From ARMs to Brains

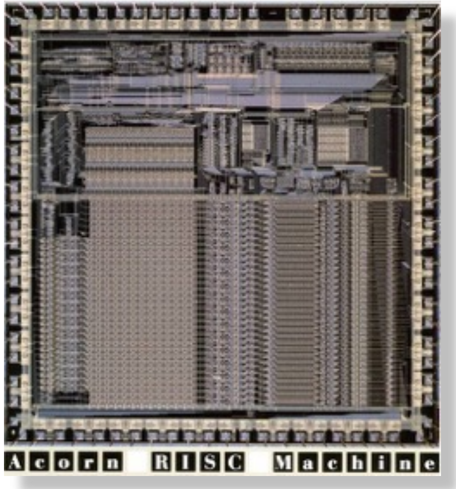
Steve Furber CBE FRS FREng

ICL Professor of Computer Engineering

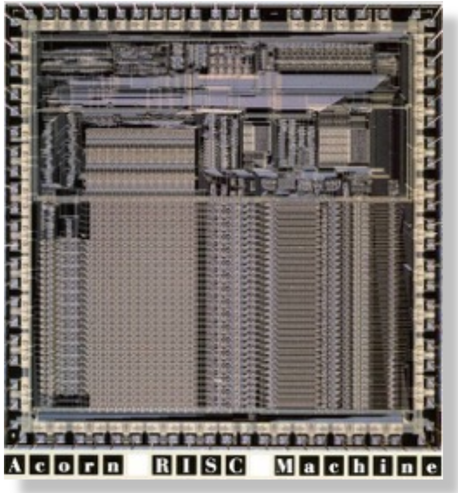
The University of Manchester

MANCHESTER 1824

The University of Manchester



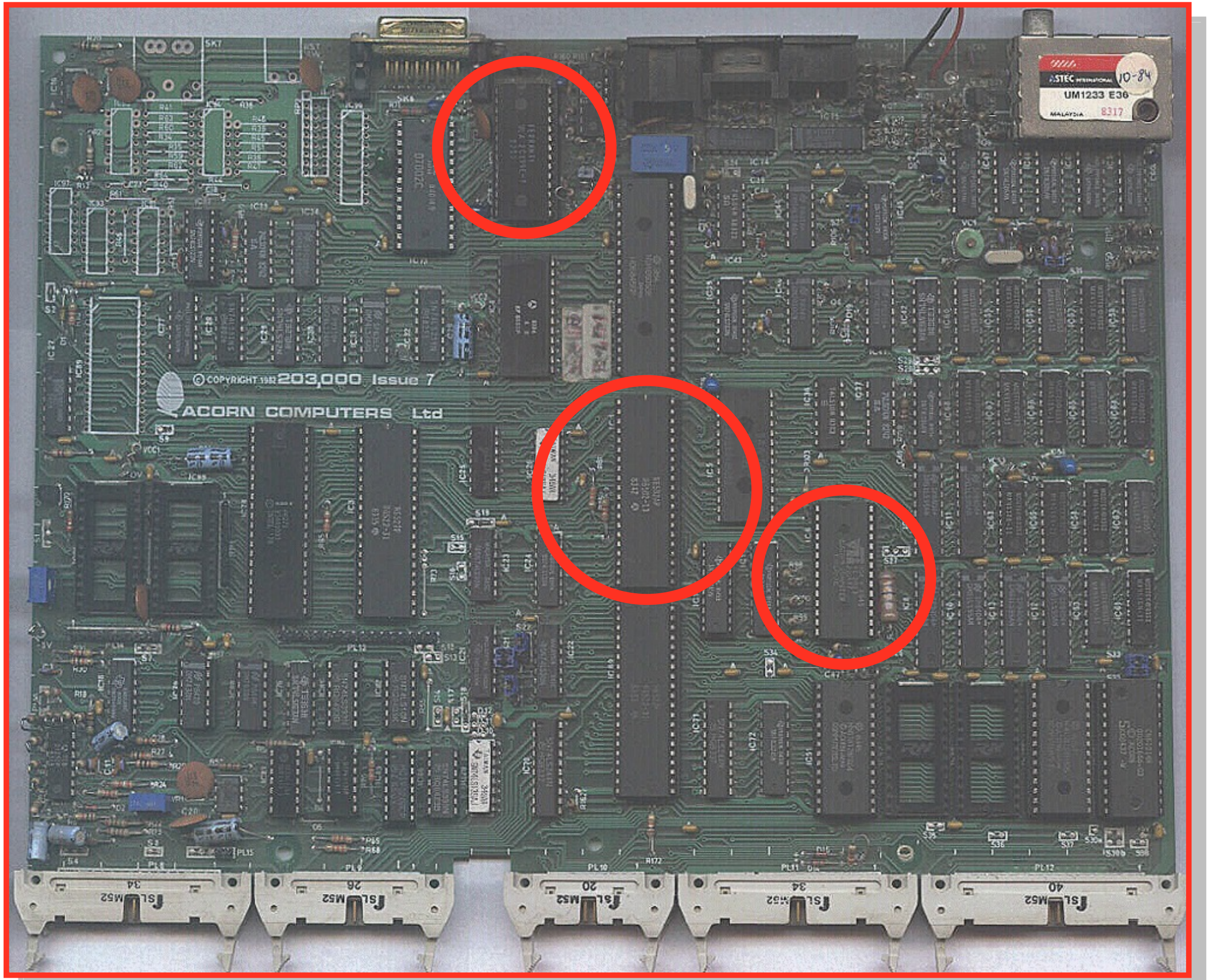
Outline

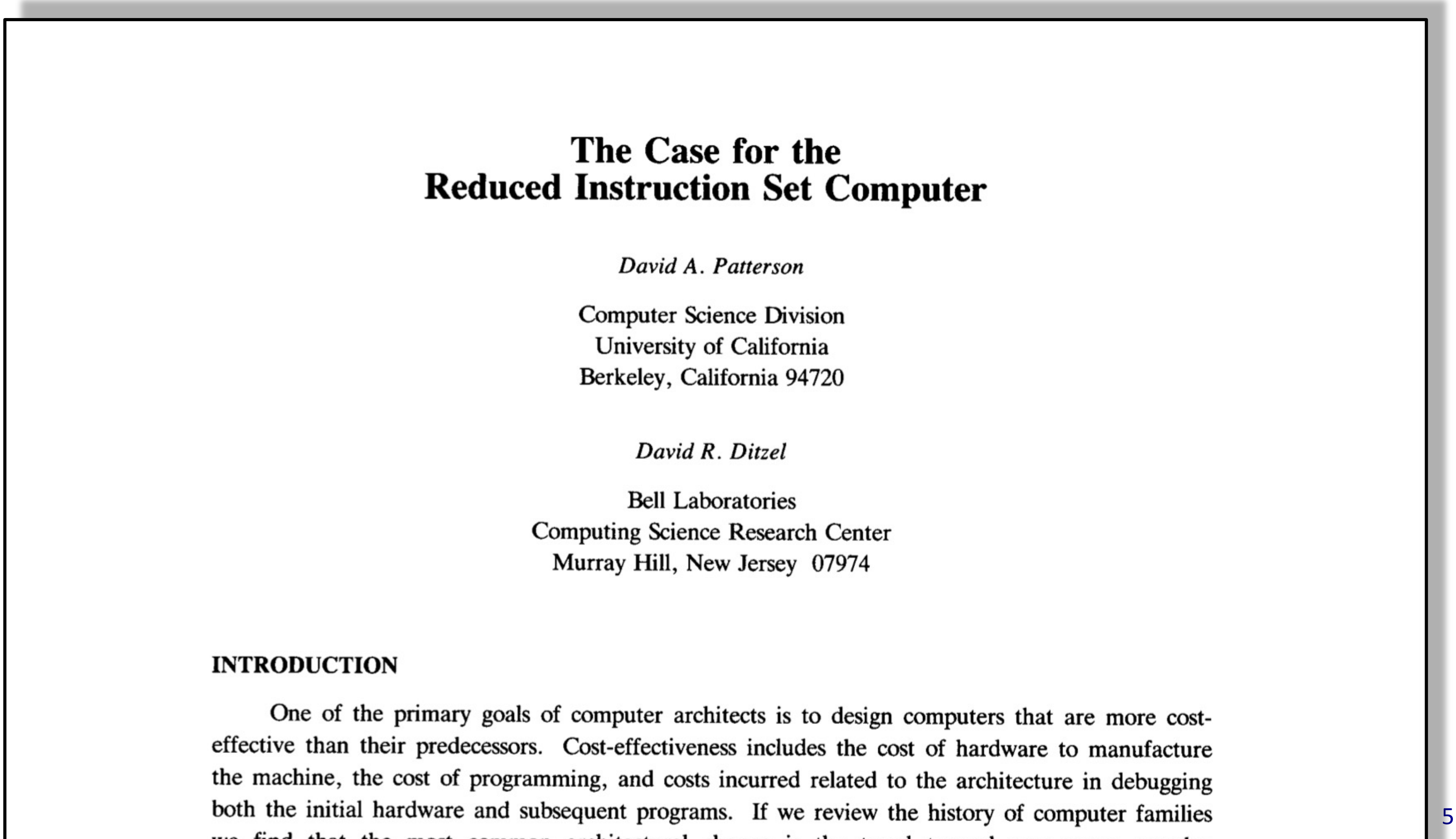


- from little Acorns…
- building brains
- 40 years of Moore’s Law
- how it started… how it’s going

