SpiNNaker:

biologically-inspired massively-parallel computing

Steve Furber

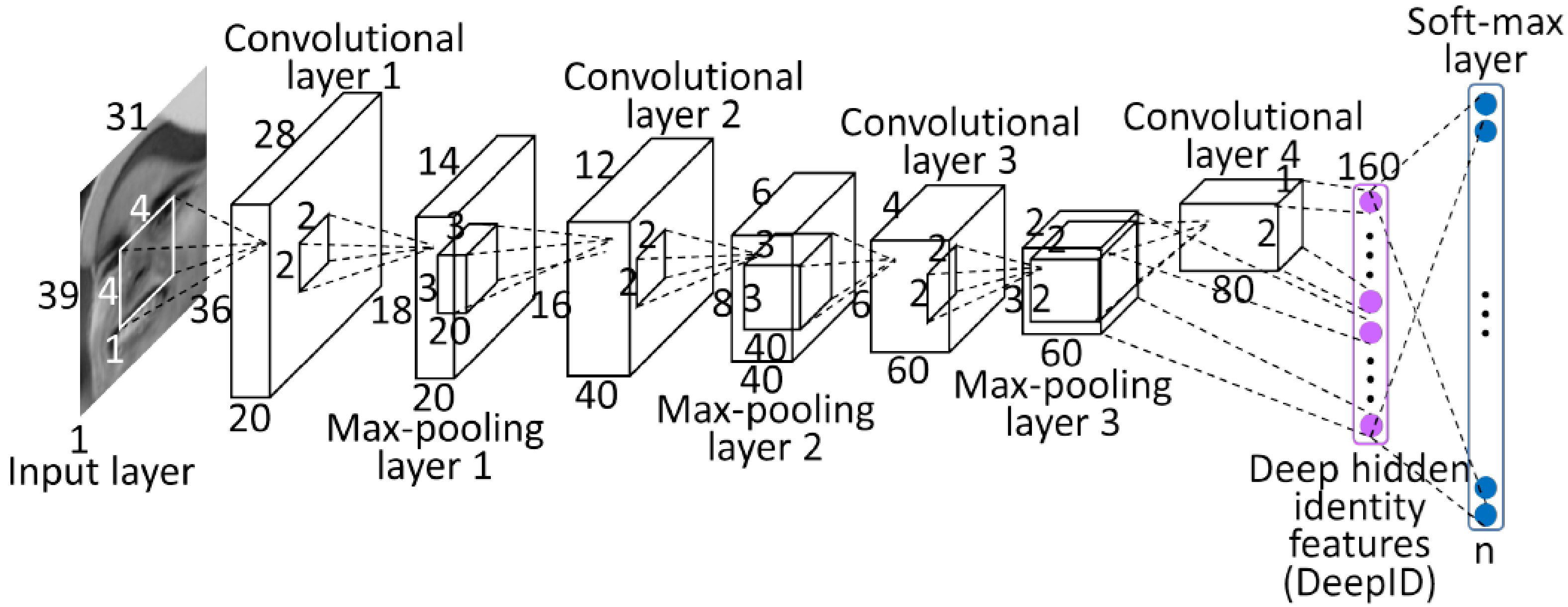
The University of Manchester

steve.furber@manchester.ac.uk

Bio-inspiration

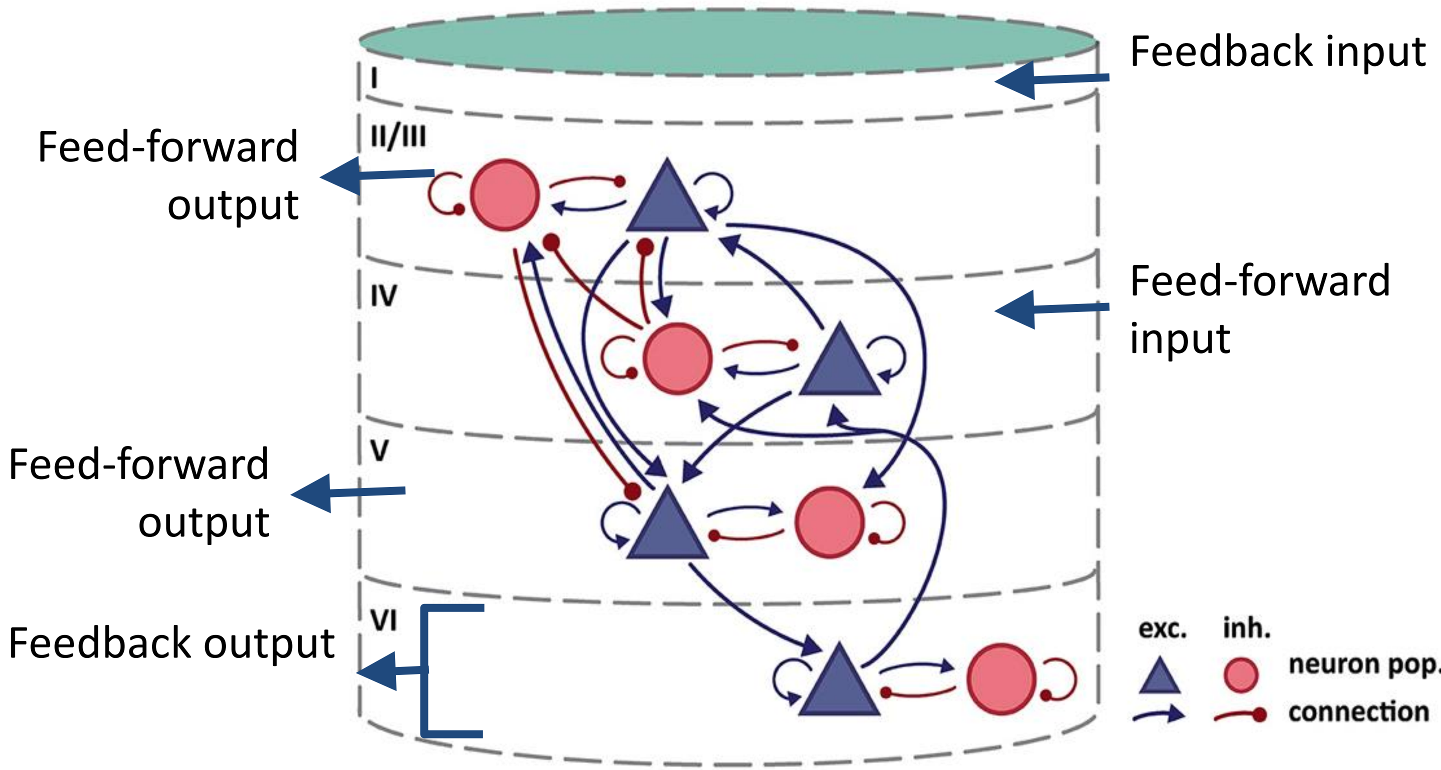
- Can massively-parallel computing resources accelerate our understanding of brain function?
- Can our growing understanding of brain function point the way to more efficient parallel, fault-tolerant computation?

ConvNets - structure



- Dense convolution kernels
- Abstract neurons
- Only feed-forward connections
- Trained through backpropagation

