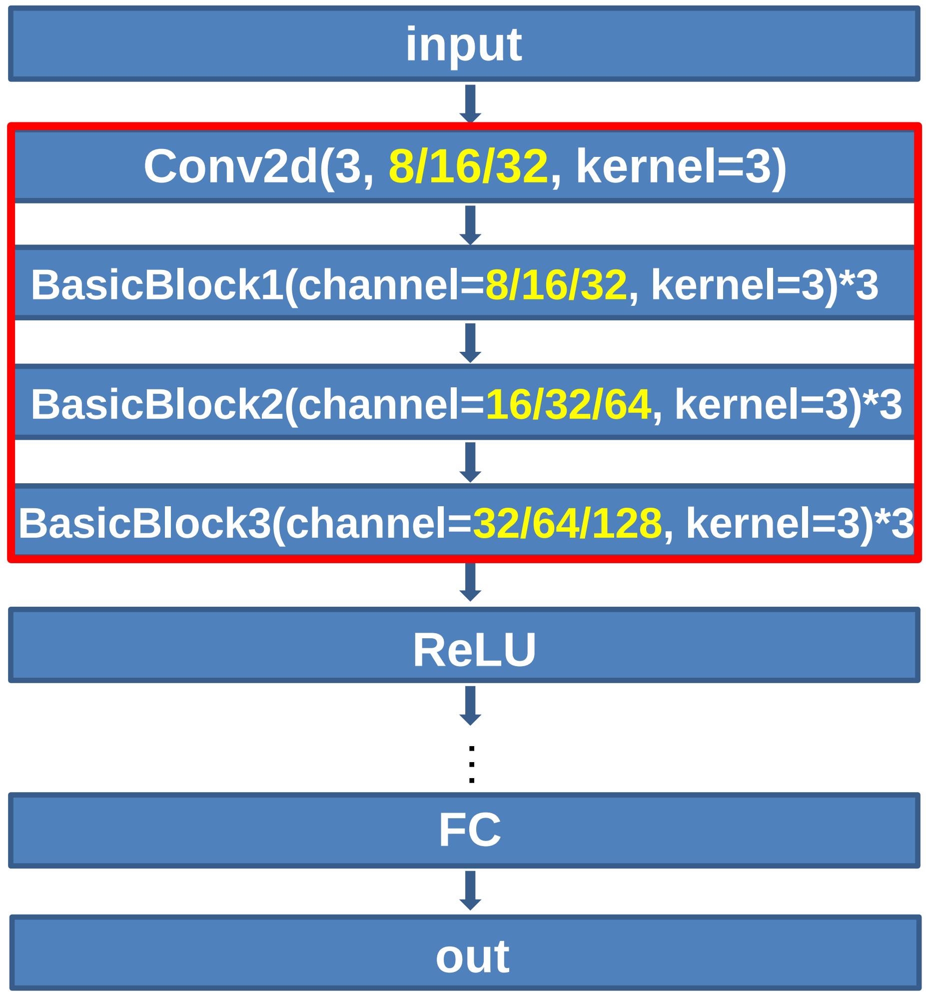
Independent Variables of Our Simulation

- Kernel Size
  - Kernel = 1; Padding = 0
  - Kernel = 3; Padding = 1
  - Kernel = 5; Padding = 2



Independent Variables of Our Simulation

- Channel Size
  - 8→16→32
  - 16→32→64
  - 32→64→128



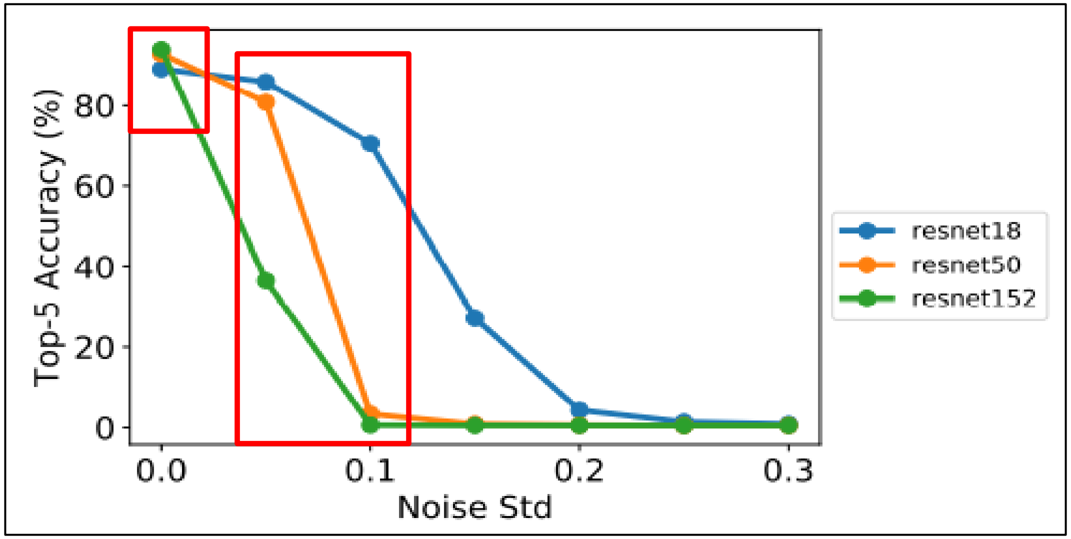
Outline

- Sensitivity Analysis of CIM Accuracy
  - Related Work: IEDM(2019)
  - Experimental Setting
  - Experimental Results Analysis
- Multi-Exit Architecture
- Summary & Future Work

Accu. of Deeper Models Decrease Faster

- Recently, the VGG-like DNNs are less frequently used and replaced by DNNs that are **deeper** and **narrower**.

- The ideal accuracy increases as the depth increases. However, the accuracy of **deeper** DNNs **decreases faster** with increased noise and eventually becomes lower than that of shallower DNNs.



