CLAIMS

1. A neural inference chip comprising a neural core, the neural core comprising:

a vector-matrix multiplier adapted to receive a weight matrix having a weight matrix precision, receive an input activation vector having an input activation vector precision, and compute a partial sum vector by multiplying the input activation vector by the weight matrix, the partial sum vector having a partial sum vector precision;

a vector processor adapted to receive one or more partial sum vector from one or more vector source, the one or more vector source including the vector-matrix multiplier, and perform one or more vector function on the one or more partial sum vector to yield a vector processor output vector, the vector processor output vector having a precision equal to the partial sum vector precision; and

an activation unit operatively coupled to the vector processor and adapted to apply an activation function to the vector processor output vector, yielding an output activation vector having an output activation precision, wherein the vector-matrix multiplier, vector processor, and/or activation unit is adapted to operate at variable precision,

wherein the vector-matrix multiplier is further adapted to vary the weight matrix precision and dimension and/or the input activation vector precision and dimension while maintaining constant bandwidth.

2. The neural inference chip of claim 1, further comprising:

at least one network interconnecting the neural core with at least one additional neural core, the at least one network adapted to deliver synaptic weights and/or input activations to the neural cores at variable precision.

3. The neural inference chip of claim 2, wherein the at least one network is further adapted to vary the weight matrix precision and dimension, input activation vector precision and dimension, and/or the output activation vector precision and dimension while maintaining constant bandwidth.

4. The neural inference chip of any of the preceding claims, wherein the neural core further comprises:

at least one memory, the at least one memory being adapted to store weight matrices, input activation vectors, and/or output activation vectors at variable precision.

5. The neural inference chip of claim 4, wherein the at least one memory is further adapted to vary the weight matrix precision and dimension, input activation vector precision and dimension, and/or the output activation vector precision and dimension while maintaining constant storage utilization.

6. The neural inference chip of claim 1, wherein the vector-matrix multiplier is further adapted to compute a variable number of multiplications per cycle at variable precision, wherein the variable number of multiplications per cycle and variable precision are inversely proportional.

7. The neural inference chip of any of the preceding claims, wherein the activation function is adapted to re-range the vector processor output vector.

8. The neural inference chip of claim 7, wherein applying the activation function comprises applying a saturating function.

9. The neural inference chip of claim 8, wherein the saturating function has as least one bound corresponding to the output activation precision.

10. The neural inference chip of claim 7, wherein applying the activation function comprises truncating one or more least significant bits.

11. The neural inference chip of any of the preceding claims, wherein the variable precision is selected from 2 bit, 4 bit, 8 bit, 16 bit, and 32 bit.

12. The neural inference chip of any of the preceding claims, wherein the variable precision is selectable at runtime.

13. The neural inference chip of any of the preceding claims, wherein the variable precision is selectable for each layer of a neural network.

14. The neural inference chip of any of the preceding claims, wherein the weight matrix precision is equal to the activation vector precision.

15. The neural inference chip of claim 14, wherein the partial sum vector precision is not equal to the output activation precision.

16. The neural inference chip of any of the preceding claims, wherein the partial sum vector precision is higher than the weight matrix precision and/or the activation vector precision.

17. The neural inference chip of claim 14, wherein the output activation precision is equal to the weight matrix precision.

18. A method of operating a neural inference chip, the neural inference chip comprising a neural core, the method comprising:

receiving, by a vector-matrix multiplier of the neural inference chip, a weight matrix having a weight matrix precision, receiving an input activation vector having an input activation vector precision, and computing, by a vector processor of the neural inference chip, a partial sum vector by multiplying the input activation vector by the weight matrix, the partial sum vector having a partial sum vector precision;

receiving one or more partial sum vector from one or more vector source, the one or more vector source including the vector-matrix multiplier, and performing one or more vector function on the one or more partial sum vector to yield a vector processor output vector, the vector processor output vector having a precision equal to the partial sum vector precision; and

applying, by an activation unit of the neural inference chip, an activation function to the vector processor output vector, yielding an output activation vector having an output activation precision, wherein the vector processor, and/or activation unit is adapted to operate at variable precision, wherein the vector-matrix multiplier is adapted to vary the weight matrix precision and dimension and/or the input activation vector precision and dimension while maintaining constant bandwidth.