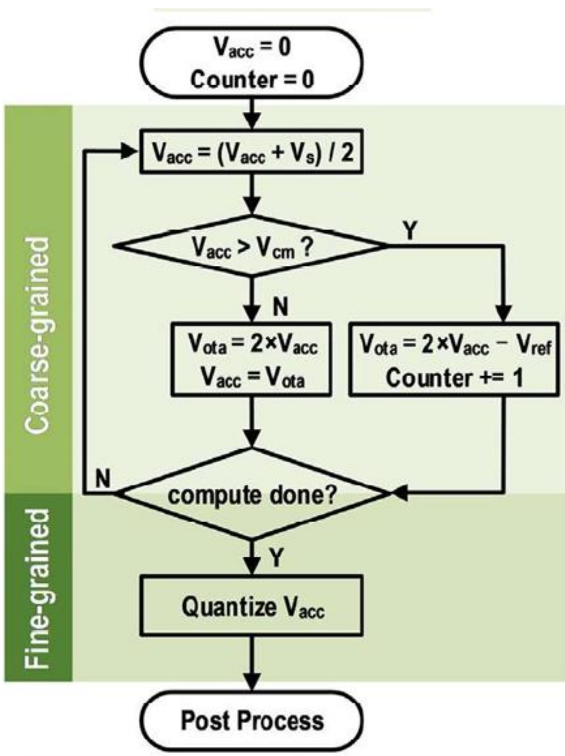
Parasitic Effects on Dual-Granularity Cooperative Quantization (DGCQ) System

B11901027 王仁軒

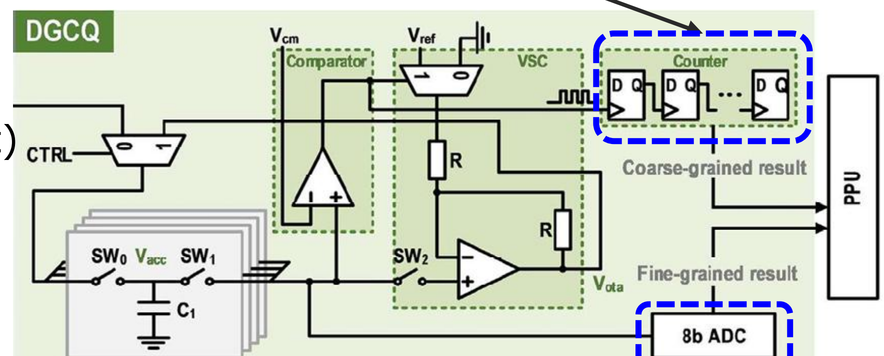
Recall

- Dual-granularity cooperative quantizer
  - When quantizing analog voltage, use another ckt to calculate an approximated result



Low precision

Vs
(MAC result)



