Cheap as Chips!

Steve Furber

ICL Professor of Computer Engineering

The University of Manchester

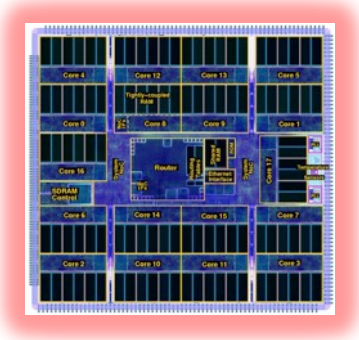
Human Brain Project

HBP Education Programme

Human Brain Project

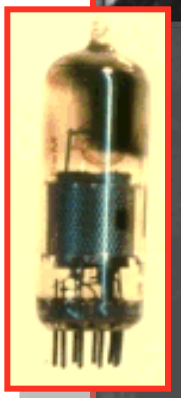
Outline

- a Manchester perspective
- 63 years of progress
- CMOS power consumption
- the future of chip design
- conclusions



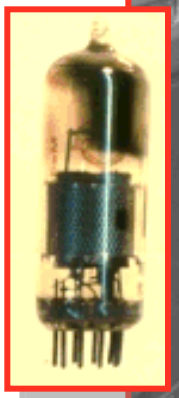
Human Brain Project

Baby (1948)



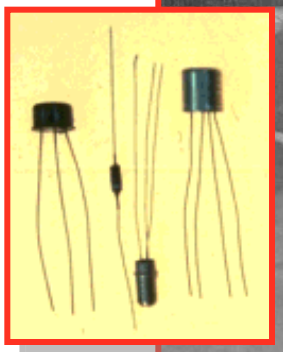
Human Brain Project

The Mark1 (1950s)



Human Brain Project

Atlas (1960s)

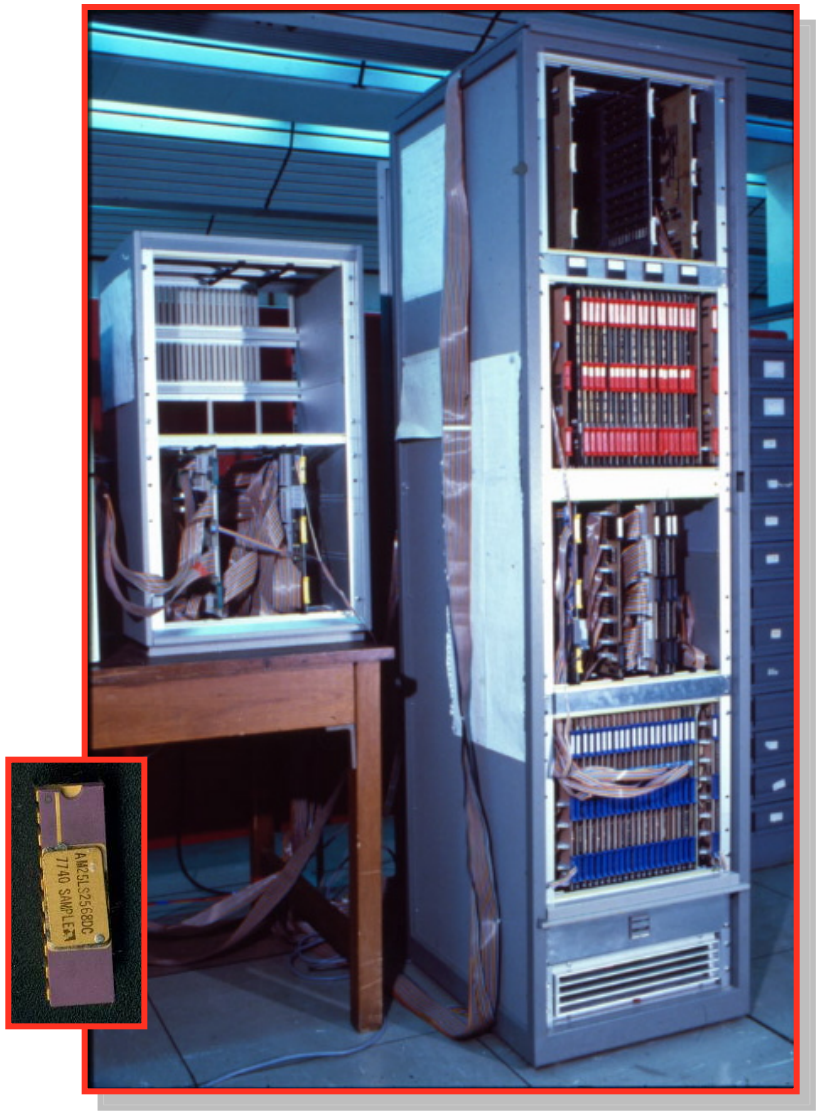


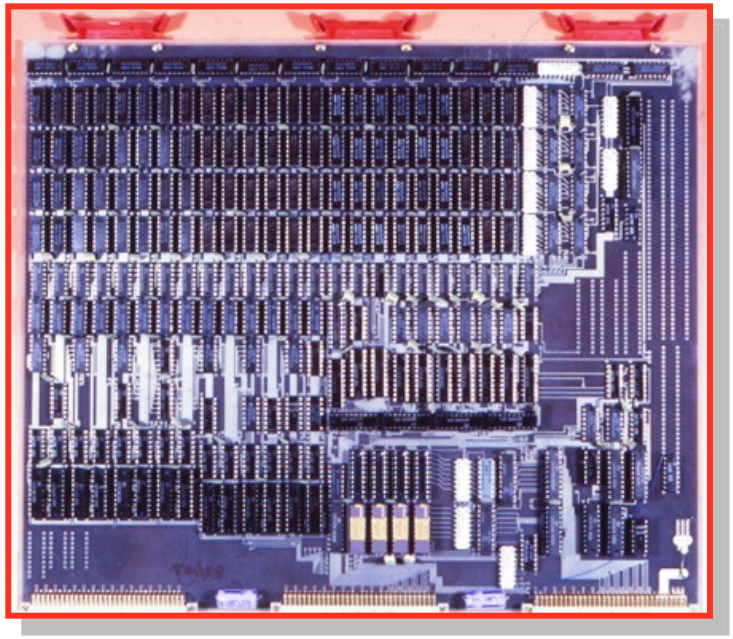
Human Brain Project

MU5 (1970s)

