

CLAIMS

1. A neural inference chip for computing neural activations, the neural inference chip comprising:
 - at least one neural core;
 - a memory array operatively coupled to the at least one neural core, the memory array comprising a plurality of elements, each element comprising a memory and a horizontal buffer, the horizontal buffer of each element of the memory array being in communication with either the horizontal buffer of another of the elements of the memory array or to the at least one neural core;
 - an instruction buffer in communication with the memory array, the instruction buffer having a position corresponding to each of the plurality of elements of the memory array;
 - an instruction memory in communication with the instruction buffer, wherein
 - the instruction memory is adapted to provide at least one instruction to the instruction buffer,
 - the instruction buffer is adapted to advance the at least one instruction between positions in the instruction buffer,
 - the instruction buffer is adapted to provide the at least one instruction to at least one of the plurality of elements of the memory array from its associated position in the instruction buffer when the memory of the at least one of the plurality of elements contains data associated with the at least one instruction,
 - each of the plurality of elements of the memory array is adapted to provide a data block from its memory to its horizontal buffer in response to the arrival of an associated instruction from the instruction buffer,
 - the horizontal buffer of each element of the memory array is adapted to provide a data block to the horizontal buffer of another of the elements of the memory array or to the at least one neural core.
2. The neural inference chip of claim 1, wherein:
 - the instruction buffer is adapted to advance instructions between positions in the instruction buffer at a rate of one position per cycle,
 - the horizontal buffer of each element of the memory array is adapted to provide a data block to the horizontal buffer of another of the elements of the memory array or to the at least one neural core at a rate of one data block per cycle.
3. The neural inference chip of claim 1, comprising an array of neural cores, the array of neural cores comprising the at least one neural core and having a plurality of rows.
4. The neural inference chip of claim 1, wherein the memory array is one-dimensional, the plurality of elements of the memory array being arranged in one row and a plurality of columns.

5. The neural inference chip of claim 1, wherein the memory array is two-dimensional, the plurality of elements of the memory array being arranged in a plurality of rows and a plurality of columns.
6. The neural inference chip of claim 5, wherein each element of the memory array further comprises a vertical buffer, the vertical buffer of each element of the memory array being in communication with the vertical buffer of another element of the memory array.
7. The neural inference chip of claim 6, wherein:
 - each of the plurality of elements of the memory array is adapted to provide a data block from its memory to its vertical buffer in response to the arrival of an associated instruction from the instruction buffer,
 - each of the plurality of elements of the memory array is adapted to provide the data block from its vertical buffer to its horizontal buffer,
 - the vertical buffer of each element of the memory array is adapted to provide the data block to the vertical buffer of another of the elements of the memory array.
8. The neural inference chip of claim 7, wherein:
 - the horizontal buffer of each element of the memory array is adapted to provide a data block to the horizontal buffer of another of the elements of the memory array or to the at least one neural core at a rate of one data block per cycle,
 - the vertical buffer of each element of the memory array is adapted to provide a data block to the vertical buffer of another of the elements of the memory array at a rate of one data block per cycle.
9. The neural inference chip of claim 6, wherein each element of the memory array further comprises a layover buffer, the layover buffer of each element of the memory array being in communication with the horizontal buffer and the vertical buffer of that element of the memory array.
10. The neural inference chip of claim 9, wherein:
 - each of the plurality of elements of the memory array is adapted to provide a data block from its memory to its vertical buffer in response to the arrival of an associated instruction from the instruction buffer,
 - each of the plurality of elements of the memory array is adapted to provide the data block from its vertical buffer to its layover buffer,
 - each of the plurality of elements of the memory array is adapted to provide the data block from its layover buffer to its horizontal buffer,
 - the vertical buffer of each element of the memory array is adapted to provide the data block to the vertical buffer of another of the elements of the memory array.

11. The neural inference chip of claim 10, wherein:
the horizontal buffer of each element of the memory array is adapted to provide a data block to the horizontal buffer of another of the elements of the memory array or to the at least one neural core at a rate of one data block per cycle,
the vertical buffer of each element of the memory array is adapted to provide a data block to the vertical buffer of another of the elements of the memory array at a rate of one data block per cycle,
the layover buffer of each element of the memory array is adapted to provide a data block to the horizontal buffer of that element of the memory array at a rate of one data block per cycle.
12. The neural inference chip of claim 10, wherein:
the instruction memory is adapted to provide a plurality of instructions to the instruction buffer per cycle, each position of the instruction buffer is adapted to store a plurality of instructions, and
the instruction buffer is adapted to advance a plurality of instructions between positions in the instruction buffer per cycle.
13. A neural inference chip for computing neural activations, the neural inference chip comprising:
at least one neural core;
a memory array operatively coupled to the at least one neural core, the memory array comprising a plurality of elements, each element comprising a memory, a horizontal buffer, and a vertical buffer,
the horizontal buffer of each element of the memory array being in communication with either the horizontal buffer of another of the elements of the memory array or to the at least one neural core and
the vertical buffer of each element of the memory array being in communication with the vertical buffer of another element of the memory array;
a plurality of instruction buffers in communication with the memory array, each of plurality of instruction buffers having a position corresponding to one of the plurality of elements of the memory array;
a plurality of instruction memories, each in communication with one of the plurality of instruction buffers, wherein
each instruction memory is adapted to provide at least one instruction to its instruction buffer,
each instruction buffer is adapted to advance the at least one instruction between positions in that instruction buffer,
each instruction buffer is adapted to provide the at least one instruction to at least one of the plurality of elements of the memory array from its associated position in that instruction buffer when the memory of the at least one of the plurality of elements contains data associated with the at least one instruction,
each of the plurality of elements of the memory array is adapted to provide a data block from its memory to its vertical buffer in response to the arrival of an associated instruction from the instruction buffer,