High precision

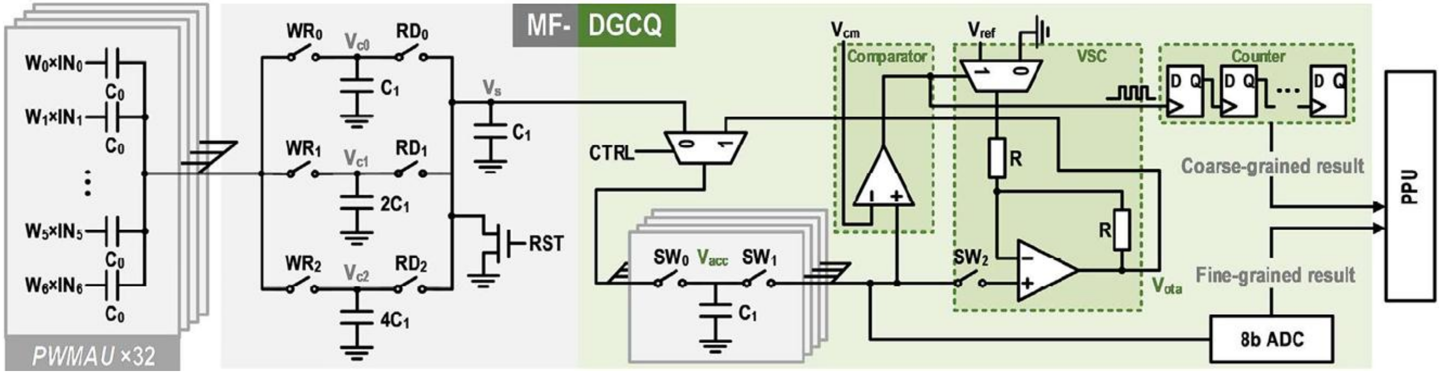
Total : N iterations

Total V : N · (Vs/2)

Counter : N · (Vs/2) / Vref

Quantizing MAC Voltage

- Coarse-grained method
  - Takes 2^n iterations to get n bit precision
  - Area is roughly a fixed value
- SAR ADC
  - Takes n iterations to get n bit precision
  - Area is proportional to 2^n (dominated by caps)



8 x 32 x 8

$$0.\underbrace{011}_{\text{Coarse-grained}}\underbrace{00011010}_{\text{Fine-grained}} \times \text{VDD}$$

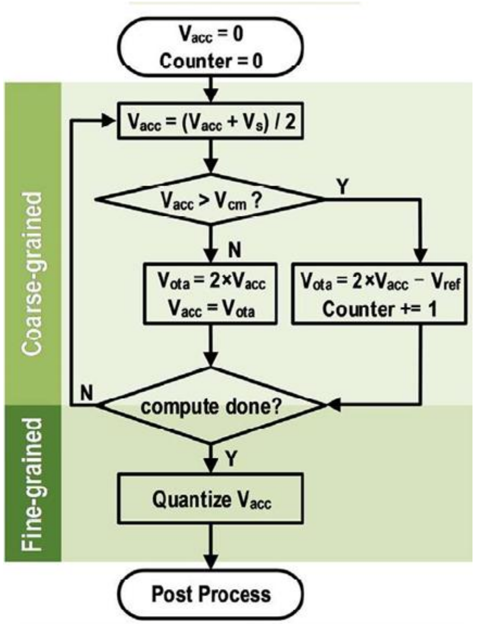
Quantizing MAC Voltage

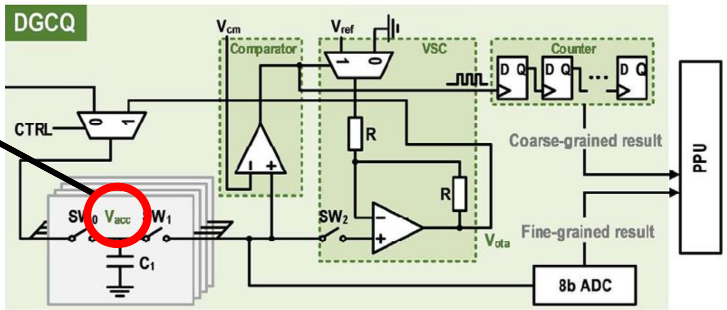
- Partition of coarse/fine grained result
  - Suppose Vref=VDD

$$0.\underbrace{011}_{\text{Coarse-grained}}\underbrace{00011010}_{\text{Fine-grained}} \quad \text{x VDD}$$

