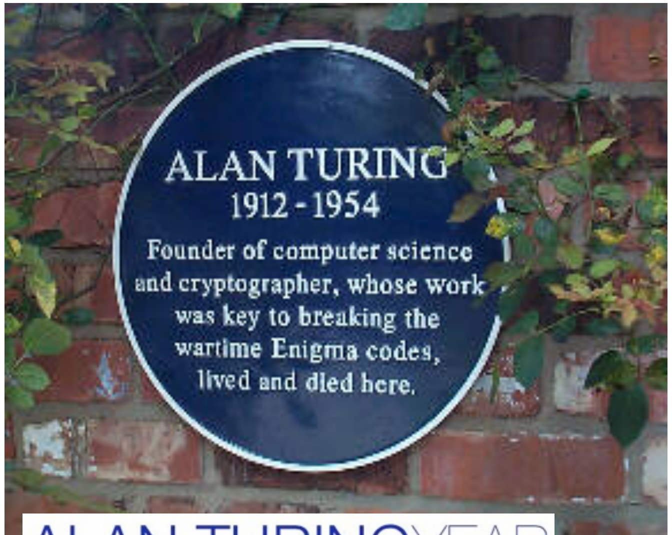
Biologically-Inspired Massively-Parallel Computation

Steve Furber

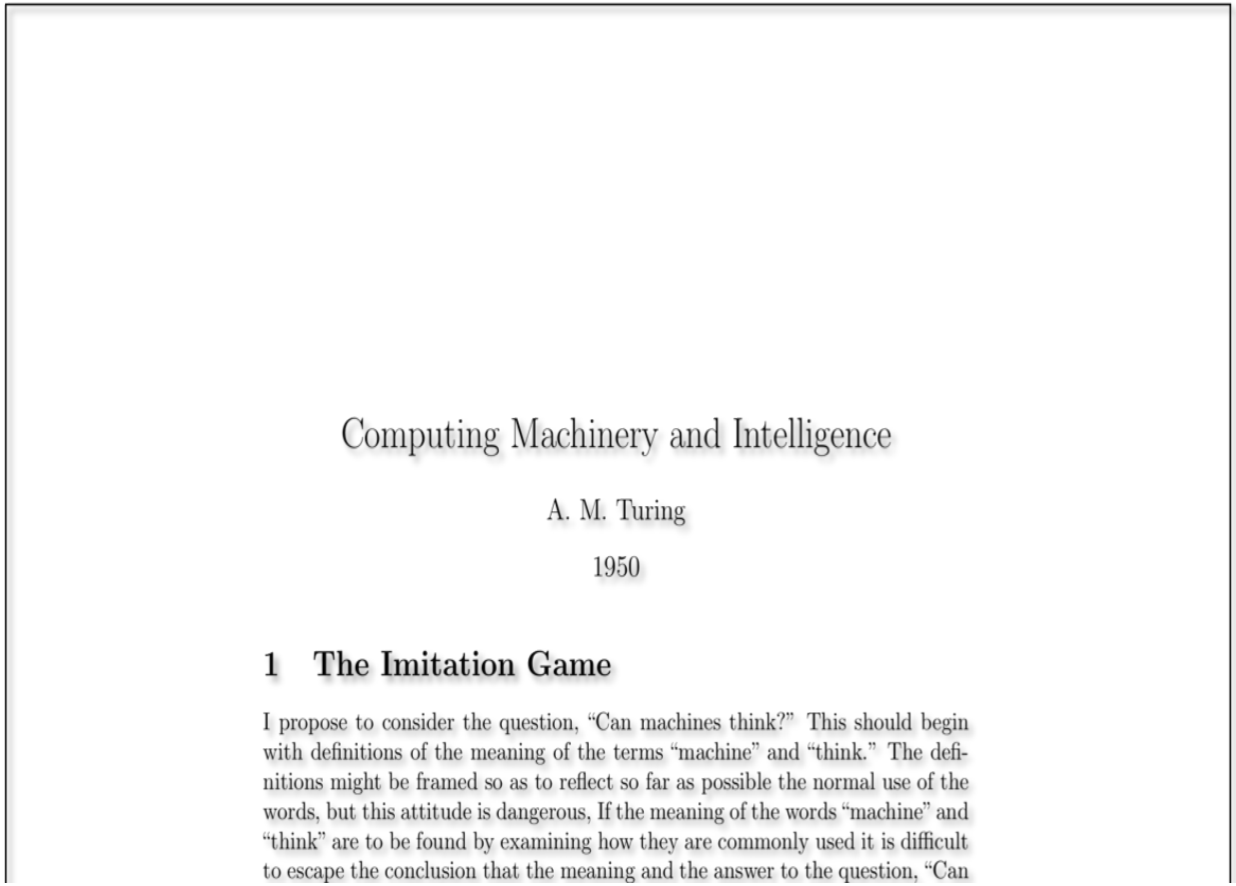
ICL Professor of Computer Engineering

The University of Manchester

Turing Centenary



Turing in Manchester



Computing Machinery and Intelligence

A. M. Turing

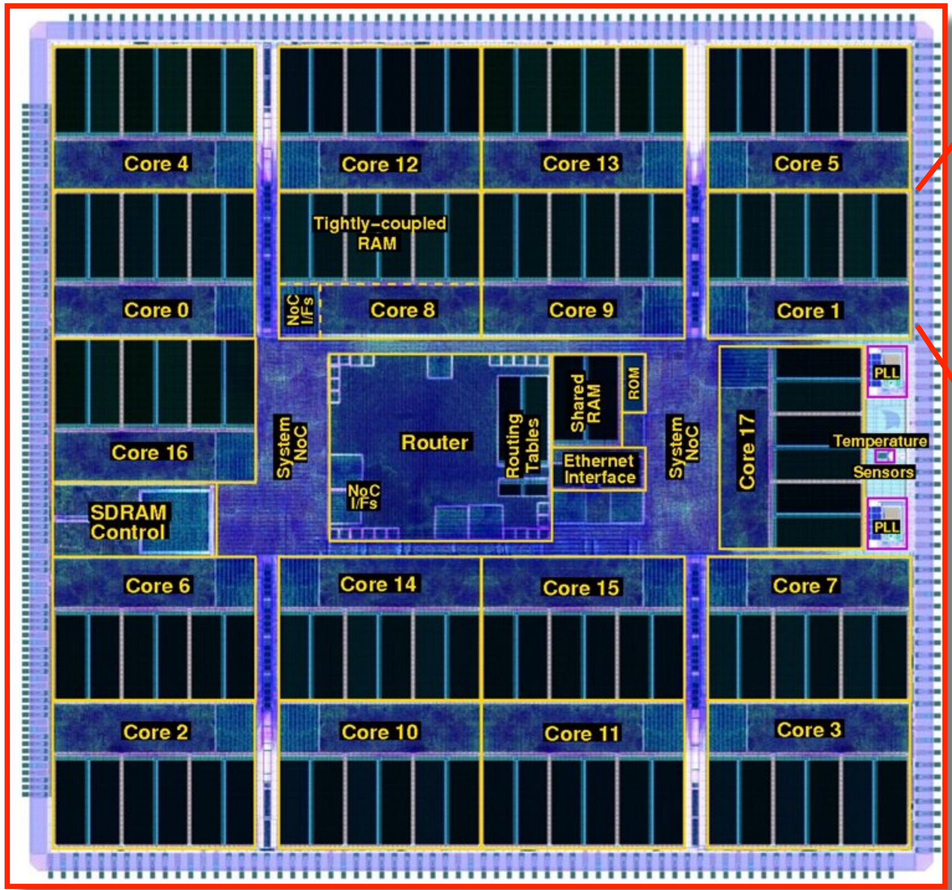
1950

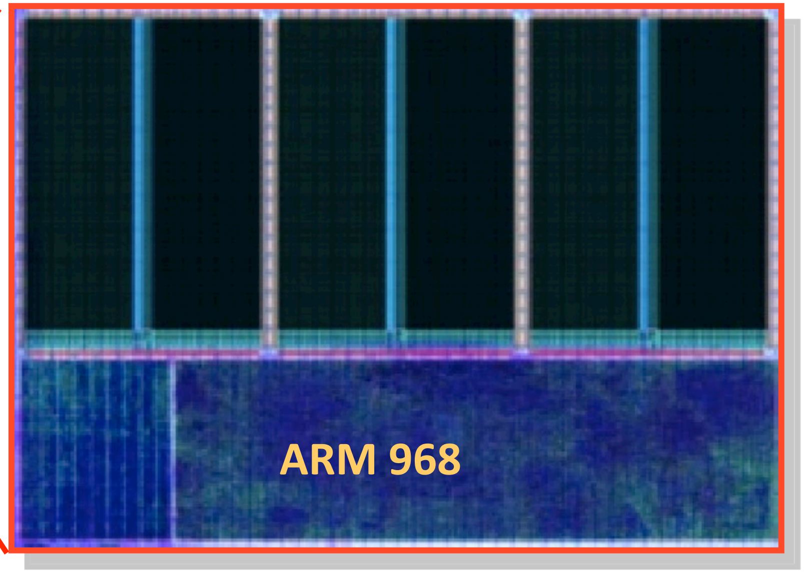
1 The Imitation Game

I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous, If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, "Can

Manchester Baby (1948)

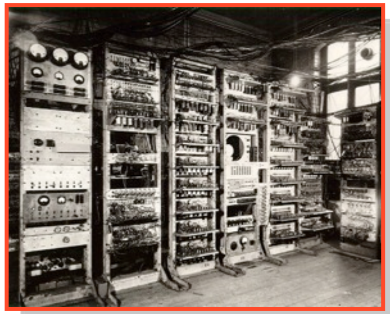
SpiNNaker CPU (2011)



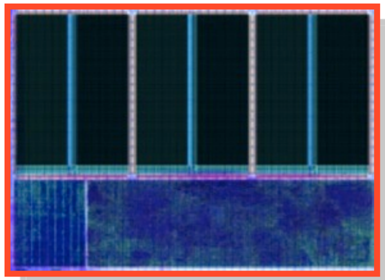


63 years of progress

- ***Baby:***
  - filled a medium-sized room
  - used 3.5 kW of electrical power
  - executed 700 instructions per second



- ***SpiNNaker ARM968 CPU node:***
  - fills ~3.5mm^2 of silicon (130nm)
  - uses 40 mW of electrical power
  - executes 200,000,000 instructions per second



Energy efficiency

- Baby:
  - 5 Joules per instruction
- SpiNNaker ARM968:
  - 0.000 000 000 2 Joules per instruction

25,000,000,000 times better than Baby!

(James Prescott Joule born Salford, 1818)

Bio-inspiration

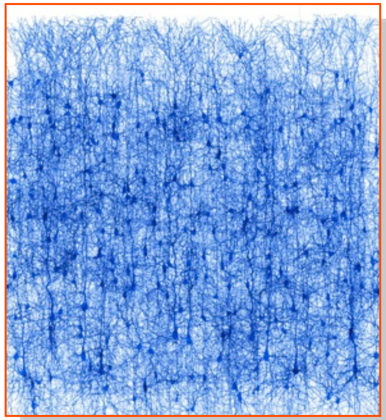
- Can massively-parallel computing resources accelerate our understanding of brain function?
- Can our growing understanding of brain function point the way to more efficient parallel, fault-tolerant computation?