The cortex - structure



- Spiking neurons
- Two-dimensional structure
- Sparse connectivity

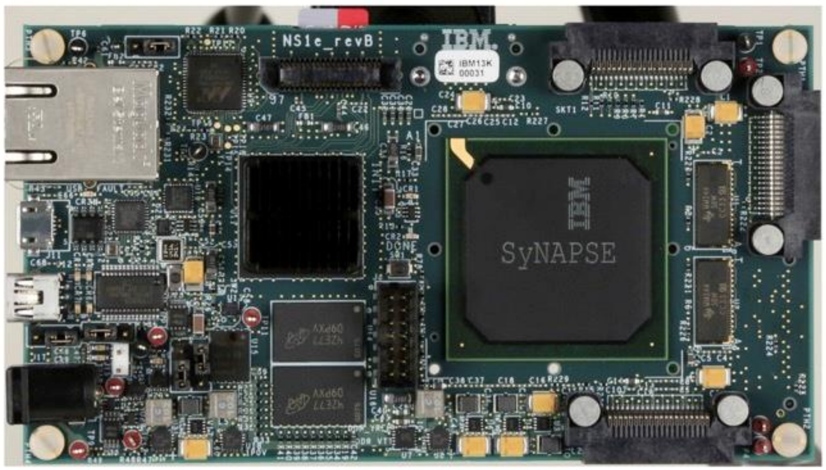
ConvNets - GPUs

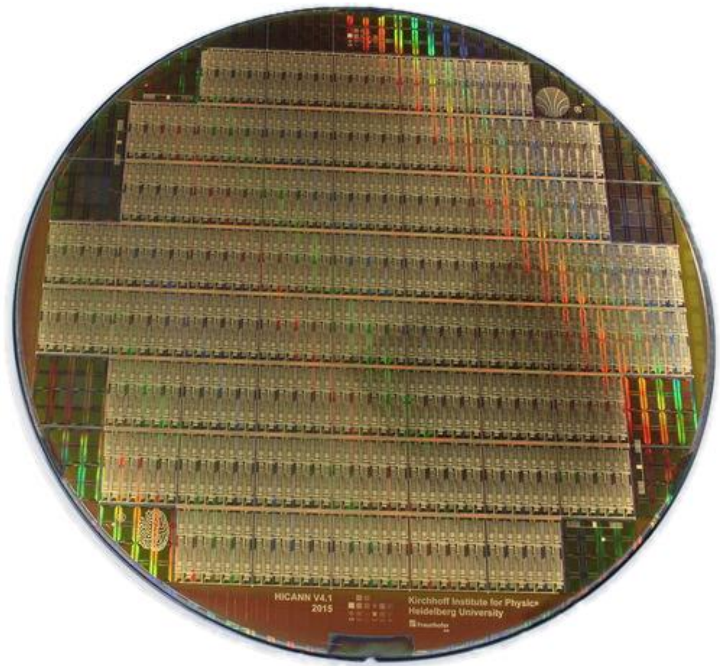
- Dense matrix multiplications
- 3.2kW
- Low precision

Cortical models - Supercomputers

- Sparse matrix operations
- Efficient communication of spikes
- 2.3MW

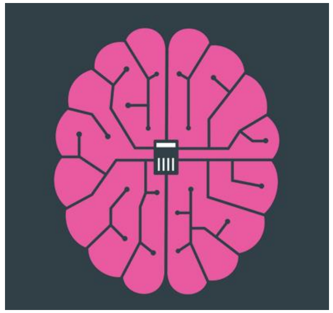
Cortical models - Neuromorphic hardware





- Memory local to computation
- Low-power
- Real time
- 62mW

SpiNNaker
Biologically Inspired Massively Parallel Architectures



MIT Technology Review

10 BREAKTHROUGH TECHNOLOGIES

Agricultural Drones
Relatively cheap drones with advanced sensors and imaging capabilities are giving farmers new ways to increase yields and reduce crop damage.

Ultraprivate Smartphones
New models built with security and privacy in mind reflect the Zeitgeist of the Snowden era.

Brain Mapping
A new map, a decade in the works, shows structures of the brain in far greater detail than ever before, providing neuroscientists with a guide to its immense complexity.

Neuromorphic Chips
Microprocessors configured more like brains than traditional chips could soon make computers far more astute about what's going on around them.

Genome Editing
The ability to create primates with intentional mutations could provide powerful new ways to study complex and genetically baffling brain disorders.

Microscale 3-D Printing
Inks made from different types of materials, precisely applied, are greatly expanding the kinds of things that can be printed.

Mobile Collaboration
The smartphone era is finally getting the productivity software it needs.

Oculus Rift
Thirty years after virtual-reality goggles and immersive virtual worlds made their debut, the technology finally seems poised for widespread use.

Agile Robots
Computer scientists have created machines that have the balance and agility to walk and run across rough and uneven terrain, making them far more useful in navigating human environments.

Smart Wind and Solar Power
Big data and artificial intelligence are producing ultra-accurate forecasts that will make it feasible to integrate much more renewable energy into the grid.

Neuromorphic Chips

Microprocessors configured more like brains than traditional chips could soon make computers far more astute about what's going on around them.

Breakthrough

An alternative design for computer chips that will enhance artificial intelligence.

Why It Matters

Traditional chips are reaching fundamental performance limits.

Key Players

+ Qualcomm
+ IBM
+ HRL Laboratories
+ Human Brain Project

http://www.technologyreview.com/featuredstory/526506/neuromorphic-chips/

Top 10 emerging technologies of 2015

WORLD ECONOMIC FORUM

The 2015 list is:

1. Fuel cell vehicles
2. Next-generation robotics
3. Recyclable thermoset plastics
4. Precise genetic engineering techniques
5. Additive manufacturing
6. Emergent artificial intelligence
7. Distributed manufacturing
8. 'Sense and avoid' drones
9. Neuromorphic technology
10. Digital genome

9. Neuromorphic technology

Computer chips that mimic the human brain

Even today's best supercomputers cannot rival the sophistication of the human brain. Computers are linear, moving data back and forth between memory chips and a central processor over a high-speed backbone. The brain, on the other hand, is fully interconnected, with logic and memory intimately cross-linked at billions of times the density and diversity of that found in a modern computer. Neuromorphic chips aim to process information in a fundamentally different way from traditional hardware, mimicking the brain's architecture to deliver a huge increase in a computer's thinking and responding power.

Miniaturization has delivered massive increases in conventional computing power over

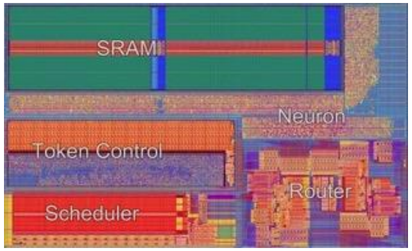
https://agenda.weforum.org/2015/03/top-10-emerging-technologies-of-2015-2/

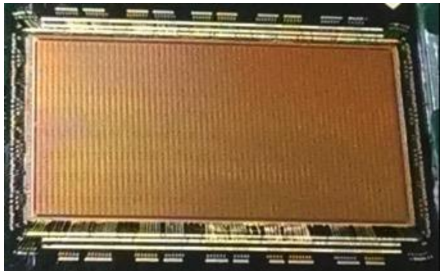
Table A1. Selected component technology vectors that have a high probability of deployment in ten years ***(* denotes more speculative deployment within this timeframe)***

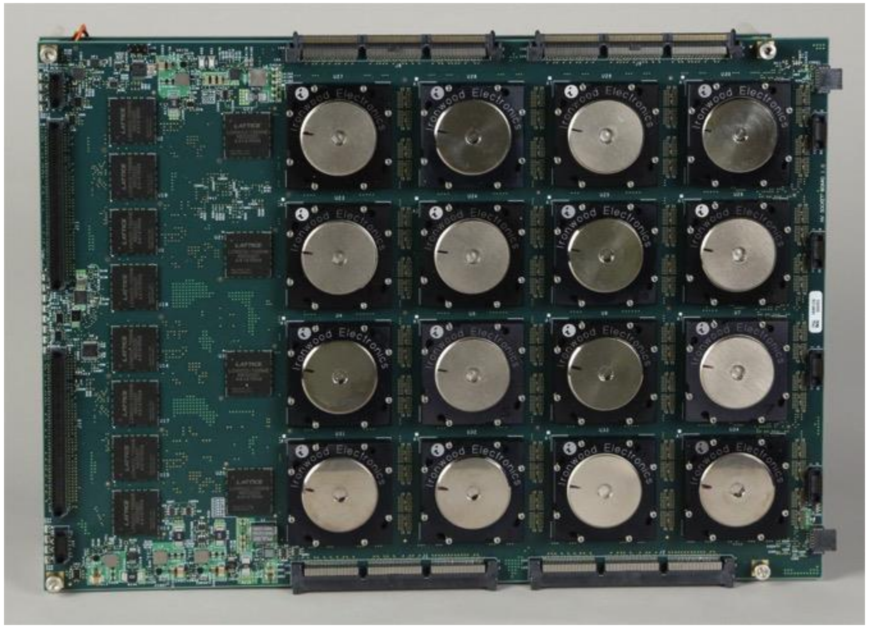
Component technology vector	Time-frame to first commercial products	Approach to achieving and retaining competitive advantage
Neuromorphic Computing	Available now	Continued R&D into new architectures coupled with 3D technologies and new materials, Deep Learning accelerators (for mobile and data center applications), and applications for true brain-inspired computing
Photonics	Available now	Foundries for tools and materials R&D; integrate photonics with CMOS and other materials
Sensors	Available now	Foundries for tools and materials R&D; integrate new types/classes of sensors with CMOS and other materials
CMOS (sub 7nm node size or new 3D structures)*	Advances in thermal management available with new process nodes	Deep understanding of transistor physics and chipset architecture and related design know-how; foundries and labs for transistor and materials R&D
Magnetics	1-2 years (MRAM as eFlash), 3 years (as DRAM), 5-7 years (as SRAM)	Foundries for tools and materials R&D; integrate magnetics with CMOS and other materials
3D	2-3 years (wafer-to-wafer stacking), 4-5 years (die-to-wafer stacking), 5-7 (Monolithic 3D)	Deep understanding of applications space and benefits associated use of 3D technologies and design know-how; foundries for tools and materials R&D; design automation tool R&D
Data-flow based architectures	3-4 years	Continued architecture R&D, coupled with materials, integration, and manufacturing; build an ecosystem for solutions using data-flow based architectures
Ultra-high performance wireless systems	3 years (5G), 10-12 years (6G)	Continued R&D in new materials and processes, antenna design advances, chipset manufacturing, and integration
Advanced non-volatile memory as SRAM	5+ years	Deep understanding of applications space and chipset architectures
Carbon nanotubes	5-7 years	Foundries/labs for materials R&D for hardware architectures;