After 8 iteration, $V_{acc} = 8V_s - \lfloor 8V_s \rfloor$

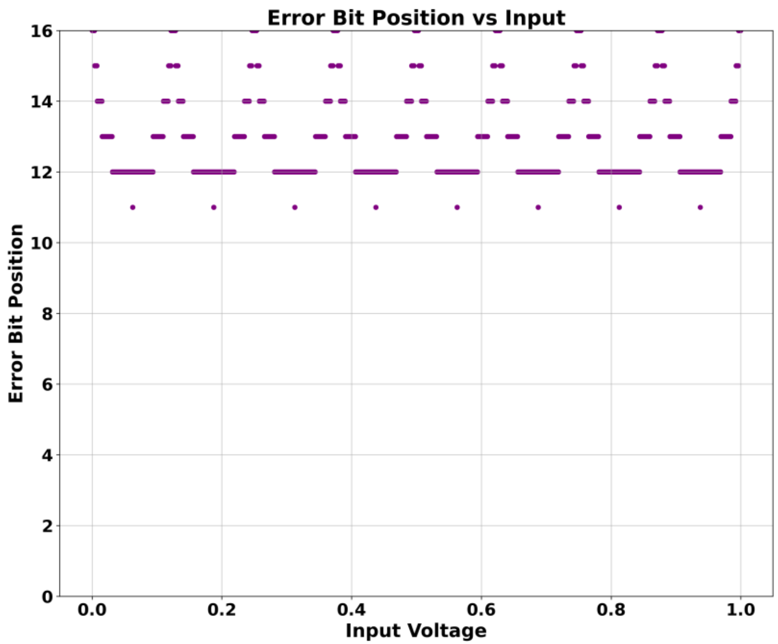
$$8V_s = 011.00011010 \quad \text{x VDD}$$
$$\lfloor 8V_s \rfloor = 011.00000000 \quad \text{x VDD}$$
$$V_{acc} = 000.00011010 \quad \text{x VDD}$$

