Memory Wall Free Architecure for Neural Network Computing
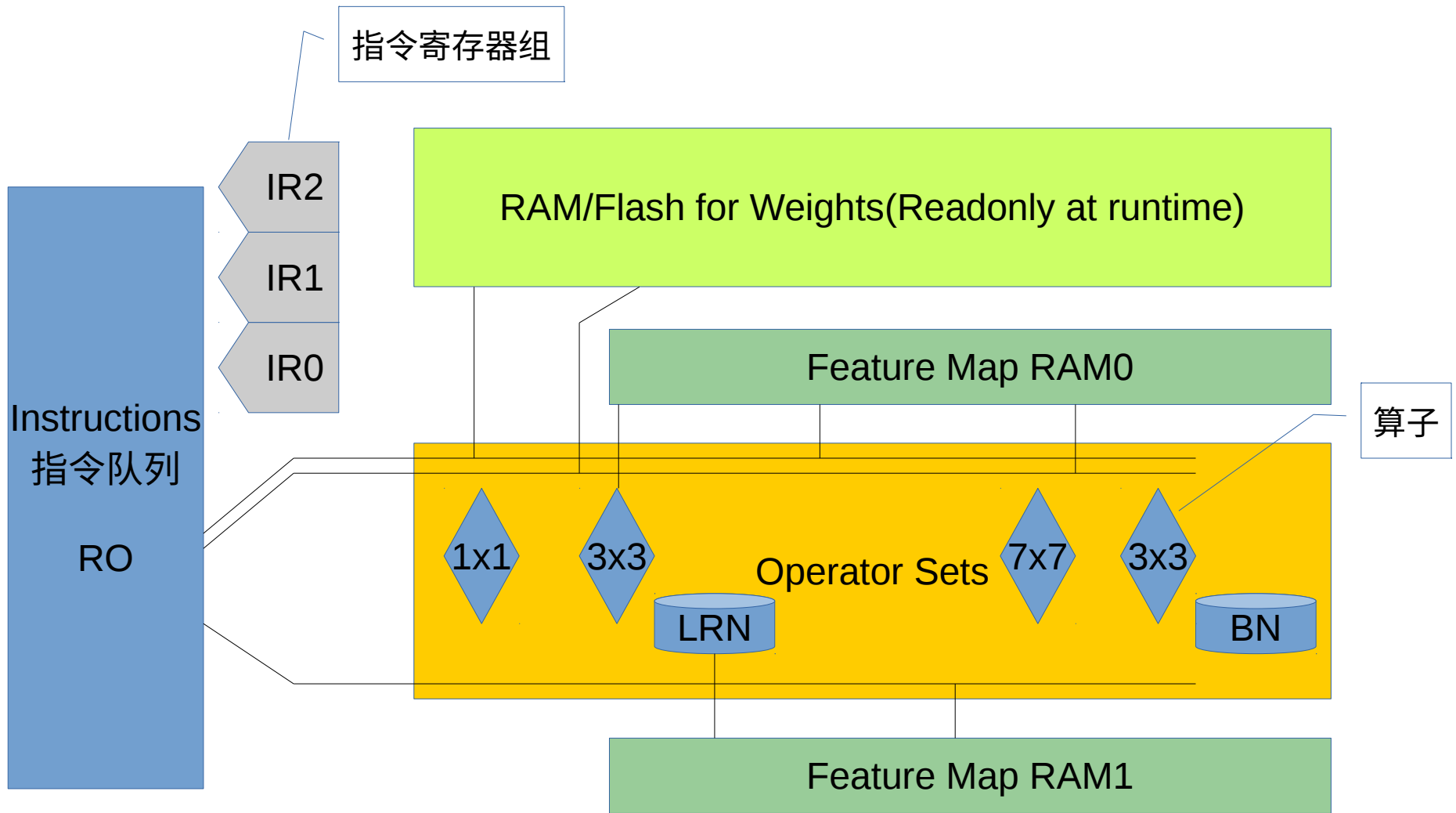
指令寄存器组

IR2

IR1

IR0

RAM/Flash for Weights(Readonly at runtime)