[1]

Simulation Results – DNN Depth (1/4)

- Without applying variation at the beginning, all the Resnet models have an accuracy about **90%**. With the variation becomes higher, the accuracies decrease faster in the cases of deeper sizes.

- The table below also shows that Resnet-44 and Resnet-56 have lower accuracies than others with **σ** = 0.2 though they are the relatively higher two cases without conductance variation.

ADC(10b)	variation **(σ)**				
depth	0	0.05	0.1	0.15	0.2
res-20	88.3	85.5	70.2	41.1	20
res-32	89.4	85.6	65.1	44.6	21.6
res-44	89.7	86	64.7	28.9	14.7
res-56	91.1	88.6	73.4	34.1	17

ideal case

Simulation Results – DNN Depth (1/4)

- Without applying variation at the beginning, all the Resnet models have an accuracy about **90%**. With the variation becomes higher, the accuracies decrease faster in the cases of deeper sizes.

- The table below also shows that Resnet-44 and Resnet-56 have lower accuracies than others with **σ** = 0.2 though they are the relatively higher two cases without conductance variation.

ADC(10b)	variation **(σ)**				
depth	0	0.05	0.1	0.15	0.2
res-20	88.3	85.5	70.2	41.1	20
res-32	89.4	85.6	65.1	44.6	21.6
res-44	89.7	86	64.7	28.9	14.7
res-56	91.1	88.6	73.4	34.1	17

↓29.1

↓39.3

Simulation Results – DNN Depth (2/4)

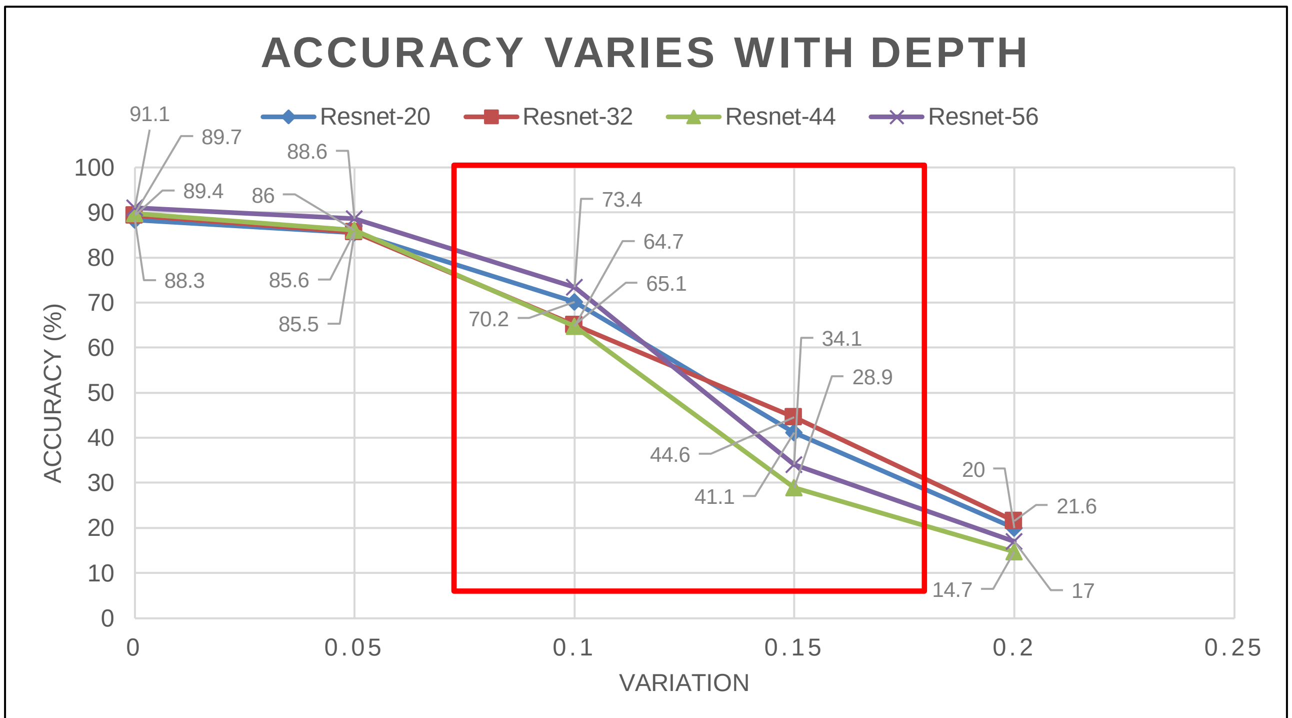
- Without applying variation at the beginning, all the Resnet models have an accuracy about **90%**. With the variation becomes higher, the accuracies decrease faster in the cases of deeper sizes.

- The table below also shows that Resnet-44 and Resnet-56 have lower accuracies than others with **σ** = 0.2 though they are the relatively higher two cases without conductance variation.

ADC(10b)	variation **(σ)**				
depth	0	0.05	0.1	0.15	0.2
res-20	88.3	85.5	70.2	41.1	20
res-32	89.4	85.6	65.1	44.6	21.6
res-44	89.7	86	64.7	28.9	14.7
res-56	91.1	88.6	73.4	34.1	17