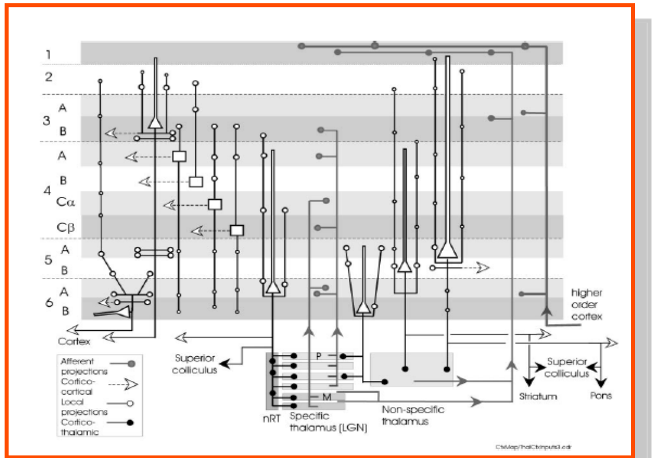
Building brains

- Brains demonstrate
  - massive parallelism (10^{11} neurons)
  - massive connectivity (10^{15} synapses)
  - excellent power-efficiency
    - much better than today's microchips
  - low-performance components (~ 100 Hz)
  - low-speed communication (~ metres/sec)
  - adaptivity – tolerant of component failure
  - autonomous learning

Building brains

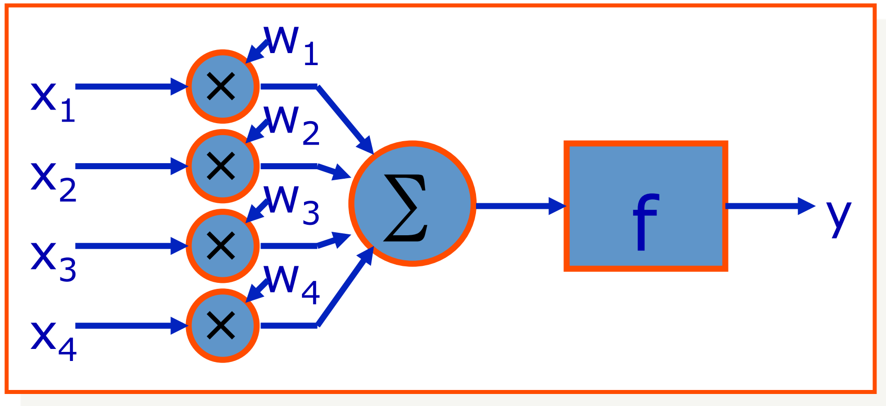
- Neurons
  - multiple inputs, single output (c.f. logic gate)
  - useful across multiple scales (10^2 to 10^{11})
- Brain structure
  - regularity
  - e.g. 6-layer cortical ‘microarchitecture’





Neural Computation

- To compute we need:
  - *Processing*
  - *Communication*
  - *Storage*
- Processing: abstract model
  - linear sum of weighted inputs
    - ignores non-linear processes in dendrites
  - non-linear output function
  - learn by adjusting synaptic weights



Processing

- Leaky integrate-and-fire model
  - inputs are a series of spikes
  - total input is a weighted sum of the spikes
  - neuron activation is the input with a "leaky" decay
  - when activation exceeds threshold, output fires
  - habituation, refractory period, …?

$$x_i = \sum_k \delta(t - t_{ik})$$

$$I = \sum_i w_i x_i$$

$$\dot{A} = -A/\tau_A + I$$

$$if\ A > \vartheta_A\ fire$$

$$\&\ set\ A = 0$$

Processing

- Izhikevich model
  - two variables, one fast, one slow:

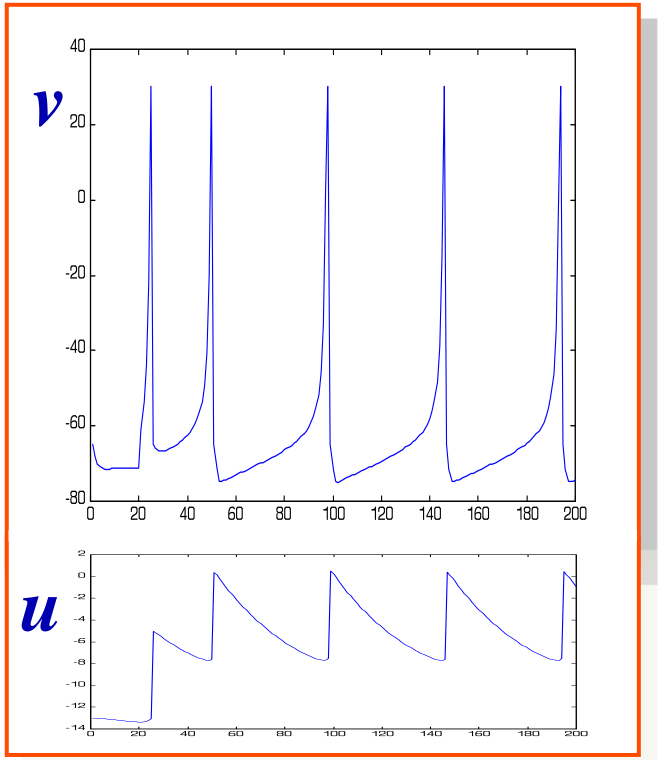
$$\dot{v} = 0.04v^2 + 5v + 140 - u + I$$
$$\dot{u} = a \cdot (bv - u)$$

  - neuron fires when

    $v > 30$; then:

$$v = c$$
$$u = u + d$$

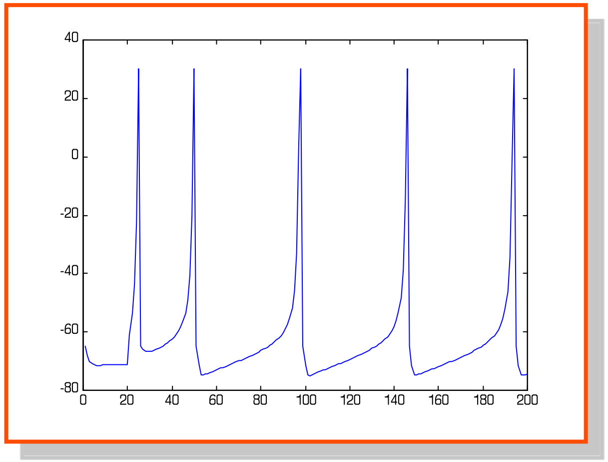
  - a, b, c & d select behaviour



(www.izhikevich.com)

Communication

- Spikes
  - biological neurons communicate principally via ‘spike’ events
  - asynchronous
  - information is only:
    - which neuron fires, and
    - when it fires
  - ‘Address Event’ Representation (AER)



Storage

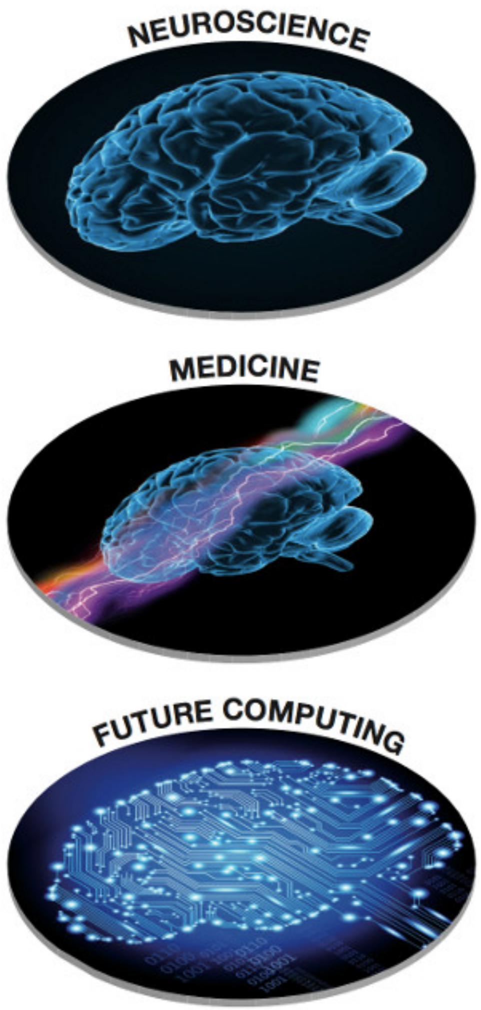
- Synaptic weights
  - stable over long periods of time
    - with diverse decay properties?
  - adaptive, with diverse rules
    - Hebbian, anti-Hebbian, LTP, LTD, ...
- Axon ‘delay lines’
- Neuron dynamics
  - multiple time constants
- Dynamic network states

The Human Brain Project

- An EU ICT Flagship project
  - headline €1B budget
    - €54M initial funding
      - 1st October 2013 to 31st March 2016
      - ~€900k to UoM
    - next 7.5 years funded under H2020
      - subject to review of ramp-up phase after 18 months
  - 80 partner institutes, 150 PIs & Cis
    - Open Call extended this
  - led by Henry Markram, EPFL

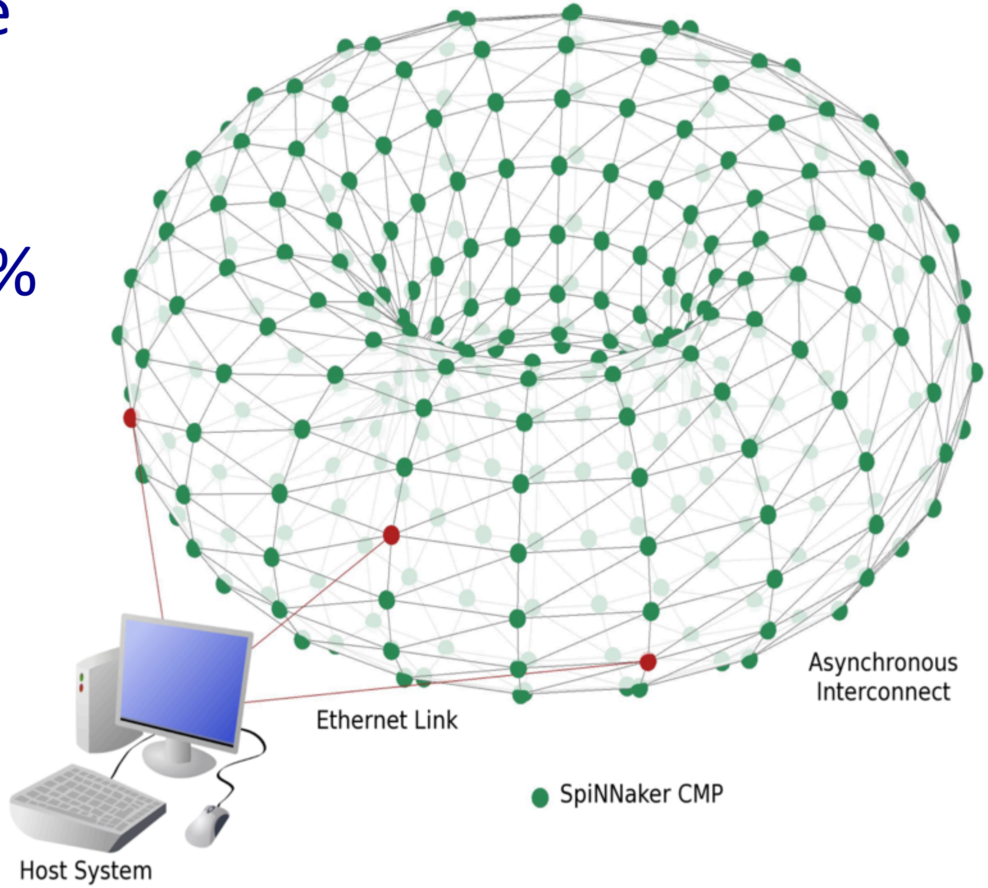
The Human Brain Project

- Research areas:
- Neuroscience
  - neuroinformatics
  - brain simulation
- Medicine
  - medical informatics
  - early diagnosis
  - personalized treatment
- Future computing
  - interactive supercomputing
  - neuromorphic computing



