CLAIMS

1. A neural inference chip wherein the layout of the chip uses silicon 3D die and wafer stacking including a plurality of silicon tiers, comprising:

a first silicon 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 silicon 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, the communication network being configured to provide a dedicated bus for each neural core of the plurality of neural cores,

each dedicated bus being separate from the plurality of neural cores,

each dedicated bus being configured to provide the synaptic weights from the first neural network model memory to each neural core of the plurality of neural cores such that the synaptic weights are provided directly from the second silicon tier to the first silicon tier;

wherein the communication network comprises a plurality of through-silicon vias providing connectivity between the first silicon tier and the second silicon tier, and wherein the communication network has at least three dimensions, a first of the at least three dimensions extending between silicon tiers of the neural inference chip, and a second of the at least three dimensions extending within a silicon tier of the neural inference chip.

2. The neural inference chip of claim 1, further comprising:

at least one additional silicon 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.

3. The neural inference chip of claim 2, wherein a neural network model is stored across the first neural network model memory and the at least one additional neural network model memory.

4. The neural inference chip of claim 2, 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.

5. 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.

6. The neural inference chip of claim 1, further comprising:

a third silicon 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.

7. The neural inference chip of claim 1, further comprising:

a third silicon 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.

8. The neural inference chip of claim 1, further comprising:

a third silicon tier comprising a plurality of neural cores, wherein

the communication network is operatively coupled to the third silicon 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 silicon tier.

9. The neural inference chip of claim 8, configured to provide a first neural network model to both the first and third silicon tiers.

10. The neural inference chip of claim 8, configured to provide different neural network models to each of the first and third silicon tiers.

11. The neural inference chip of claim 1, wherein the communication network is adapted to provide the same synaptic weights to each of the cores.

12. 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.