CLAIMS

What is claimed is:

1. A neural network processor system comprising:

a neural processing unit comprising a plurality of neural processing cores;

a router network comprising a plurality of routers communicatively coupled to the plurality of neural processing cores, respectively; and

a host processing unit communicatively coupled to the neural processing unit based on the router network and configured to coordinate the neural processing unit for performing neural network computations, wherein

each neural processing core of the plurality of neural processing cores comprises:

a control register block configured to receive and store partial sum configuration information from the host processing unit; and

a partial sum interface communicatively coupled to the control register block and configured to transmit a first partial sum neural packet generated by the neural processing core to a first another neural processing core of the plurality of neural processing cores and/or receive a second partial sum neural packet generated by a second another neural processing core of the plurality of neural processing cores, based on the partial sum configuration information stored in the neural processing core, and

a first set of neural processing cores of the plurality of neural processing cores are combinable based on the partial sum configuration information respectively stored therein to form a first neurosynaptic column chain, the first neurosynaptic column chain comprising:

a very first neural processing core of the plurality of neural processing cores, the very first neural processing core being a partial sum transmitter;

a very last neural processing core of the plurality of neural processing cores, the very first neural processing core being a partial sum receiver; and

one or more intermediate neural processing cores of the plurality of neural processing cores, each of the one or more intermediate neural processing cores being a partial sum transceiver configured to receive an incoming partial sum neural packet from a predecessor neural processing core of the first neurosynaptic column chain, add the partial sum configuration information stored in the intermediate neural processing core to the incoming partial sum neural packet received from the predecessor neural processing core to produce an accumulated partial sum, and

generate an outgoing partial sum neural packet comprising the accumulated partial sum to a next neural processing core of the first neurosynaptic column chain.

2. The neural network processor system according to claim 1, wherein the partial sum configuration information respectively stored in the first set of neural processing cores are collectively configured to combine the first set of neural processing cores to form the first neurosynaptic column chain.

3. The neural network processor system according to claim 1, wherein for each neural processing core of the first neurosynaptic column chain except the very last neural processing core thereof, the partial sum interface of the neural processing core is configured to transmit the first partial sum neural packet generated to the first another neural processing core of the first neurosynaptic column chain based on relative core addressing information included in the partial sum configuration information stored in the neural processing core.

4. The neural network processor system according to claim 3, wherein

the relative core addressing information of the partial sum configuration information comprises directional data indicating a direction relative to the neural processing core at which the first another neural processing core of the first neurosynaptic column chain is located, and

the first another neural processing core is immediately succeeding the neural processing core in the first neurosynaptic column chain.

5. The neural network processor system according to claim 4, wherein

the first partial sum neural packet generated by the neural processing core comprises an operation field comprising operation data indicating that the first partial sum neural packet is a partial sum neural packet, a payload field comprising partial sum data computed by the neural processing core and a destination field comprising destination data corresponding to the directional data stored in the neural processing core, and/or

the second partial sum neural packet generated by the second another neural processing core comprises an operation field comprising operation data indicating that the second partial sum neural packet is a partial sum neural packet, a payload field comprising partial sum data computed by the second another neural processing core and a destination

field comprising destination data corresponding to the directional data stored in the second another neural processing core.

6. The neural network processor system according to claim 1, wherein the first set of neural processing cores is configured to perform partial summations in parallel.

7. The neural network processor system according to claim 1, wherein

for said each neural processing core of the plurality of neural processing cores,

the control register block of the neural processing core is further configured to receive and store axon input retransmission configuration information from the host processing unit; and

the neural processing core further comprises an axon input retransmission interface communicatively coupled to the control register block and configured to transmit a duplicate axon input neural packet generated by the neural processing core to another neural processing core of the plurality of neural processing cores, based on the axon input retransmission configuration information stored in the neural processing core, and

a second set of neural processing cores of the plurality of neural processing cores are combinable based on the axon input retransmission configuration information respectively stored therein to form a first neurosynaptic row chain.

