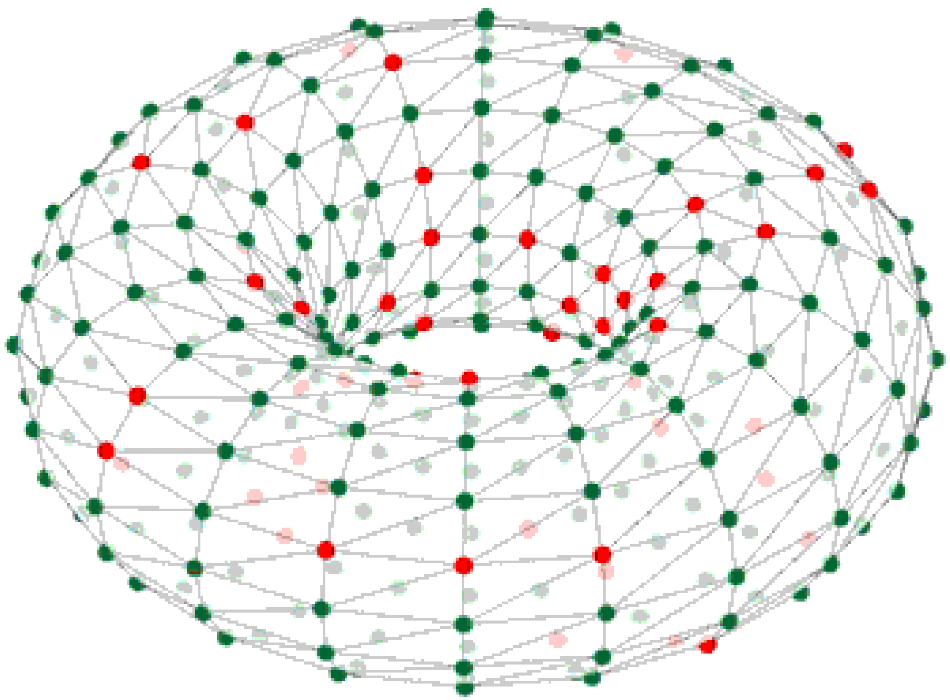
SpiNNaker – a Neuromorphic Supercomputer



Steve Temple – University of Manchester, UK

SOS21 - 21 Mar 2017

Outline of talk

- Introduction
- Modelling neurons
- Architecture and technology
- Principles of operation
- Summary / Future plans
- [Short (2 min) video]

So what is ...

- Neuromorphic computing / engineering?
  - Has its origins in the work of Carver Mead at Caltech in late 1980s
  - Originally coined for VLSI analogue systems to mimic neural biology
  - Latterly, any brain-inspired hardware computation system or technology
- SpiNNaker?
  - An experimental neuromorphic computing platform for modelling spiking neurons
  - Based around a custom multi-core ASIC
  - Scalable from a few chips to around 1 million cores
  - Targets large-scale brain models

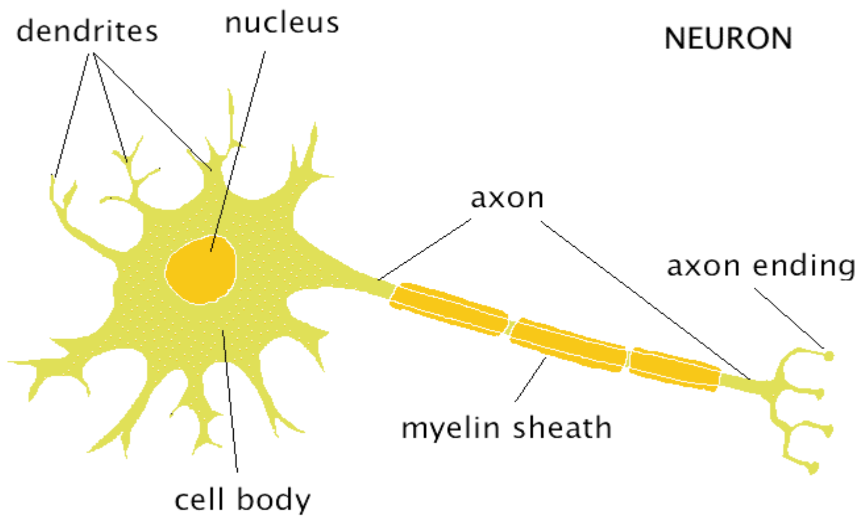
MANCHESTER 1824

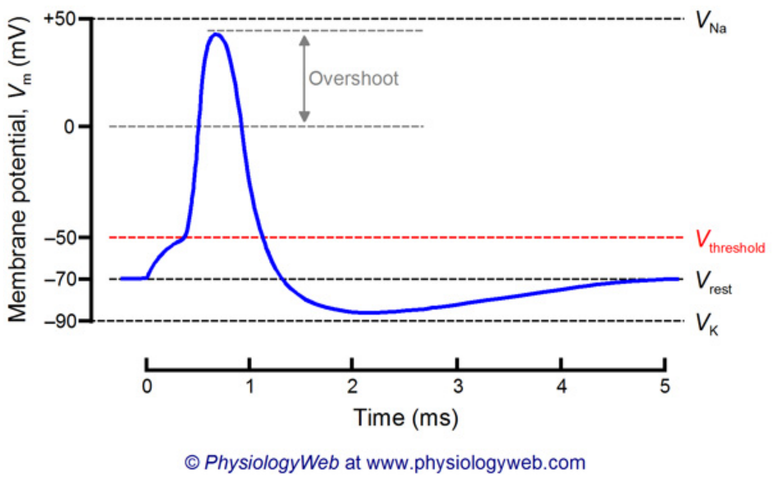
SpiNNaker Motivation

- 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?

Neurons

- Neurons are the single-celled building blocks of the brain (cf logic gates)
- Multiple inputs (dendrites)
- Single output (axon)
- High fanout (average 7000 in human brain)
- Communicate using spikes (or action potentials)

