CLAIMS

1. A system comprising: multiple sets of hardware computing units for accelerating inference computations for a plurality of layers of a neural network, wherein each set of hardware computing units comprises:

a first computing unit configured to:

receive instructions for performing inference computations for a first layer of the plurality of layers of the neural network, layer inputs for the first layer, and a respective set of weights for the first layer; and

perform at least a subset of the inference computations for the first layer based on execution of a first loop nest to access the layer inputs for the first layer and the respective set of weights for the first layer; and

a second computing unit configured to:

receive instructions for performing inference computations for a second layer of the plurality of layers of the neural network, layer inputs for the second layer, and the respective set of weights for the second layer; and

perform at least a subset of the inference computations for the second layer based on execution of a second loop nest to access the layer inputs for the second layer and the respective set of weights for the second layer.

2. The system of claim 1, wherein execution of the first loop nest comprises: executing, by a processor of the first computing unit, a plurality of nested loops included in the first loop nest; and accessing, memory of the first computing unit, to retrieve data corresponding to elements of a tensor, wherein the data includes at least one of: the layer inputs for the first layer or weights for the first layer.

3. The system of claim 2, wherein the first computing unit comprises: at least one traversal unit that uses the first loop nest to access the elements of the tensor; wherein a structure of the first loop nest indicates a manner in which the at least one traversal unit traverses dimensions of the tensor.

4. The system of claim 1, wherein execution of the second loop nest comprises: executing, by a processor of the second computing unit, a plurality of nested loops included in the second loop nest; accessing, memory of the second computing unit, to retrieve data corresponding to elements of a tensor, wherein the data includes at least one of: the layer inputs for the second layer or weights for the second layer.

5. The system of claim 4, wherein the second computing unit comprises: at least one traversal unit that uses the second loop nest to access particular elements of the tensor; wherein a structure of the second loop nest indicates a manner in which the at least one traversal unit traverses dimensions of the tensor.

6. The system of claim 1, wherein: the first layer is a neural network layer of a first layer type; and the second layer is a neural network layer of a second layer type that is different than the first layer type.

7. The system of claim 1, further comprising: a data communications instruction bus configured to: receive one or more instructions from an external source; provide, to the first computing unit, the instructions for performing the subset of inference computations for the first layer; and provide, to the second computing unit, the instructions for performing the subset of inference computations for the second layer.

8. The system of claim 7, further comprising: a data communications ring bus configured to: receive multiple inputs and multiple weights from an external source; provide, to the first computing unit, the layer inputs for the first layer, and the respective set of weights for the first layer; and provide, to the second computing unit, the layer inputs for the second layer, and the respective set of weights for the second layer.

9. A method of accelerating inference computations for a plurality of layers of a neural network using a system comprising multiple sets of hardware computing units, the method comprising:

receiving, by a first computing unit, instructions for performing inference computations for a first layer of the plurality of layers of the neural network, layer inputs for the first layer, and a respective set of weights for the first layer;

performing, by the first computing unit, at least a subset of the inference computations for the first layer based on execution of a first loop nest to access the layer inputs for the first layer and the respective set of weights for the first layer;

receiving, by a second computing unit, instructions for performing inference computations for a second layer of the plurality of layers of the neural network, layer inputs for the second layer, and a respective set of weights for the second layer; and

performing, by the second computing unit, at least a subset of the inference computations for the second layer based on execution of a second loop nest to access the layer inputs for the second layer and the respective set of weights for the second layer.

10. The method of claim 9, wherein execution of the first loop nest comprises: executing, by a processor of the first computing unit, a plurality of nested loops included in the first loop nest; and accessing, memory of the first computing unit, to retrieve data corresponding to elements of a tensor, wherein the data includes at least one of: the layer inputs for the first layer or weights for the first layer.

11. The method of claim 10, wherein the first computing unit comprises: at least one traversal unit that uses the first loop nest to access the elements of the tensor; wherein a structure of the first loop nest indicates a manner in which the at least one

traversal unit traverses one or more dimensions of the tensor.

12. The method of claim 9, wherein execution of the second loop nest comprises: executing, by a processor of the second computing unit, a plurality of nested loops included in the second loop nest; accessing, memory of the second computing unit, to retrieve data corresponding to elements of a tensor, wherein the data includes at least one of: the layer inputs for the second layer or weights for the second layer.

13. The method of claim 12, wherein the second computing unit comprises: at least one traversal unit that uses the second loop nest to access particular elements of the tensor; wherein a structure of the second loop nest indicates a manner in which the at least one traversal unit traverses dimensions of the tensor.

14. The method of claim 9, wherein: the first layer is a neural network layer of a first layer type; and the second layer is a neural network layer of a second layer type that is different than the first layer type.

15. The method of claim 9, further comprising: receiving, at an instruction bus, one or more instructions from an external source, wherein the instruction bus is configured to provide data communications to the multiple hardware computing units; providing, to the first computing unit and by the instruction bus, the instructions for performing the subset of inference computations for the first layer; and providing, to the second computing unit and by the instruction bus, the instructions for performing the subset of inference computations for the second layer.

16. The method of claim 15, further comprising: receiving, at a ring bus, multiple inputs and multiple weights from an external source, wherein the ring bus is configured to provide data communications to the multiple hardware computing units; providing, to the first computing unit and by the ring bus, the layer inputs for the first layer, and the respective set of weights for the first layer; and providing, to the second computing unit and by the ring bus, the layer inputs for the first layer, and the respective set of weights for the second layer.

17. A non-transitory computer-readable storage medium having instructions stored thereon, which, when executed by a system comprising multiple sets of hardware computing units, cause the system to perform operations comprising:

receiving, by a first computing unit, instructions for performing inference computations for a first layer of a plurality of layers of the neural network, layer inputs for the first layer, and a respective set of weights for the first layer;

performing, by the first computing unit, at least a subset of the inference computations for the first layer based on execution of a first loop nest to access the layer inputs for the first layer and the respective set of weights for the first layer;

receiving, by a second computing unit, instructions for performing inference computations for a second layer of the plurality of layers of the neural network, layer inputs for the second layer, and a respective set of weights for the second layer; and

performing, by the second computing unit, at least a subset of the inference computations for the second layer based on execution of a second loop nest to access the layer inputs for the second layer and the respective set of weights for the second layer.

18. The computer-readable storage medium of claim 17, wherein execution of the first loop nest comprises: executing, by a processor of the first computing unit, a plurality of nested loops included in the first loop nest; and accessing, memory of the first computing unit, to retrieve data corresponding to elements of a tensor, wherein the data includes at least one of: the layer inputs for the first layer or weights for the first layer.

19. The computer-readable storage medium of claim 18, wherein the first computing unit comprises: at least one traversal unit that uses the first loop nest to access the elements of the tensor; wherein a structure of the first loop nest indicates a manner in which the at least one traversal unit traverses dimensions of the tensor.

20. The computer-readable storage medium of claim 17, wherein execution of the second loop nest comprises: executing, by a processor of the second computing unit, a plurality of nested loops included in the second loop nest; accessing, memory of the second computing unit, to retrieve data corresponding to elements of a tensor, wherein the data includes at least one of: the layer inputs for the second layer or weights for the second layer.