IBM TrueNorth

- 4,096 digital neurosynaptic cores
  - one million configurable neurons
  - 256 million programmable synapses
  - ~70mW
  - over 400 Mbits of embedded SRAM
  - 5.4 billion transistors
- 16 TrueNorth Chips assembled into a 4x4 mesh
  - 16 million neurons and 4 billion synapses.

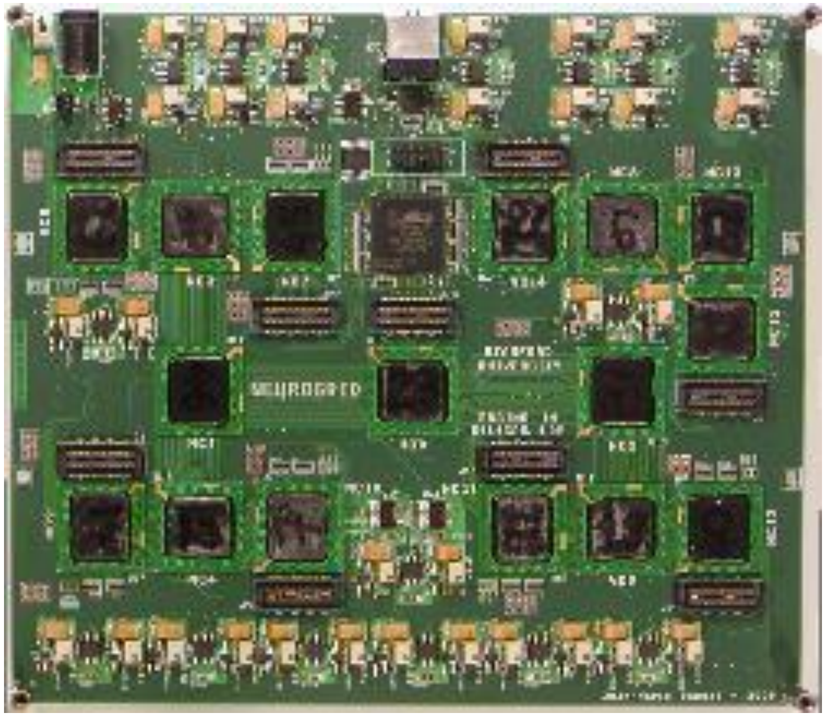






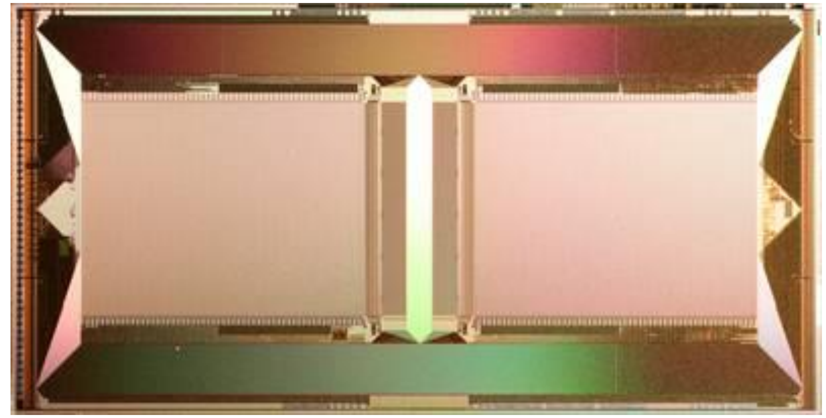
Stanford Neurogrid

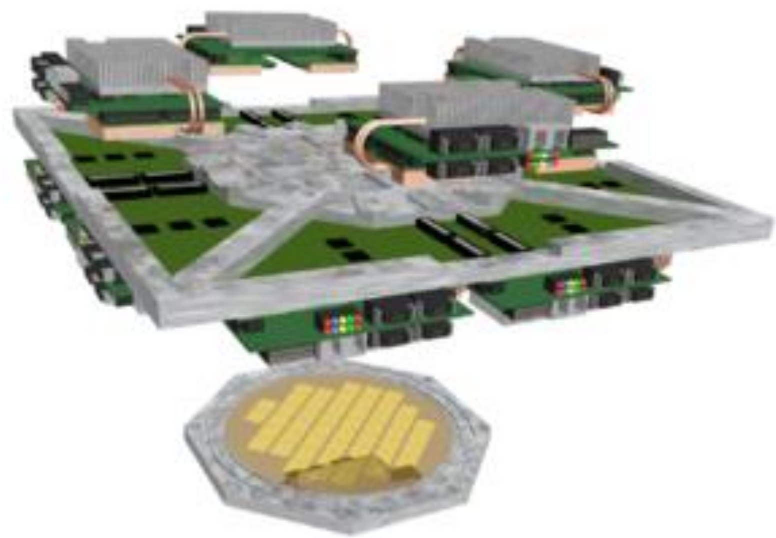
- Neurocore Chip
  - 65k neurons
  - each with two compartments and a set of configurable silicon ion channels
- Sixteen Neurocores are assembled on a board
  - million-neuron Neurogrid



Heidelberg HiCANN

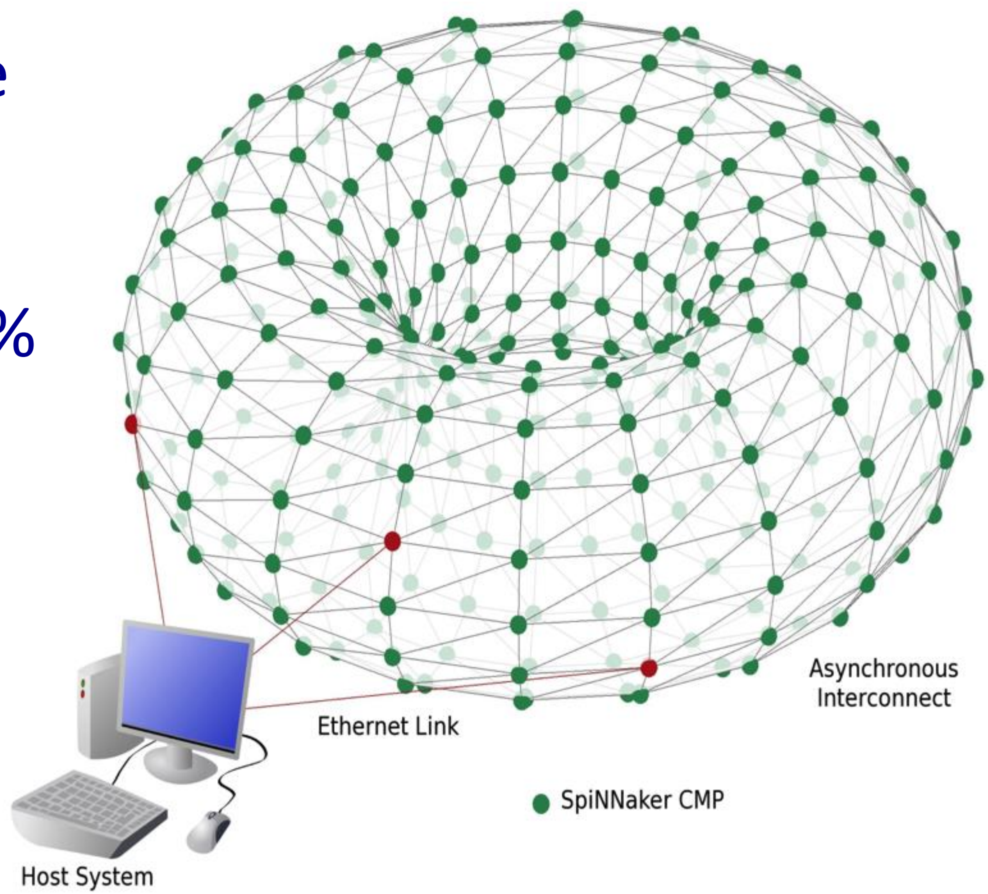
- Wafer-scale analogue neuromorphic system
- 8” 180nm wafer:
  - 200,000 neurons
  - 50M synapses
  - 10^4x faster than biology