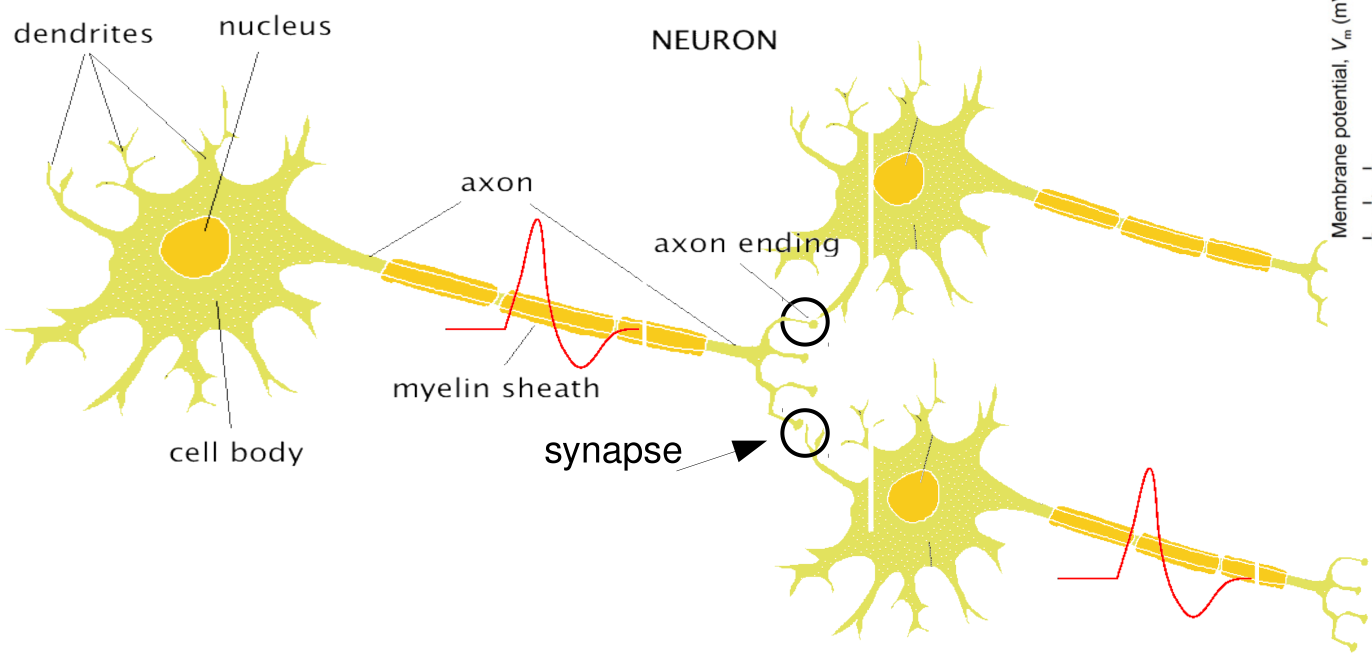


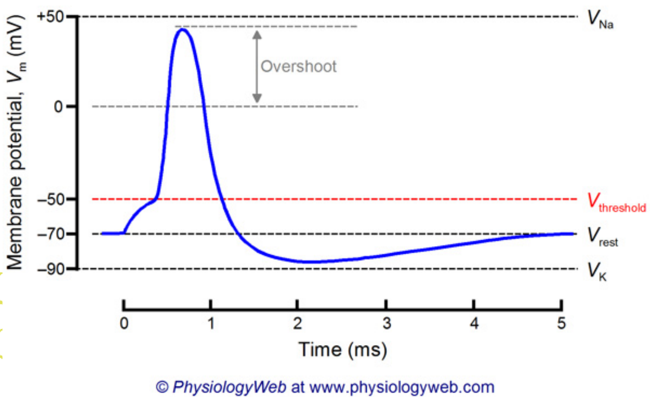


Brains

- Brain behaviour
  - Adaptive – tolerant of component failure
  - Perform tasks that conventional computers find hard
    - Autonomous learning
    - Complex pattern recognition
    - Inference
- Brain characteristics
  - Massive parallelism (10^{11} neurons in human)
  - Massive connectivity (10^{15} synapses in human)
  - Excellent power efficiency (~ 25W in human)
  - Low speed components (~ 10 Hz)
  - Low speed communication (~ metres/sec)

Neural Computation





- The cell body contains short-term state (membrane potential). A spike is output when a threshold is reached
- The membrane potential is modified by the arrival of spikes
- Neuron-to-neuron connection takes place at a “synapse”
- The effect of each spike is modulated by a “weight” in each synapse. The weights can change over time (learning!)
- There is a (N ms) delay as the spike traverses the axon

MANCHESTER 1824

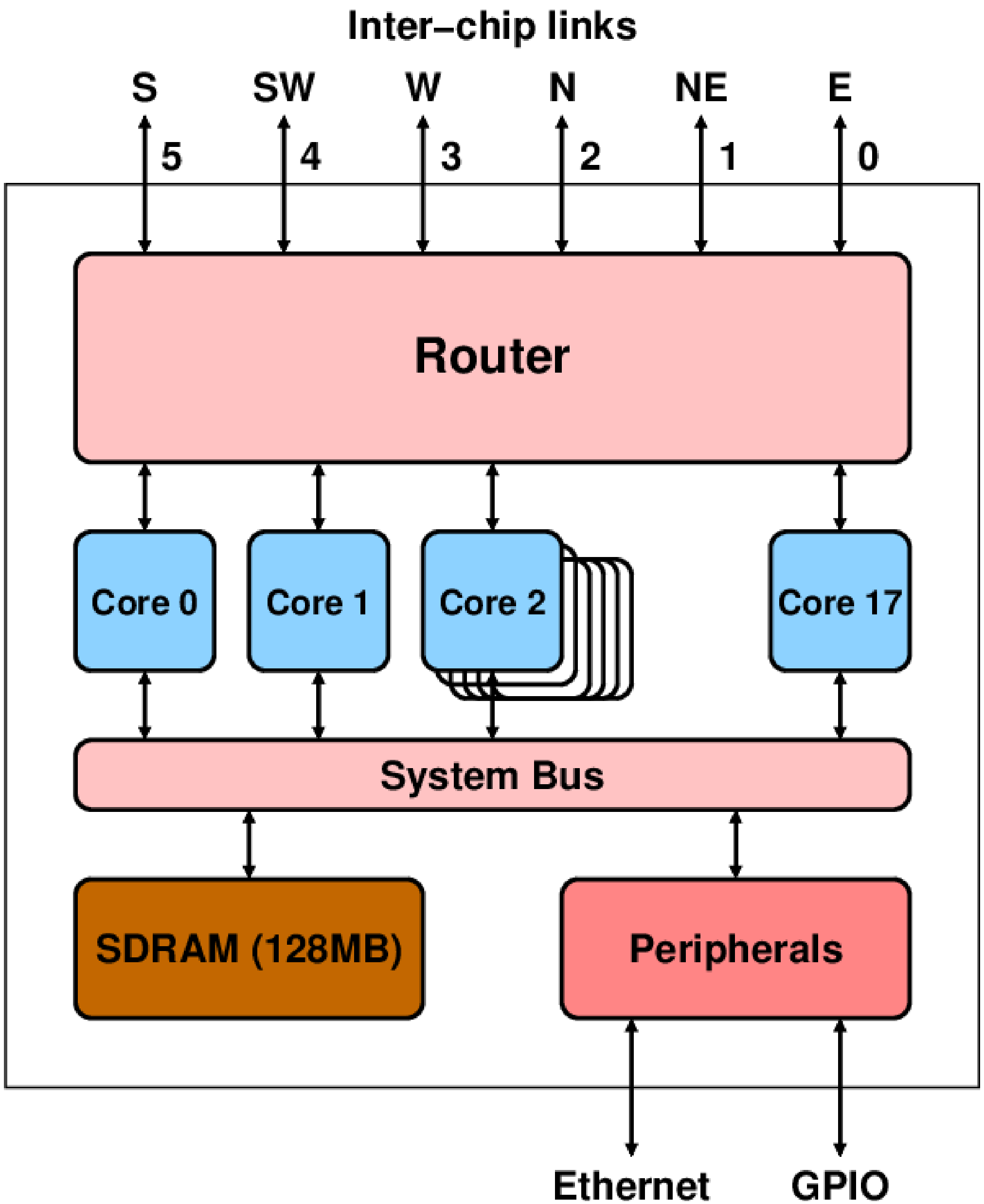
SpiNNaker is ...