SpiNNaker project

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

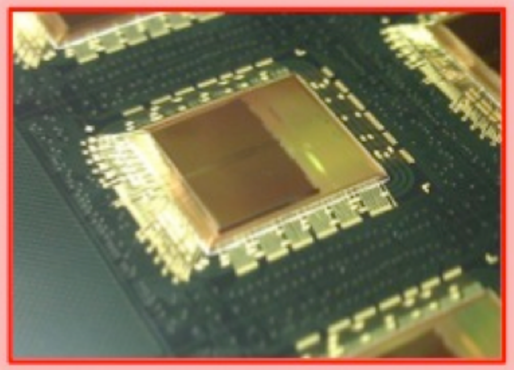


EPSRC

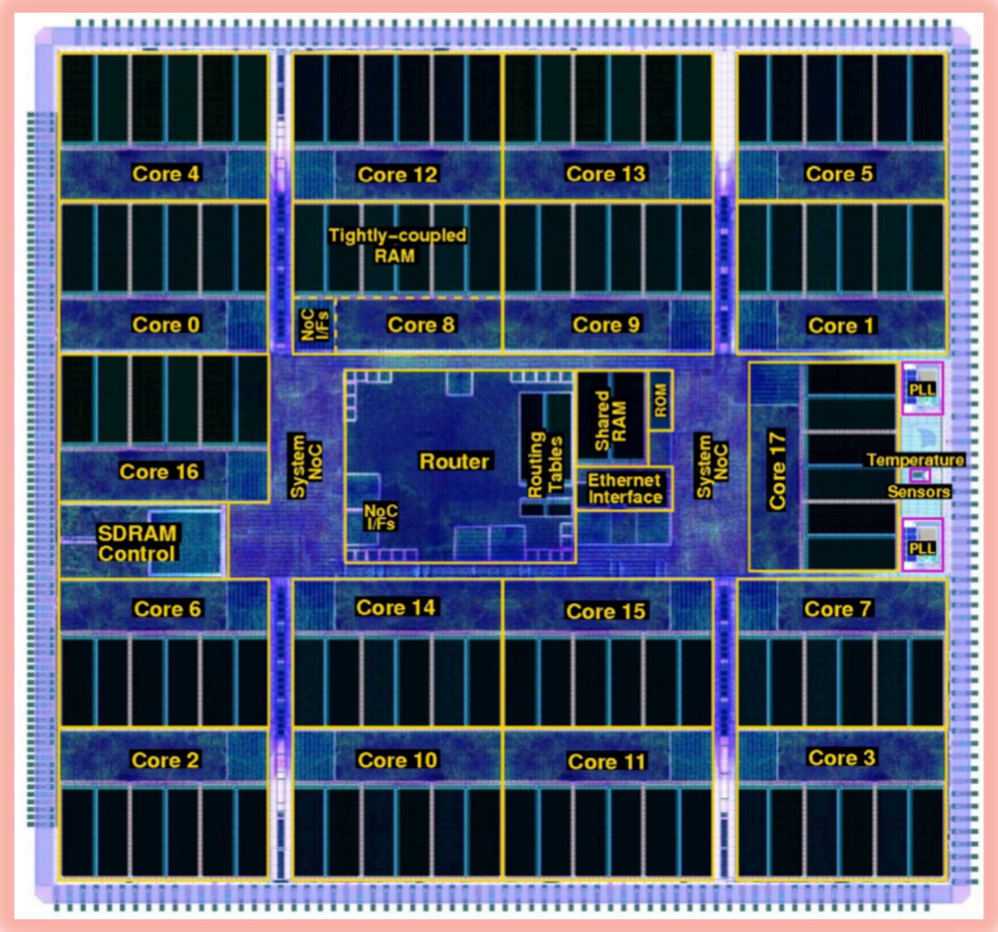
Design principles

- *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

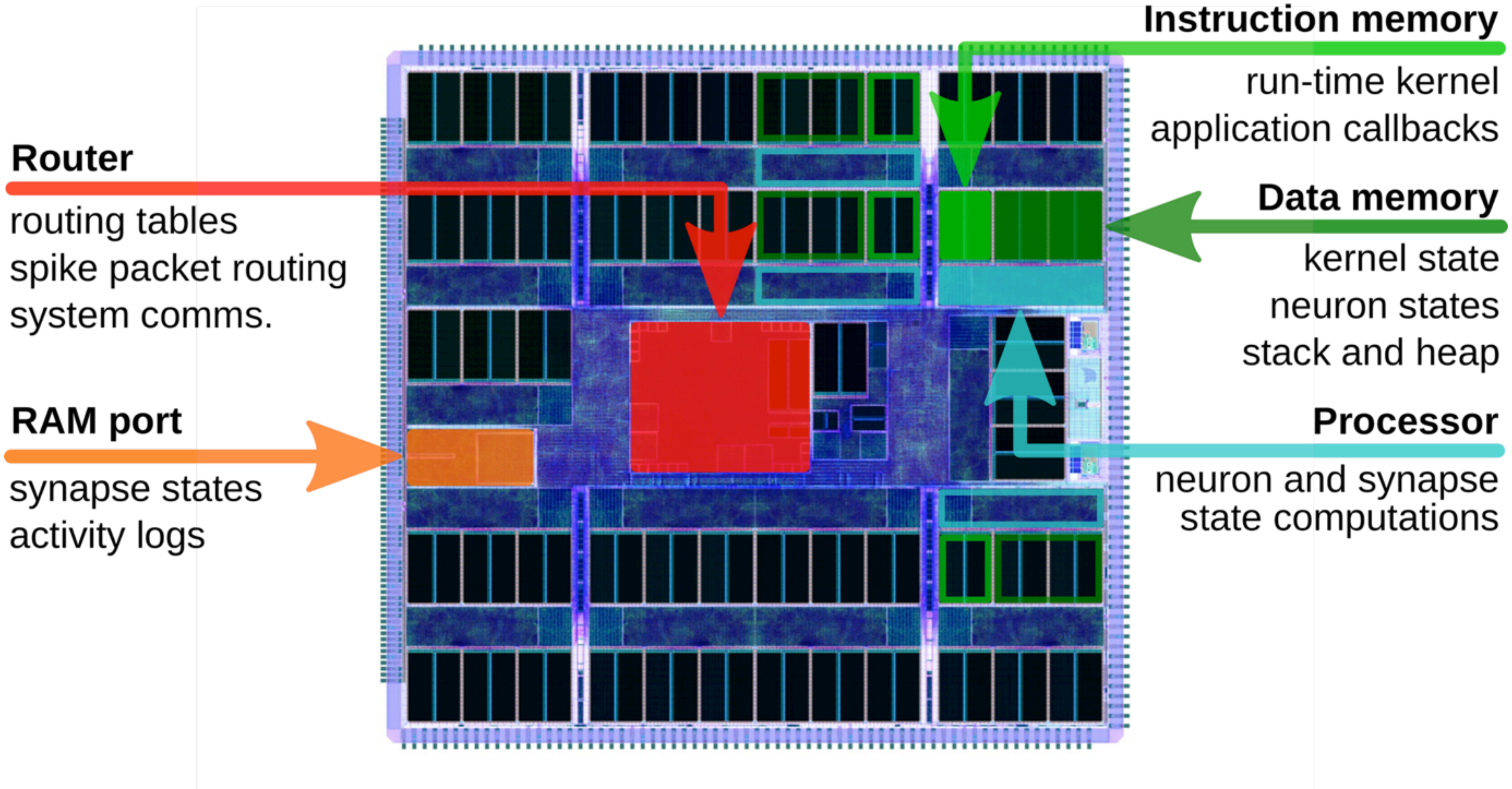
SpiNNaker chip



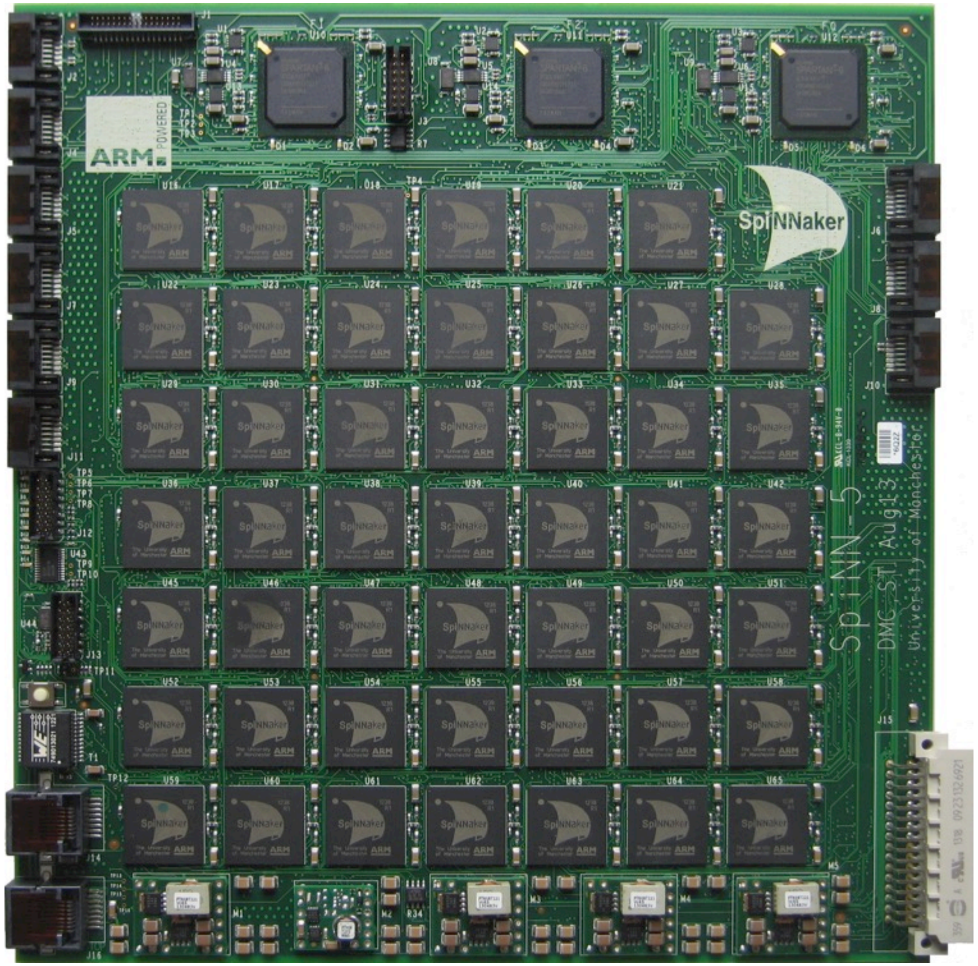
Multi-chip packaging by UNISEM Europe