8. The neural network processor system according to claim 7, wherein the axon input retransmission configuration information respectively stored in the second set of neural processing cores are collectively configured to combine the second set of neural processing cores to form the first neurosynaptic row chain.

9. The neural network processor system according to claim 7, wherein for each neural processing core of the first neurosynaptic row chain except a last neural processing core thereof, the axon input retransmission interface of the neural processing core is configured to transmit the duplicate axon input neural packet generated to said another neural processing core of the first neurosynaptic row chain based on relative core addressing information included in the axon input retransmission configuration information stored in the neural processing core.

10. The neural network processor system according to claim 9, wherein

the relative core addressing information of the axon input retransmission configuration information comprises directional data corresponding to a direction relative to the neural processing core at which said another neural processing core of the first neurosynaptic row chain is located, and

said another neural processing core is immediately succeeding the neural processing core in the first neurosynaptic row chain.

11. The neural network processor system according to claim 10, wherein the duplicate axon input neural packet comprises an operation field comprising operation data indicating that the duplicate axon input neural packet is an axon input neural packet, a payload field comprising duplicated axon input row data of an axon input neural packet received by the neural processing core and a destination field comprising destination data corresponding to the directional data stored in the neural processing core.

12. The neural network processor system according to claim 1, wherein

the neural processing unit further comprises a neural cache block,

the router network further comprises a router communicatively coupled to the neural cache block,

the host processing unit is further communicatively coupled to the neural cache block based on the router network,

the neural cache block comprises:

a control register block configured to receive and store axon input retransmission configuration information from the host processing unit; and

an axon input retransmission configuration interface configured to transmit a duplicate axon input neural packet generated by the neural cache block to another one or more neural processing cores of the plurality of neural processing cores based on the axon input retransmission configuration information stored in the neural cache block, and

a second set of neural processing cores of the plurality of neural processing cores are combinable based on the axon input retransmission configuration information stored in the neural cache block to form a first neurosynaptic row chain.

13. The neural network processor system according to claim 7, wherein

for said each neural processing core of the plurality of neural processing cores,

the control register block of the neural processing core is further configured to receive and store core truncation configuration information from the host processing unit; and

the neural processing core further comprises a core truncator communicatively coupled to the control register block and configured to modify a neurosynaptic row count and/or a neurosynaptic column count of the neural processing core based on the core truncation configuration information stored in the neural processing core.

14. The neural network processor system according to claim 7, wherein

the first neurosynaptic column chain is symmetrical or asymmetrical, and

the first neurosynaptic row chain is symmetrical or asymmetrical.

15. The neural network processor system according to claim 7, wherein a third set of neural processing cores of the plurality of neural processing cores are combinable based on the partial sum configuration information respectively stored in the first set of neural processing cores and the axon input retransmission configuration information respectively stored in the second set of neural processing cores to form a first combined neural processing core comprising the first set of neural processing cores forming the first neurosynaptic column chain and the second set of neural processing cores forming the first neurosynaptic row chain.

16. The neural network processor system according to claim 15, wherein the first combined neural processing core further comprises one or more first additional sets of neural processing cores forming one or more additional neurosynaptic column chains and one or more second additional sets of neural processing cores forming one or more additional neurosynaptic row chains.

17. A method of operating a neural network processor system,

the neural network processor system comprising:

a neural processing unit comprising a plurality of neural processing cores;

a router network comprising a plurality of routers communicatively coupled to the plurality of neural processing cores, respectively; and

a host processing unit communicatively coupled to the neural processing unit based on the router network and configured to coordinate the neural processing unit for performing neural network computations, wherein

each neural processing core of the plurality of neural processing cores comprises:

a control register block configured to receive and store partial sum configuration information from the host processing unit; and

a partial sum interface communicatively coupled to the control register block and configured to transmit a first partial sum neural packet generated by the neural processing core to a first another neural processing core of the plurality of neural processing cores and/or receive a second partial sum neural packet generated by a second another neural processing core of the plurality of neural processing cores, based on the partial sum configuration information stored in the neural processing core, and