- A scalable hardware/software architecture for modelling spiking neural networks in real-time
- Based around a custom chip ►
- Scalable – to 64K chips (> 10^6 cores)

- Special packet-routing hardware for modelling spiking networks

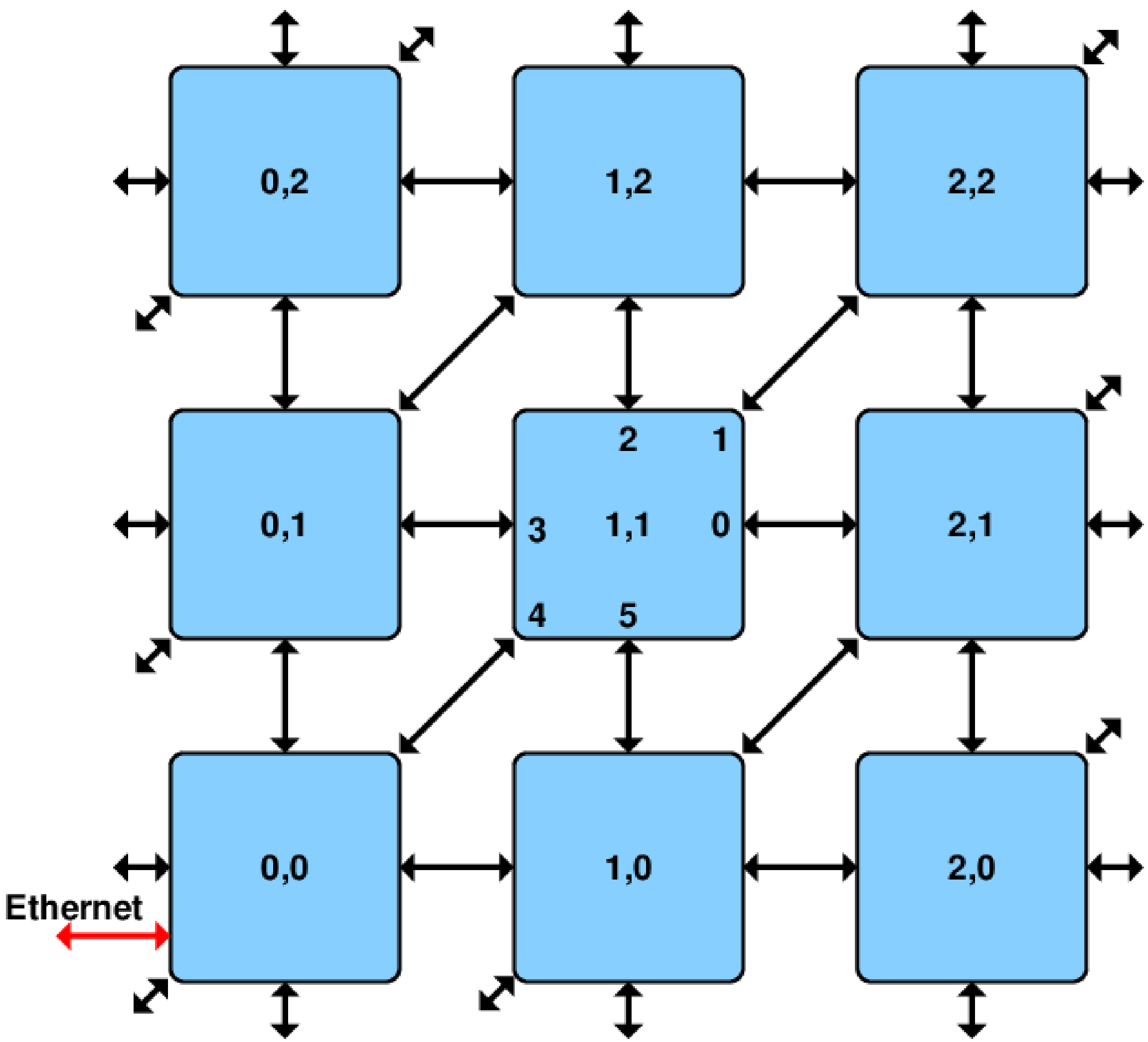
- Low power (around 1W per chip)
- Real-time – eg for robotic applications
- Flexible – mostly software – not tied to a particular modelling style or level

Chip Architecture



- 6 inter-chip links for packet communication (~6 M spikes/sec each)
- Spike packet router
- 18 ARM968 cores each with 96K SRAM and clocked at 200MHz
- 128MB SDRAM for neural weight storage
- Ethernet for system I/O