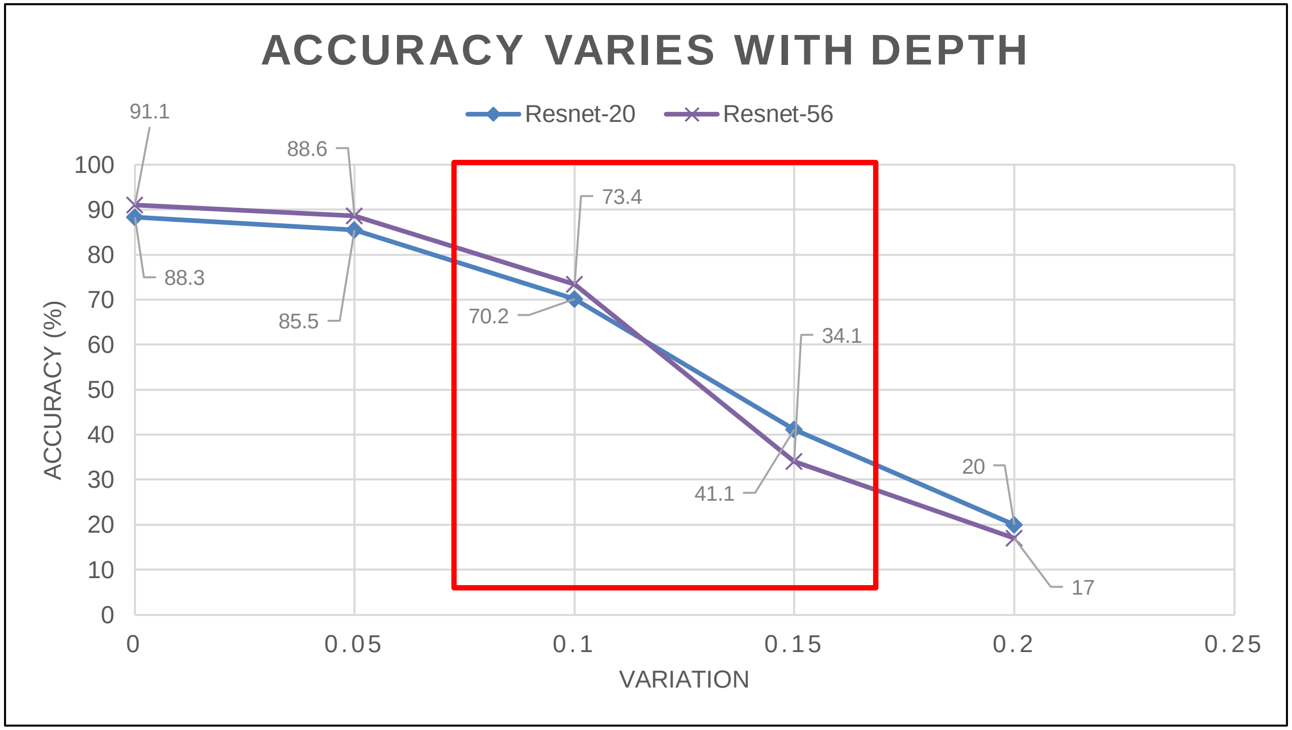
Simulation Results – DNN Depth (3/4)

- With variation = 0 and 0.05, Resnet-44 and Resnet-56 have higher accuracies, while Resnet-20 and Resnet-32 have higher accuracies with variation = 0.15 and 0.2.



Simulation Results – DNN Depth (4/4)

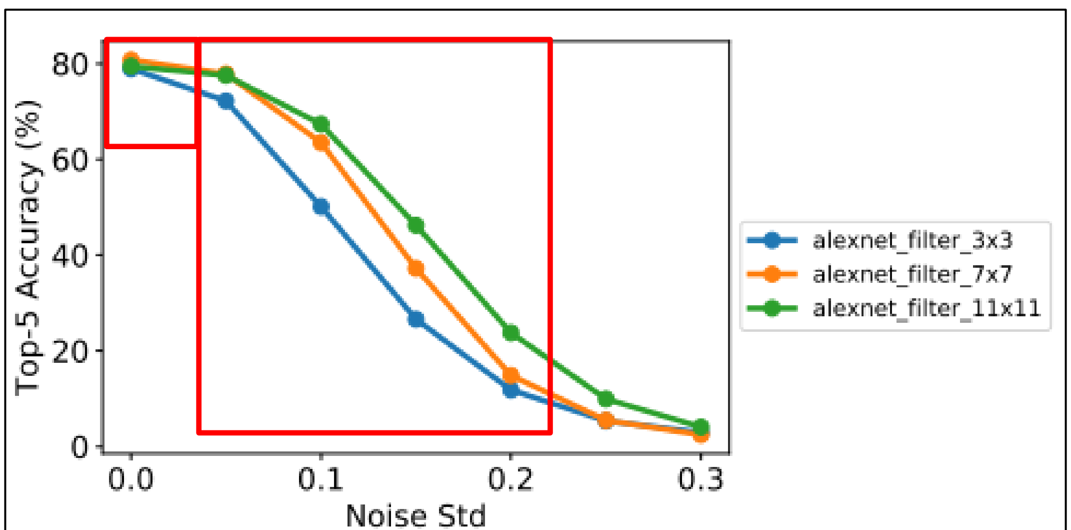
- We can notice that deeper models' accuracies **decrease faster**.



Accu. of Models with Smaller Kernel Size Decrease Faster

- The accuracies of models with **smaller** filters **decreases faster** with noise.

- The recent trend of making a DNN deeper with smaller layers may not be the most suitable for CIM.



[1]