Human Brain Project

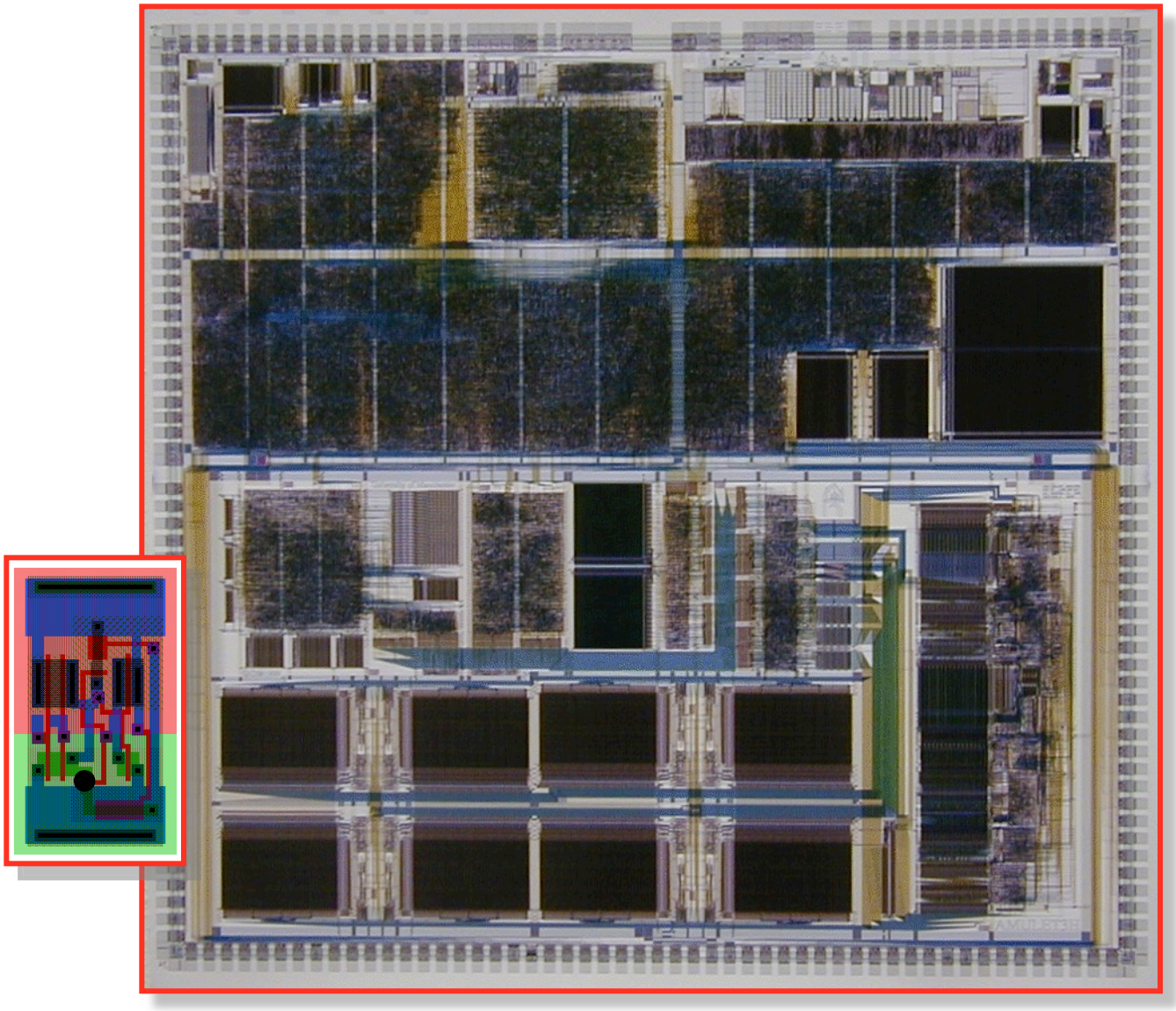
Dataflow (1980s)





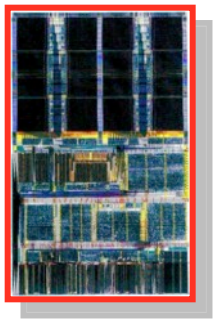
Human Brain Project

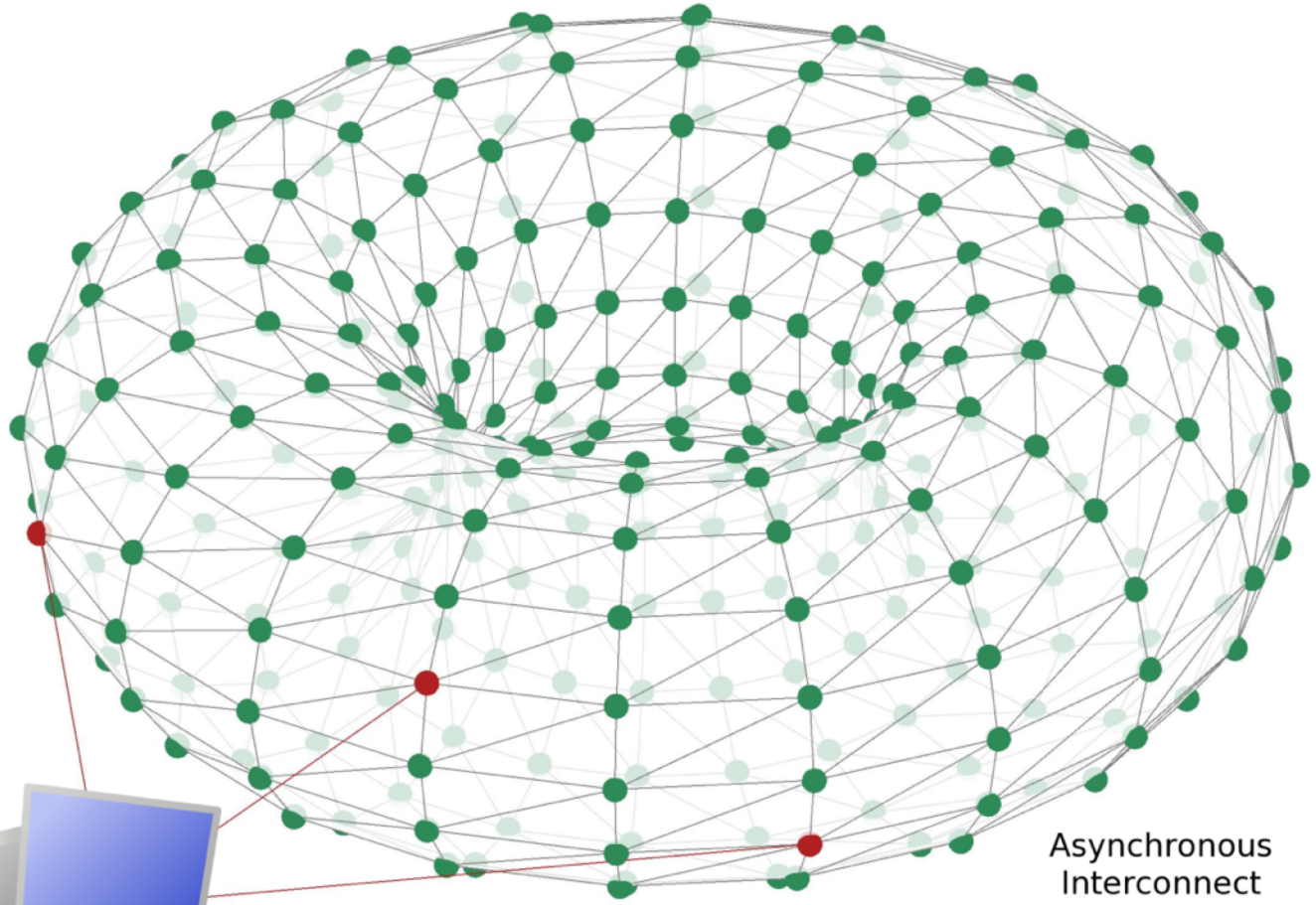
Amulet (1990s)