Feature Map RAM0

算子

Instructions
指令队列

RO

Operator Sets
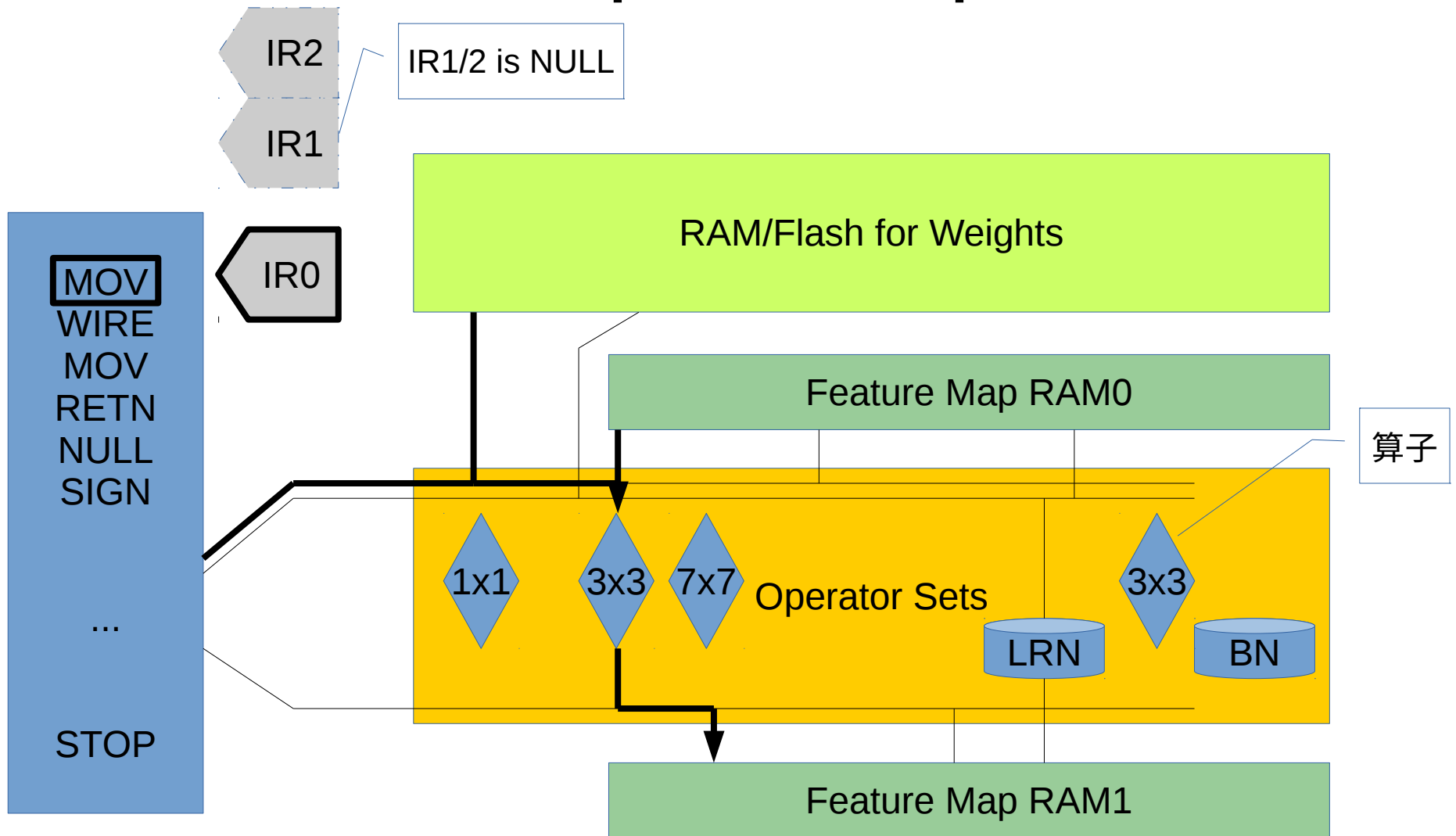
1x1 3x3 7x7 3x3

LRN BN

Feature Map RAM1

Graph Compute Architecture Compile

Caffe

PyTorch

Compiler

每一层权重多大

layer0	Layer1	?		

特征图大小及复用

MOV
WIRE
MOV
RETN
NULL
SIGN

...

STOP

数据流动时如何开关
经过权重特征和算子
的连线

Operator Sets

? ? ? ? ? ?

Layer7/9			

Graph Compute Architecture
Compute Step 1

IR2

IR1

IR1/2 is NULL

MOV
WIRE
MOV
RETN
NULL
SIGN

...