The BBC Microcomputer (1982)

Acorn


©COPYRIGHT 1982 203,000 Issue 7
ACORN COMPUTERS Ltd
ASTEC INTERNATIONAL
UM1233 E36
MALAYSIA
8317
10-84



The Case for the Reduced Instruction Set Computer

David A. Patterson

Computer Science Division
University of California
Berkeley, California 94720

David R. Ditzel

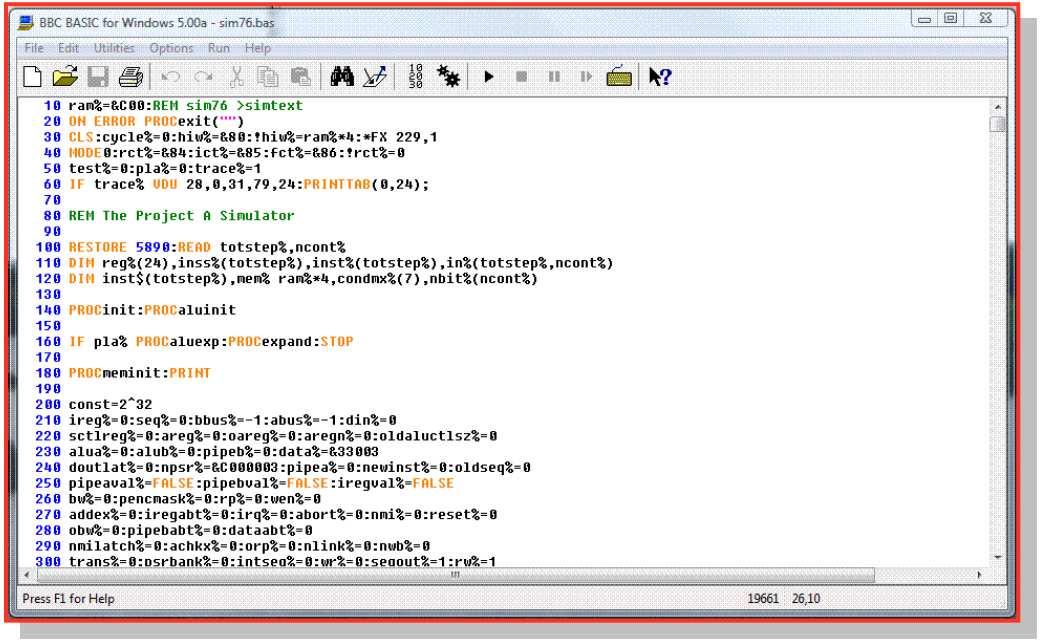
Bell Laboratories
Computing Science Research Center
Murray Hill, New Jersey 07974