Chip Interconnection

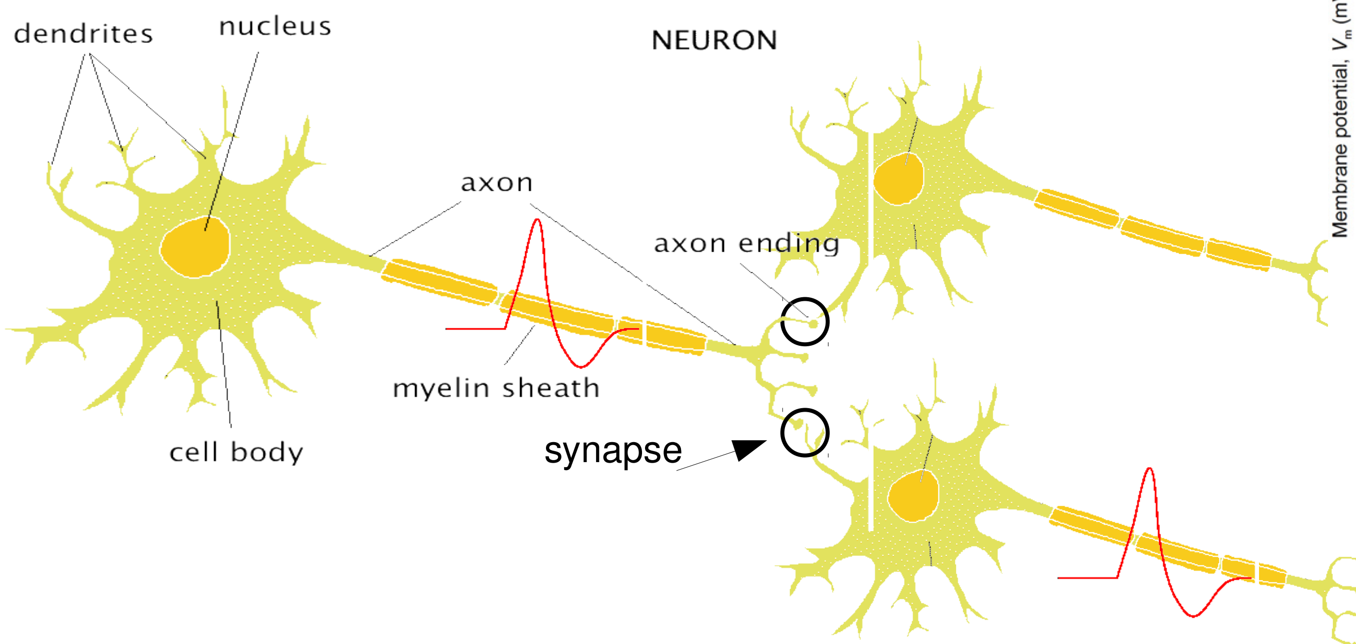


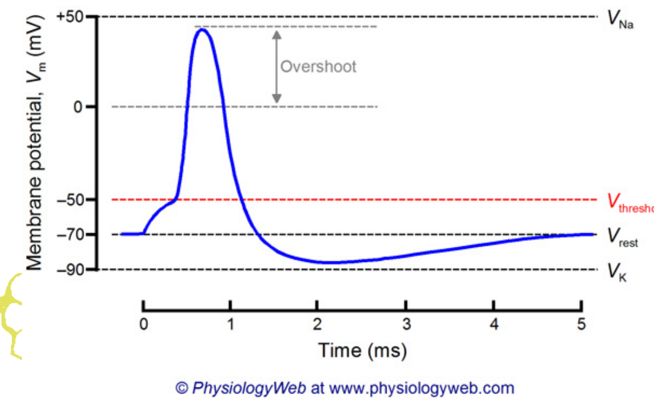
- Chips laid out in 2D grid and connected to 6 neighbours
- Scalable to 64K chips
- Edges can be wrapped to make a toroid
- Packets (spikes) pass from chip-to-chip via interchip links

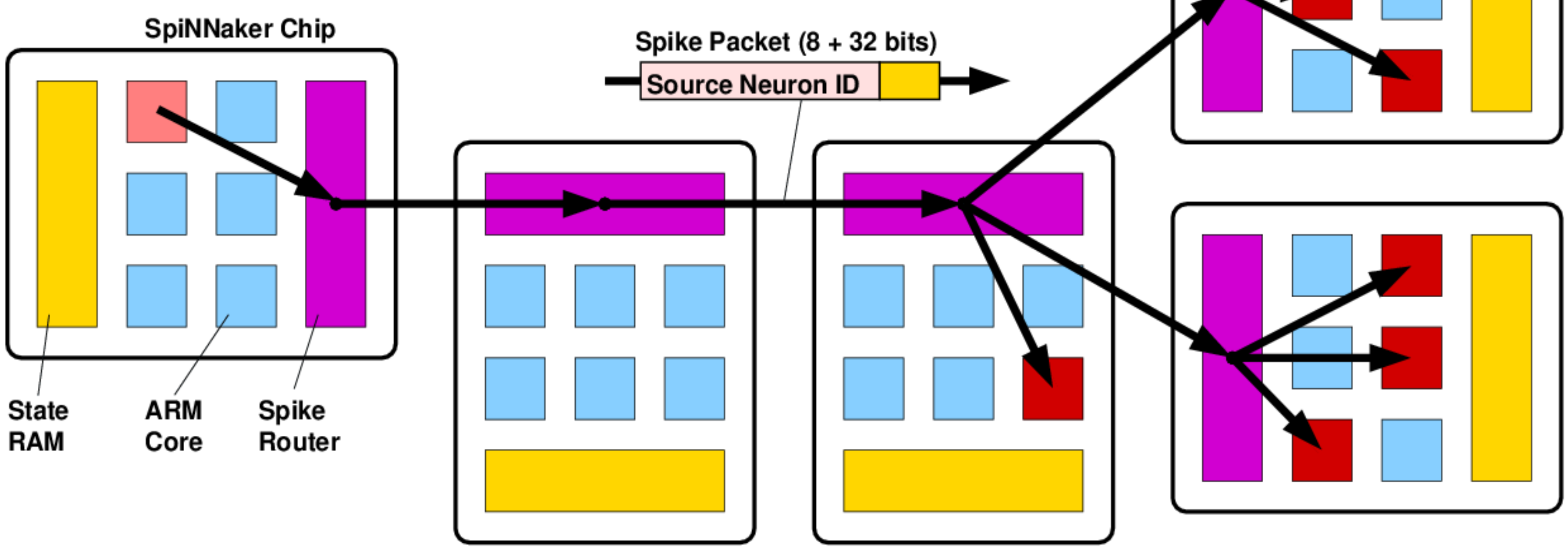
SpiNNaker Software Architecture

- Each SpiNNaker core models a number of neurons (typically 10 – 1000)
- A neuron model which "spikes" sends a packet containing the source ID of that neuron (source routing)
- A spike/packet is routed to all cores modelling neurons which connect to the spiking neuron
- Biological delays are modelled in software
- Only simple, abstract models can be efficiently implemented
- All state is kept in RAM
- All coding is done in C (and a little assembler)