STOP

IR0

RAM/Flash for Weights

Feature Map RAM0

算子

Operator Sets

1x1 3x3 7x7 3x3

LRN BN

Feature Map RAM1
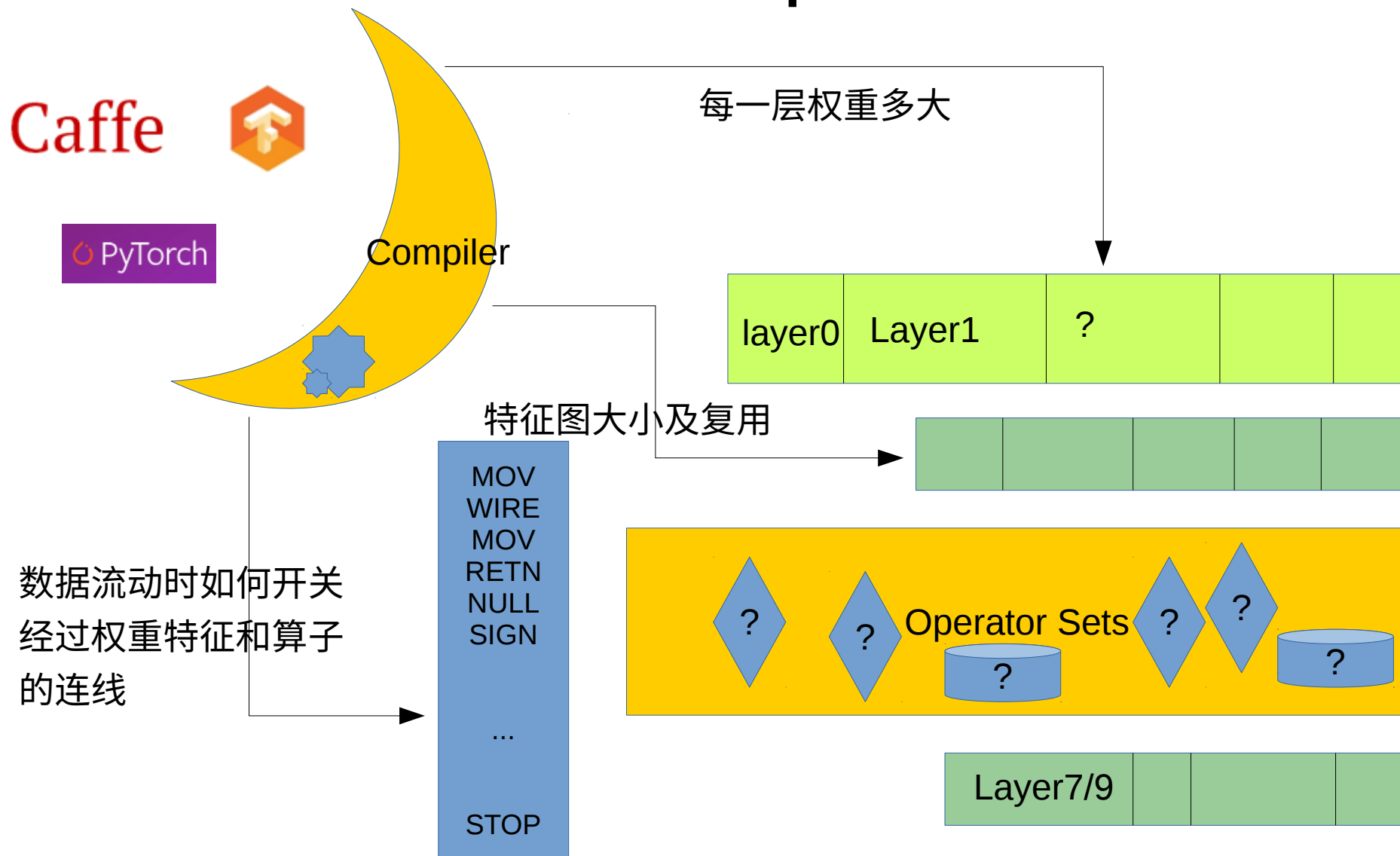
Graph Compute Architecture Computer Step 2

IR2

2 is NULL

2nd image

RAM/Flash for Weights

IR1

MOV
WIRE
MOV
RETN
NULL
SIGN

IR0

Feature Map RAM0

算子

...