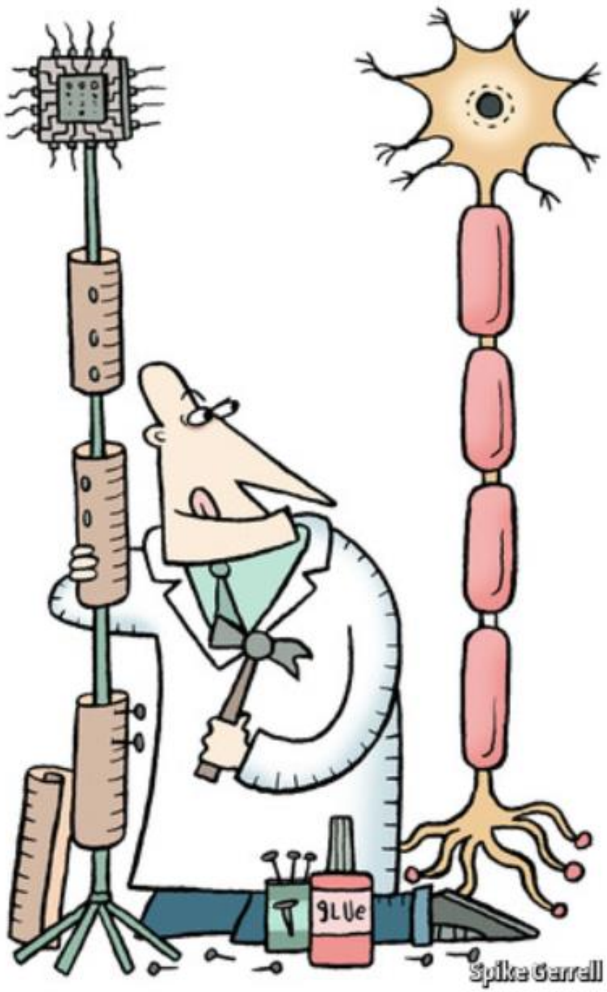
SpiNNaker project

- A million mobile phone processors in one computer
- Able to model about 1% of the human brain…
- …or 10 mice!



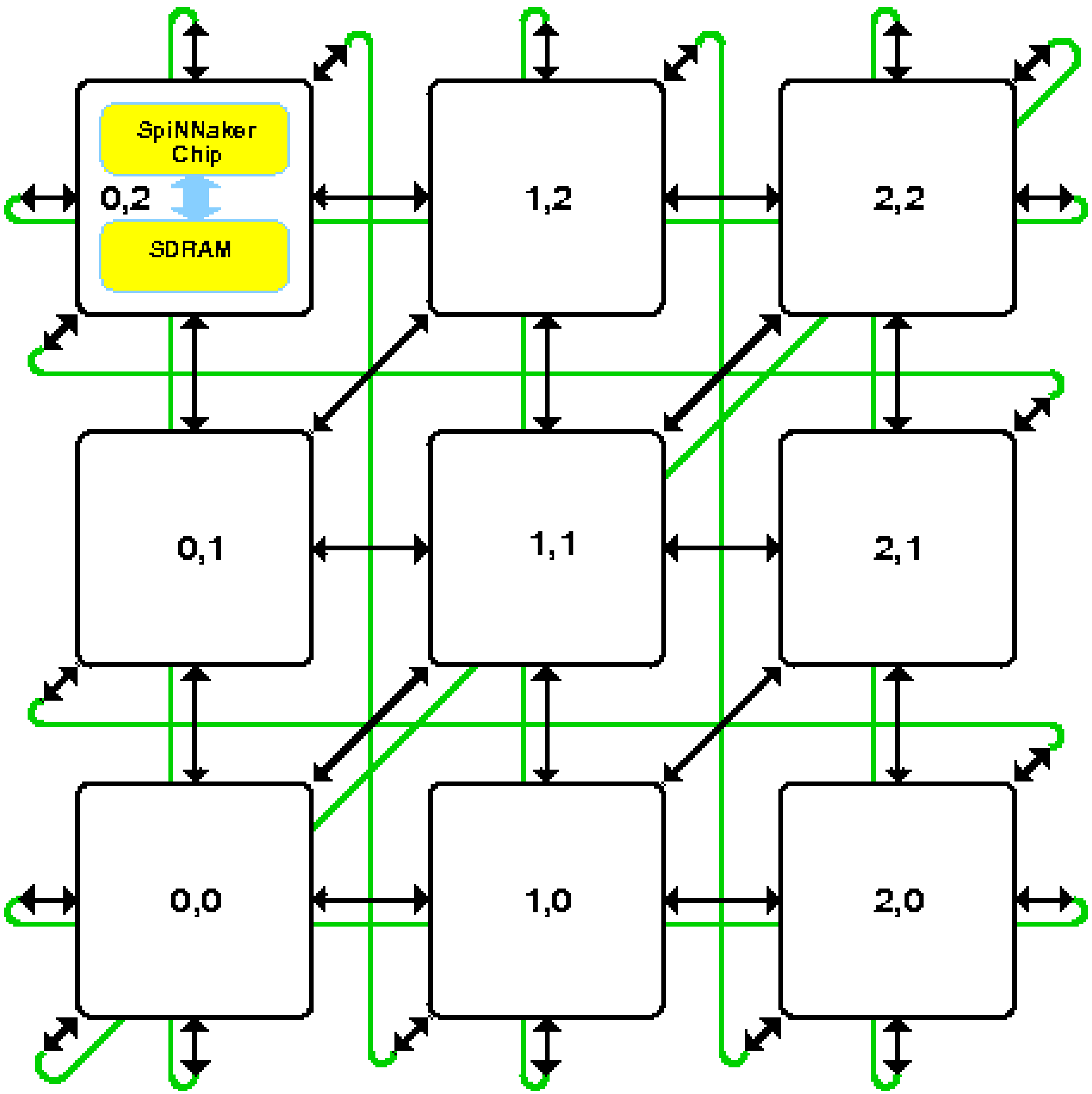
SpiNNaker
Biologically
Inspired
Massively
Parallel
Architectures

Design principles

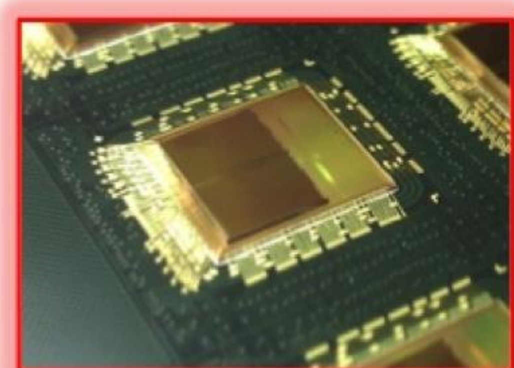


- *Virtualised topology*
  - physical and logical connectivity are decoupled
- *Bounded asynchrony*
  - time models itself
- *Energy frugality*
  - processors are free
  - the real cost of computation is energy