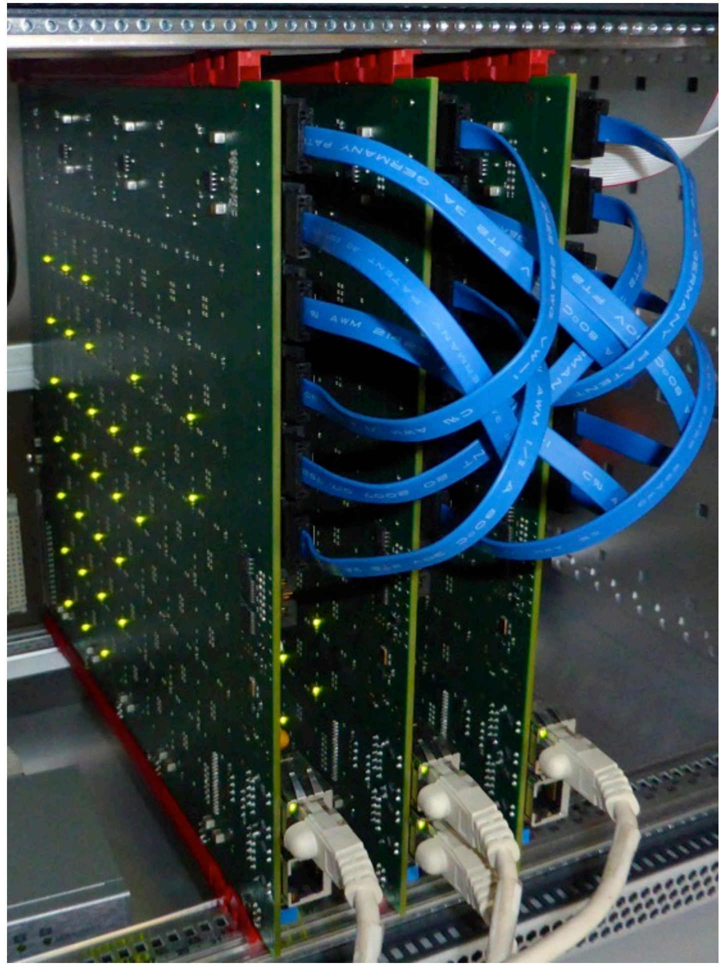
Chip resources



48-node PCB



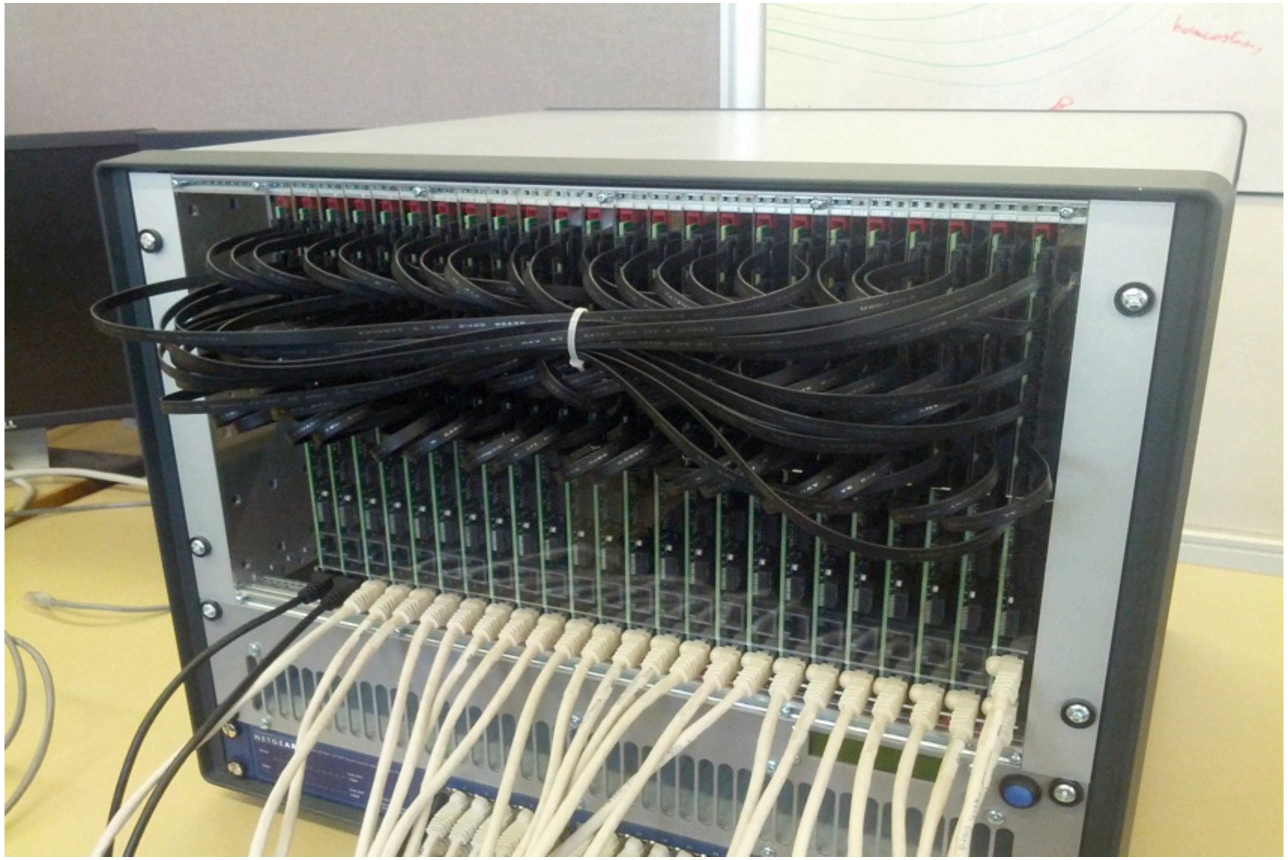
864 cores

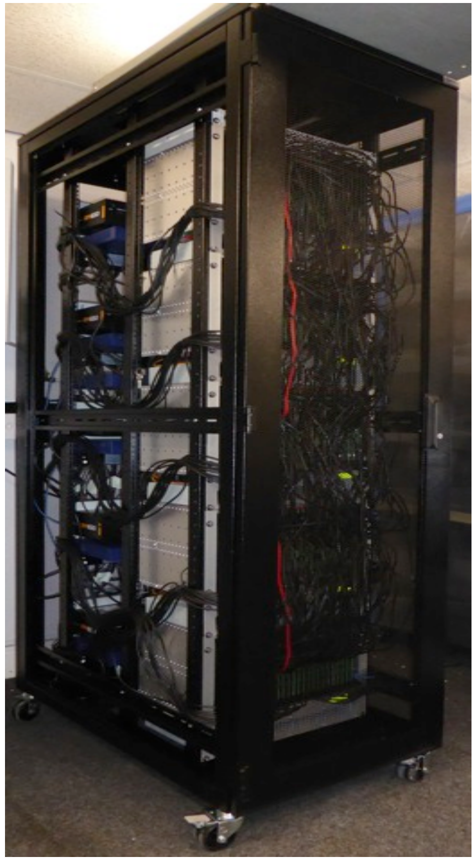


2,592 cores

SpiNNaker
Biologically Inspired Massively Parallel Architectures

SpiNNaker machines

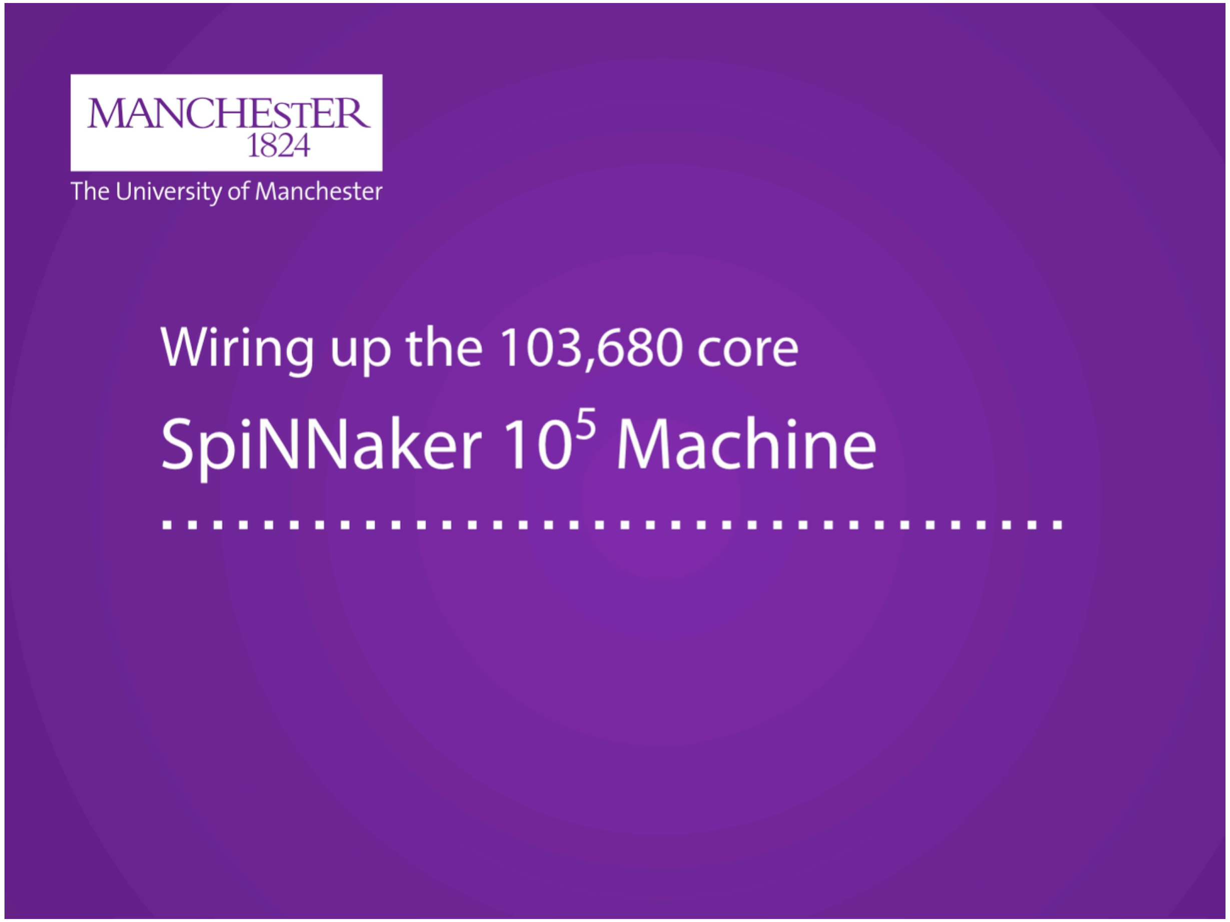




20,000 cores

100,000 cores

Building the 105 machine



Network – packets