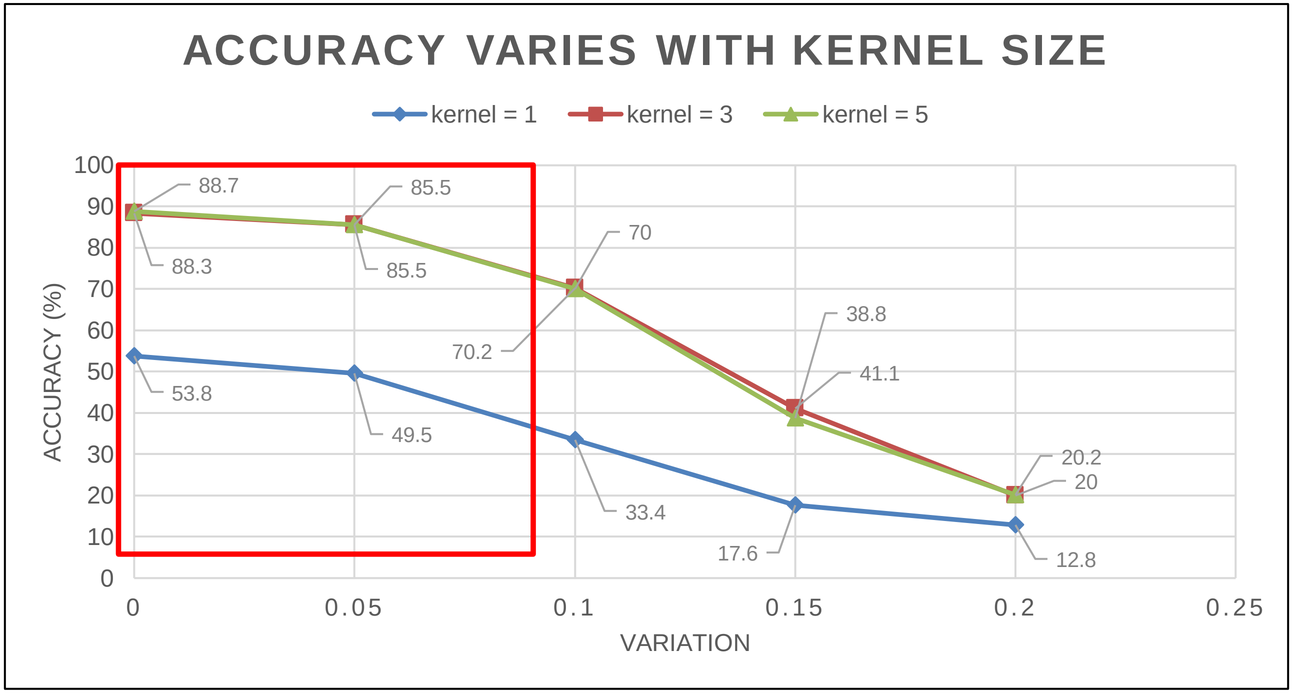
Simulation Results – DNN Kernel Size (1/2)

- Compared to Resnet-20 (default kernel = 3), larger kernel size (5) leads to slightly greater accuracy with **σ** = 0.2.

ADC(10b)	variation **(σ)**				
kernel	0	0.05	0.1	0.15	0.2
1	53.8	49.5	33.4	17.6	12.8
3	88.3	85.5	70.2	41.1	20
5	88.7	85.5	70	38.8	20.2

Simulation Results – DNN Kernel Size (2/2)

- The case of kernel = 1 has lower accuracy than models with kernel size = 3 and 5.
- The case of kernel = 1 **decreases slightly faster** than that of kernel = 3 and kernel = 5.



Simulation Results – DNN Channel Size (1/3)

- Narrower models **decrease faster** with increasing conductance variation.

- Without variation, all three cases have accuracy higher than **80%**, however, accuracy of model with channel = 8 decreases to **12.3%** while model with channel = 32 retains an accuracy of **75.4%**.

ADC(10b)	variation **(σ)**				
channel	0	0.05	0.1	0.15	0.2
8	82.8	74.7 ↓**29.5**	45.2	21.2	12.3
16	88.3	85.5 ↓**15.3**	70.2	41.1	20
32	90.9	90.3 ↓**1.7**	88.6	84.5	75.4

Simulation Results – DNN Channel Size (2/3)

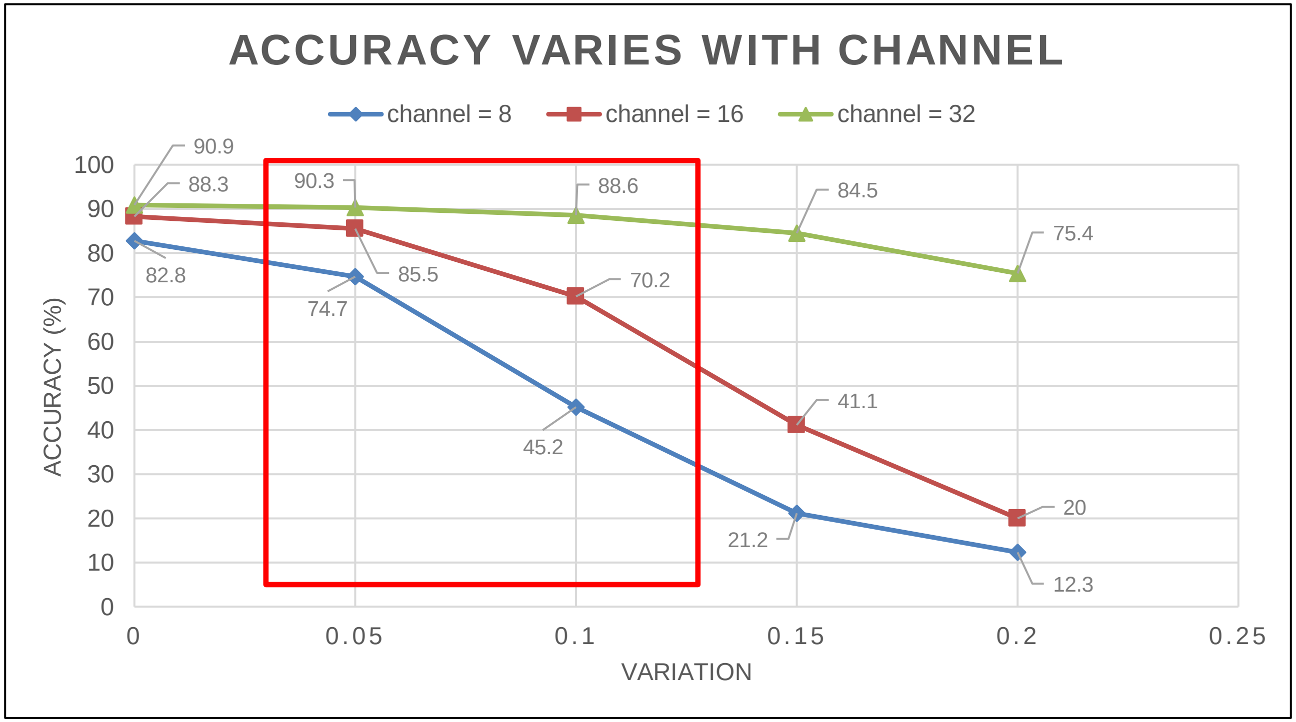
- Narrower models **decrease faster** with increasing conductance variation.

