ECG Classification Accelerator

Workflow

1. Algorithm
    1. Data
    2. Model
    3. Evaluation
    4. Quantization
    5. Evaluation again
    6. Fixed-point conversion
    7. Evaluation again
2. Hardware
    1. Submodule
        1.
        2. 16*16 Systolic Array

Model

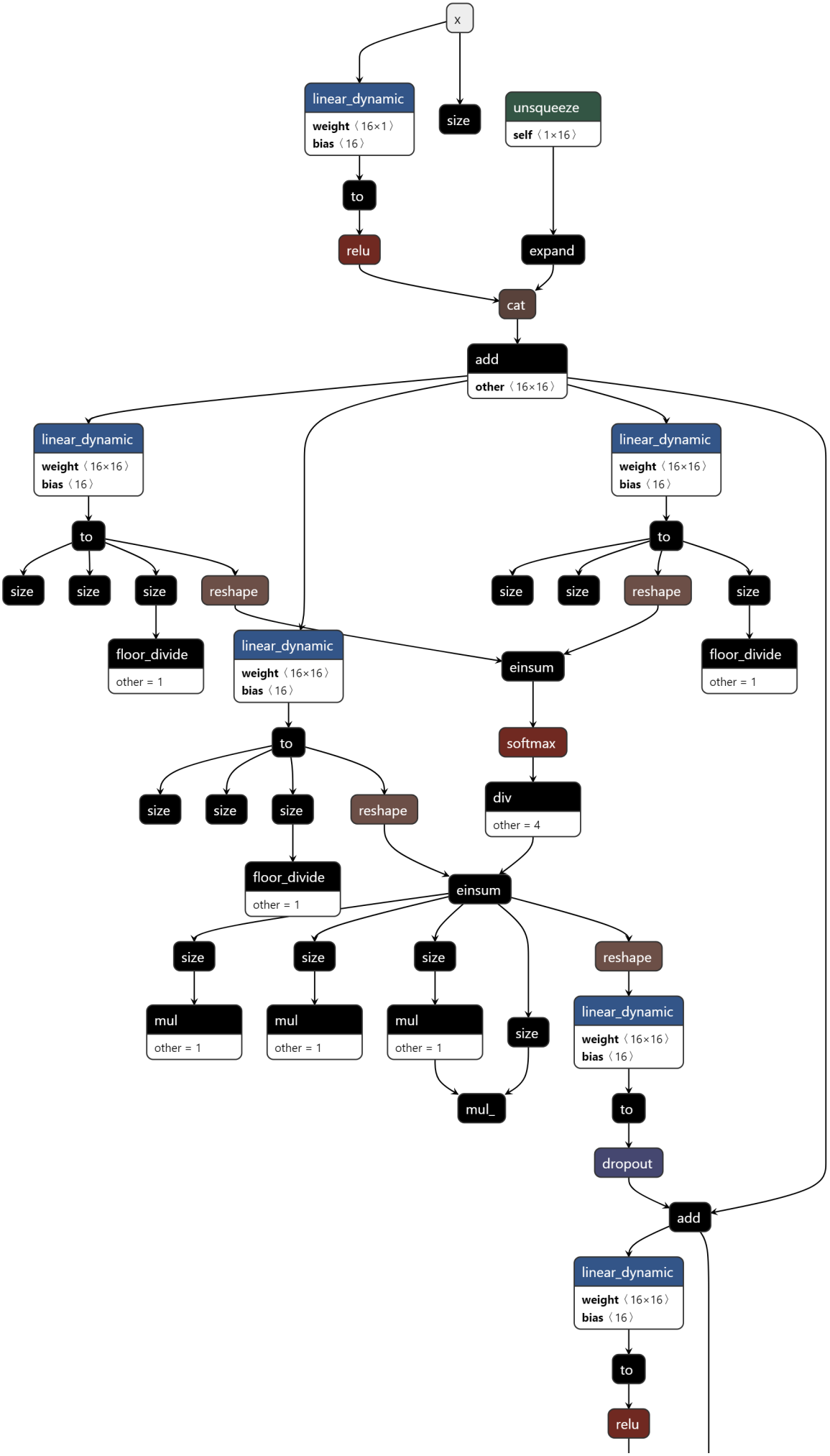
Data

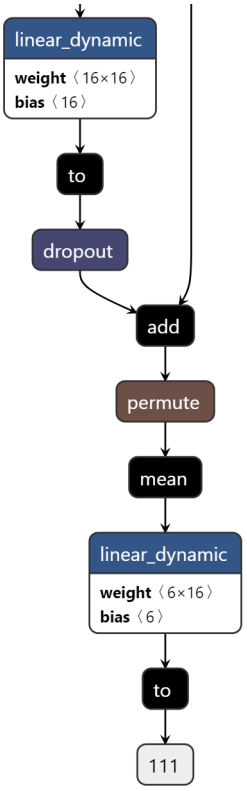
- Signal Length: First 15 samples
- Data: ECG Heartbeat Classification: A Deep Transferable Representation