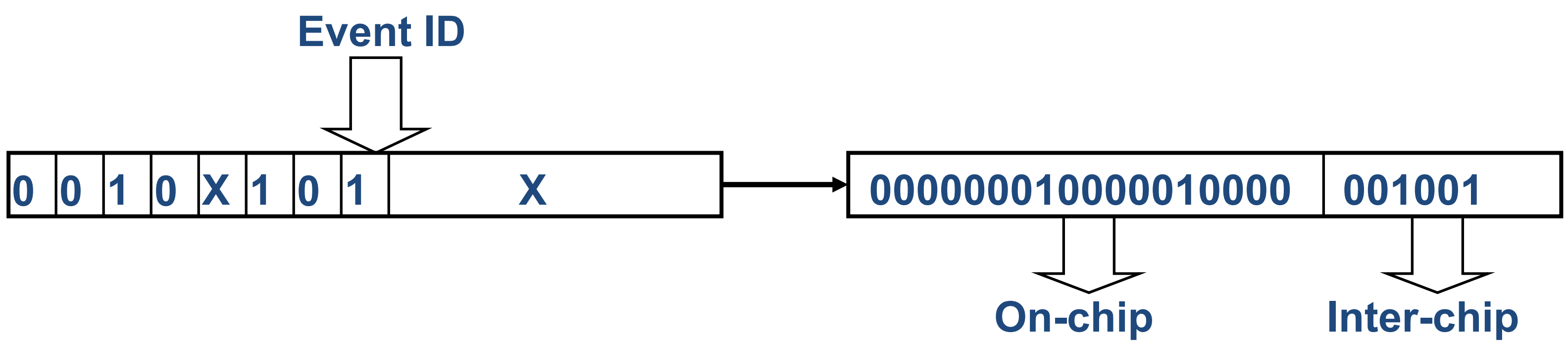
- Four packet types
  - MC (multicast): source routed; carry events (spikes)
  - P2P (point-to-point): used for bootstrap, debug, monitoring, etc
  - NN (nearest neighbour): build address map, flood-fill code
  - FR (fixed route): carry 64-bit debug data to host
- Timestamp mechanism removes errant packets
  - which could otherwise circulate forever