- Without variation, all three cases have accuracy higher than **80%**, however, accuracy of model with channel = 8 decreases to **12.3%** while model with channel = 32 retains an accuracy of **75.4%**.

ADC(10b)	variation **(σ)**				
channel	0	0.05	0.1	0.15	0.2
8	82.8	74.7	45.2	21.2	12.3
16	88.3	85.5	70.2	41.1	20
32	90.9	90.3	88.6	84.5	75.4

Simulation Results – DNN Channel Size (3/3)

- The graph below shows that accuracies of narrower models **decrease faster** with increasing conductance variation.

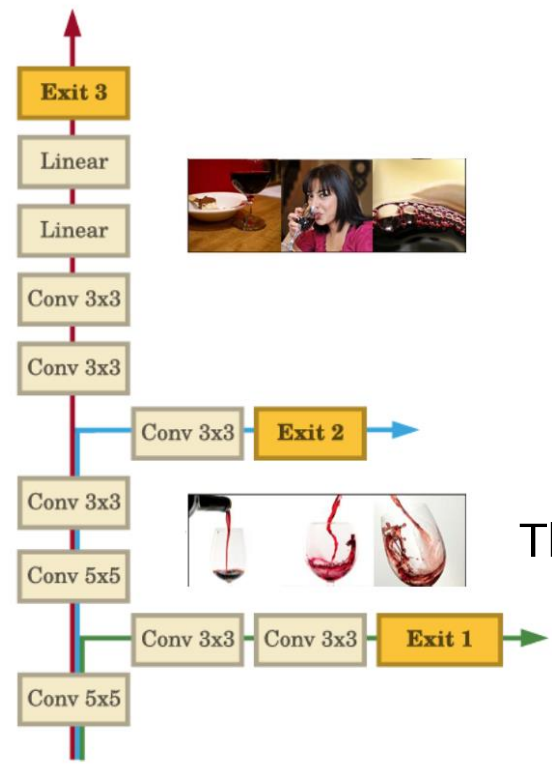


Outline

- Sensitivity Analysis of CIM Accuracy
  - Related Work: IEDM(2019)
  - Experimental Setting
  - Experimental Results Analysis
- Multi-Exit Architecture
- Summary & Future Work

Multi-Exit Architecture

- BranchyNet [4] :
  - Produce multiple sub-models by adding multiple exits.
  - The most suitable exit might be different on each CIM-based accelerator due to varying **σ** of conductance variation → The model is more flexible.

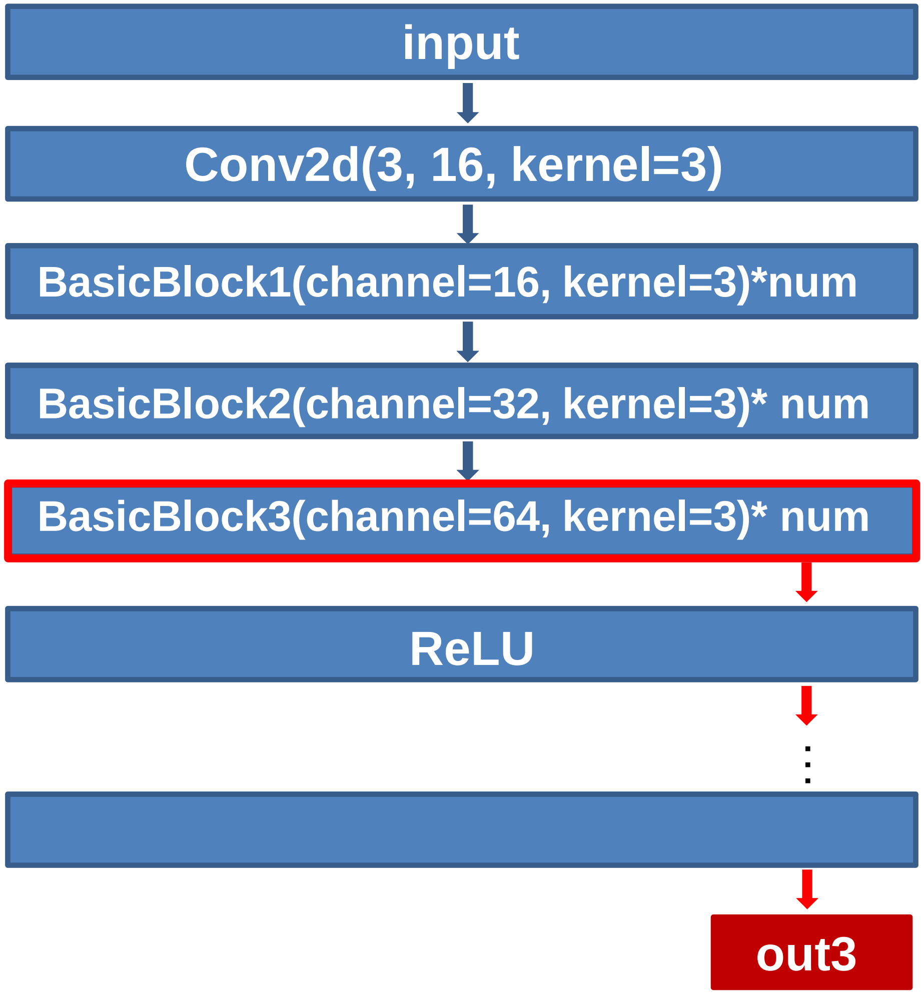


The entropy is already low.
(< threshold value)