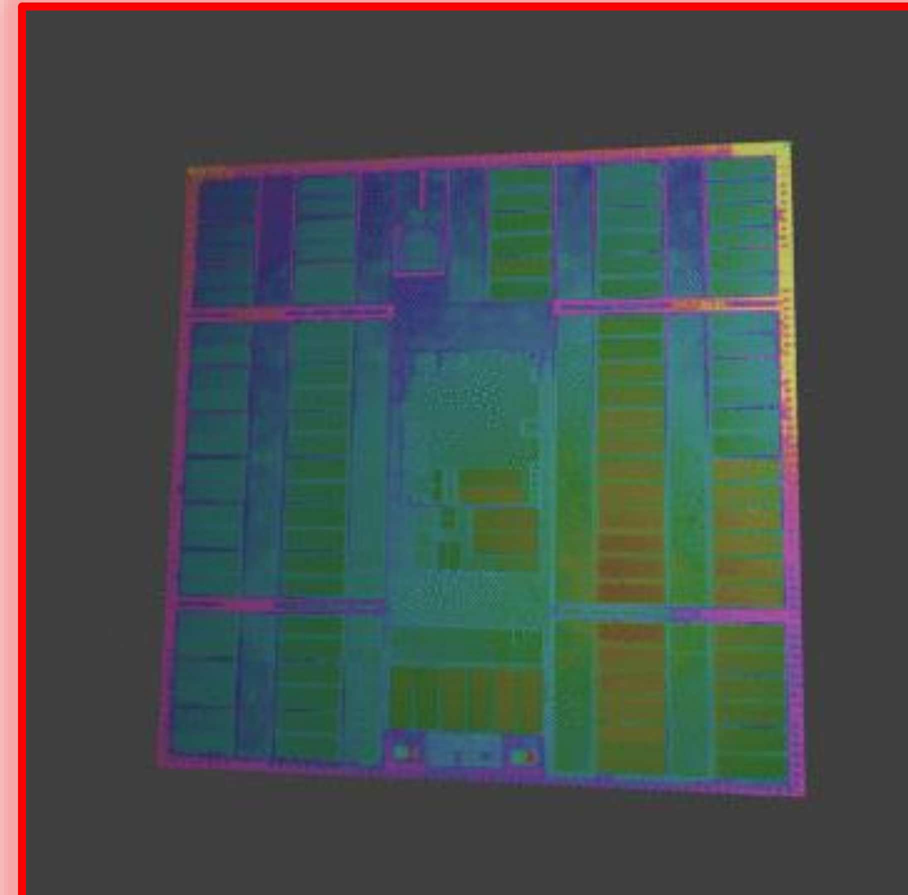
SpiNNaker system



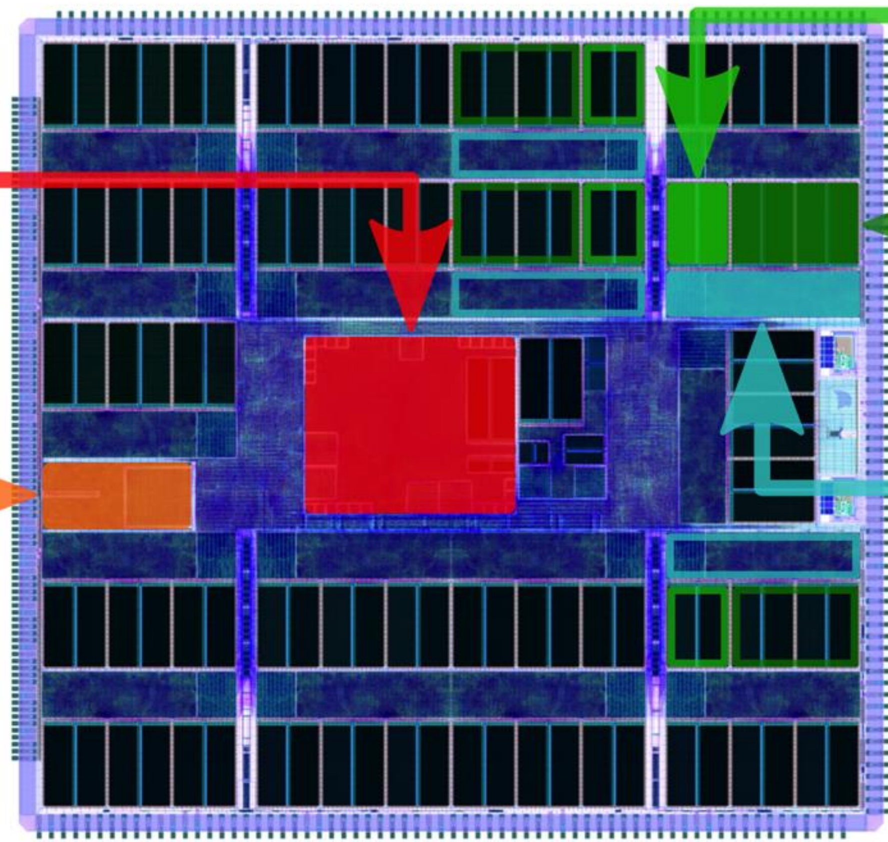
SpiNNaker chip



Multi-chip packaging by UNISEM Europe



Chip resources



SpiNNaker
Biologically
Inspired
Massively
Parallel
Architectures

SpiNNaker machines

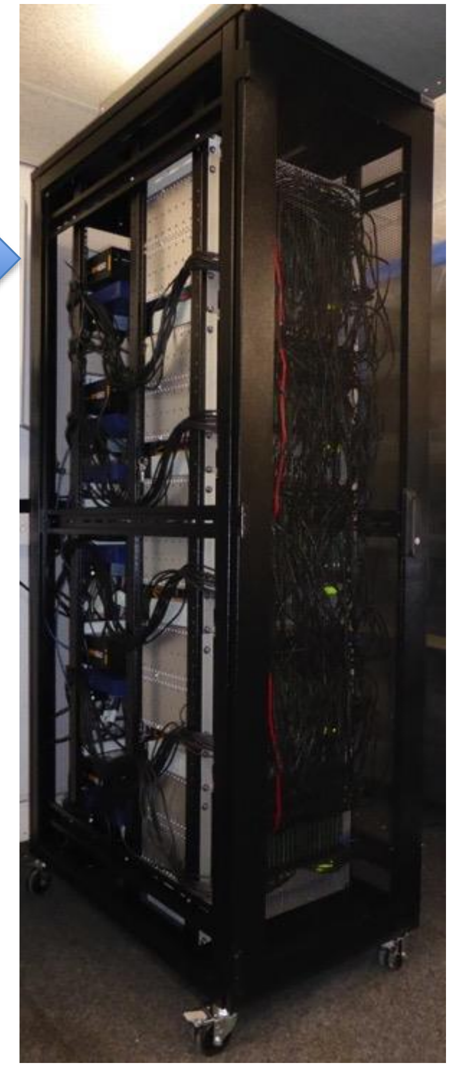
103

864 cores
- drosophila scale

104

20,000 cores
– frog scale