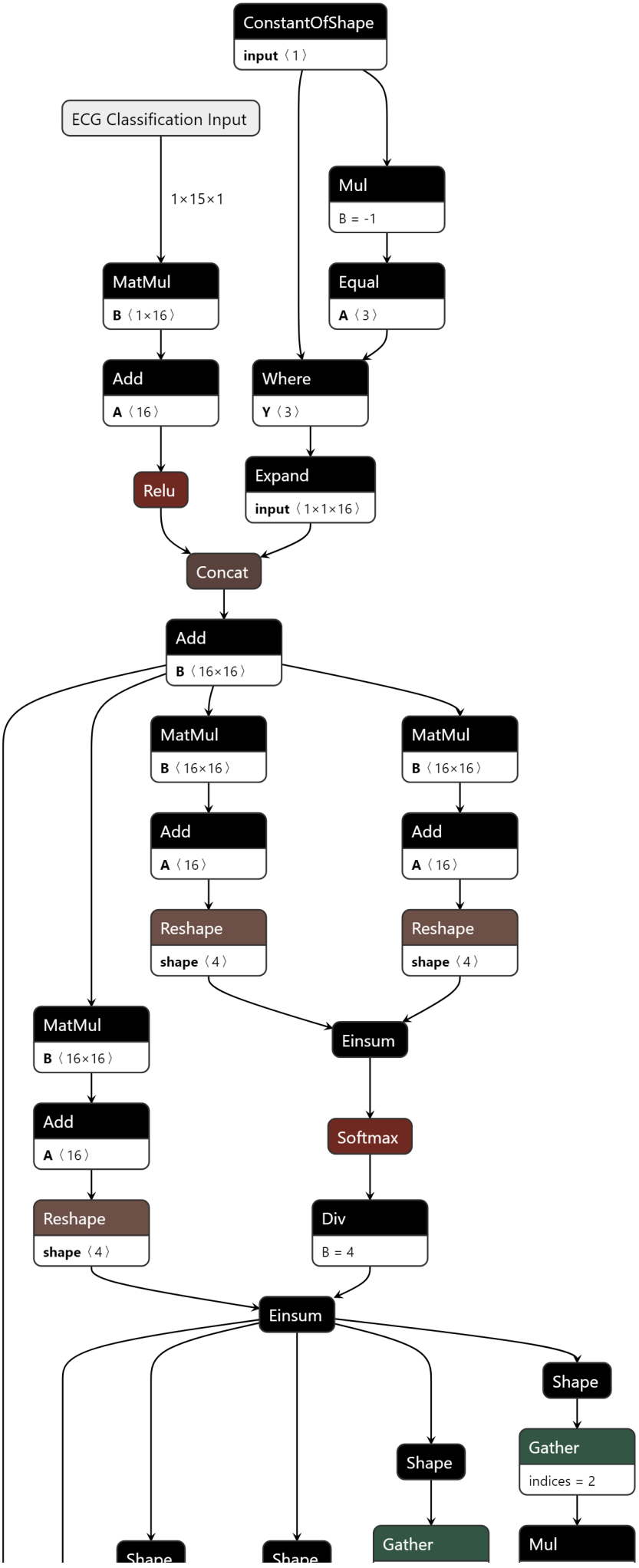
Label	Output
N	0
S	1
V	2
F	3
Q	4
MI	5

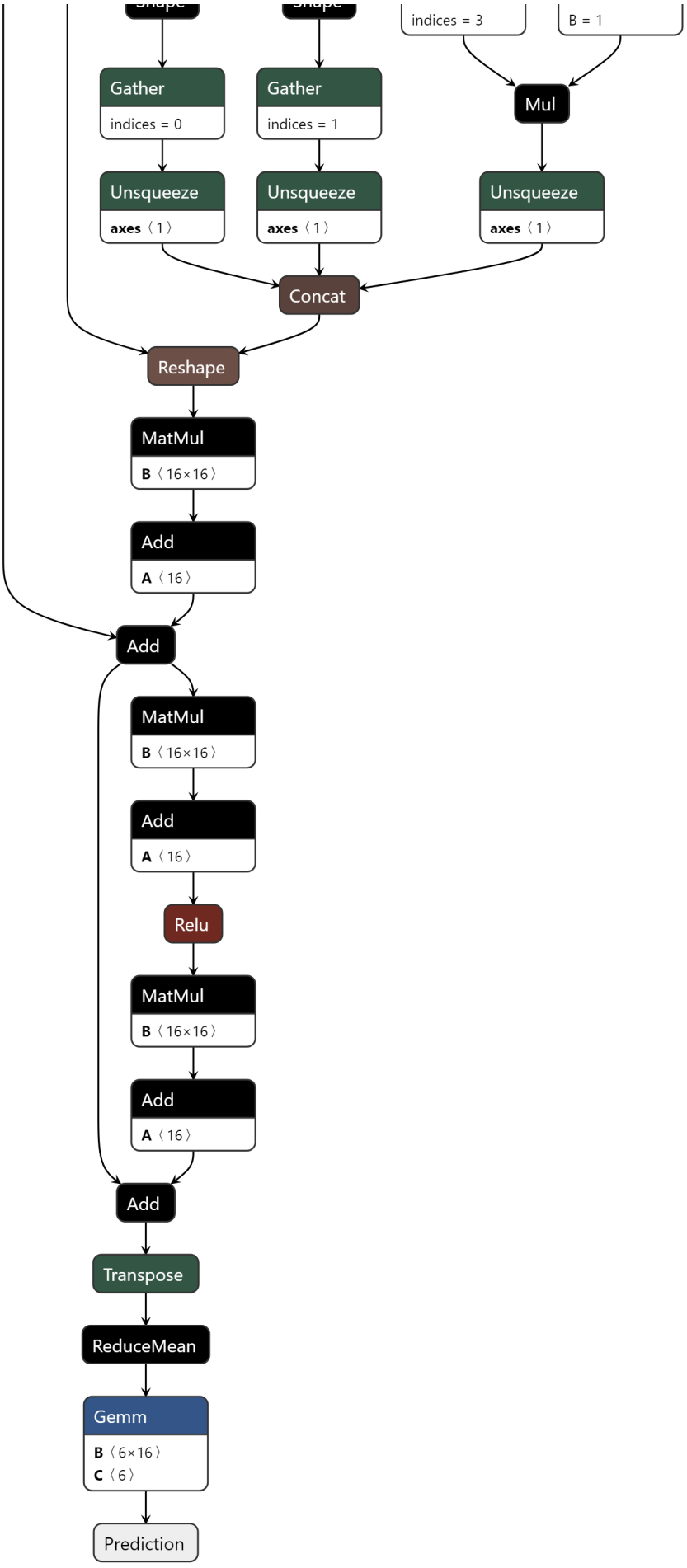
Visualization

Netron Jit traced


x
linear_dynamic
weight〈16×1〉
bias〈16〉
size
unsqueeze
self〈1×16〉
to
relu
expand
cat
add
other〈16×16〉
linear_dynamic
weight〈16×16〉
bias〈16〉
linear_dynamic
weight〈16×16〉
bias〈16〉
to
to
size
size
size
reshape
size
size
reshape
size
floor_divide
other = 1
linear_dynamic
weight〈16×16〉
bias〈16〉
einsum
floor_divide
other = 1
to
softmax
size
size
size
reshape
div
other = 4
floor_divide
other = 1
einsum
size
size
size
reshape
mul
other = 1
mul
other = 1
mul
other = 1
size
linear_dynamic
weight〈16×16〉
bias〈16〉
mul_
to
dropout
add
linear_dynamic
weight〈16×16〉
bias〈16〉
to
relu