INTRODUCTION

One of the primary goals of computer architects is to design computers that are more cost-effective than their predecessors. Cost-effectiveness includes the cost of hardware to manufacture the machine, the cost of programming, and costs incurred related to the architecture in debugging both the initial hardware and subsequent programs. If we review the history of computer families

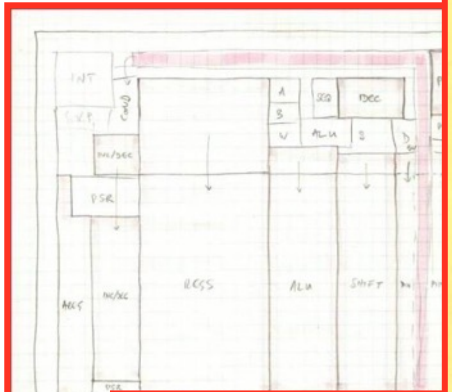
ARM development

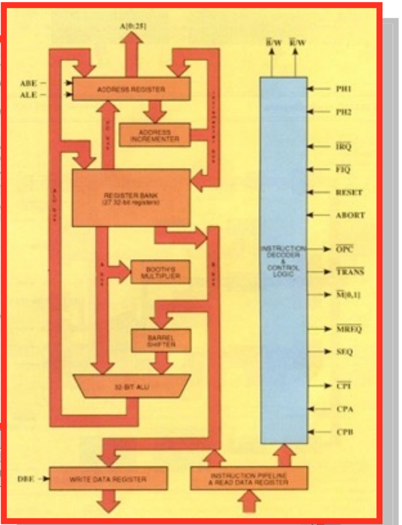
- Instruction set
  - Sophie Wilson
- BBC Basic reference model
  - 808 lines of code
    - it’s that simple!
- Instruction set validation suite
  - software team
- Block specs
  - implemented by VLSI group

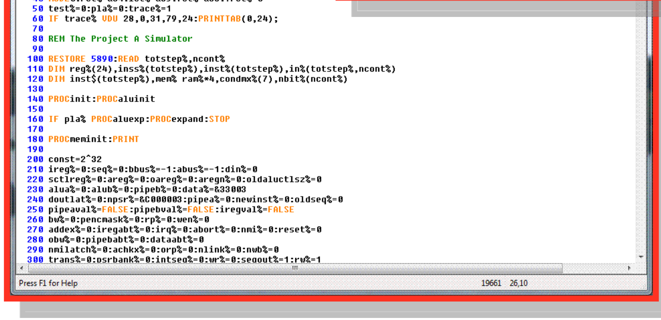


```
 10 ram%=&C00:REM sim76 >simtext
 20 ON ERROR PROCexit("")
 30 CLS:cycle%=0:hiw%=&80:!hiw%=ram%*4:*FX 229,1
 40 MODE0:rct%=&84:ict%=&85:fct%=&86:!rct%=0
 50 test%=0:pla%=0:trace%=1
 60 IF trace% VDU 28,0,31,79,24:PRINTTAB(0,24);
 70
 80 REM The Project A Simulator
 90
100 RESTORE 5890:READ totstep%,ncont%
110 DIM reg%(24),inss%(totstep%),inst%(totstep%),in%(totstep%,ncont%)
120 DIM inst$(totstep%),mem% ram%*4,condmx%(7),nbit%(ncont%)
130
140 PROCinit:PROCaluinit
150
160 IF pla% PROCaluexp:PROCexpand:STOP
170
180 PROCmeminit:PRINT
190
200 const=2^32
210 ireg%=0:seq%=0:bbus%=-1:abus%=-1:din%=0
220 sctlreg%=0:areg%=0:oareg%=0:aregn%=0:oldaluctlsz%=0
230 alua%=0:alub%=0:pipeb%=0:data%=&33003
240 doutlat%=0:npsr%=&C000003:pipea%=0:newinst%=0:oldseq%=0
250 pipeaval%=FALSE:pipebval%=FALSE:iregval%=FALSE
260 bw%=0:pencmask%=0:rp%=0:wen%=0
270 addex%=0:iregabt%=0:irq%=0:abort%=0:nmi%=0:reset%=0
280 obw%=0:pipebabt%=0:dataabt%=0
290 nmilatch%=0:achkx%=0:orp%=0:nlink%=0:nwb%=0
300 trans%=0:psrbank%=0:intseq%=0:wr%=0:seqout%=1:rw%=1
```