a first set of neural processing cores of the plurality of neural processing cores are combinable based on the partial sum configuration information respectively stored therein to form a first neurosynaptic column chain, the first neurosynaptic column chain comprising:

a very first neural processing core of the plurality of neural processing cores, the very first neural processing core being a partial sum transmitter;

a very last neural processing core of the plurality of neural processing cores, the very first neural processing core being a partial sum receiver; and

one or more intermediate neural processing cores of the plurality of neural processing cores, each of the one or more intermediate neural processing cores being a partial sum transceiver configured to receive an incoming partial sum neural packet from a predecessor neural processing core of the first neurosynaptic column chain, add the partial sum configuration information stored in the intermediate neural processing core to the incoming partial sum neural packet received from the predecessor neural processing core to produce an accumulated partial sum, and generate an outgoing partial sum neural packet comprising the accumulated partial sum to a next neural processing core of the first neurosynaptic column chain, and

the method comprising:

executing, by the host processing unit, one or more neural network applications;

assigning, by the host processing unit, one or more neural network operations associated with the one or more neural network applications to the neural processing unit, comprising transmitting the respective partial sum configuration information to the control register block of each neural processing core of the first set of neural processing cores for combining the first set of neural processing cores to form the first neurosynaptic column chain; and

performing, by the neural processing unit, neural network computations to obtain computation results in relation to the one or more neural network operations.

18. The method according to claim 17, wherein

said assigning, by the host processing unit, the one or more neural network operations associated with the one or more neural network applications to the neural processing unit, further comprising transmitting axon input retransmission configuration information to the neural processing unit for combining a second set of neural processing cores to form a first neurosynaptic row chain.

19. A method of forming a neural network processor system, the method comprising:

providing a neural processing unit comprising a plurality of neural processing cores;

providing a router network comprising a plurality of routers communicatively coupled to the plurality of neural processing cores, respectively; and

providing a host processing unit communicatively coupled to the neural processing unit based on the router network and configured to coordinate the neural processing unit for performing neural network computations, wherein

each neural processing core of the plurality of neural processing cores comprises:

a control register block configured to receive and store partial sum configuration information from the host processing unit; and

a partial sum interface communicatively coupled to the control register block and configured to transmit a first partial sum neural packet generated by the neural processing core to a first another neural processing core of the plurality of neural processing cores and/or receive a second partial sum neural packet generated by a second another neural

processing core of the plurality of neural processing cores, based on the partial sum configuration information stored in the neural processing core, and

a first set of neural processing cores of the plurality of neural processing cores are combinable based on the partial sum configuration information respectively stored therein to form a first neurosynaptic column chain, the first neurosynaptic column chain comprising:

a very first neural processing core of the plurality of neural processing cores, the very first neural processing core being a partial sum transmitter;

a very last neural processing core of the plurality of neural processing cores, the very first neural processing core being a partial sum receiver; and

one or more intermediate neural processing cores of the plurality of neural processing cores, each of the one or more intermediate neural processing cores being a partial sum transceiver configured to receive an incoming partial sum neural packet from a predecessor neural processing core of the first neurosynaptic column chain, add the partial sum configuration information stored in the intermediate neural processing core to the incoming partial sum neural packet received from the predecessor neural processing core to produce an accumulated partial sum, and generate an outgoing partial sum neural packet comprising the accumulated partial sum to a next neural processing core of the first neurosynaptic column chain.

20. The method of according to claim 19, wherein

for said each neural processing core of the plurality of neural processing cores,

the control register block of the neural processing core is further configured to receive and store axon input retransmission configuration information from the host processing unit; and

the neural processing core further comprises an axon input retransmission interface communicatively coupled to the control register block and configured to duplicate axon input row data of an axon input neural packet received by the neural processing core to generate a duplicate axon input neural packet and to transmit the duplicate axon input neural packet to another neural processing core of the plurality of neural processing cores, based on the axon input retransmission configuration information stored in the neural processing core, and

a second set of neural processing cores of the plurality of neural processing cores are combinable based on the axon input retransmission configuration information respectively stored therein to form a first neurosynaptic row chain.