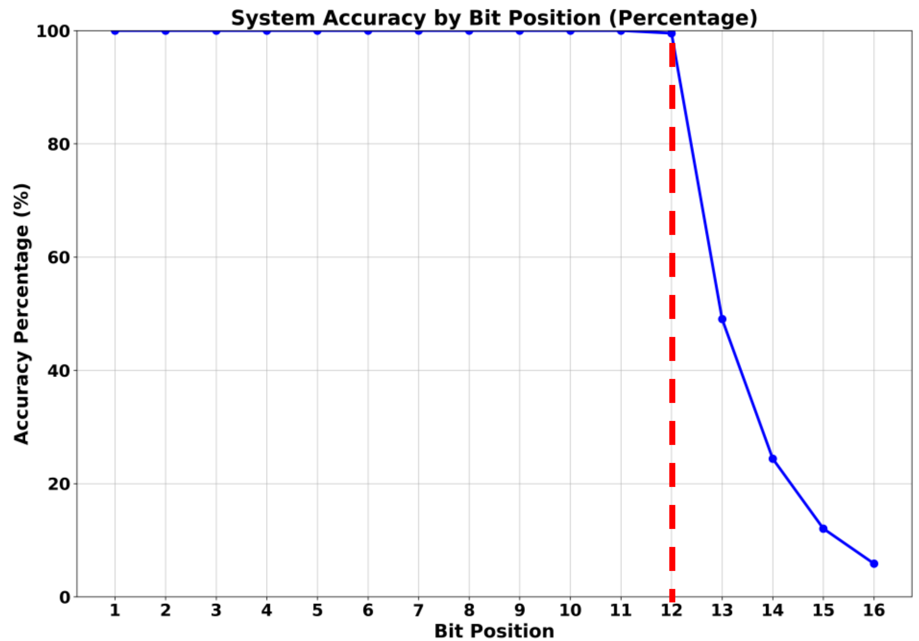




Quantizing MAC Voltage

- 3-bit coarse, 8-bit fine

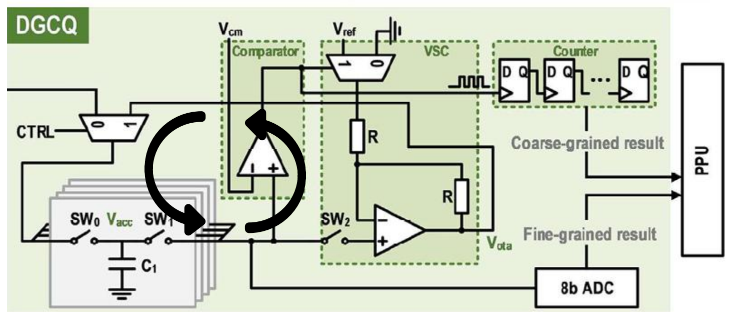


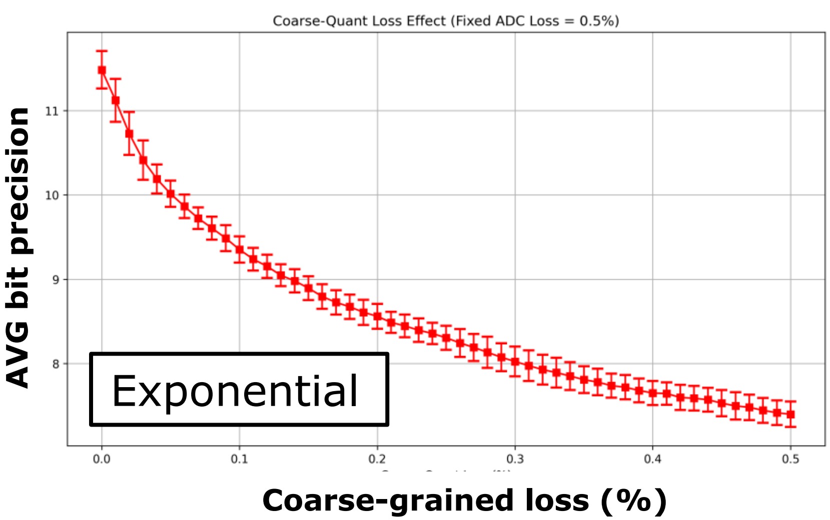


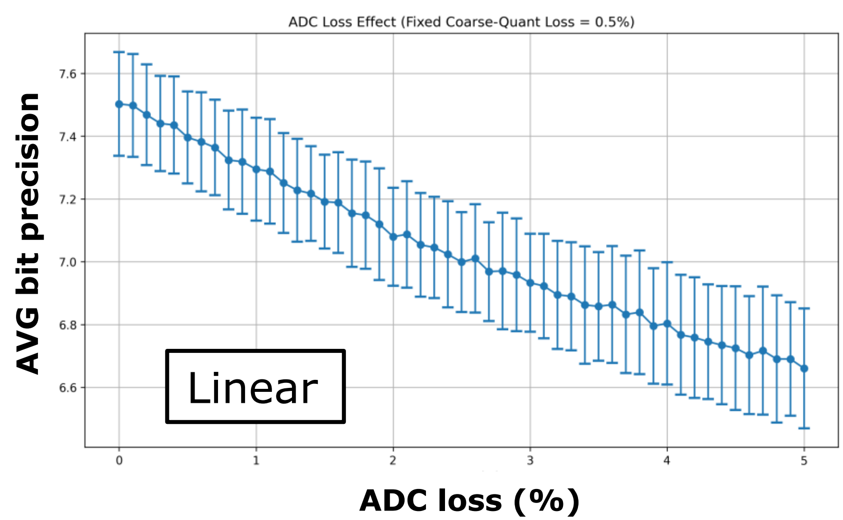
Bit precision can reach 12 bit in most situation (lossless)

Parasitic Effect

- Coarse-grained method is iterative
  - Loss in each iteration accumulates
  - ADC is also a source of noise