Header (8 bits)						Event ID (32 bits)
T	ER	TS	0	-	P	

Header (8 bits)						Address (16+16 bits)		Payload (32 bits)
T	SQ	TS	1	-	P	Dest	Srce	

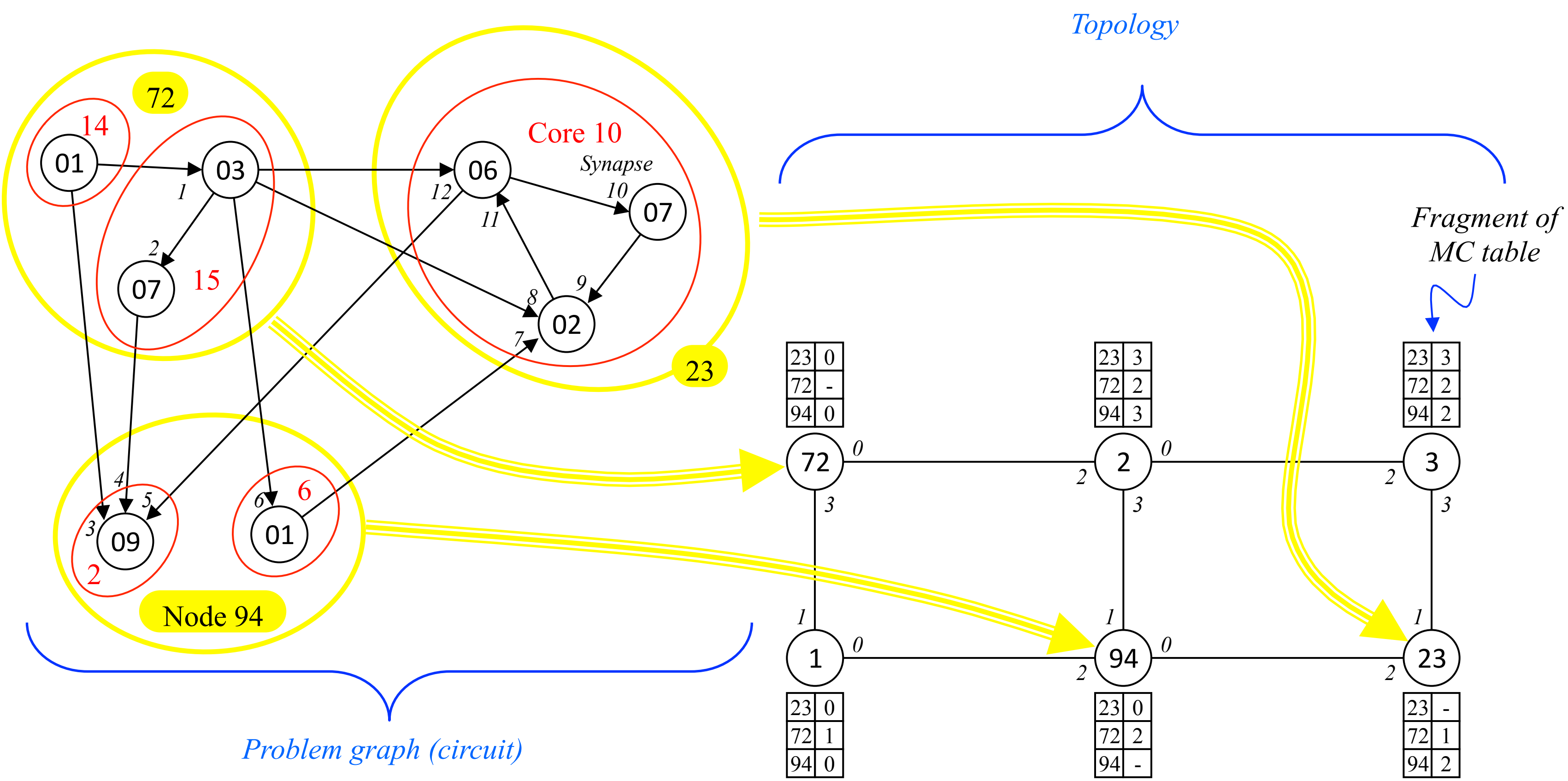
Network – MC Router

- All MC spike event packets are sent to a router
- Ternary CAM keeps router size manageable at 1024 entries (but careful network mapping also essential)
- CAM 'hit' yields a set of destinations for this spike event
  - automatic multicasting
- CAM 'miss' routes event to a 'default' output link



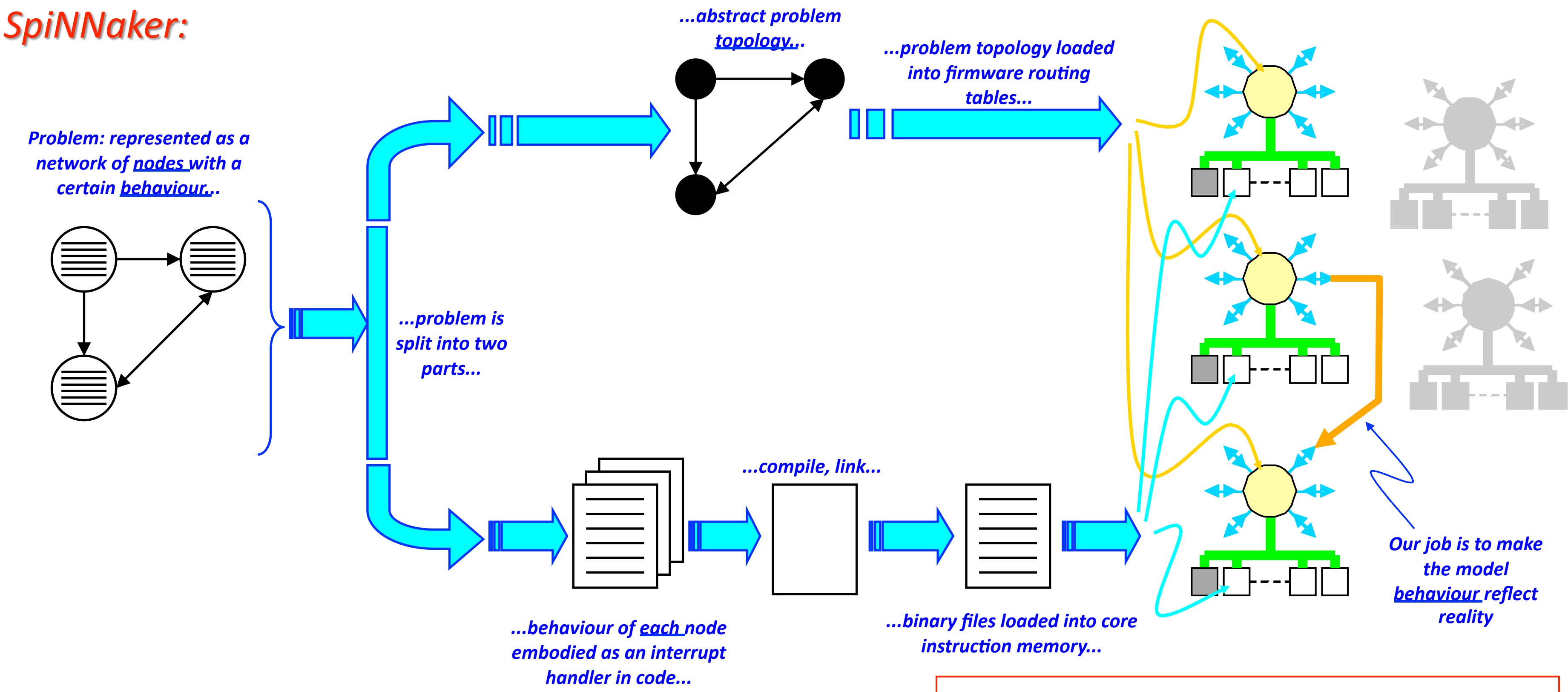
SpiNNaker
Biologically Inspired Massively Parallel Architectures