each of the plurality of elements of the memory array is adapted to provide the data block from its vertical buffer to its horizontal buffer,

the vertical buffer of each element of the memory array is adapted to provide the data block to the vertical buffer of another of the elements of the memory array

the horizontal buffer of each element of the memory array is adapted to provide a data block to the horizontal buffer of another of the elements of the memory array or to the at least one neural core.

14. The neural inference chip of claim 13, wherein:

each instruction buffer is adapted to advance instructions between positions therein at a rate of one position per cycle,

the horizontal buffer of each element of the memory array is adapted to provide a data block to the horizontal buffer of another of the elements of the memory array or to the at least one neural core at a rate of one data block per cycle.

15. The neural inference chip of claim 13, comprising an array of neural cores, the array of neural cores comprising the at least one neural core and having a plurality of rows.

16. The neural inference chip of claim 13, wherein the memory array is two-dimensional, the plurality of elements of the memory array being arranged in a plurality of rows and a plurality of columns.

17. The neural inference chip of claim 16, wherein each element of the memory array further comprises a vertical buffer, the vertical buffer of each element of the memory array being in communication with the vertical buffer of another element of the memory array.

18. The neural inference chip of claim 17, wherein:

each of the plurality of elements of the memory array is adapted to provide a data block from its memory to its vertical buffer in response to the arrival of an associated instruction from the instruction buffer,

each of the plurality of elements of the memory array is adapted to provide the data block from its vertical buffer to its horizontal buffer,

the vertical buffer of each element of the memory array is adapted to provide the data block to the vertical buffer of another of the elements of the memory array.

19. The neural inference chip of claim 18, wherein:

the horizontal buffer of each element of the memory array is adapted to provide a data block to the horizontal buffer of another of the elements of the memory array or to the at least one neural core at a rate of one data block per cycle,

the vertical buffer of each element of the memory array is adapted to provide a data block to the vertical buffer of another of the elements of the memory array at a rate of one data block per cycle.

20. The neural inference chip of claim 17, wherein each element of the memory array further comprises a layover buffer, the layover buffer of each element of the memory array being in communication with the horizontal buffer and the vertical buffer of that element of the memory array.

21. The neural inference chip of claim 20, wherein:

each of the plurality of elements of the memory array is adapted to provide a data block from its memory to its vertical buffer in response to the arrival of an associated instruction from the instruction buffer,

each of the plurality of elements of the memory array is adapted to provide the data block from its vertical buffer to its layover buffer,

each of the plurality of elements of the memory array is adapted to provide the data block from its layover buffer to its horizontal buffer,

the vertical buffer of each element of the memory array is adapted to provide the data block to the vertical buffer of another of the elements of the memory array.

22. The neural inference chip of claim 21, wherein:

the horizontal buffer of each element of the memory array is adapted to provide a data block to the horizontal buffer of another of the elements of the memory array or to the at least one neural core at a rate of one data block per cycle,

the vertical buffer of each element of the memory array is adapted to provide a data block to the vertical buffer of another of the elements of the memory array at a rate of one data block per cycle,

the layover buffer of each element of the memory array is adapted to provide a data block to the horizontal buffer of that element of the memory array at a rate of one data block per cycle.

23. The neural inference chip of claim 21, wherein:

the instruction memory is adapted to provide a plurality of instructions to the instruction buffer per cycle, each position of the instruction buffer is adapted to store a plurality of instructions, and

the instruction buffer is adapted to advance a plurality of instructions between positions in the instruction buffer per cycle.

24. A method comprising:

providing at least one instruction to an instruction buffer from an instruction memory;

advancing the at least one instruction between positions in the instruction buffer;

providing the at least one instruction from the instruction buffer to at least one of a plurality of elements of a memory array when the memory of the at least one of the plurality of elements contains data associated with the at least one instruction,

the memory array comprising the plurality of elements, each element comprising a memory and a horizontal buffer, the horizontal buffer of each element of the memory array being in communication with either the horizontal buffer of another of the elements of the memory array or to the at least one neural core

providing a data block from the memory to the horizontal buffer of the at least one of a plurality of elements in response to the arrival of the at least one instruction from the instruction buffer;

providing the data block from the horizontal buffer of the at least one of a plurality of elements to the horizontal buffer of another of the elements of the memory array or to at least one neural core.