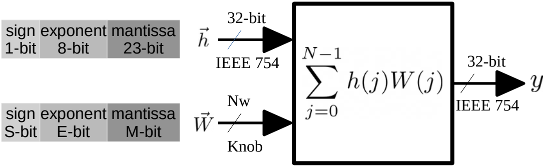
Trans-Precision Neural Network Deployment for Low-Power Embedded Systems

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

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

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