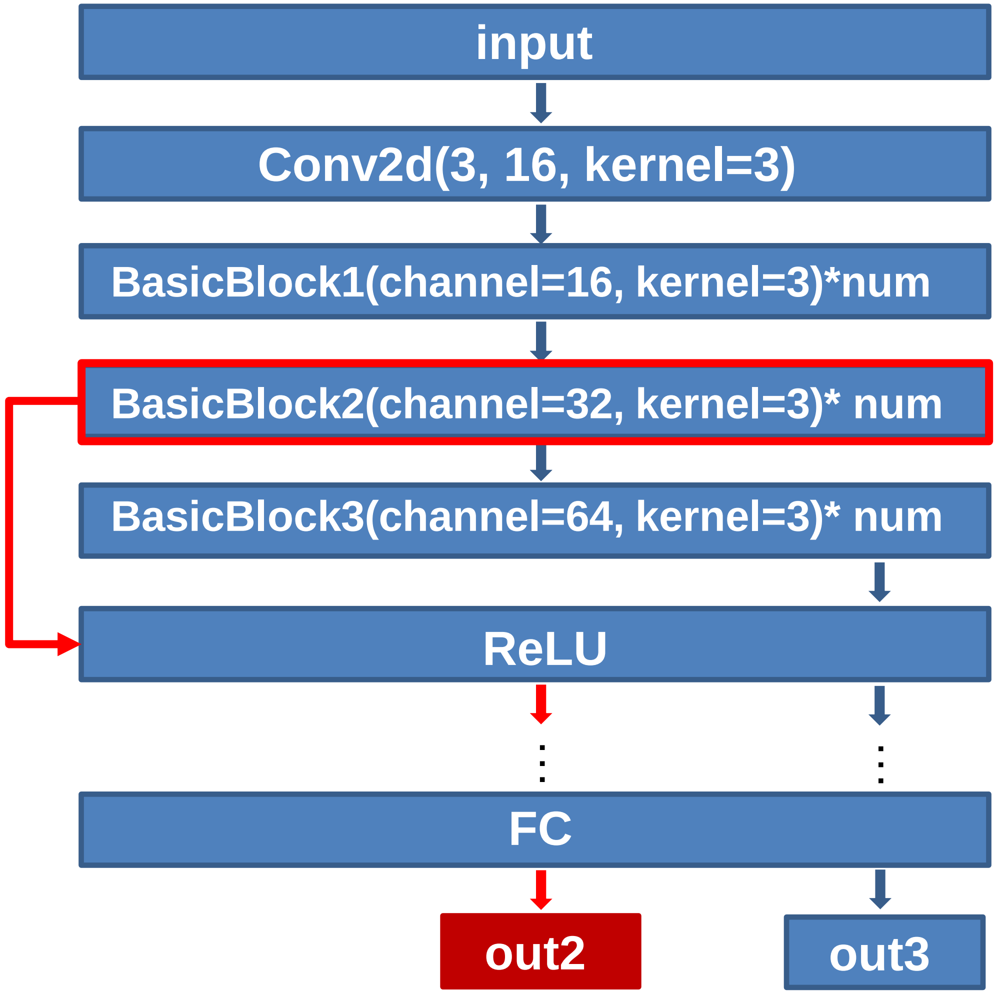
Resnet with Multi-Exits

input
↓
Conv2d(3, 16, kernel=3)
↓
BasicBlock1(channel=16, kernel=3)*num
↓
BasicBlock2(channel=32, kernel=3)* num
↓
BasicBlock3(channel=64, kernel=3)* num
↓
ReLU
↓
⋮
FC
↓
out

