

What is claimed is:

1. A neural network accelerator tile for exploiting input sparsity defining a set of weight lanes and a set of activation lanes, each weight lane corresponding to an activation lane, the tile comprising:

a weight memory to supply each weight lane of the set of weight lanes with a weight and a weight selection metadata;

an activation selection unit to receive a set of input activation values and rearrange the set of input activation values to supply each activation lane with a set of rearranged activation values;

a set of multiplexers, the set of multiplexers including at least one multiplexer per pair of activation and weight lanes, each multiplexer configured to select a combination activation value for the activation lane from the activation lane set of rearranged activation values based on the weight lane weight selection metadata; and

a set of combination units, the set of combination units including at least one combination unit per multiplexer, each combination unit configured to combine the activation lane combination value with the weight lane weight to output a weight lane product.
2. The tile of claim 1, further comprising an activation memory to supply the set of input activation values to the activation selection unit.
3. The tile of claim 1, wherein each multiplexer of the set of multiplexers is configured to select the combination activation value from the corresponding set of rearranged activation values and from a set of additional lane activation values, the set of additional lane activation values formed of at least one rearranged activation value of at least one additional activation lane.
4. The tile of claim 1, further comprising an adder tree to receive at least two eight lane products.

5. The tile of claim 1, wherein the weight lane weights of the set of weight lanes define at least one neural network filter.
6. The tile of claim 1, wherein the combination unit is one of a multiplier, an adder, and a shifter.
7. A neural network accelerator comprising at least two tiles of claim 1.
8. The tile of claim 1, wherein each set of rearranged activation values includes a standard weight activation value and at least one lookahead activation value.
9. The tile of claim 1, implemented on an activation efficiency exploiting accelerator structure.
10. The tile of claim 1, wherein the set of initial activation values are activation bits.
11. The tile of claim 1, wherein the set of initial activation values are signed powers of two.
12. The tile of claim 3, wherein the set of multiplexers is a set of multiplexers of a uniform size.
13. The tile of claim 12, wherein the uniform size is a power of two.
14. The tile of claim 13, wherein the size of the set of rearranged activation values is larger than the size of the set of additional lane activation values.
15. The tile of claim 12, wherein the set of rearranged activation values and the set of additional lane activation values ~~for-form~~ form a combined set of activation values, and the combined set of activation values contains 8 activations.
16. The tile of claim 3, wherein the set of additional lane activation values is formed of at least one rearranged activation value from each of at least two additional activation lanes.
17. The tile of claim 16, wherein the at least two additional activation lanes are non-contiguous activation lanes.

18. The tile of claim 1, wherein the tile is configured to receive the set of input activation values as at least one set of packed activation values stored bitwise to a required precision defined by a precision value, the tile configured to unpack the at least one set of packed activation values.
19. The tile of claim 18, wherein the at least one set of packed activation values includes a first set of packed activation values and a second set of packed activation values, the first set of packed activation values stored bitwise to a first required precision defined by a first precision value and the second set of packed activation values stored bitwise to a second required precision defined by a second precision value, the first precision value independent of the second precision value.
20. The tile of claim 18, wherein the tile is configured to receive a set of bit vectors including a bit vector corresponding to each set of packed activation values of the set of input activation values, the tile configured to unpack each set of packed activation values to insert zero values as indicated by the corresponding bit vector.
21. The tile of claim 1, wherein the tile is configured to receive the weight lane weights of the set of weight lanes as at least one set of packed weight lane weights stored bitwise to a required precision defined by a precision value, the tile configured to unpack the at least one set of weight lane weights.
22. The tile of claim 1, wherein the set of activation lanes is at least two sets of column activation lanes, each set of column activation lanes forming a column in which each activation lane corresponds to a weight lane, the tile further including at least one connection between at least two columns to transfer at least one weight lane product between the columns.
23. A system for bit-serial computation in a neural network, comprising:

one or more bit-serial tiles configured according to claim 1 for performing bit-serial computations in a neural network, each bit-serial tile receiving input neurons and synapses, the input neurons including at least one set of input activation values and the

synapses including at least one set of weights and at least one set of weight selection metadata, the one or more bit-serial tiles generating output neurons, each output neuron formed using at least one weight lane product;

an activation memory for storing neurons and in communication with the one or more bit-serial tiles via a dispatcher and a reducer,

wherein the dispatcher reads neurons from the activation memory and communicates the neurons to the one or more bit-serial tiles via a first interface,

and wherein the dispatcher reads synapses from a memory and communicates the synapses to the one or more bit-serial tiles via a second interface;

and wherein the reducer receives the output neurons from the one or more bit-serial tiles, and communicates the output neurons to the activation memory via a third interface;

and wherein one of the first interface and the second interface communicates the neurons or the synapses to the one or more bit-serial tiles bit-serially and the other of the first interface and the second interface communicates the neurons or the synapses to the one or more bit-serial tiles bit-parallelly.

24. A system for computation of layers in a neural network, comprising:

one or more tiles configured according to claim 1 for performing computations in a neural network, each tile receiving input neurons and synapses, the input neurons each including at least one offset, each offset including at least one activation value, and the synapses including at least one set of weights and at least one set of weight selection metadata, the one or more tiles generating output neurons, each output neuron formed using at least one weight lane product;

an activation memory for storing neurons and in communication with the one or more tiles via a dispatcher and an encoder,

wherein the dispatcher reads neurons from the activation memory and communicates the neurons to the one or more tiles, and wherein the dispatcher reads synapses from a memory and communicates the synapses to the one or more tiles,

and wherein the encoder receives the output neurons from the one or more tiles, encodes them and communicates the output neurons to the activation memory;

and wherein the offsets are processed by the tiles in order to perform computations on only non-zero neurons.

25. An accelerator tile, comprising:

an activation selection unit to receive a set of activation values and rearrange the set of activation values into at least one set of multiplexer input values;

a set of weight value receptors to receive at least one weight and at least one weight selection metadata;

at least one multiplexer to receive at least one of the at least one set of multiplexer input values and at least one weight selection metadata, the at least one multiplexer configured to apply the at least one weight selection metadata to select at least one combination activation value from the at least one set of multiplexer input values;

at least one combinator to apply the at least one combination activation value to the at least one weight to produce at least one product; and

at least one product output dispenser to output the at least one product.

26. A neural network accelerator comprising at least one of the tiles of claim 25.

27. The accelerator tile of claim 25, further including an activation memory to supply the set of activation values to the activation selection unit.

28. The accelerator tile of claim 25, wherein the at least one set of multiplexer input values is at least two sets of multiplexer input values and the at least one multiplexer is configured to receive at least one of the at least two sets of multiplexer input values and at least one activation value from at least one other set of multiplexer input values.
29. The accelerator tile of claim 25, wherein the combinator is at least one of a multiplier, an adder, and a shifter.
30. The accelerator tile of claim 25, wherein each set of multiplexer input values includes a standard activation value and at least one lookahead activation value.
31. The accelerator tile of claim 25, implemented on an activation efficiency exploiting accelerator structure.
32. The accelerator tile of claim 25, wherein the set of activation values are activation bits.
33. The accelerator tile of claim 25, wherein the set of activation values are signed powers of two.
34. The accelerator tile of claim 25, wherein the size of each multiplexer of the at least one multiplexer is a power of two.
35. The accelerator tile of claim 34, wherein the size of each multiplexer of the at least one multiplexer is 8.
36. Use of the neural network accelerator tile of claim 1 for training.
37. Use of the accelerator tile of claim 25 for training.
38. The tile of claim 1, wherein the weight lane weight selection metadata indexes a table that specifies a multiplexer select signal.
39. The accelerator tile of claim 25, wherein the weight selection metadata indexes a table that specifies a multiplexer select signal.