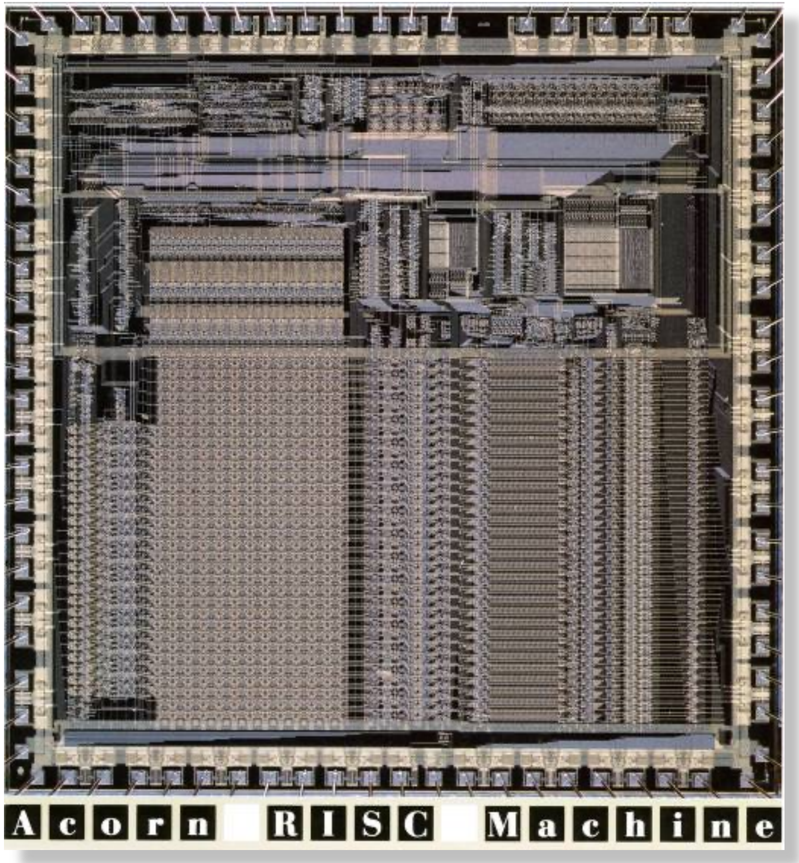
First ARM chip: 26th April 1985







```
 10 ram%=&C00:REM sim76 >simtext
 20 ON ERROR PROCexit("")
 30 CLS:cycle%=0:hiw%=&80:!hiw%=ram%*4:*FX 229,1
 40 MODE0:rct%=&84:ict%=&85:fct%=&86:!rct%=0
 50 test%=0:pla%=0:trace%=1
 60 IF trace% VDU 28,0,31,79,24:PRINTTAB(0,24);
 70
 80 REM The Project A Simulator
 90
100 RESTORE 5890:READ totstep%,ncont%
110 DIM reg%(24),inss%(totstep%),inst%(totstep%),in%(totstep%,ncont%)
120 DIM inst$(totstep%),mem% ram%*4,condmx%(7),nbit%(ncont%)
130
140 PROCinit:PROCaluinit
150
160 IF pla% PROCaluexp:PROCexpand:STOP
170
180 PROCmeminit:PRINT
190
200 const=2^32
210 ireg%=0:seq%=0:bbus%=-1:abus%=-1:din%=0
220 sctlreg%=0:areg%=0:oareg%=0:aregn%=0:oldaluctlsz%=0
230 alua%=0:alub%=0:pipeb%=0:data%=&33003
240 doutlat%=0:npsr%=&C000003:pipea%=0:newinst%=0:oldseq%=0
250 pipeaval%=FALSE:pipebval%=FALSE:iregval%=FALSE
260 bw%=0:pencmask%=0:rp%=0:wen%=0
270 addex%=0:iregabt%=0:irq%=0:abort%=0:nmi%=0:reset%=0
280 obw%=0:pipebabt%=0:dataabt%=0
290 nmilatch%=0:achkx%=0:orp%=0:nlink%=0:nwb%=0
300 trans%=0:psrbank%=0:intseq%=0:wr%=0:seqout%=1:rw%=1
```



ARM design team advantages

(according to Acorn founder, Hermann Hauser)

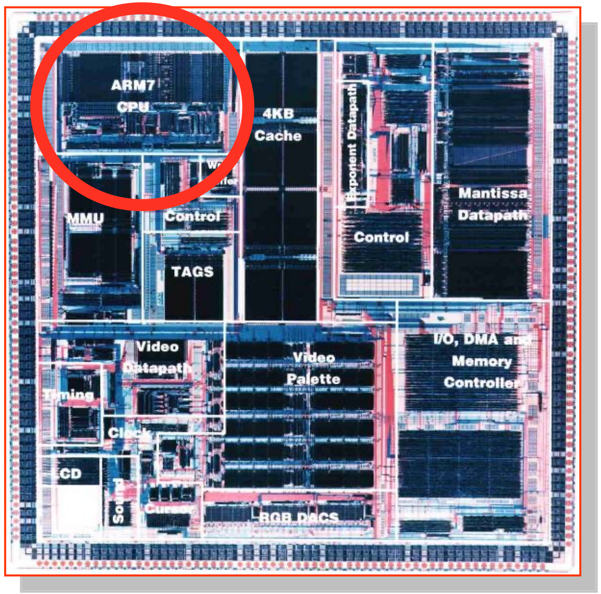
- No people
  - small team meant simplicity in design was an absolute requirement
- No money
  - everything was done in-house using simple, familiar home-grown tools
    - apart from VLSI design tools

1990: ARM Limited

- Acorn’s market was too small to support ARM development
- Apple wanted ARM for the Newton
- Joint Venture set up (with Apple & VLSI Technology) in November 1990

ARM Limited

- Systems-on-Chip
  - SoCs took off in the early 1990s
  - ARM’s simplicity
    - led to low power and small size
      - leaving room for other components
    - important features in early SoCs
      - where chip area and power were at a premium
- 2023: 250 billion ARM-powered chips shipped



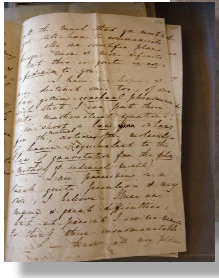
Outline

- from little Acorns…
- building brains
- 40 years of Moore’s Law
- how it started… how it’s going

200 years ago…

- Ada Lovelace, b. 10 Dec. 1815

"I have my hopes, and very distinct ones too, of one day getting cerebral phenomena such that I can put them into mathematical equations--in short, a law or laws for the mutual actions of the molecules of brain. …. I hope to bequeath to the generations a calculus of the nervous system.”