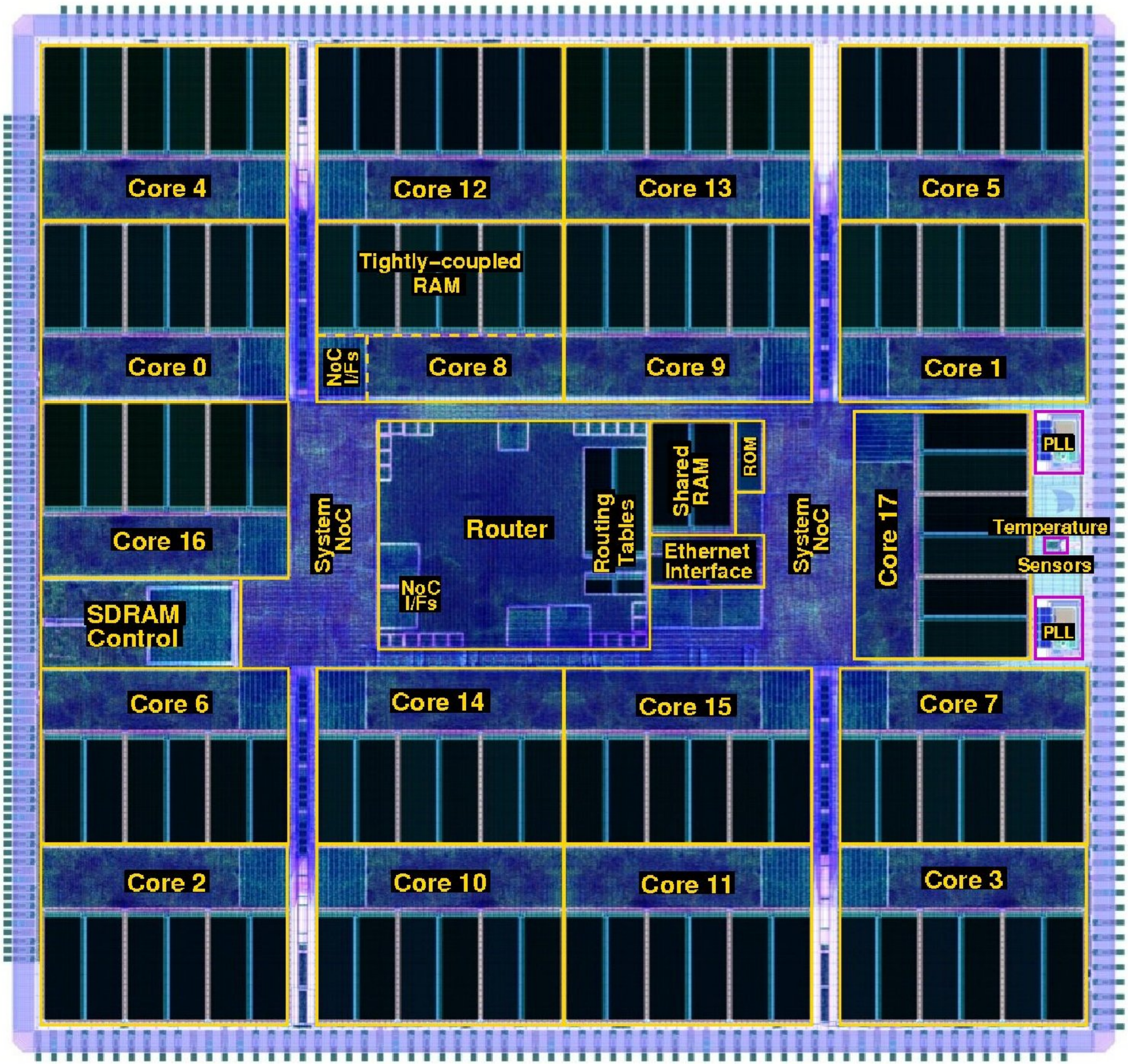
Spiking Neural Simulation





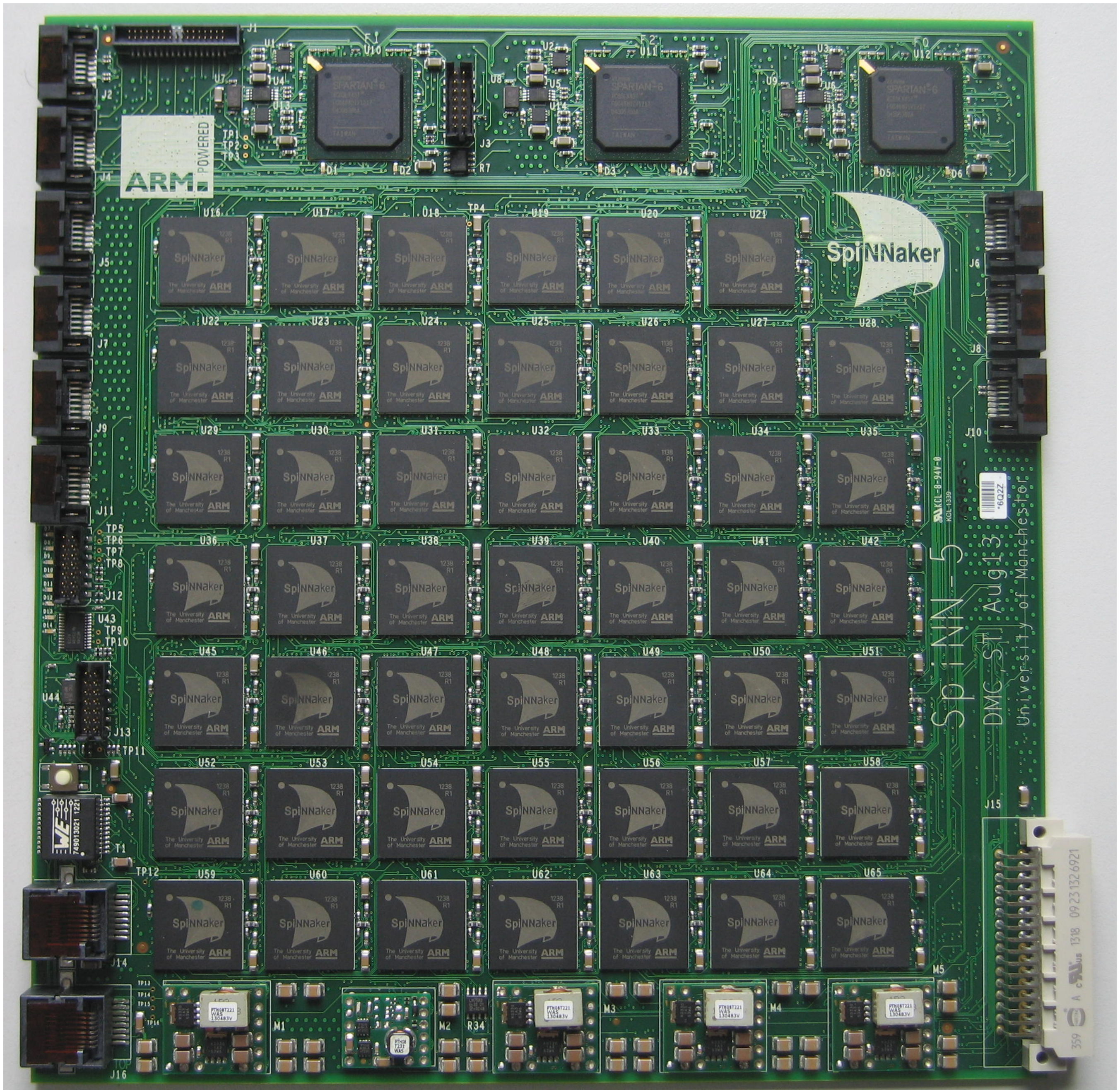


SpiNNaker Chip (Mar 2011)