105



100,000 cores
– mouse scale

102

72 cores
- pond snail scale

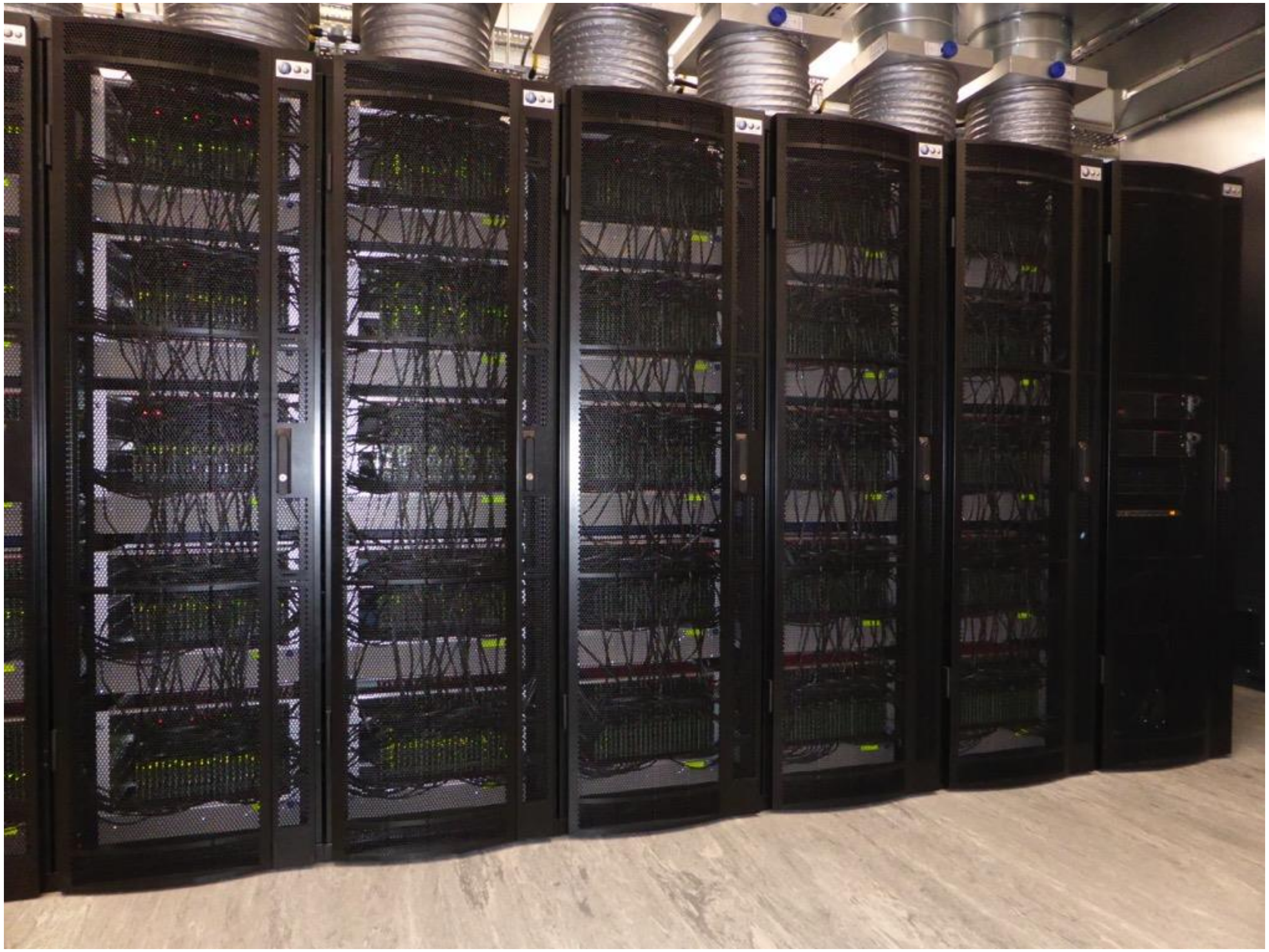
SpiNNaker
Biologically Inspired Massively Parallel Architectures

Largest *SpiNNaker* machine

EPSRC

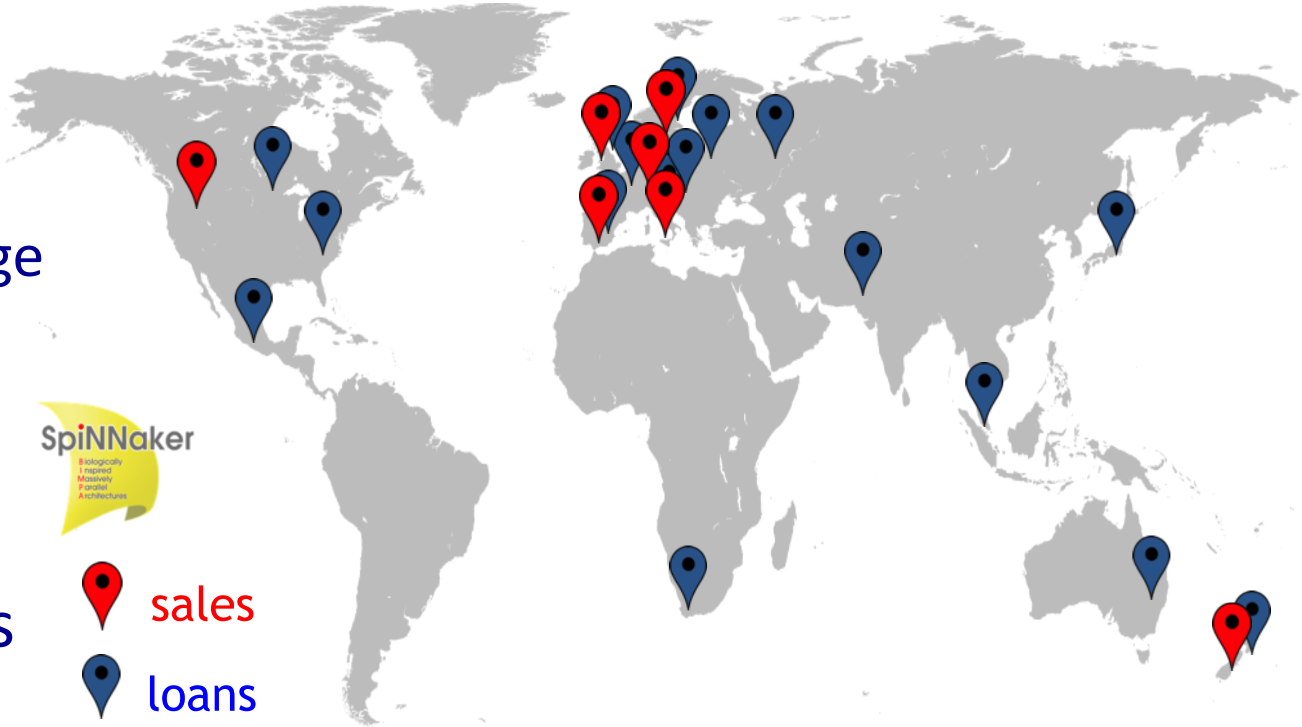
Human Brain Project

- HBP platform
  - 500,000 cores
  - 6 cabinets
    (including server)
- Launch
  - 30 March 2016