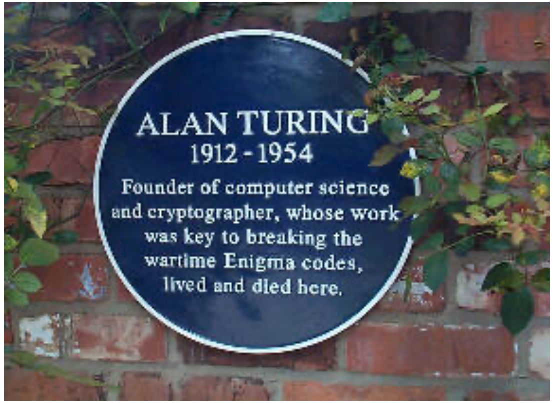
all the much that you mention you still have to communicate. My own scientific plans begin more & more definite. But this is quite in confidence to you.

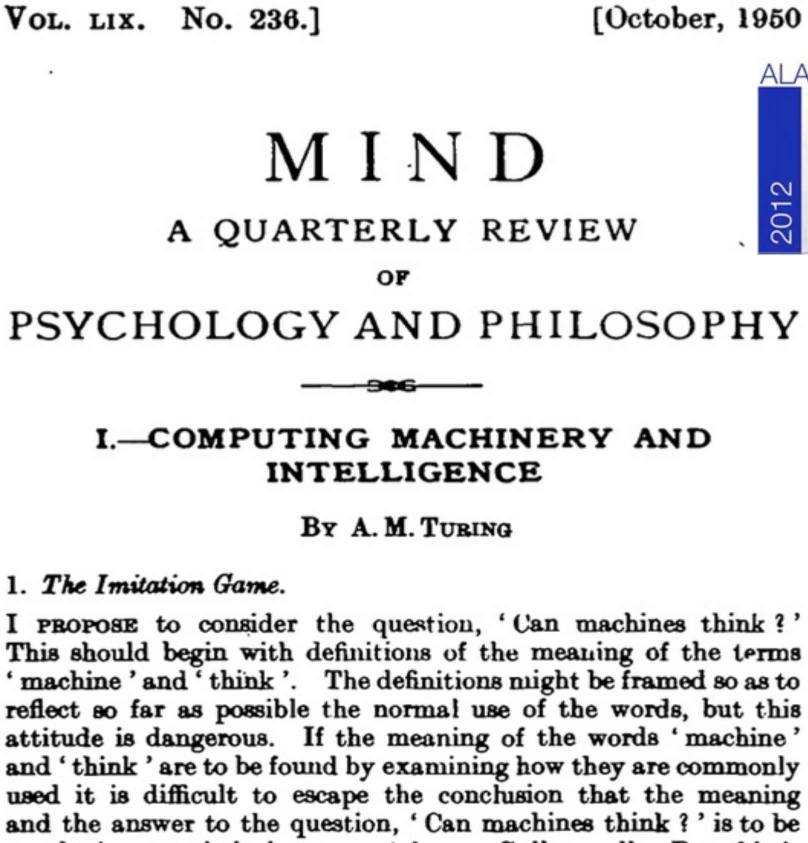
I have my hopes, & very distinct ones too, of one day getting cerebral phenomena such that I can put them into mathematical equations; in short a law, or laws for the mutual actions of the molecules of brain; (equivalent to the law of gravitation for the planetary & sidereal world). —

I am proceeding on a track quite peculiar & my own, I believe. There are many & great difficulties, but at present I see no reason to think them insurmountable.

L— knows all my plans

70 years ago…





Vol. lix. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

1. *The Imitation Game.*

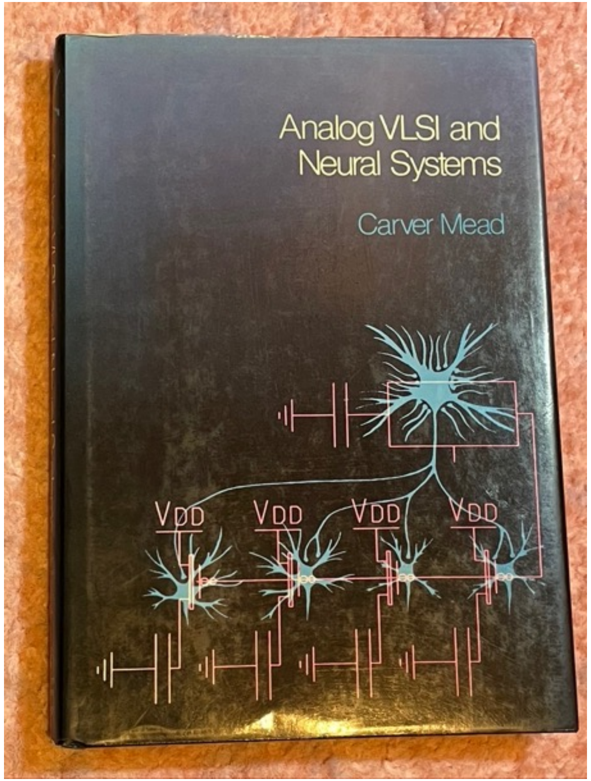
I propose to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words '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 machines think?' is to be