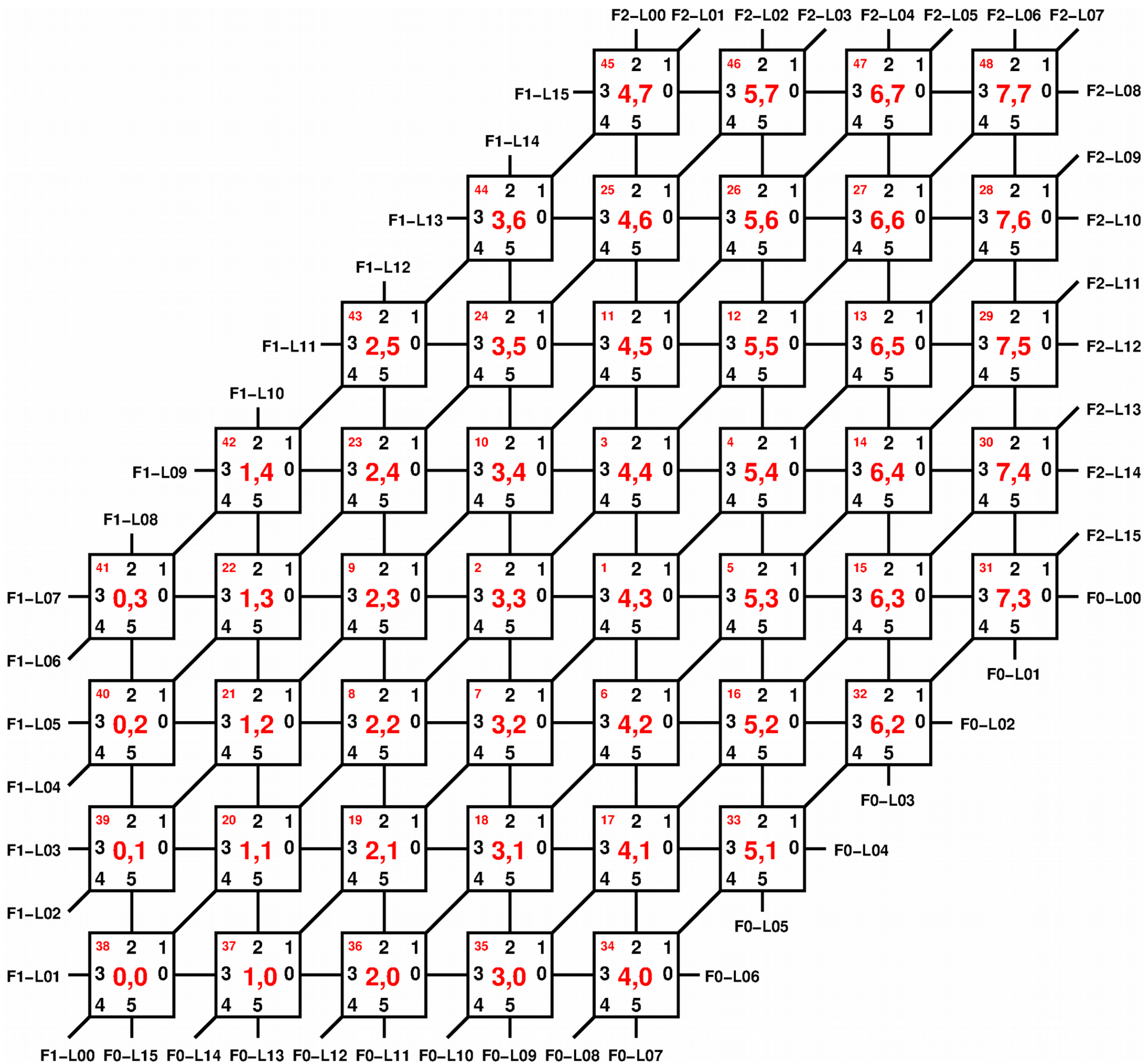
- 130nm process
- 10 x 10 mm
- 18 ARM9 cores each with 96K SRAM
- Spike packet router
- SDRAM controller
- Asynchronous GALS NoC
- External 128MB SDRAM in same package
- 300 pin plastic BGA package

SpiNN-5 PCB (Oct 2013)



- 48 SpiNNaker chips
- 3 FPGAs for inter-board links
-
- 9 * 3 Gbit/s 'SATA' links
- Backplane connector for power
- 2 * 100 Mbit/S Ethernet ports
- Board management processor (BMP)
- Max power consumption ~60W
- Standard Eurocard sizing

48 Chip Interconnect



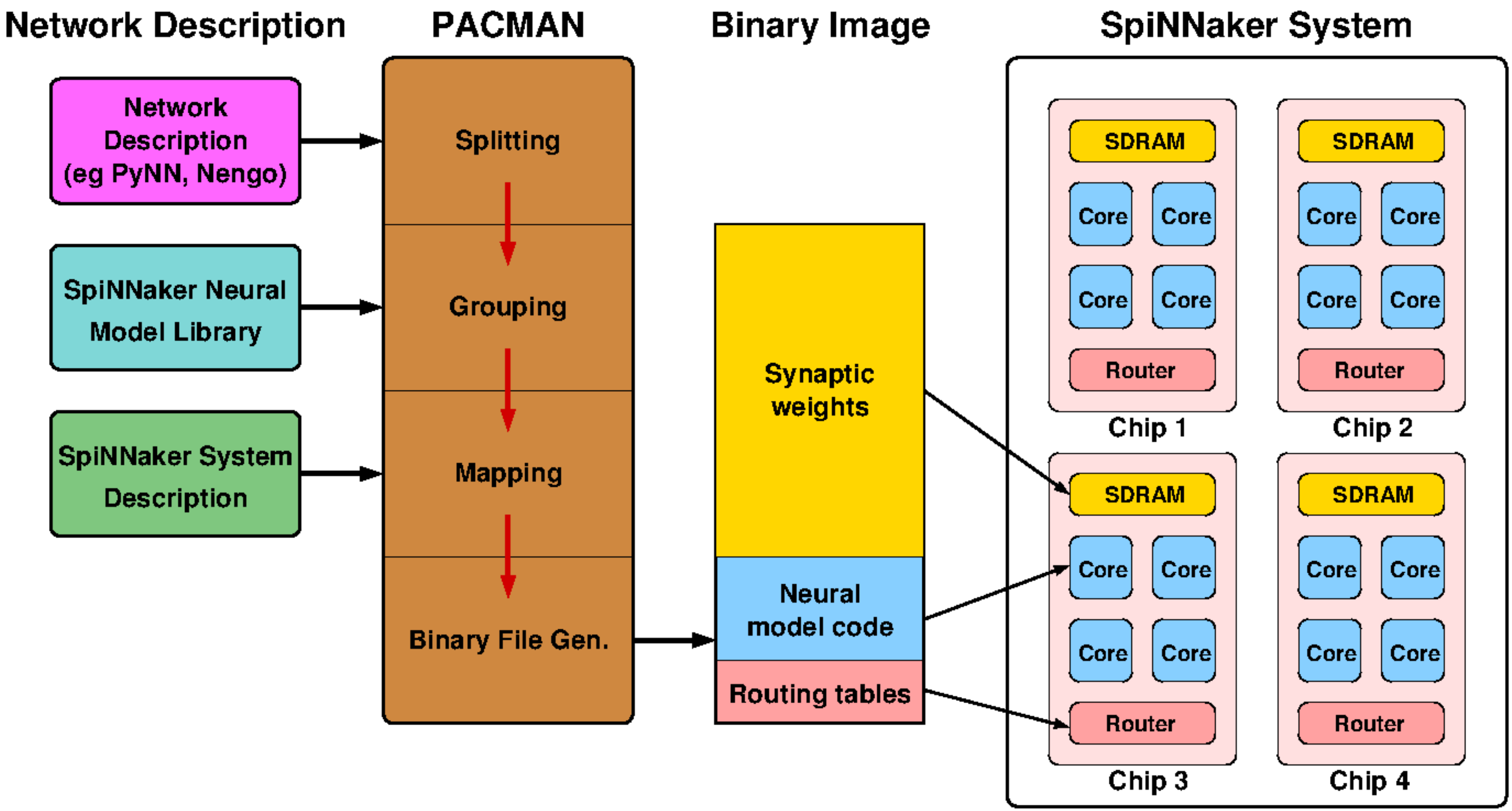
HBP System (Mar 2016)