Human Brain Project

SpiNNaker (2000s)

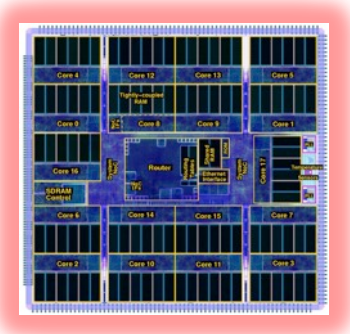




Human Brain Project

Outline

- a Manchester perspective
- 63 years of progress
- CMOS power consumption
- the future of chip design
- conclusions

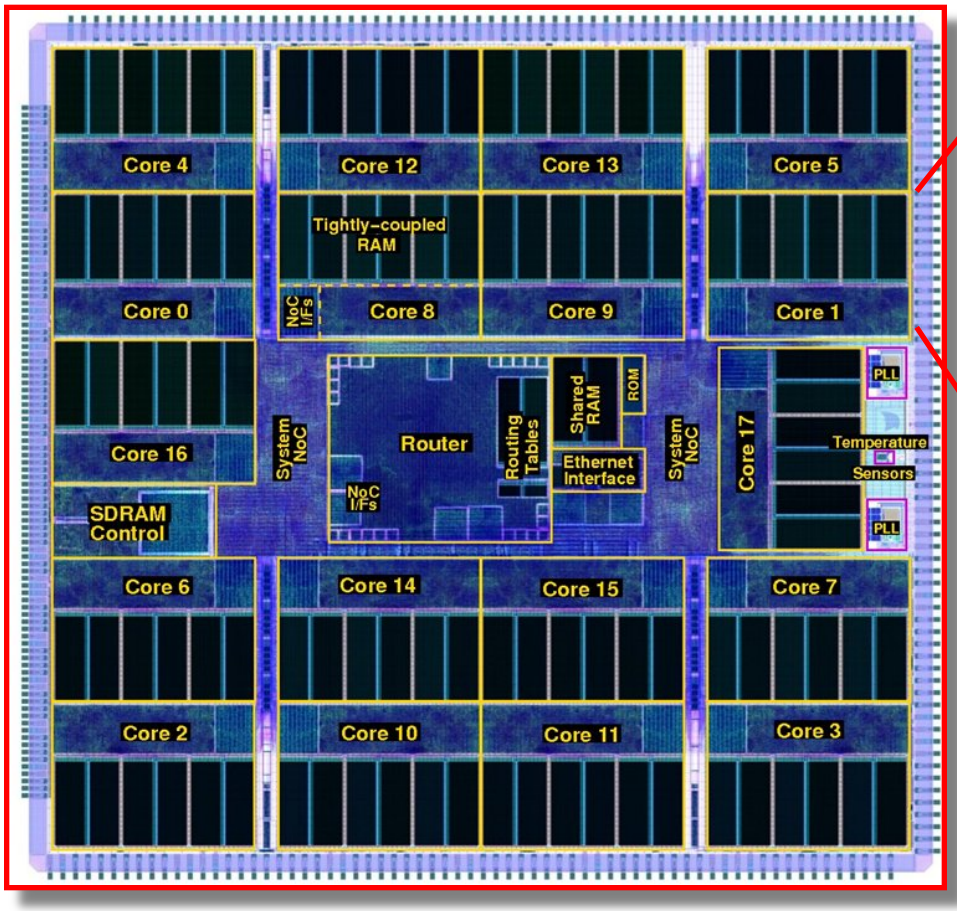


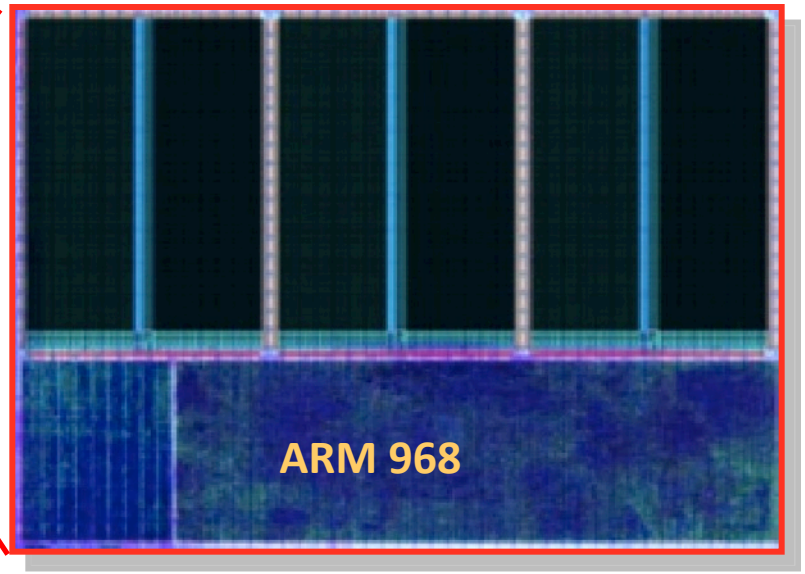
Human Brain Project

Manchester Baby (1948)

