SpiNNaker

a spiking neural network architecture

Sergio Davies
The University of Manchester
daviess@cs.man.ac.uk
http://apt.cs.man.ac.uk/people/daviess
Bio-inspiration

• How can massively parallel computing resources accelerate our understanding of brain function?
• How 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
  - 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’
**SpiNNaker project**

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

- **Virtualised topology**
  - physical and logical connectivity are decoupled
- **Bounded asynchrony**
  - time models itself
- **Energy frugality**
  - the real cost of computation is energy
  - energy for neural modelling vs. comms
SpiNNaker system
SpiNNaker chip

Mobile DDR SDRAM interface

Multi-chip packaging by UNISEM Europe
The networking challenge

- Emulate the very high connectivity of real neurons
- A spike generated by a neuron firing must be conveyed efficiently to >1,000 inputs
- On-chip and inter-chip spike communication should use the same delivery mechanism
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
• 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
Topology mapping

Problem graph (circuit)

Node 94

Core 10
Synapse 10

Fragment of MC table
Problem mapping

SpiNNNaker:

- Problem: represented as a network of nodes with a certain behaviour...
- Problem topology...
- Problem is split into two parts...
- Abstract problem topology...
- Compile, link...
- Behaviour of each node embodied as an interrupt handler in code...
- Binary files loaded into core instruction memory...
- Problem topology loaded into firmware routing tables...

Our job is to make the model behaviour reflect reality.

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
Event-driven software model
PyNN design flow
PyNN integration

- LIF

- Izhikevich
Synaptic plasticity

Output spike delay (msec)

Simulation time (sec)

Potentiation: Pre-synaptic Leads Post-synaptic

Depression: Pre-synaptic Lags Post-synaptic
SpiNNNaker robots
48-node PCB
Hexagonal PCB structure

2x 3.1 Gbps SATA links

3-board basic unit:
SpiNNaker platforms
**SpiNNaker machines**

103 machine: 864 cores, 1 PCB, 75W

105 m/c: 103,680 cores, 1 cabinet, 9kW

104 machine: 10,368 cores, 1 rack, 900W
(NB 12 PCBs for operation without aircon)

106 m/c: 1M cores, 10 cabs, 90kW
Where are we going?

- The original project design included 1 billion neuron on a massively parallel (10^6) machine;

- New 3D chip, 28 nanometers technology – under evaluation.
Conclusions

- Brains represent a significant computational challenge
  - now coming within range?

- **SpiNNaker** is driven by the brain modelling objective
  - virtualised topology, bounded asynchrony, energy frugality

- The major architectural innovation is the multicast communications infrastructure

- Self-tuning at many levels
  - hardware (for fault-tolerance), software and, most effectively, in the neurons themselves!

- We have working hardware!
Acknowledgements