- Human Brain Project deliverable and resource
- 600 SpiNN-5 PCBs
- 28,800 chips
- 518,400 cores
- 1800 SATA cables
- Max power ~ 40kW
- Forced air cooling
- Dedicated machine room
- Standard servers as front-end interface
- Accessible via HBP portal for job submission

Application Software

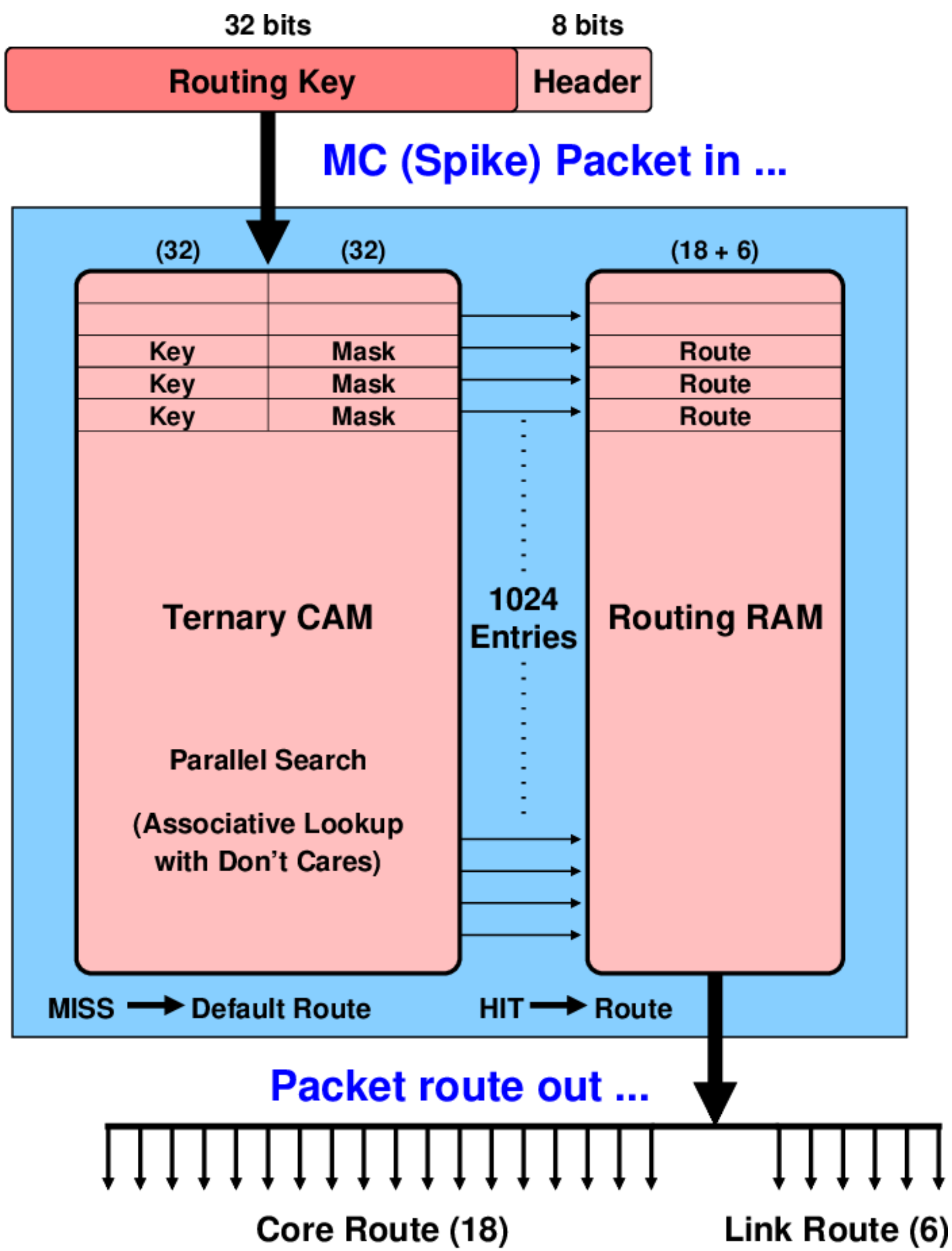
- Neural models running on SpiNNaker are coded in C
- Many neuroscientists don't do C (or any form of programming...)
- Must provide a way for them to describe a neural network and simulate it in SpiNNaker
- We provide a library of models exposed via high-level neural network descriptions
- PyNN is a standard language to describe networks and interact with neural simulators
- We map PyNN (and others) onto SpiNNaker