Neuromorphic Computing

Carver Mead

Caltech, 1980s

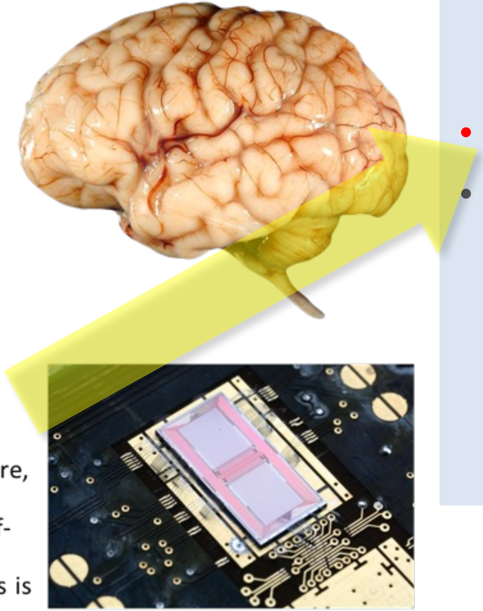
- observed analogy between ion channels in neurons and sub-threshold analogue transistor behaviour
- neuromorphic touch, hearing & vision sensors



The Human Brain Project

Why focus on the brain ? Three Reasons

- Understanding the brain (Unifying Science Goal)
  - Underpins what we are,
  - Data & knowledge are fragmented,
  - Integration is needed,
  - Large scale collaborative approach is essential.
- Understanding brain diseases (Society)
  - Costs Europe over €800 Billon/year,
  - Affects 1/3 people,
  - Number one cause of loss of economic productivity,
  - No fundamental treatments exist or are in sight
  - Pharma companies pulling out of the challenge.
- Developing Future Computing (Technology)
  - Computing underpins modern economies,
  - Traditional computing faces growing hardware, software, & energy barriers,
  - Brain can be the source of energy efficient, robust, self-adapting & compact computing technologies,
  - Knowledge driven process to derive these technologies is missing.



Neuromorphic Computing

- Neuromorphic Machines
- Algorithms and Architectures for Neuromorphic Computing
  - Theory
  - Applications

Co-funded by the European Union

The HBP Neuromorphic Computing Strategy

Next generation of NMC is more biology driven

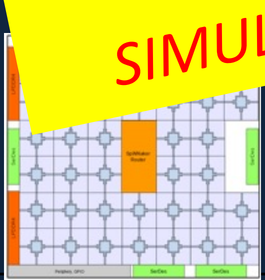
1st generation SpiNNaker-1 Machine

y-core system
on ARM cores
me simulator

Many-core Architecture

SIMULATION

Towards 2nd



152 Cortex M4F per chip
36 GIPS/Watt per chip
x10 with constant power

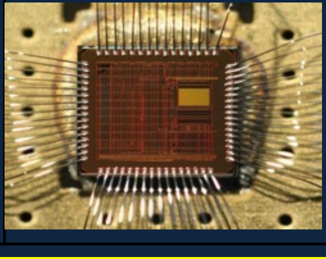
1st generation BrainScaleS-1 Machine

Physical model system
4M neurons, 1B plastic syn.
celerated emulator

Physical mode

EMULATION

Towards 2nd g



On-chip plasticity processors
Flexible hybrid plasticity
Active dendrites

Designed and built from the transistor up !

Co-designed with (theoretical) neuroscience

Neuromorphic systems worldwide

SpiNNaker
Biologically Inspired Massively Parallel Architectures

IBM
TrueNorth

intel
Loihi

BrainScaleS

Biological realism →

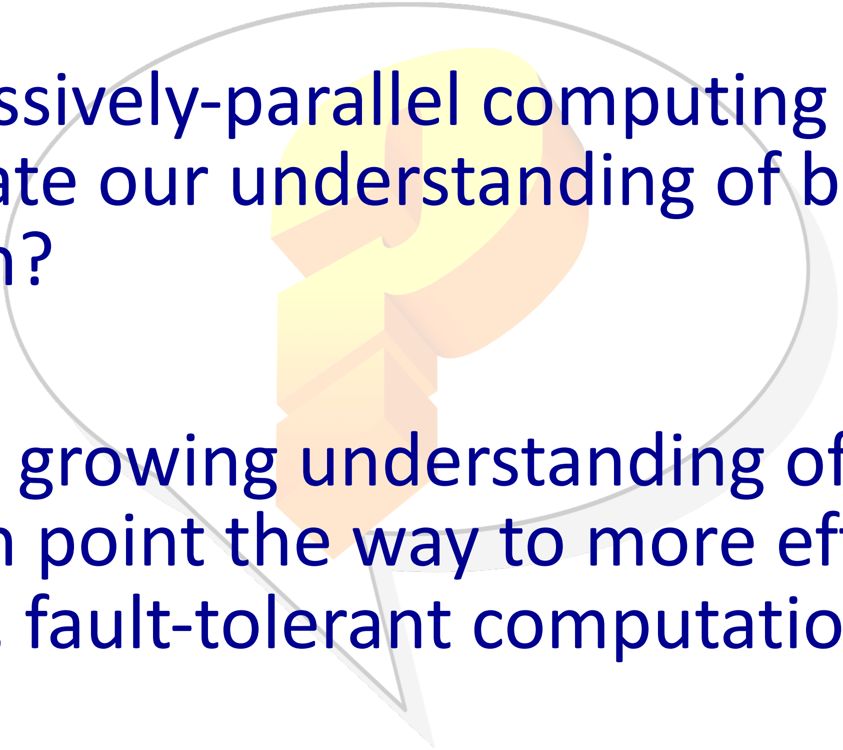
← Ease of use

Many-core (ARM) architecture
Optimized spike communication network
Programmable local learning
x0.01 real-time to x10 real-time

