

CLAIMS

1. A neural inference chip comprising:
 - a first tier comprising a plurality of neural cores, each core comprising:
 - a neural computation unit, the neural computation unit adapted to apply a plurality of synaptic weights to a plurality of input activations to produce a plurality of output activations;
 - a second tier comprising a first neural network model memory adapted to store the plurality of synaptic weights;
 - a communication network operatively coupled to the first neural network model memory and to each of the plurality of neural cores, and adapted to provide the synaptic weights from the first neural network model memory to each of the plurality of neural cores.
2. The neural inference chip of claim 1, wherein the communication network comprises a plurality of through-silicon vias.
3. The neural inference chip of claim 1, further comprising:
 - at least one additional tier comprising at least one additional neural network model memory, wherein the communication network is additionally operatively coupled to the at least one additional neural network model memory and adapted to provide synaptic weights from the at least one additional neural network model memory to each of the plurality of neural cores.
4. The neural inference chip of claim 3, wherein a neural network model is stored across the first neural network model memory and the at least one additional neural network model memory.
5. The neural inference chip of claim 3, wherein a plurality of neural network models are stored across the first neural network model memory and the at least one additional neural network model memory.
6. The neural inference chip of claim 1, wherein each core further comprises:
 - an activation memory adapted to store the input activations and the output activations;
 - a local controller, the local controller being adapted to load the input activations from the activation memory to the neural computation unit and to store the plurality of output activations from the neural computation unit to the activation memory.
7. The neural inference chip of claim 1, further comprising:
 - a third tier comprising an activation memory, wherein the communication network is additionally operatively coupled to the activation memory and adapted to provide activations from the activation memory to each of the plurality of neural cores.

8. The neural inference chip of claim 1, further comprising:
a third tier comprising an activation memory, wherein
an additional communication network is operatively coupled to the activation memory and adapted to provide activations from the activation memory to each of the plurality of neural cores.
9. The neural inference chip of claim 1, further comprising:
a third tier comprising a plurality of neural cores, wherein
the communication network is operatively coupled to the third tier and adapted to provide the synaptic weights from the first neural network model memory to each of the plurality of neural cores of the third tier.
10. The neural inference chip of claim 9, configured to provide a first neural network model to both the first and third tiers.
11. The neural inference chip of claim 9, configured to provide different neural network models to each of the first and third tiers.
12. The neural inference chip of claim 1, wherein the communication network has at least two dimensions, a first of the at least two dimensions extending between tiers of the neural inference chip.
13. The neural inference chip of claim 1, wherein the communication network has at least three dimensions, a first of the at least three dimensions extending between tiers of the neural inference chip and a second of the at least three dimensions extending within a tier of the neural inference chip.
14. The neural inference chip of claim 1, wherein the communication network is adapted to provide the same synaptic weights to each of the cores.
15. The neural inference chip of claim 1, wherein the communication network is adapted to provide the same synaptic weights to a subset of the cores.
16. The neural inference chip of claim 1, wherein the communication network is configured to provide a dedicated bus for each of the cores.
17. The neural inference chip of claim 1, wherein the communication network comprises a plurality of rows with one of the tiers, each connected to a subset of the plurality of cores across tiers, and wherein the network is adapted to provide the same synaptic weights to those cores connected to each of the plurality of rows.
18. A method comprising:

providing synaptic weights from a first neural network model memory to each of a plurality of neural cores via a communication network,

the communication network being operatively coupled to the first neural network model memory and to each of the plurality of neural cores,

the plurality of neural cores being arrayed on a first tier of a neural inference chip,

each core comprising a neural computation unit, the neural computation unit adapted to apply a plurality of synaptic weights to a plurality of input activations to produce a plurality of output activations,

the first neural network model memory being arrayed on a second tier of a neural inference chip.

19. A computer program product for neural inference processing, the computer program product comprising a computer readable storage medium having program instructions embodied therewith, the program instructions executable by a neural inference chip to cause the neural inference chip to perform a method comprising:

providing synaptic weights from a first neural network model memory to each of a plurality of neural cores via a communication network,

the communication network being operatively coupled to the first neural network model memory and to each of the plurality of neural cores,

the plurality of neural cores being arrayed on a first tier of a neural inference chip,

each core comprising a neural computation unit, the neural computation unit adapted to apply a plurality of synaptic weights to a plurality of input activations to produce a plurality of output activations,

the first neural network model memory being arrayed on a second tier of a neural inference chip.