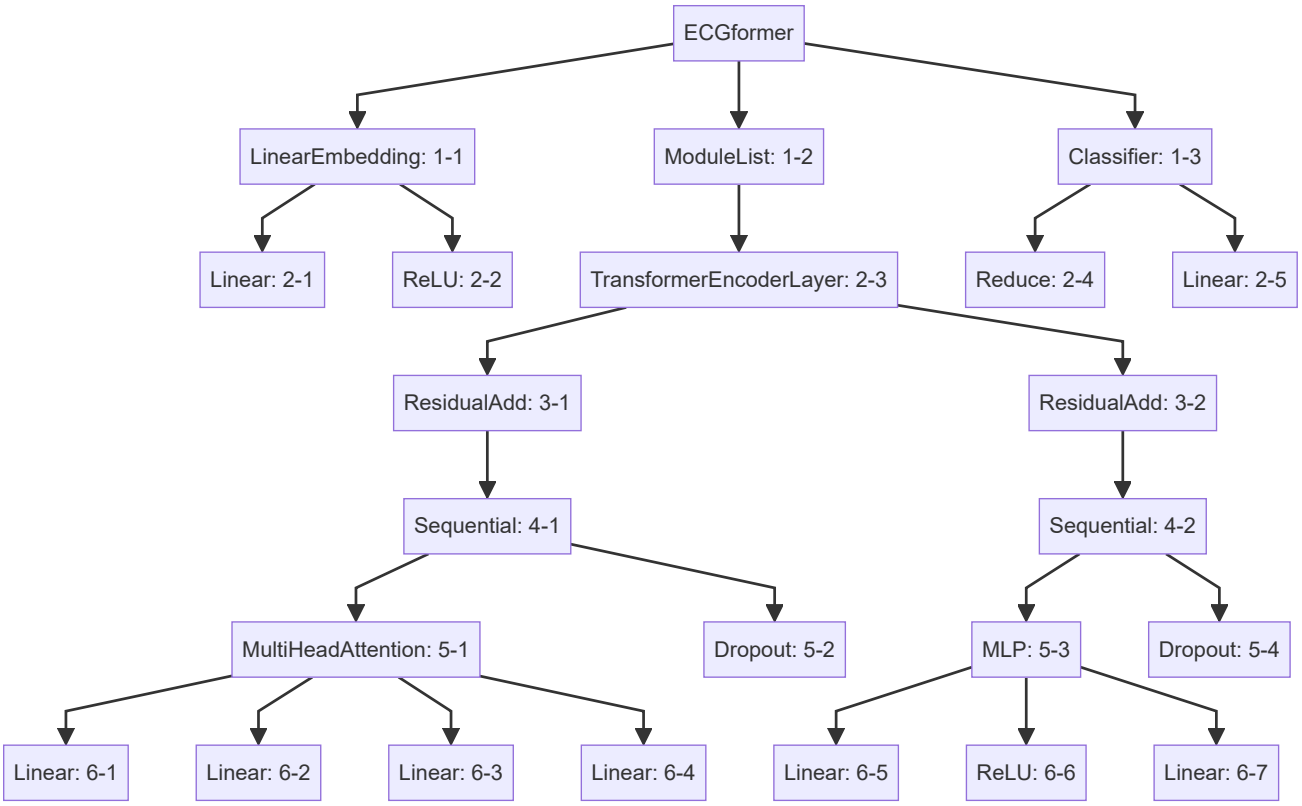

linear_dynamic
weight〈16×16〉
bias〈16〉
to
dropout
add
permute
mean
linear_dynamic
weight〈6×16〉
bias〈6〉
to
111


ConstantOfShape
input 〈1〉
ECG Classification Input
1×15×1
Mul
B = -1
MatMul
B 〈1×16〉
Equal
A 〈3〉
Add
A 〈16〉
Where
Y 〈3〉
Relu
Expand
input 〈1×1×16〉
Concat
Add
B 〈16×16〉
MatMul
B 〈16×16〉
MatMul
B 〈16×16〉
Add
A 〈16〉
Add
A 〈16〉
Reshape
shape 〈4〉
Reshape
shape 〈4〉
MatMul
B 〈16×16〉
Einsum
Add
A 〈16〉
Softmax
Reshape
shape 〈4〉
Div
B = 4
Einsum
Shape
Shape
Gather
indices = 2
Shape
Gather
Mul


indices = 3
B = 1
Gather
indices = 0
Gather
indices = 1
Mul
Unsqueeze
axes 〈1〉
Unsqueeze
axes 〈1〉
Unsqueeze
axes 〈1〉
Concat
Reshape
MatMul
B 〈16×16〉
Add
A 〈16〉
Add
MatMul
B 〈16×16〉
Add
A 〈16〉
Relu
MatMul
B 〈16×16〉
Add
A 〈16〉
Add
Transpose
ReduceMean
Gemm
B 〈6×16〉
C 〈6〉
Prediction

Torchinfo Summary[^fn1]

[fn1]: ignore drop out



Fixed Point

Layer/Parameter	Bit Width	Integer Part	Fraction Part
Linear: Scale	16-bit	1-bit	15-bit
Linear: Bias	16-bit	1-bit	15-bit
Linear: Weight	8-bit	8-bit	0-bit
cls_token	16-bit	4-bit	12-bit
positional_encoding	16-bit	4-bit	12-bit

Hardware