Human Brain Project

SpiNNaker CPU (2011)





Human Brain Project

63 years of progress

- ***Baby:***
  - used 3.5 kW of electrical power
  - executed 700 instructions per second
  - 5 Joules per instruction
- ***SpiNNaker ARM968 CPU node:***
  - uses 40 mW of electrical power
  - executes 200,000,000 instructions per second
  - 0.000 000 000 2 Joules per instruction

(James Prescott Joule born Salford, 1818)

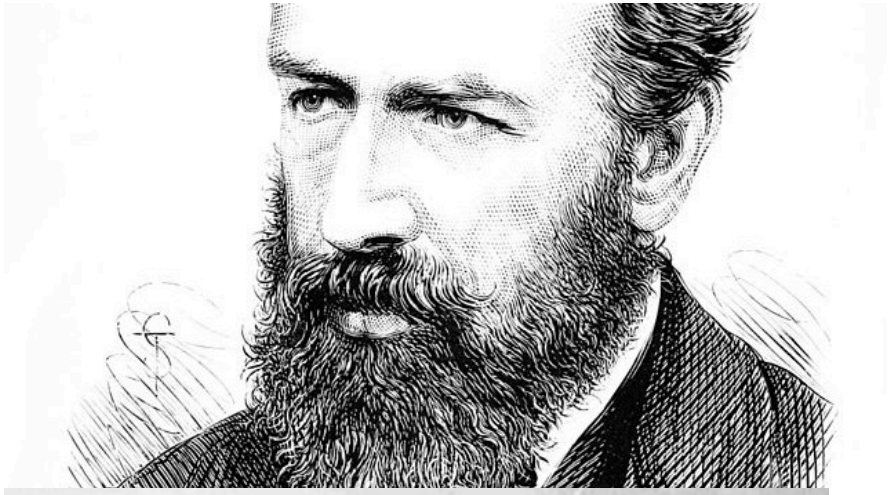
25,000,000,000 times better than Baby!