STOP

Operator Sets

1x1 3x3 7x7

3x3

LRN BN

Feature Map RAM1
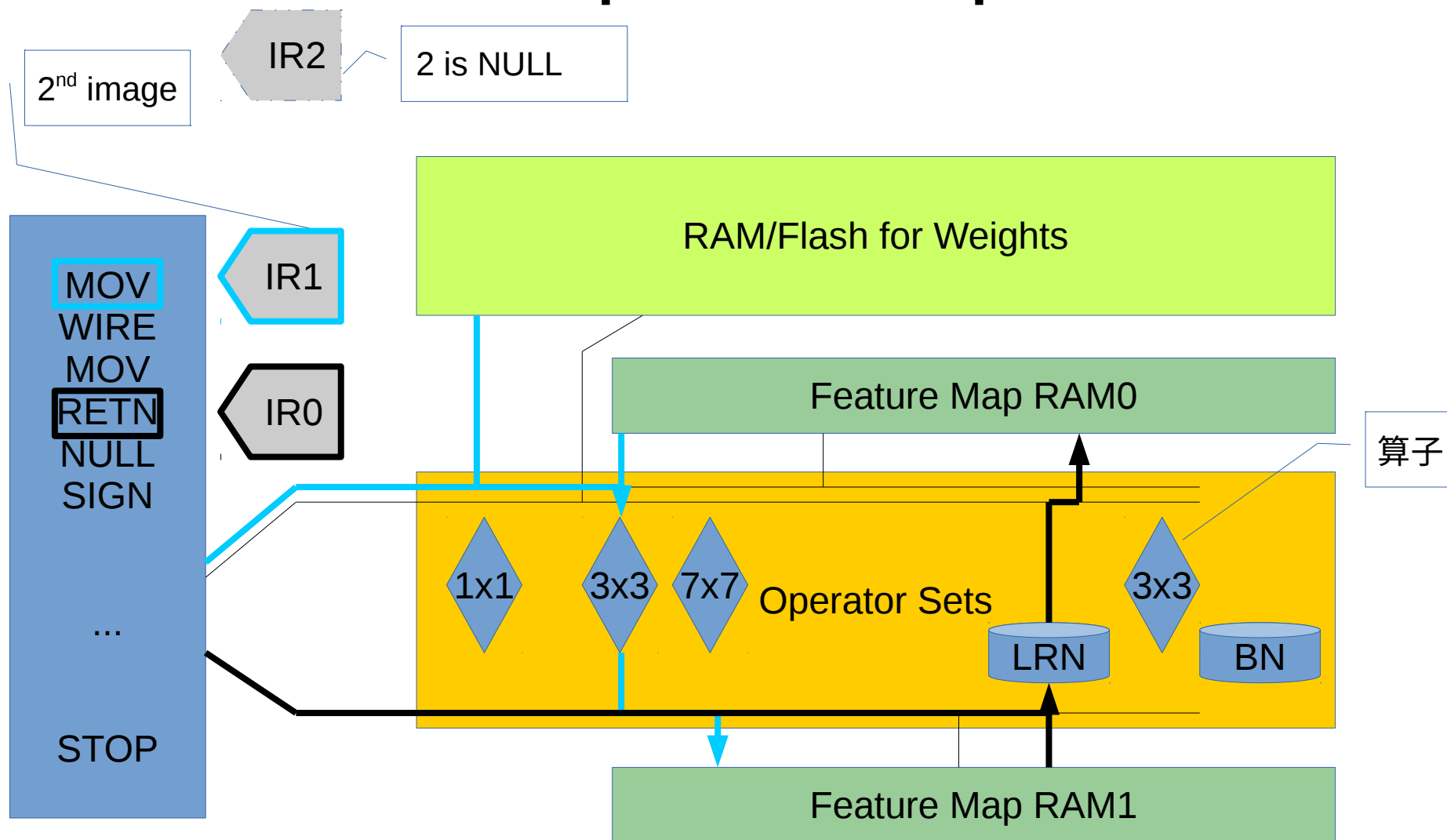
Graph Compute Architecture Low Power Consumption