PACMAN



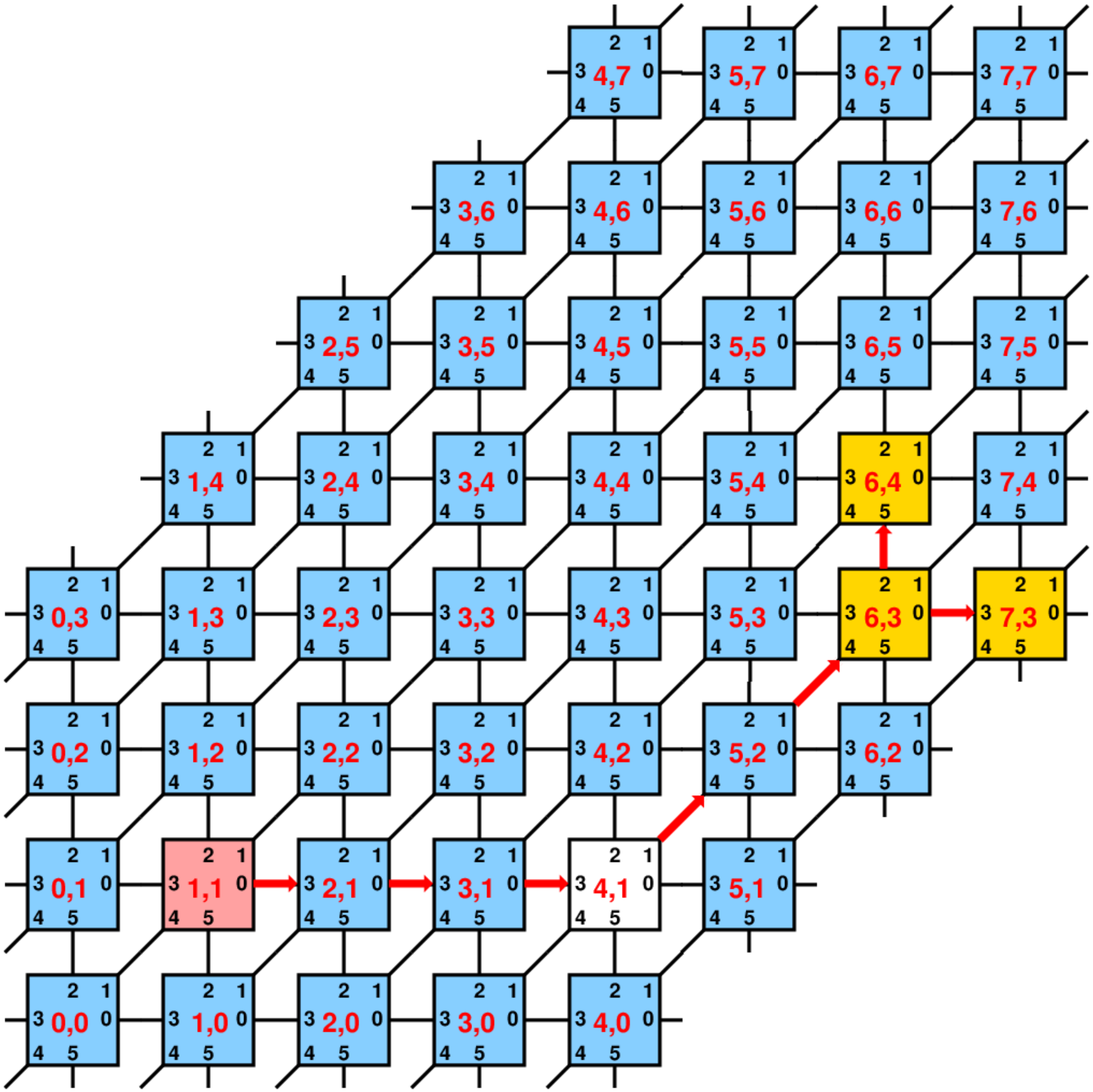
PACMAN currently runs on a workstation

Spike Packet Router



- Single input packet (spike) can replicate to 0-24 output ports
- CAM-based lookup based on routing key (source neuron ID)
- CAM implementation limited to 1024 entries
- Don't-care states in CAM bits allows output routes to be shared
- Network configuration software responsible for setting routing table
- Efficient use of CAM entries essential to stay within 1024 entries
- Routing delay ~200ns per chip

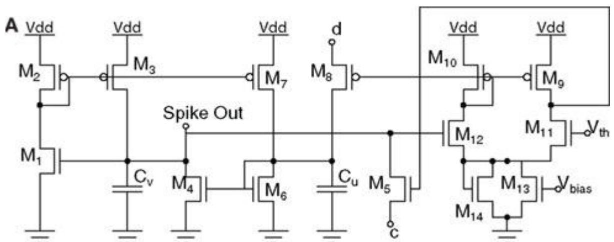
Spike Packet Routing



MANCHESTER 1824

Alternative Approaches

Analogue Neurons



+ Very low power

+ Neuron in a few tens of transistors

+ Fast

- Inflexible - model fixed in silicon

- Spike routing difficult

- Non-deterministic – hard to calibrate

Supercomputer

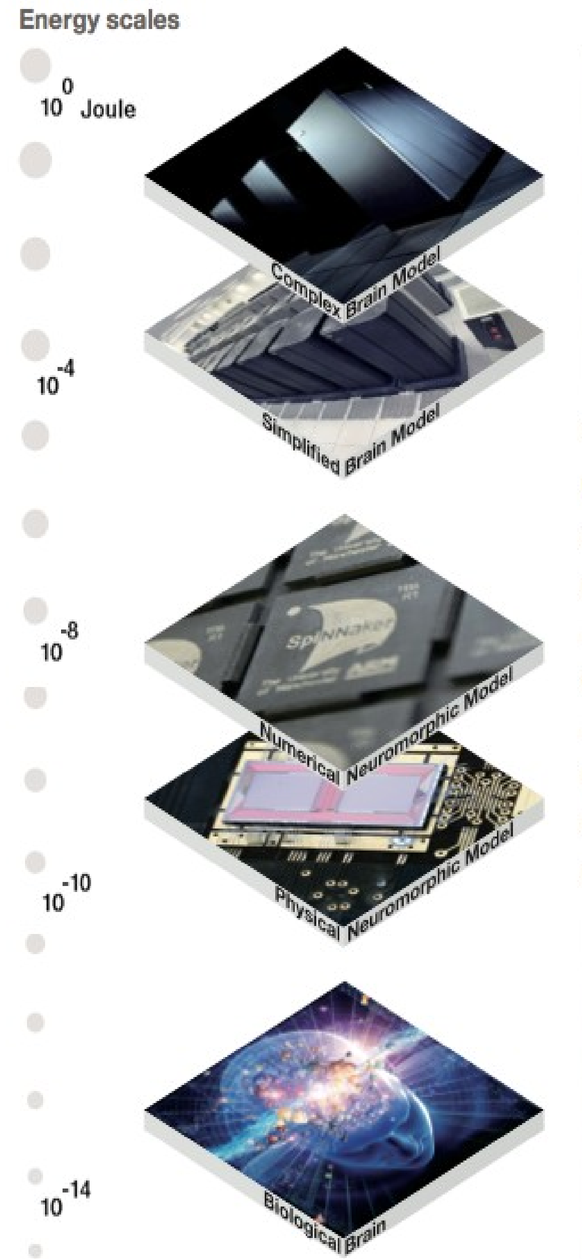
+ Flexible – model at many levels

+ Deterministic – results reproducible

- Expensive

- High power

- Spike routing slow



Summary / Future Plans

- After 10 years we have a large-scale SpiNNaker machine and over 100 smaller systems deployed worldwide
- Users are exploring various application spaces
  - Brain modelling
  - Deep networks (using spikes)
  - Real-time Spaun (U. of Waterloo brain model)
  - Mobile apps (untethered speech/image recognition)
- SpiNNaker-2 in design phase - aiming for 10x improvement

SpiNNaker Video

https://www.youtube.com/watch?v=z1_gE_ugEgE