Full-custom-digital neural circuits
No local learning (TrueNorth)
Programmable local learning (Loihi)
Exploit economy of scale
x0.01 real-time to x100 real-time

Analog neural cores
Digital spike communication
Biological local learning
Programmable local learning
x10.000 to x1000 real-time

Bio-inspiration

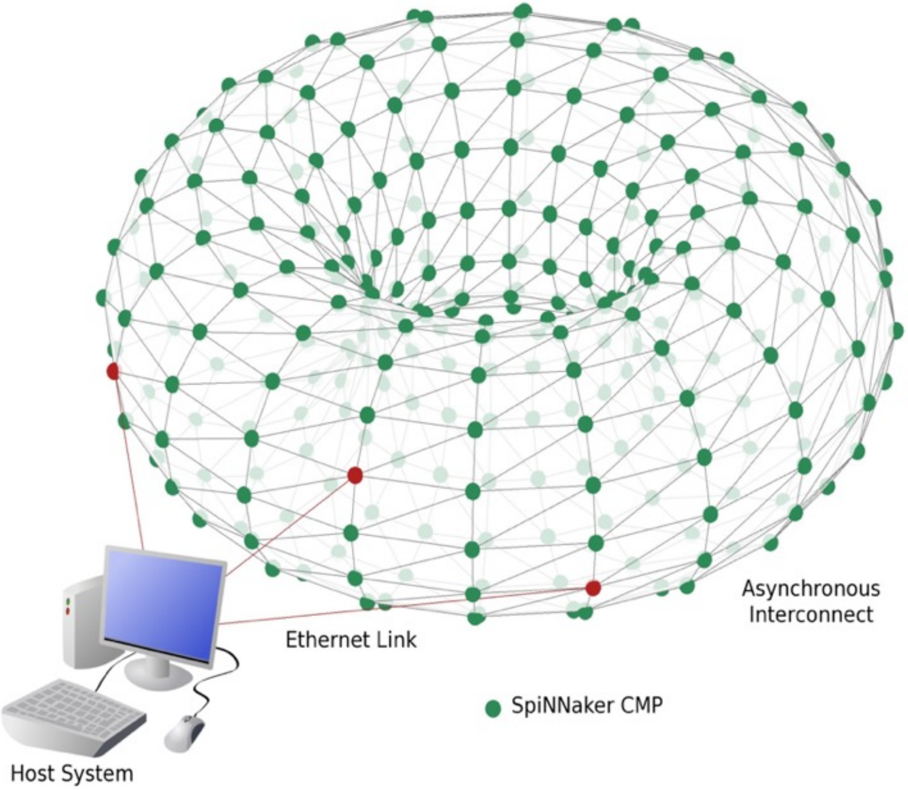


- Can massively-parallel computing resources accelerate our understanding of brain function?

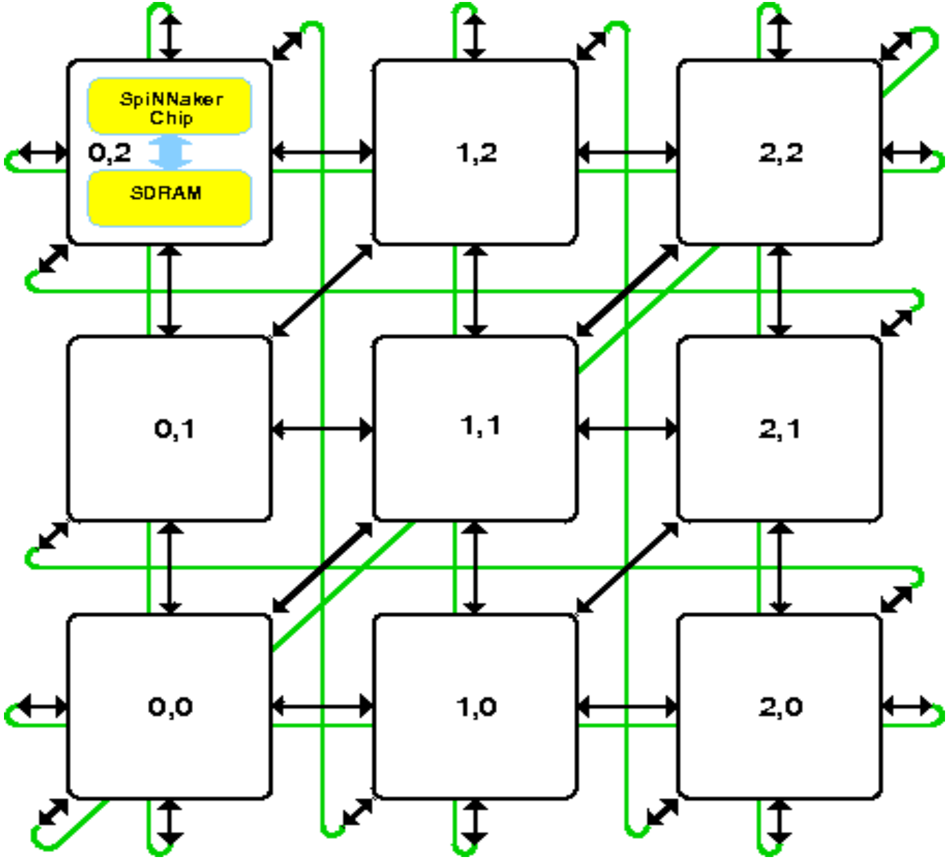
- Can our growing understanding of brain function point the way to more efficient parallel, fault-tolerant computation?