指令寄存器组

IR2

IR1

IR0

Flash for Weights
voltage-variable conductances

算子

Instructions
指令队列

RO

1x1   3x3   7x7   3x3

LRN   BN
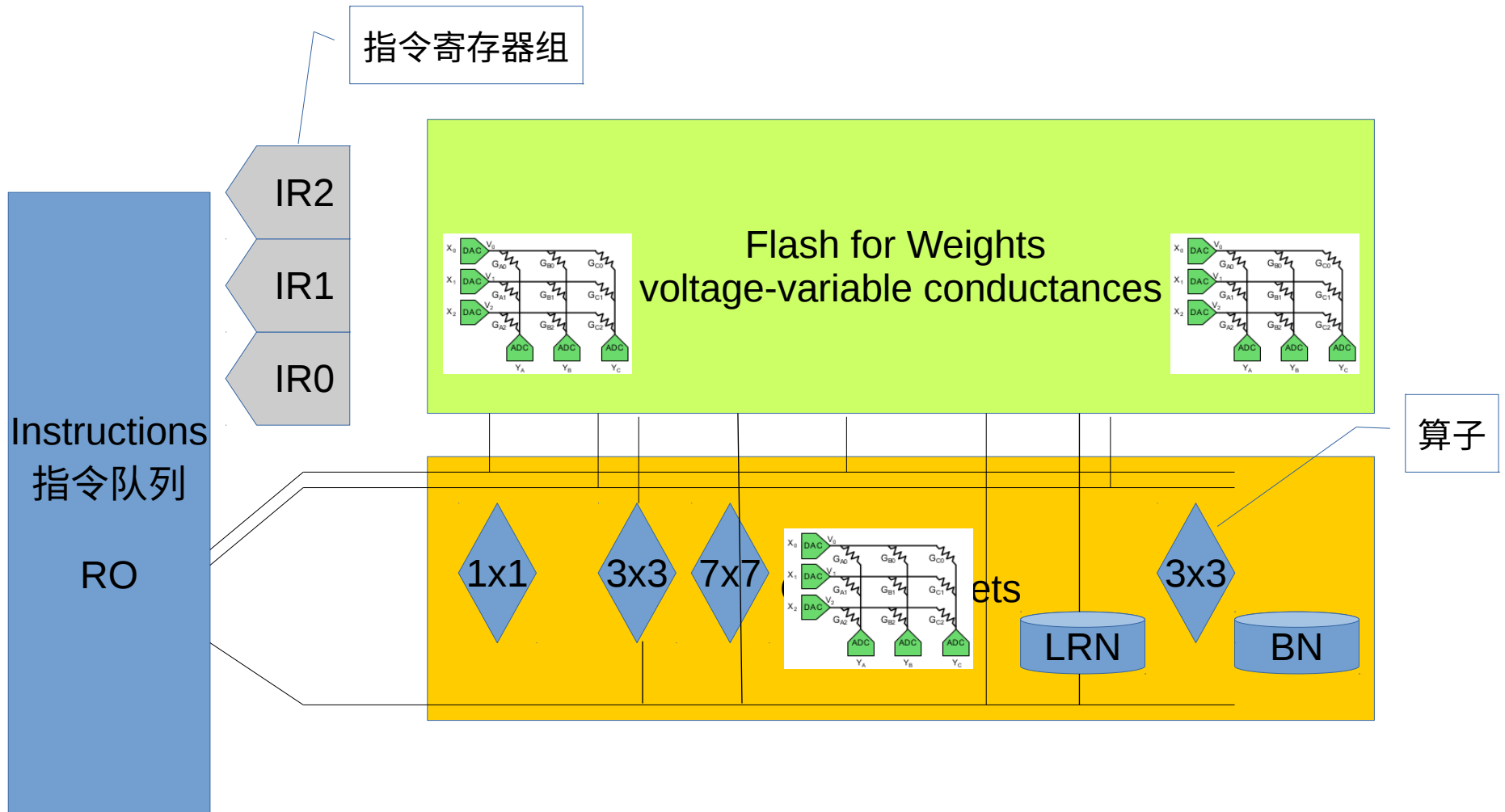
We focused on NLP
CTC-RNN and Attention LSTM

Cloud User Scenarion
Beep illegal word for live video

Edge User Scenarion 1
NLP for Office Device

Edge User Scenarion 2
NLP and Low Power for Drones

Edge User Scenarion 3
NLP for Room Device

We want do more
Word for cmd and print for Priority

This is the future
Robot can help US