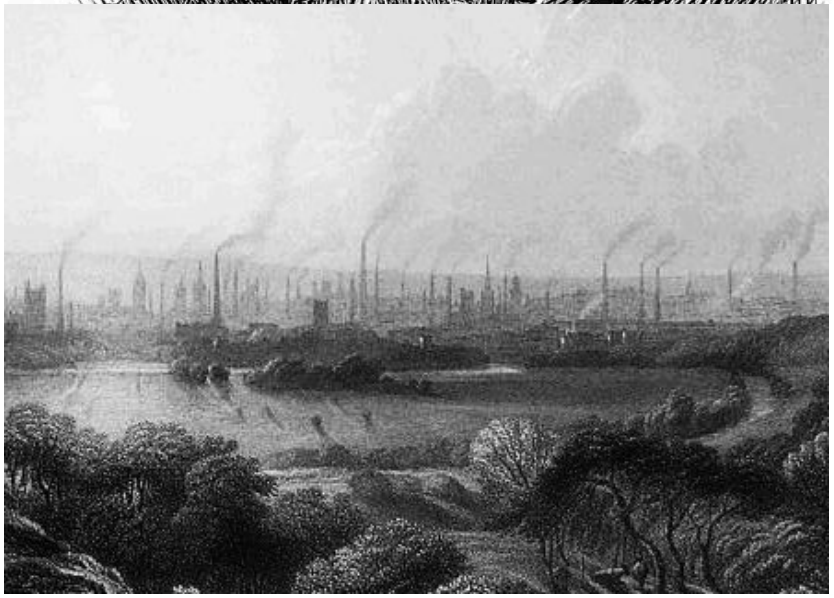
Human Brain Project

Jevons paradox

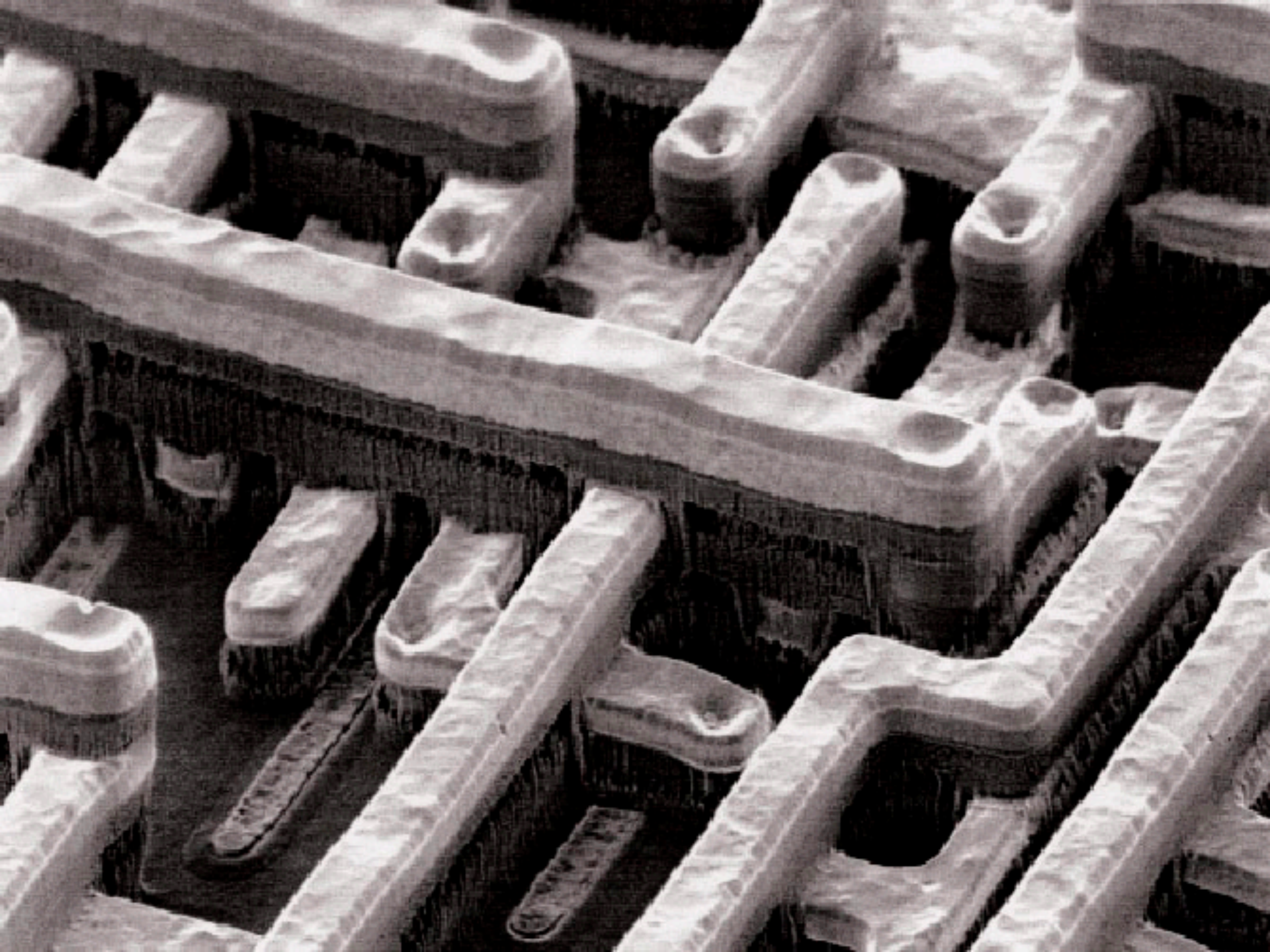
1865 “The Coal Question”

- James Watt’s coal-fired steam engine was much more efficient than Thomas Newcomen’s…
- …and coal consumption ***rose*** as a result





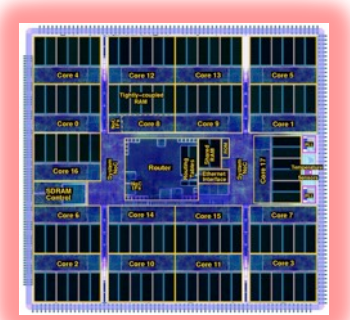
William Stanley Jevons, at Owens College (which later became the University of Manchester) 1863-75



Human Brain Project

Outline

- a Manchester perspective
- 63 years of progress
- CMOS power consumption
- the future of chip design
- bonclusions



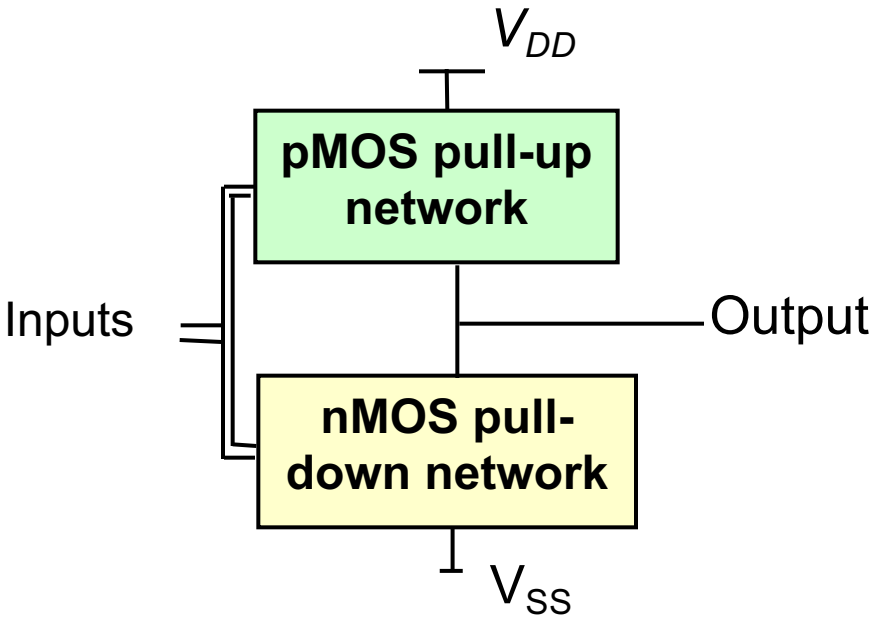
Human Brain Project

CMOS power consumption

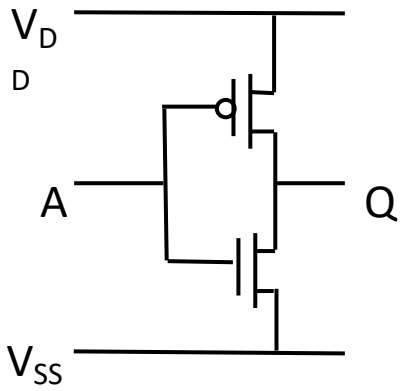
- CMOS power consumption
  - voltage change on a gate capacitance requires *charge transfer*, & therefore power consumption
  - once a gate is charged it can maintain its level without any additional charge movement
- CMOS circuitry **only** consumes power when switching states
  - well, until leakage starts to bite!