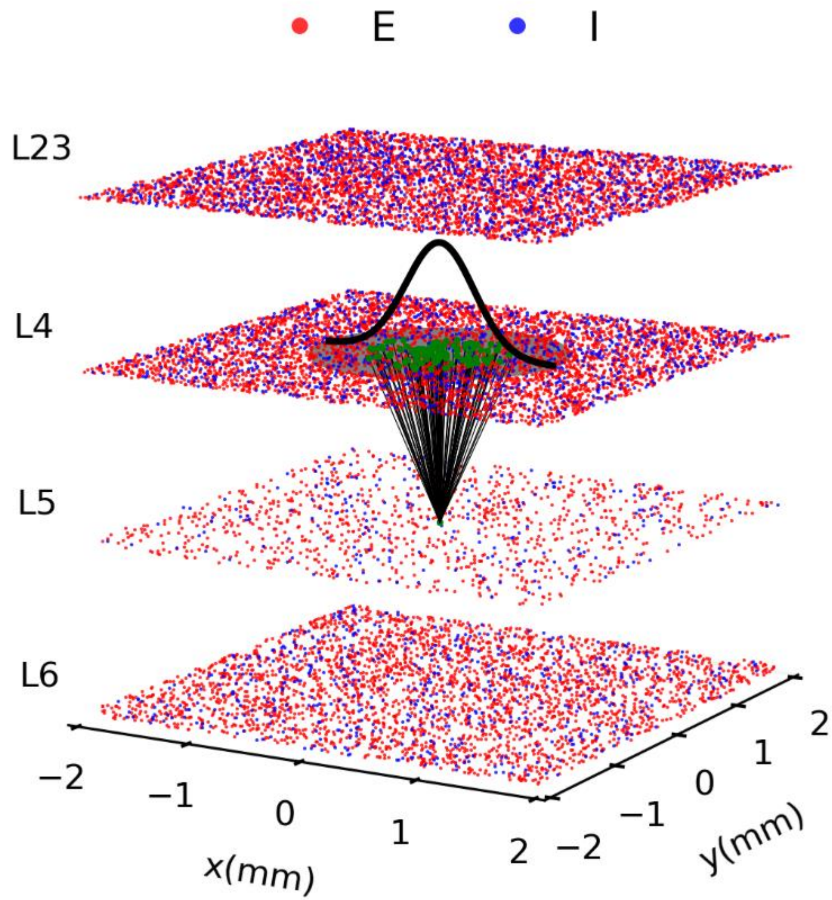
SpiNNaker machines

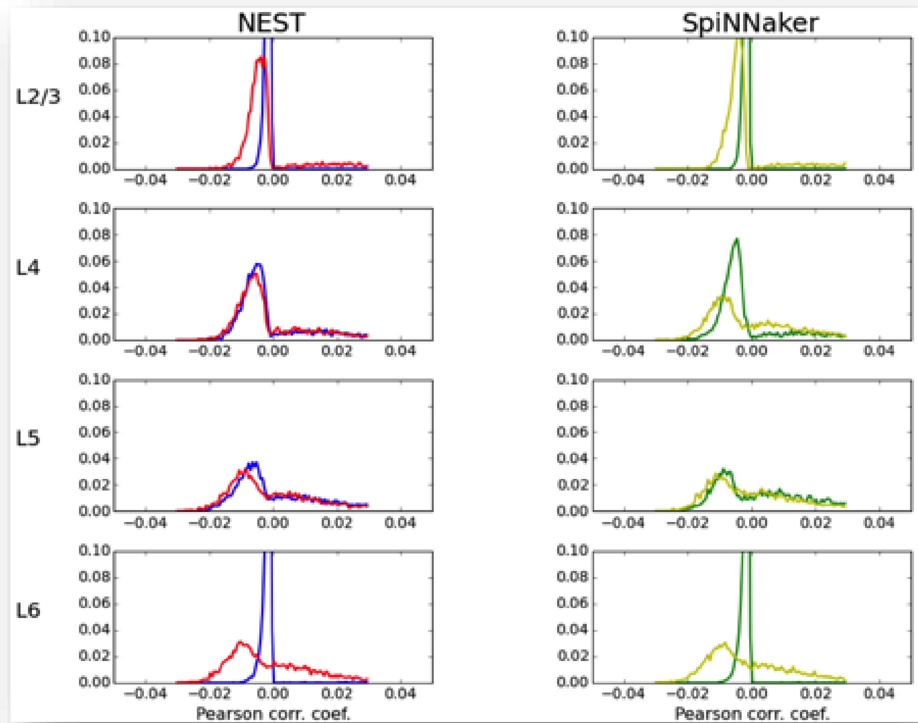
- 90 SpiNNaker systems in use
  - global coverage
- 4-node boards
  - training & small-scale robotics
- 48-node boards
  - insect-scale networks
- multi-board systems



Cortical microcolumn

- 77,169 neurons, 285M synapses, 2,901 cores





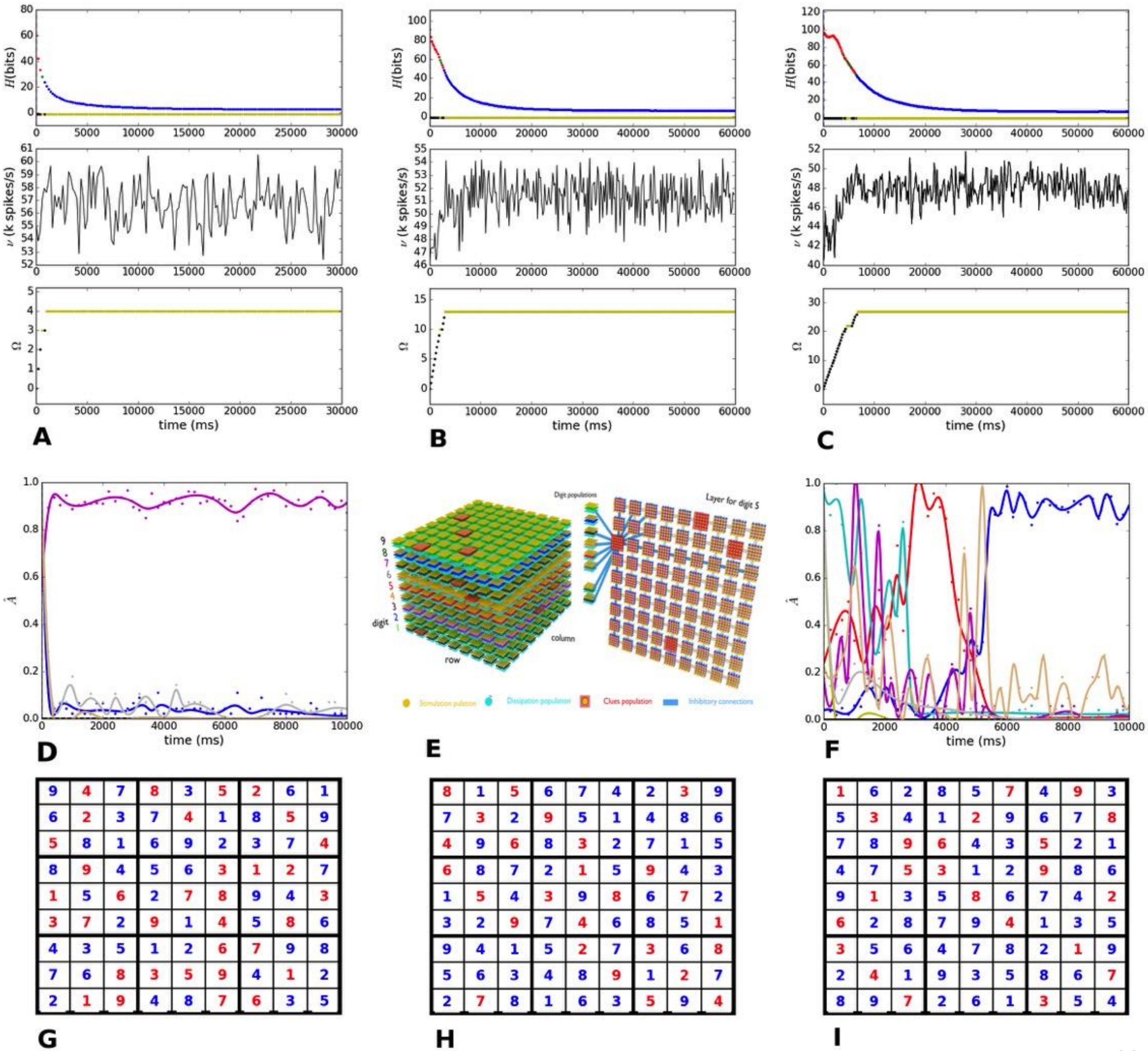
S.J. van Alba, A.G. Rowley, A. Stokes, J. Senk, M. Hopkins, M. Schmidt, D.R. Lester, M. Diesmann, S.B. Furber, "Performance comparison of the digital neuromorphic hardware SpiNNaker and the Neural network simulation software NEST for a full-scale cortical microcircuit model", Frontiers, in preparation.