Topology mapping



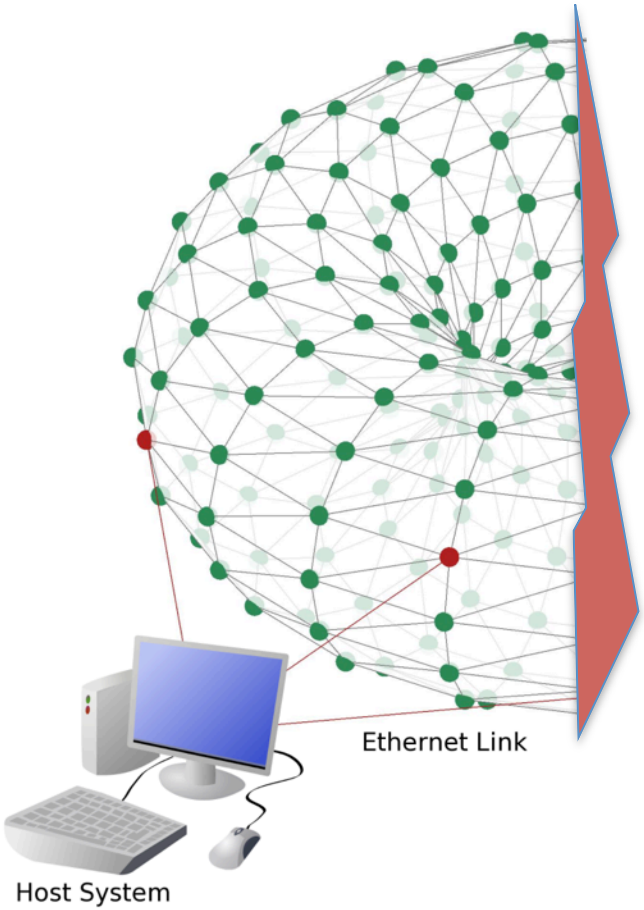
Problem mapping

SpiNNaker:

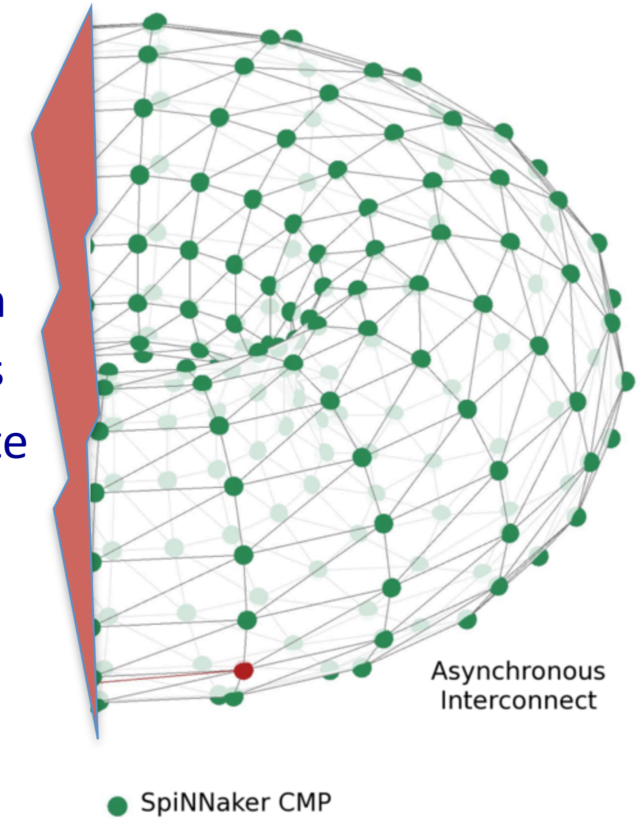


The code says "send message" but has no control where the output message goes

Bisection performance

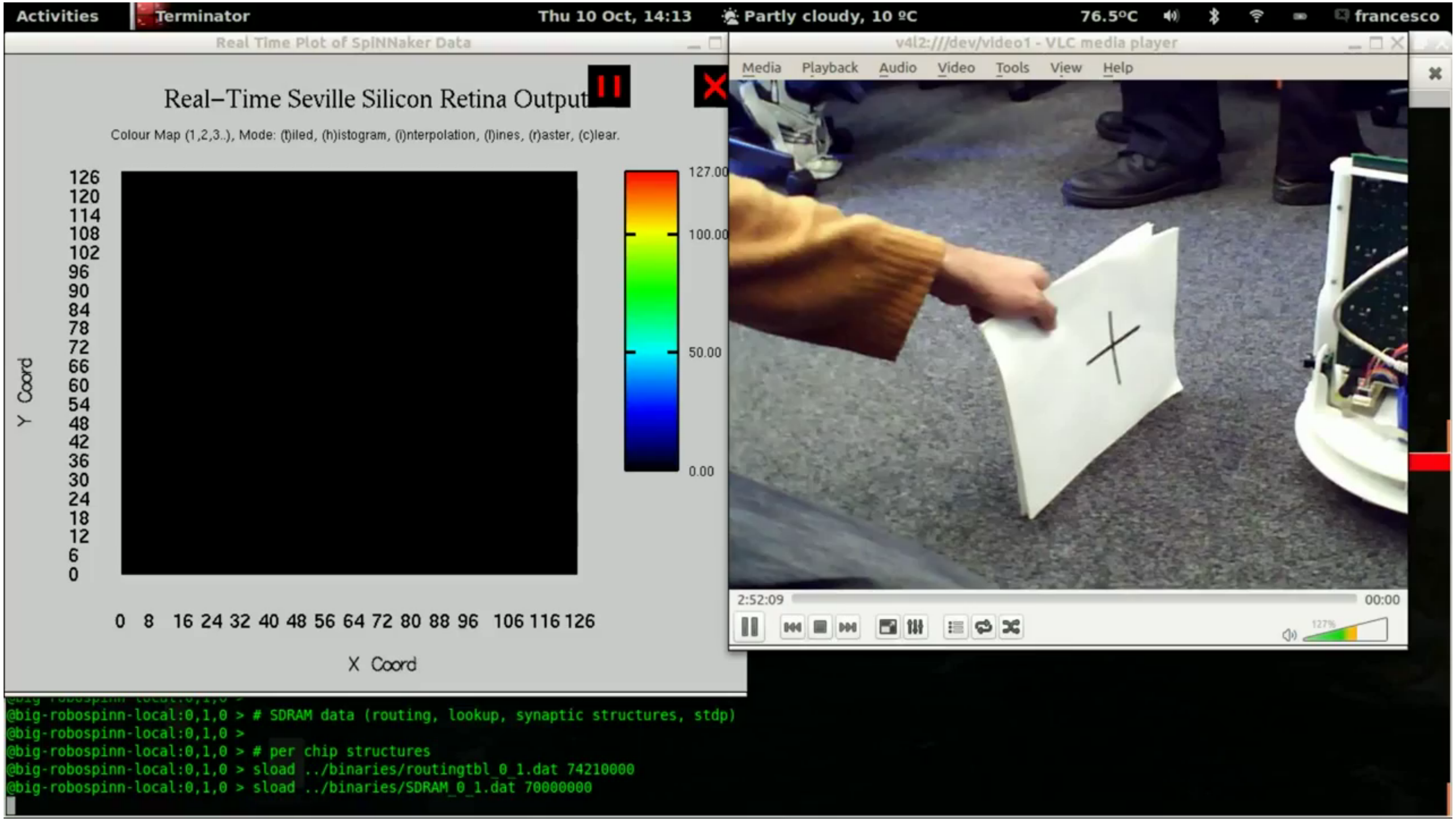


- 1,024 links
  - in each direction
- ~10 billion packets/s
- 10Hz mean firing rate
- 250 Gbps bisection bandwidth



SpiNNaker
Biologically Inspired Massively Parallel Architectures

SpiNNaker robot control



Conclusions

- We have come a long way in 60 years…
  - x10^{10} improvement in efficiency
- We still don't have the computer power to model the human brain
  - but we are getting there!
- Manchester is still building interesting machines…