Human Brain Project

CMOS circuits

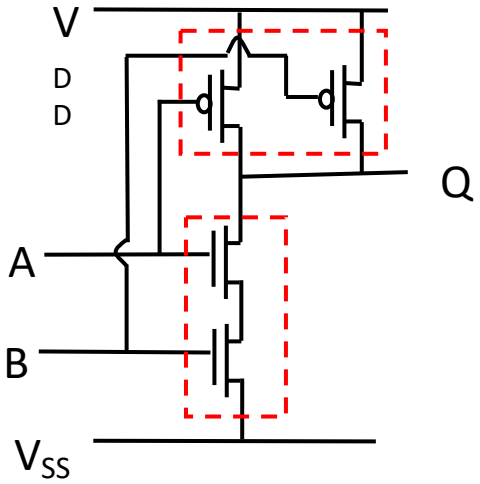


In general



An inverter

A 2-input NAND gate



- Normally Vss = 0V (GND)
- When not switching (static) only pull-up or pull-down network conducts
  - not both

Human Brain Project

Dynamic power consumption

$$P = 1/2 \times f_{clock} \times V_{DD}^2 \times \sum_{\text{all gates}} \alpha_g C_g \quad \text{Watts}$$

where:

f_{clock} = switching frequency of device clock

V_{DD} = supply voltage (assuming V_{SS}=0)

C_g = capacitance load on gate g

α_g = 'activity' on gate g:

= mean number of transitions per clock cycle

= 2 for a clock signal, ≈ 0.1 otherwise

Human Brain Project

Dynamic power consumption (simplified expression)

$$P = 1/2 \times C_{total} \times f_{clock} \times V_{DD}^2 \times \alpha$$

where:

C_{total} = total node capacitance

f_{clock} = switching frequency of device clock

V_{DD} = supply voltage

α = mean overall activity:

= mean number of transitions per clock cycle

= 2 for gates connected to a clock

Human Brain Project

Reducing dynamic power consumption: 1

$$P = 1/2 \times C_{total} \times f_{clock} \times V_{DD}^2 \times \alpha$$

- Reducing f_{clock} reduces P
- But consider effect on energy for running a given program
  - time to complete computation $\propto 1 / f_{clock}$
  - power $\propto f_{clock}$
  - so energy to run a program remains the same
  - number of instructions per Joule independent of f_{clock}
  - reducing f_{clock} is only a good idea if it allows lower V_{DD}

Reducing dynamic power consumption: 2

$$P = 1/2 \times C_{total} \times f_{clock} \times V_{DD}^2 \times \alpha$$

- Reducing V_{DD} greatly reduces P
  - but it also decreases the current than can be supplied by each transistor when it is switched on.
- Lower supply voltage means lower current.
  - load capacitances will charge more slowly.
  - gate switching will become slower
  - maximum possible value of f_{clock} will reduce
  - programs may take longer to run
- DVFS: *Dynamic Voltage and Frequency Scaling*
  - modern circuits have multiple V_{DD} and f_{clock} settings
- Can use parallelism to offset increases in circuit delay.

Human Brain Project

Reducing dynamic power consumption: 3

$$P = 1/2 \times C_{total} \times f_{clock} \times V_{DD}^2 \times \alpha$$

- Reducing C_{total} will clearly reduce P
- How can we do this?
  - use smaller, simpler circuits
    - e.g. ARM core rather than Pentium
  - do not over-size gates and buffers
    - in particular, reduce drive off critical path
  - use on-chip rather than off-chip memories
    - off-chip capacitances >> on-chip

Human Brain Project

Reducing dynamic power consumption: 4

$$P = 1/2 \times C_{total} \times f_{clock} \times V_{DD}^2 \times \alpha$$

How to reduce activity factor(s) α ?

- design circuits that do not switch more than is necessary
- use gates to avoid unnecessary distribution of clock signals
- turn off processor when it has nothing to do
  - don't make it sit in an idle loop!
- use an 'event-driven' style of design
  - in the limit, use asynchronous design (globally or locally)

Human Brain Project

Leakage

- Transistor off current is not zero!

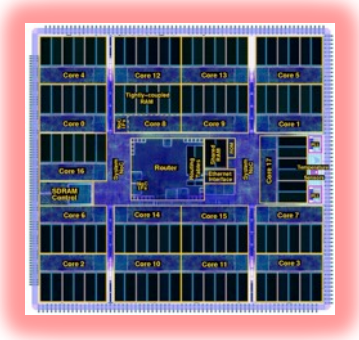
$$I_{off} \propto 10^{(-V_t/100mV)}$$

  - V_t is the transistor threshold
- In my day, when V_{DD} = 5 V, V_t = 0.7 V, I_{off} ~ pA
  - x 1,000,000 transistors = 1 μA (not much to worry about)
- In deep submicron CMOS V_{dd} is lower
  - e.g. 130 nm, V_{DD} = 1.2 V, V_t = 0.3 V, I_{off} ~ 10 nA
  - x 100,000,000 transistors = 1 A
- This is a big problem!
  - leads to unacceptable standby power in mobile systems

Human Brain Project

Outline

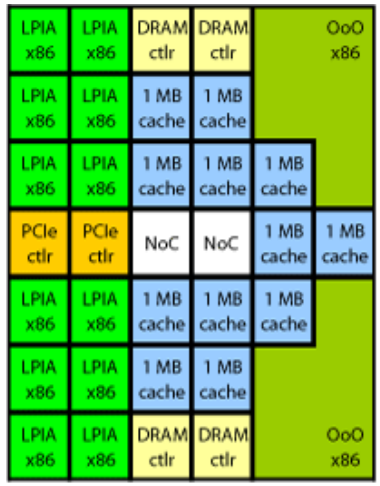
- a Manchester perspective
- 63 years of progress
- CMOS power consumption
- the future of chip design
- conclusions

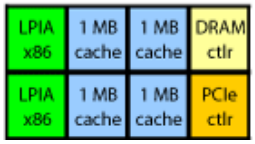


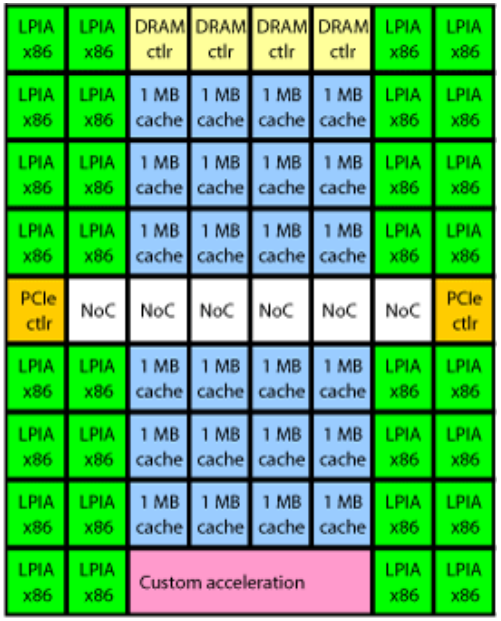
Human Brain Project

So what are the options?