Sudoku solver

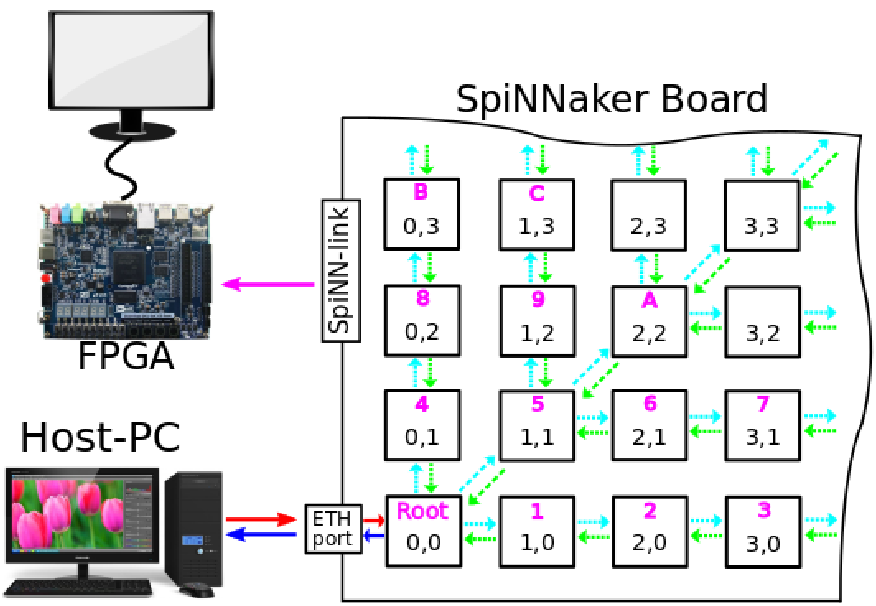
- 37k neurons
- 86M synapses

work by: Gabriel Fonseca Guerra

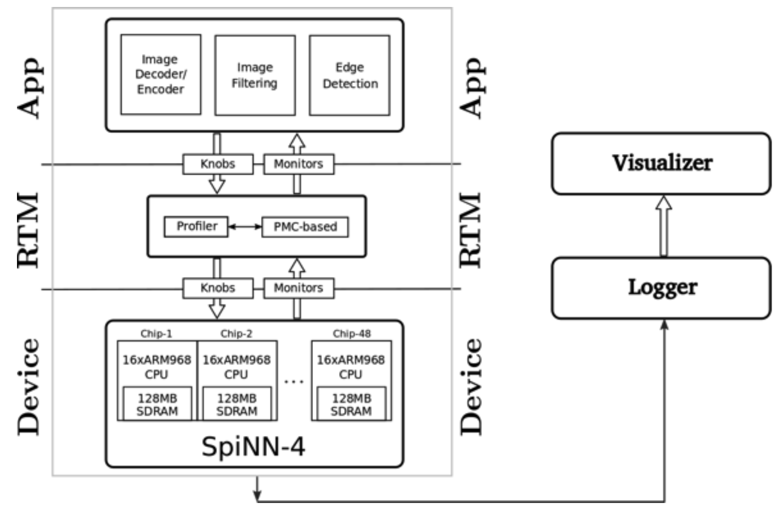
S. Habenschuss, Z. Jonke, and W. Maass, "Stochastic computations in cortical microcircuit models", PLOS Computational Biology, 9(11):e1003311, 2013.

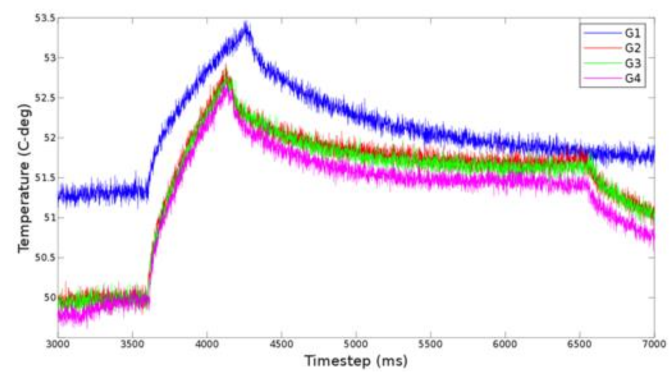


Run-Time Management on *SpiNNaker*



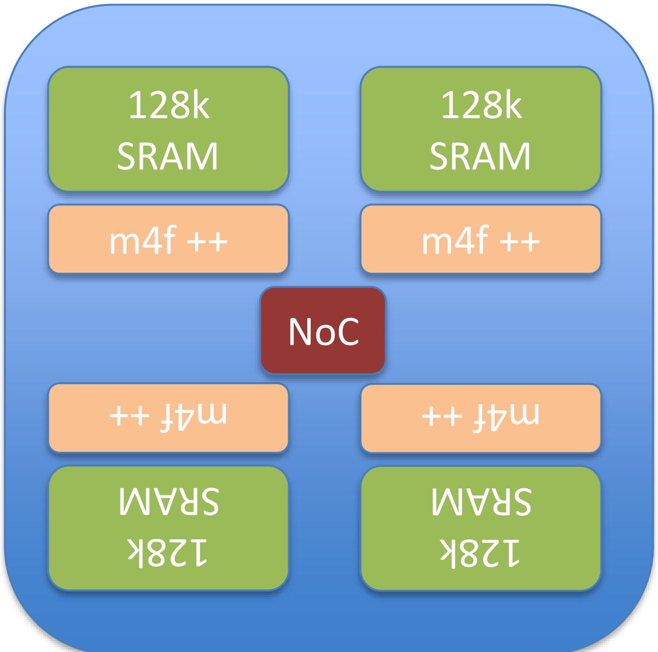
Sending/receiving images via SDP (through ethernet)
Broadcasting images or propagating result via MCPL
Sending images via FR (through SpiNN-link)





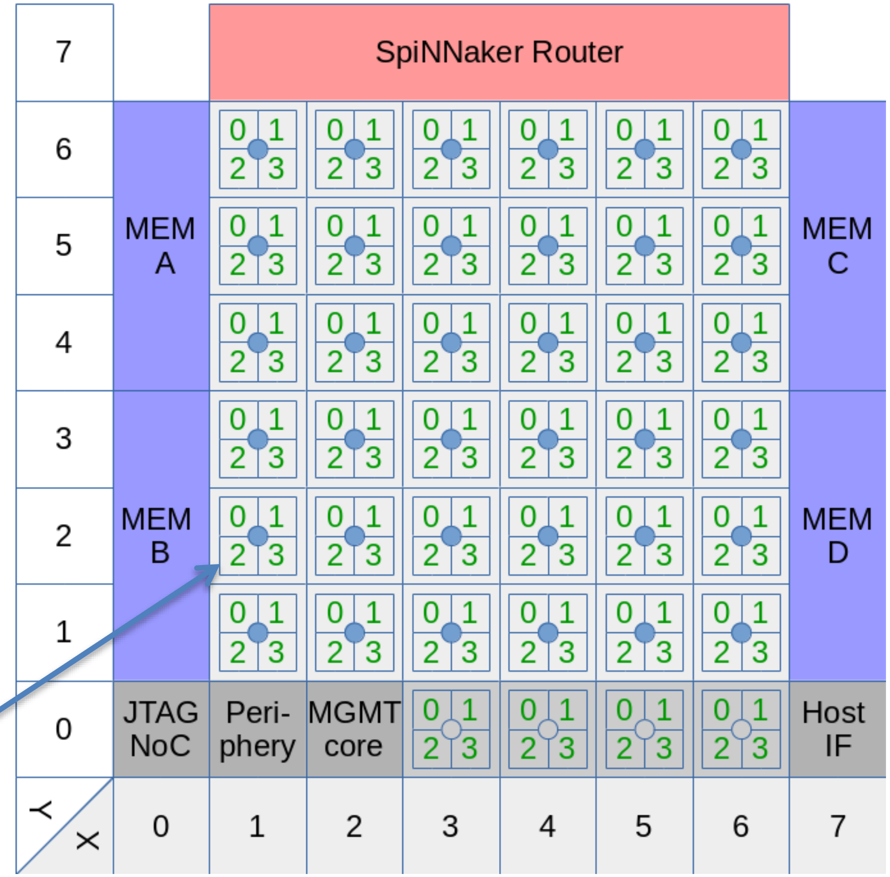
"Software-defined PMC-based Runtime Power Management System for a Many-core Neuromorphic Platform", Indar Sugiarto, Delong Shang, Steve Furber, Amit Kumar Singh, Bassem Ouni, Geoff Merrett, Bashir Al- Hashimi, Proc. ICCES 2017, Cairo, Egypt, 19-20 Dec 2017. ***Best Paper Award!***

SpiNNaker-2



4-core QPE (Quad Processing Element)

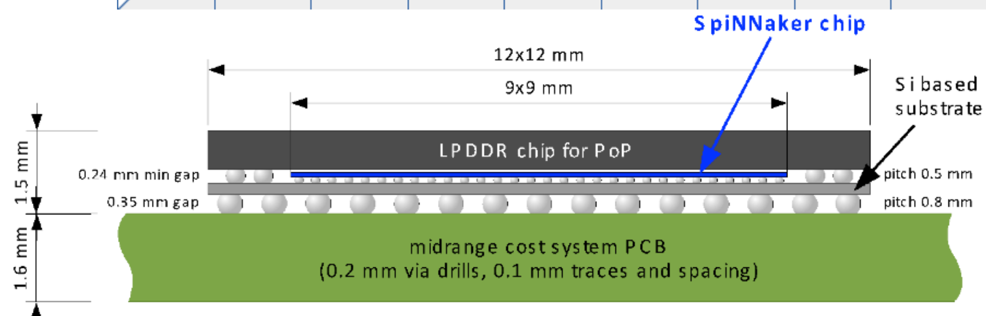
36 QPEs on a 70mm² die: ~30 GFLOPS/w



Global Foundries 22nm FDSOI

~4Gbytes LPDDR4 SDRAM

ML: ~5 TOPS/W



Conclusions

- ***SpiNNaker*:**
  - has been 20 years in conception…
  - …and 10 years in construction,
    - and is now ready for action!
  - ~90 boards with groups around the world
  - 500,000 core machines built
- **PRiME & Graceful:**
  - have used SpiNNaker as an exemplar many-core machine
  - demonstrating power management andoptimisation, run-time monitoring & recovery, and run-time management