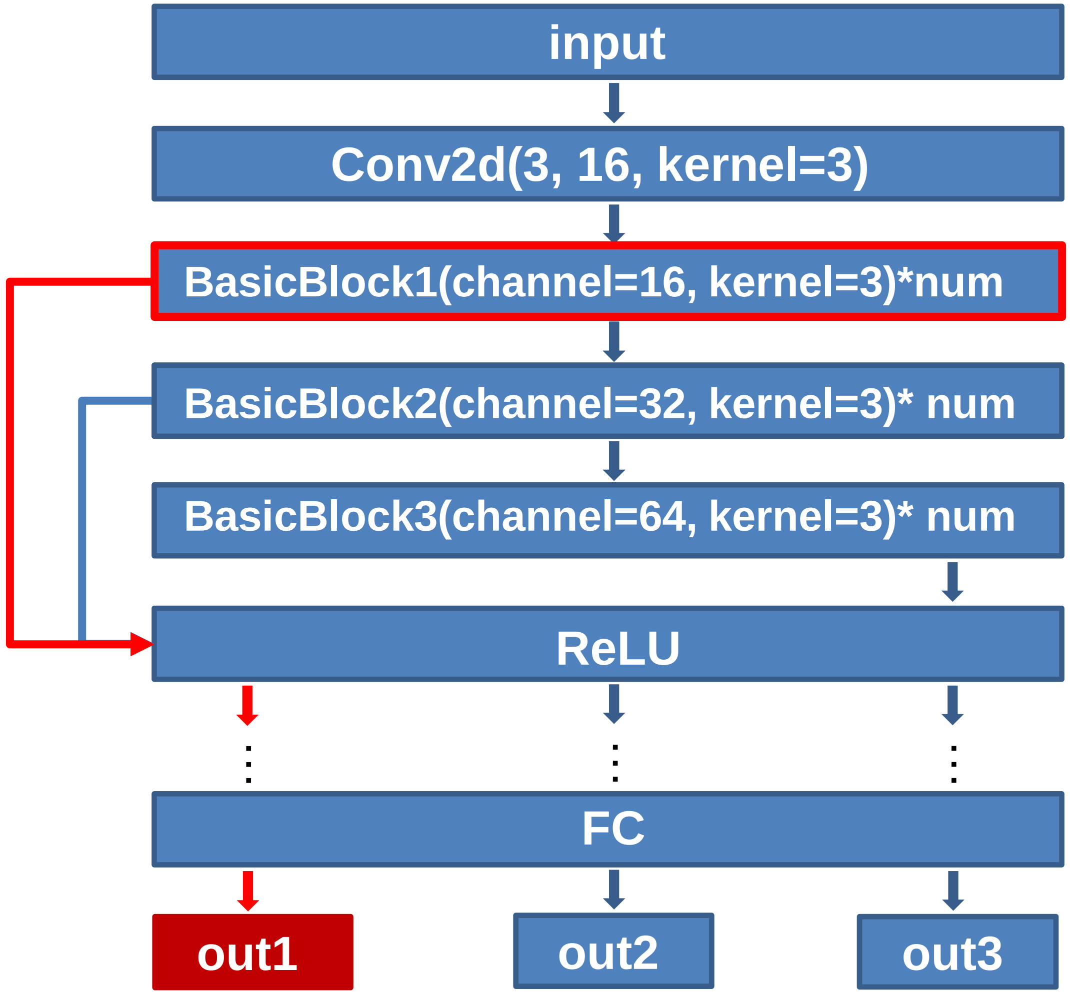
Resnet with Multi-Exits



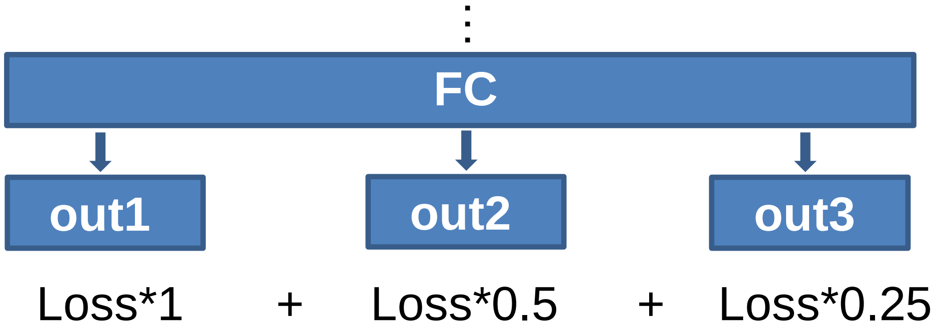
Resnet with Multi-Exits



Resnet with Multi-Exits



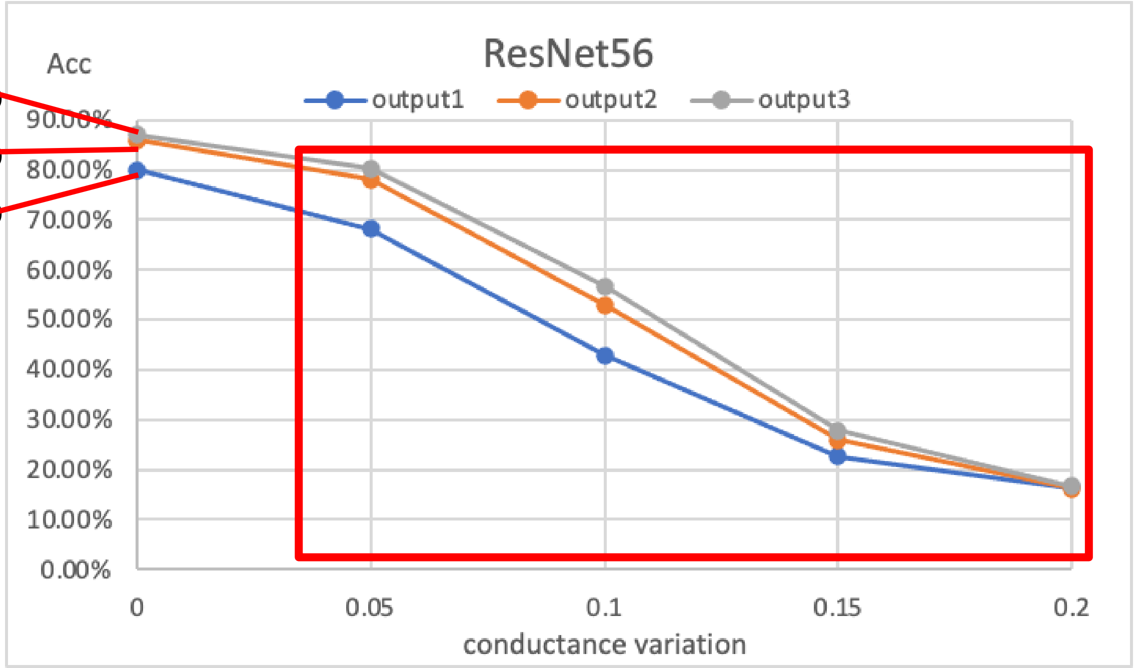
Train Resnet with Multi-Exits



Assign **shallower exit's loss with bigger weight** to increase the accuracy of shallower exits.
→ Higher probability of outputs exiting from shallower exits.

Inference Results (ResNet56)