- Many-core
  - homogeneous or heterogeneous?
- Accelerators
  - GPGPUs & similar
  - application-specific, e.g. neuromorphic
- Network on chip
- Memory
  - big, high-bandwidth, off chip
    - but in (3D) package?
  - more than just cache on chip
- Dark silicon
  - can't turn it all on at any time!





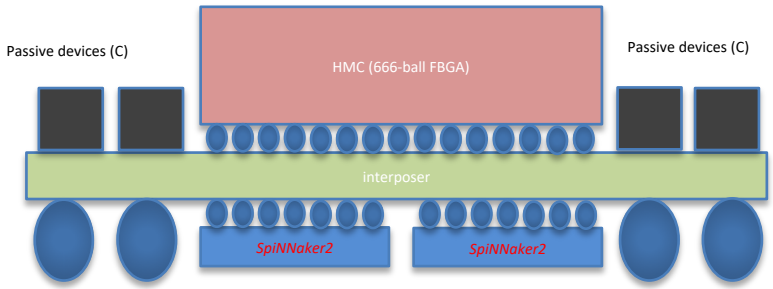


(Example client, server and embedded processors – J.L. Manferdelli, CTWatch Quarterly 3(1), Feb 2007)

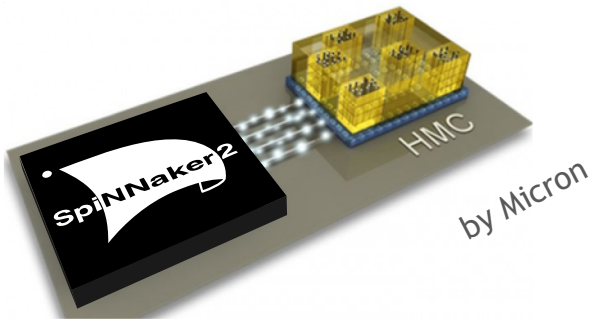
Human Brain Project

A personal view…

- 3D packaging has more to offer than Moore's Law
  - energy = distance x bits moved
  - all memory is package-local
- Few fat cores
  - for the code that just won't parallelize – if needed!
- Many thin cores
  - more flexible than GPGPU accelerators
- Selective accelerators
- Very powerful run-time management layer
  - to manage 'dark silicon' power constraints
- All for 1-2W per package
  - you can't let the memory get hot!



(SpiNNaker2 packaging concept by Sebastian Höppner, TU Dresden)



