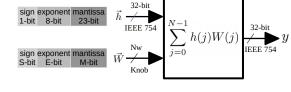# Trans-Precision Neural Network Deployment for Low-Power Embedded Systems

The methodology efficiently deploys and accelerates floating-point neural networks on embedded systems, optimizing performance, energy consumption, and hardware utilization.

Abstraction levels:

1. Model deployment

2. System infrastructure

3. Streamlined acceleration

4. **Optimized processing**

Multiply-Accumulate Unit
Hybrid Custom floating-point computation

sign	exponent	mantissa
1-bit	8-bit	23-bit

$\vec{h}$ 32-bit, IEEE 754

sign	exponent	mantissa
S-bit	E-bit	M-bit

$\vec{W}$ Nw, Knob

$$\sum_{j=0}^{N-1} h(j)W(j)$$

32-bit, IEEE 754

y