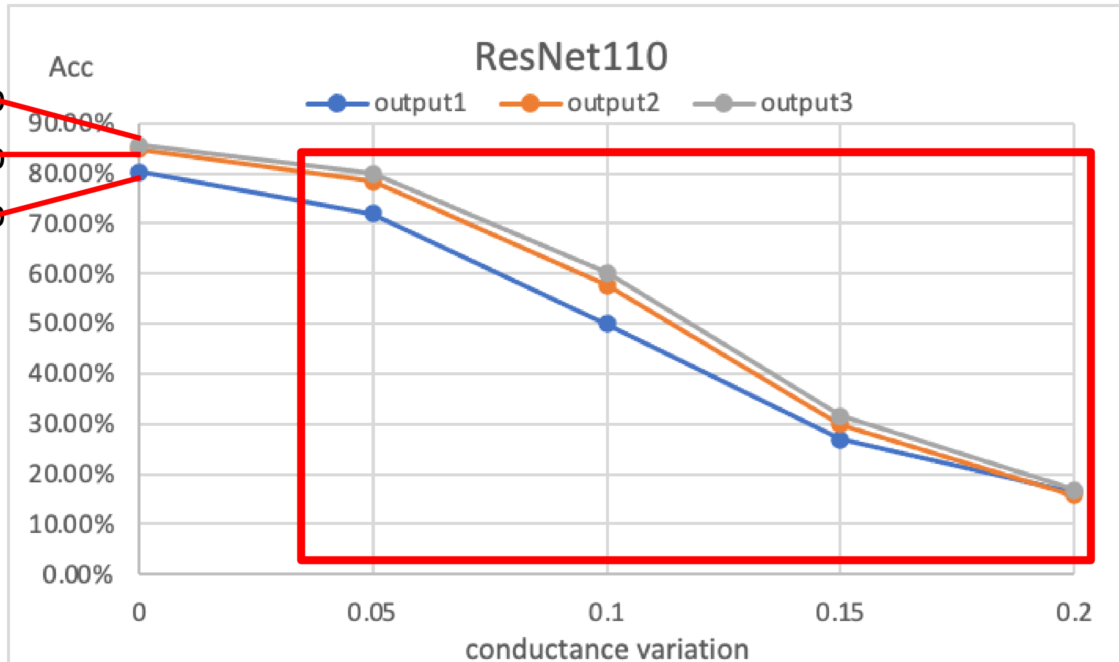
output3: 87.0%
output2: 85.9%
output1: 79.9%



Deeper exit's accuracy drops faster but is still higher.

Inference Results (ResNet110)

output3: 85.6%
output2: 84.8%
output1: 80.3%



Deeper exit’s accuracy drops faster but is still higher.

Shallower Exit Performs Better in Some Cases

- Bird
  - Without variation:
    Exit1: 84.8%, **Exit2: 87.9%**, Exit3: 84.8%
  - With σ = 0.1:
    Exit1: 51.5%, Exit2: 48.5%, Exit3: 48.5%

- Plane
  - Without variation:
    Exit1: 58.6%, Exit2: 69.0%, **Exit3: 75.9%**
  - With σ = 0.1:
    Exit1: 37.9%, **Exit2: 55.2%**, Exit3: 51.7%

Outline

- Sensitivity Analysis of CIM Accuracy
  - Related Work: IEDM(2019)
  - Experimental Setting
  - Experimental Results Analysis
- Multi-Exit Architecture
- Summary & Future Work

Summary

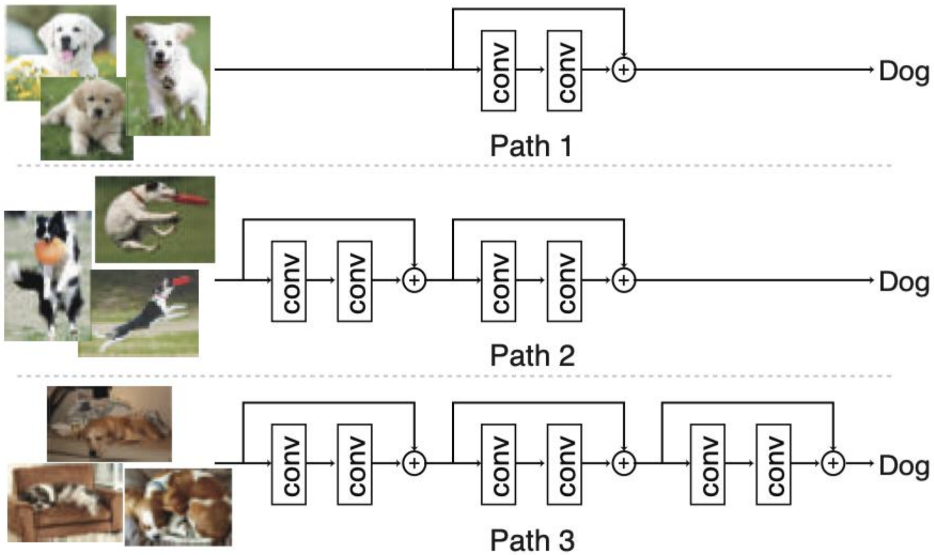
- Shallower Resnet models (less layers) are more robust with conductance variation.
- Wider Resnet models (more channels) are more robust with conductance variation.
- Resnet models with bigger kernel size aren’t visibly more robust with conductance variation.
- Multi-exits model enables users to choose the most suitable exit on different devices.

Future Work

- **Try models other than Resnet with different kernel sizes on another dataset(e.g. ImageNet) with conductance variation.**

- **BlockDrop: Dynamic Inference Paths in Residual Networks [5]**

  Learn a policy to select the most suitable configuration of blocks to correctly classify a given input image on CIM-based accelerators.



Reference & Resource

[1] **Design Considerations for Efficient Deep Neural Networks on Processing-in-Memory Accelerators**
Tien-Ju Yang, Vivienne Sze (Massachusetts Institute of Technology)

[2] **DNN+NeuroSim: An End-to-End Benchmarking Framework for Compute-in-Memory Accelerators with Versatile Device Technologies, IEEE International Electron Devices Meeting (IEDM), 2019.**
X. Peng, S. Huang, Y. Luo, X. Sun and S. Yu (Georgia Institute of Technology)

[3] **Algorithm-Accelerator Co-Design for Deep Learning Specialization**
Zhiru Zhang, School of ECE (Cornell University)

[4] **BranchyNet: Fast Inference via Early Exiting from Deep Neural Networks**
Surat Teerapittayanon, Bradley McDanel, H.T. Kung (Harvard University)

[5] **BlockDrop: Dynamic Inference Paths in Residual Networks**
Zuxuan Wu, Tushar Nagarajan, Abhishek Kumar, Steven Rennie Larry S. Davis1, Kristen Grauman, Rogerio Feris
(UMD, UT Austin, IBM Research, Fusemachines Inc.)

Thanks for listening !