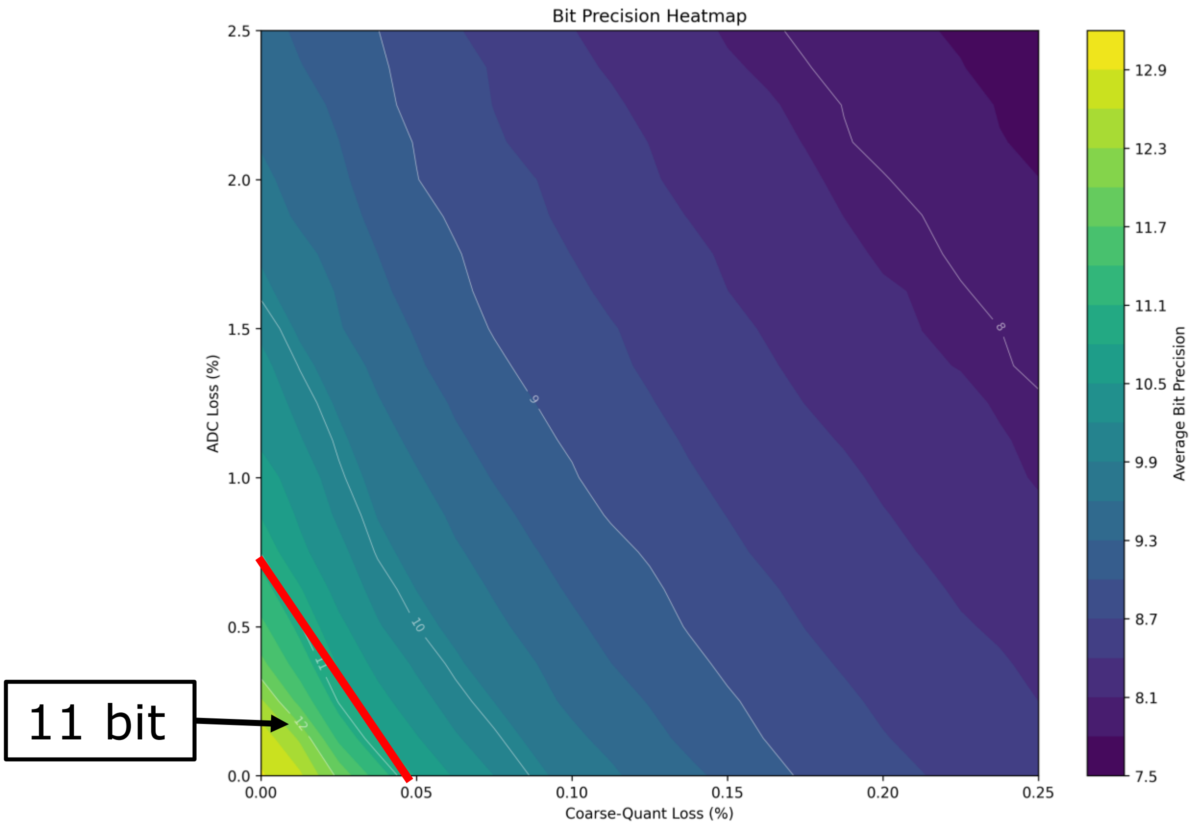






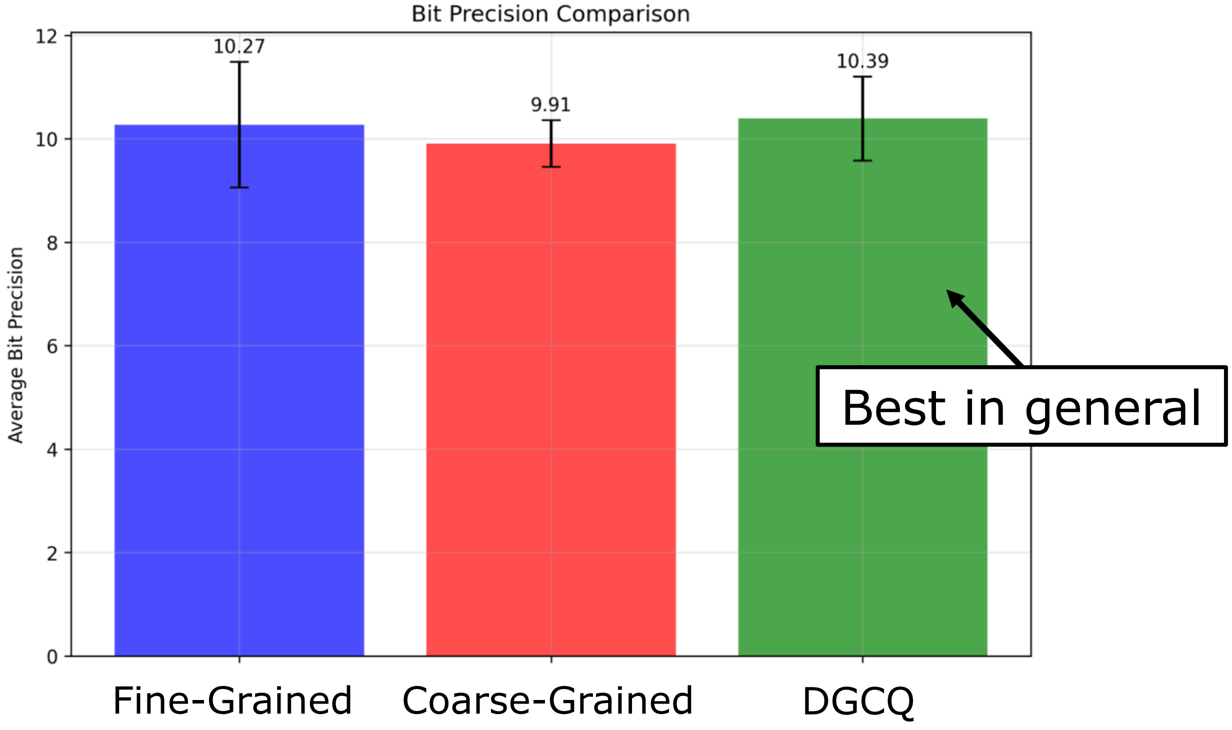
Parasitic Effect

- Coarse-grained method is iterative
  - Loss in each iteration accumulates
  - ADC is also a source of noise



Effect of DGCQ (Lossy Case)

- ADC : Large area, larger loss
- Coarse-grained : Time-consuming
- Dual-granularity : Less area, low loss, fast



Effect of DGCQ on ML Tasks

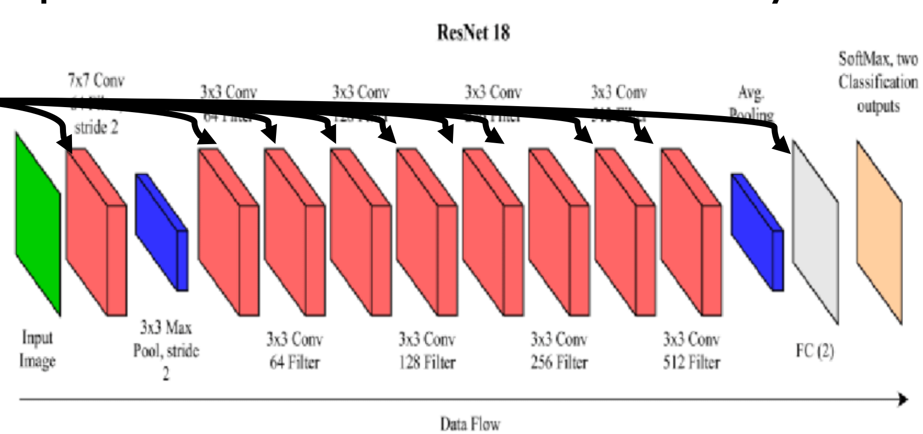
- The dataset used in the paper is ImageNet
  - ImageNet is not implemented in torch
  - Use a simpler dataset CIFAR-10

Model	ResNet-18
Dataset	ImageNet
Data precision	INT8
Task	Classification
Metric	Accuracy

- ResNet-18
  - Trained using fp32 precision
  - Quantize to int8 after training
  - Compute MAC result for each layer and apply loss

-LOSS

Gaussian(-Loss,0.01)



Effect of DGCQ on ML Tasks

- Comparison between coarse/fine-grained value

DGCQ : Merge 2 types of granularity