CLAIMS

1. A computing unit (200) for accelerating tensor computations of a neural network, the computing unit comprising:

a first memory bank (210) having a first data width for storing at least one of input activations or output activations;

a second memory bank (212) having a second data width that is larger than the first data width for storing one or more parameters used in performing computations;

a direct memory access traversal unit configured to:

access at least one memory location of the first memory bank to obtain at least one input activation, or at least one output activation, based on instructions received from a source that is external to the computing unit;

provide a control signal for writing the parameters to the second memory bank; and

provide a control signal for writing the input activations to the first memory bank;

at least one cell comprising at least one multiply accumulate, MAC, operator (215) that receives parameters from the second memory bank and performs computations;

a tensor traversal unit (226) in data communication with at least the first memory bank, the tensor traversal unit configured to provide a control signal to the first memory bank to cause an input activation to be provided to a data bus (216) accessible by the MAC operator; and

wherein the computing unit performs one or more computations associated with at least one element of a data array, the one or more computations being performed by the MAC operator and comprising, in part, a multiply operation of the input activation received from the data bus and a parameter received from the second memory bank.

2. The computing unit of claim 1, wherein the computing unit performs one or more of the computations by executing a loop nest comprising a plurality of loops, wherein a structure of the loop nest includes respective loops that are used by the tensor traversal unit to traverse one or more dimensions of the data array.

3. The computing unit of claim 2, wherein the one or more computations are performed based, in part, on a tensor operation provided by the tensor traversal unit, the tensor operation including a loop nest structure for accessing one or more elements of the data array.

4. The computing unit of claim 1, wherein the computing unit includes a non-linear unit and a first portion of the computations comprises producing one or more output activations based on the multiply operation and a second portion of the computations comprises applying, by the non-linear unit, a non-linear function to the one or more output activations.

5. The computing unit of claim 4, wherein the one or more computations performed by the computing unit comprises using a shift register to shift the output activations to the first memory bank.

6. The computing unit of claim 1, further comprising a portion of a ring bus that extends outside of the computing unit, wherein the ring bus provides a data path between the first memory bank and a memory bank of another adjacent computing unit, and between the second memory bank and a memory bank of another adjacent computing unit.

7. The computing unit of claim 1, wherein the second memory bank is configured to store at least one of partial sums or one or more pooling layer inputs.

8. A computer-implemented method for accelerating tensor computations of a neural network, the method comprising:

accessing, by a direct memory access traversal unit, at least one memory location of the first memory bank to obtain at least one input activation, or at least one output activation, based on instructions received from a source that is external to the computing unit;

providing, by the direct memory access traversal unit, a control signal for writing parameters to the second memory bank;

providing, by the direct memory access traversal unit, a control signal for writing input activations to the first memory bank;

sending, by the first memory bank having a first data width, a first input activation in response to the first memory bank receiving a control signal from a tensor traversal unit, the first memory bank being disposed in a computing unit, and wherein the first input activation is provided by a data bus that is accessible by at least one cell of the computing unit;

receiving, by the at least one cell, one or more parameters from a second memory bank having a second data width that is larger than the first data width, and wherein the at least one cell comprises at least one multiply accumulate, MAC, operator; and

performing, by the MAC operator, one or more computations associated with at least one element of a data array, wherein the one or more computations comprise, in part, a multiply operation of at least the first input activation accessed from the data bus and at least one parameter received from the second memory bank.

9. The computer-implemented method of claim 8, wherein the one or more computations are performed based, in part, on the computing unit executing a loop nest comprising a plurality of loops, wherein a structure of the loop nest includes respective loops that are used by the tensor traversal unit to traverse one or more dimensions of the data array.

10. The computer-implemented method of claim 9, further comprising, providing, by the tensor traversal unit, a tensor operation that includes a loop nest structure for accessing one or more elements of the data array.

11. The computer-implemented method of claim 8, further comprising, performing, a first portion of the one or more computations by producing at least one output activation based on the multiply operation.

12. The computer-implemented method of claim 11, further comprising, performing, a second portion of the one or more computations by applying a non-liner function to the one or more output activations.

13. A non-transitory computer-readable storage medium comprising instructions executable by one or more processors which, upon such execution, causes the one or more processors to perform operations for accelerating tensor computations of a neural network, the operations comprising:

accessing, by a direct memory access traversal unit, at least one memory location of the first memory bank to obtain at least one input activation, or at least one output activation, based on instructions received from a source that is external to the computing unit;

providing, by the direct memory access traversal unit, a control signal for writing the parameters to the second memory bank;

providing, by the direct memory access traversal unit, a control signal for writing input activations to the first memory bank;

sending, by the first memory bank having a first data width, a first input activation in response to the first memory bank receiving a control signal from a tensor traversal unit, the first memory bank being disposed in a computing unit, and wherein the first input activation is provided by a data bus that is accessible by at least one cell of the computing unit;

receiving, by the at least one cell, one or more parameters from the second memory bank having a second data width that is larger than the first data width, and wherein the at least one cell comprises at least one multiply accumulate, MAC, operator; and

performing, by the MAC operator, one or more computations associated with at least one element of a data array, wherein the one or more computations comprise, in part, a multiply operation of at least the first input activation accessed from the data bus and at least one parameter received from the second memory bank.

14. The non-transitory computer-readable medium of claim 13, wherein the one or more computations are performed based, in part, on the computing unit executing a loop nest comprising a plurality of loops, wherein a structure of the loop nest includes respective loops that are used by the tensor traversal unit to traverse one or more dimensions of the data array.

15. The non-transitory computer-readable medium of claim 14, further comprising, providing, by the tensor traversal unit, a tensor operation that includes a loop nest structure for accessing one or more elements of the data array.

16. The non-transitory computer-readable medium of claim 13, further comprising, performing, a first portion of the one or more computations by producing at least one output activation based on the multiply operation.

17. The non-transitory computer-readable medium of claim 16, further comprising, performing, a second portion of the one or more computations by applying a non-liner function to the one or more output activations.