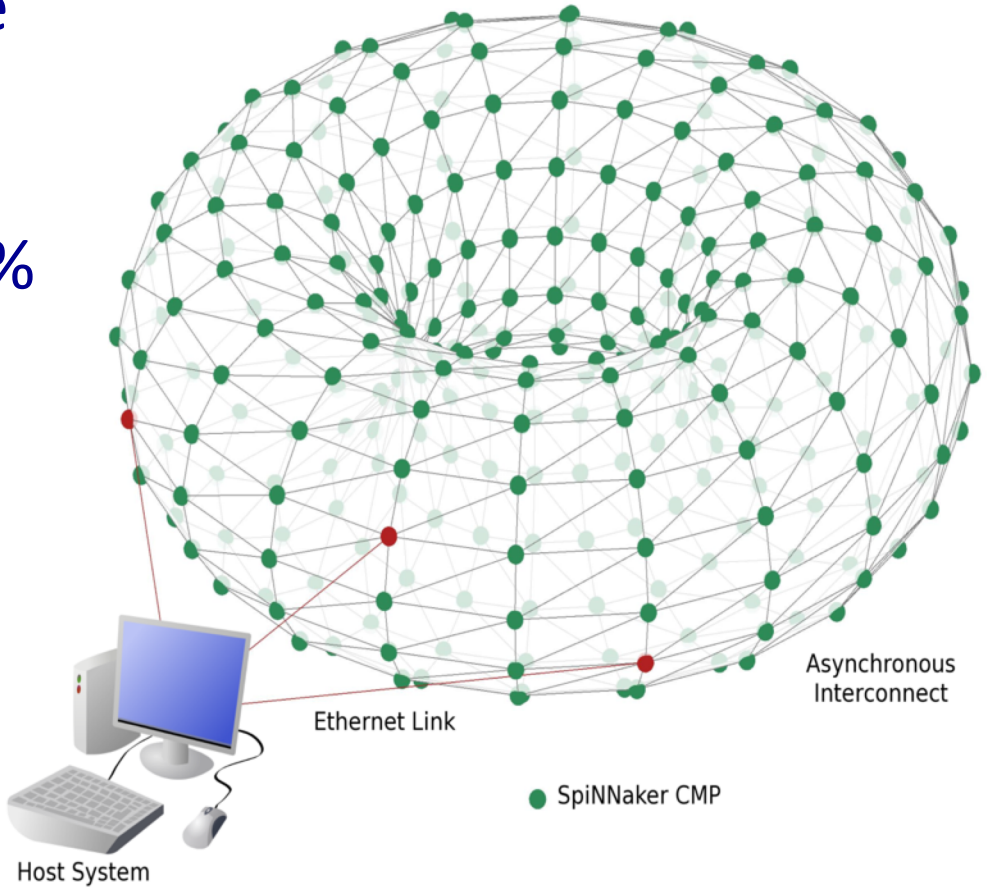
Human Brain Project

SpiNNaker project

SpiNNaker
Biologically Inspired Massively Parallel Architectures

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

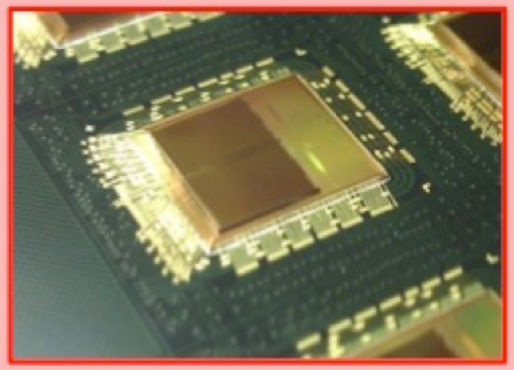
EPSRC



Human Brain Project

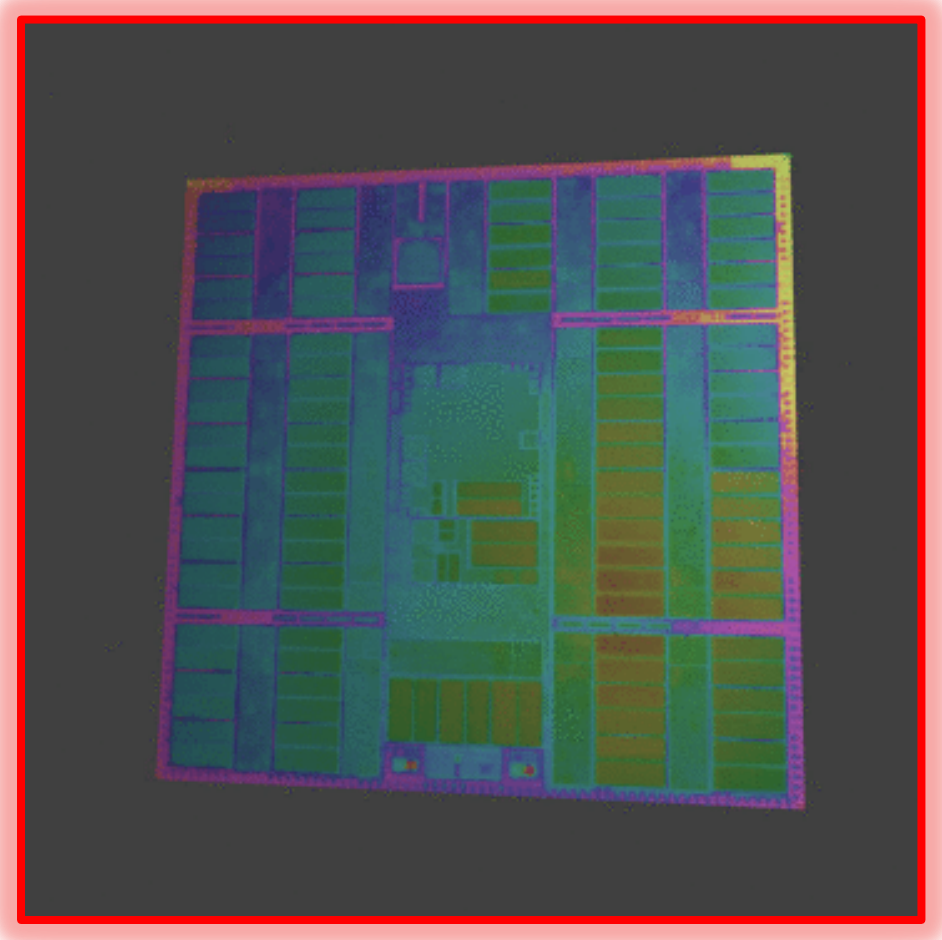
SpiNNaker chip

SpiNNaker
Biologically Inspired Massively Parallel Architectures



Multi-chip packaging by UNISEM Europe

unisem

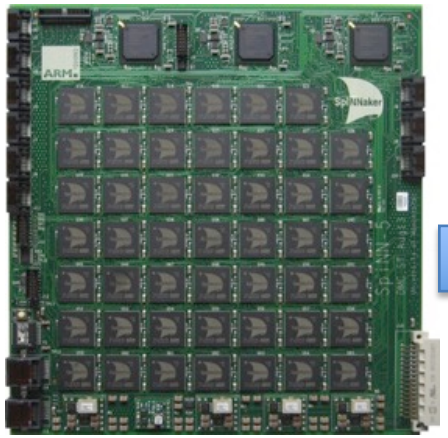


Human Brain Project

SpiNNaker machines

SpiNNaker
Biologically Inspired Massively Parallel Architectures

103

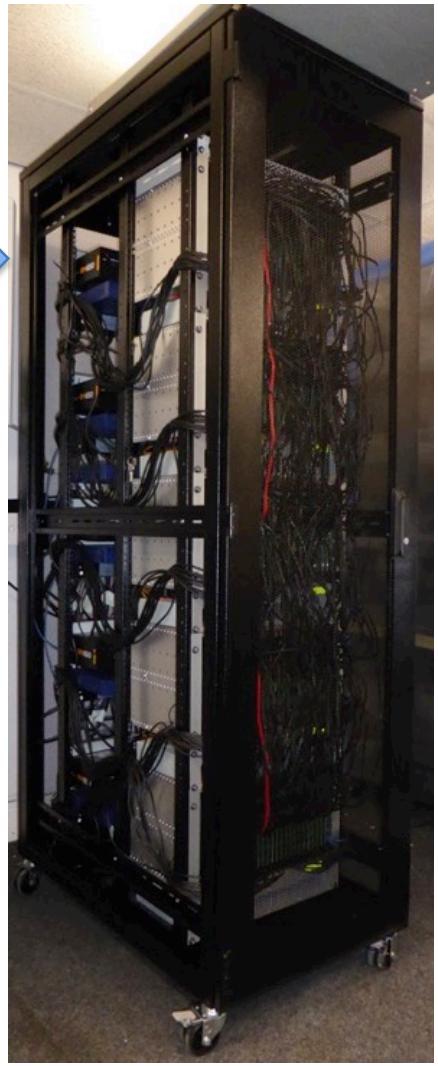


864 cores
- drosophila scale

104

20,000 cores
– frog scale

105



100,000 cores
– mouse scale

102

72 cores
- pond snail scale

Human Brain Project

SpiNNaker machines

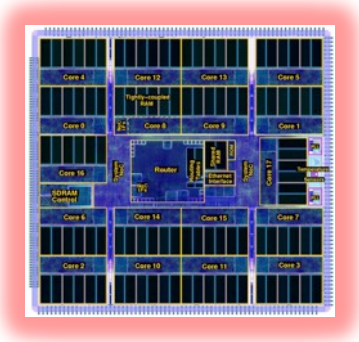
SpiNNaker
Biologically Inspired Massively Parallel Architectures

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

Human Brain Project

Outline

- a Manchester perspective
- 63 years of progress
- CMOS power consumption
- the future of chip design
- conclusions



Human Brain Project

Cheap as Chips!

- We have come a long way in 60 years…
  - x 10^{11} improvement in power-efficiency
- Heterogeneous many-core architectures are the future
  - with hybrid accelerators
  - including neuromorphic cores?
- Mobile processors and bio-inspired architectures may point the way
  - but there is still a lot more to do!