SpiNNaker project

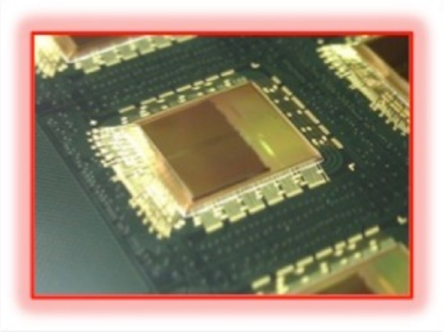
- A million ARM processors in one computer
- Able to model about 1% of the human brain…
- …or 10 mice!



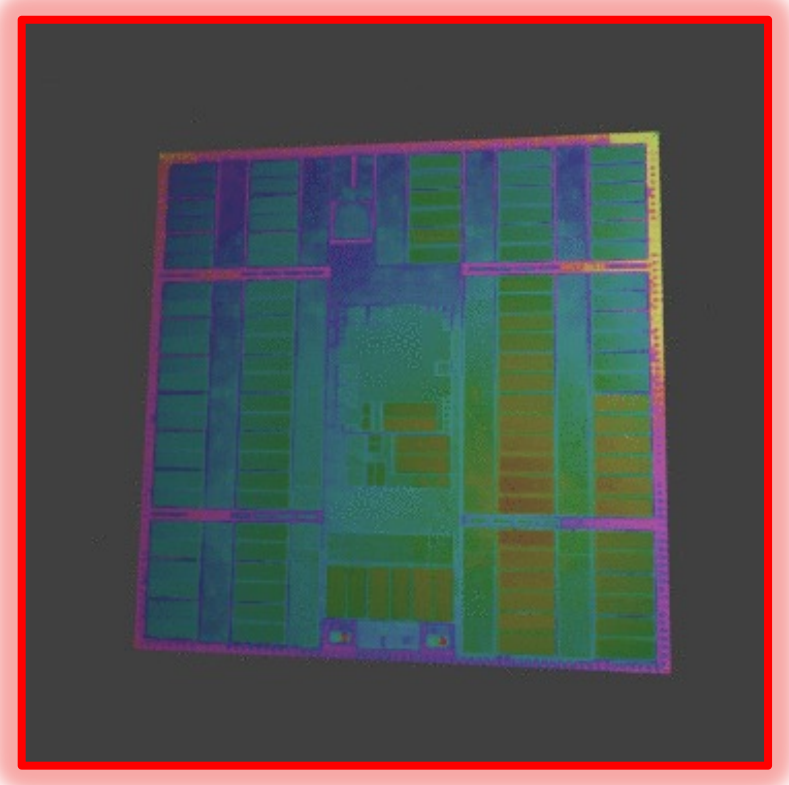
SpiNNaker system



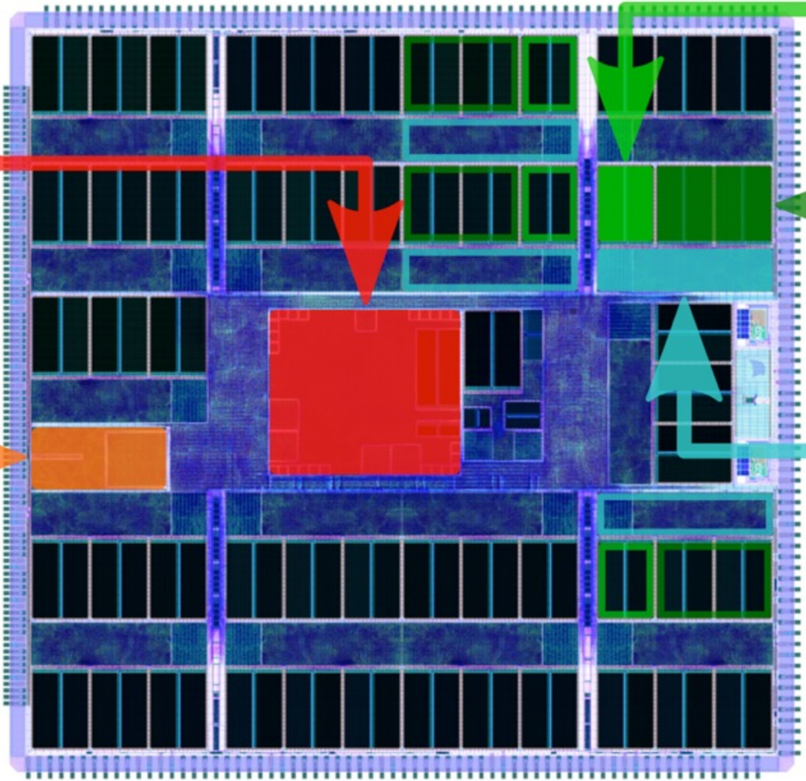
SpiNNaker chip



Multi-chip packaging by UNISEM



Chip resources



Router

routing tables
spike packet routing
system comms.

RAM port

synapse states
activity logs

Instruction memory

run-time kernel
application callbacks

Data memory

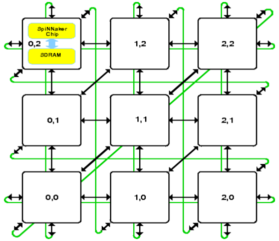
kernel state
neuron states
stack and heap

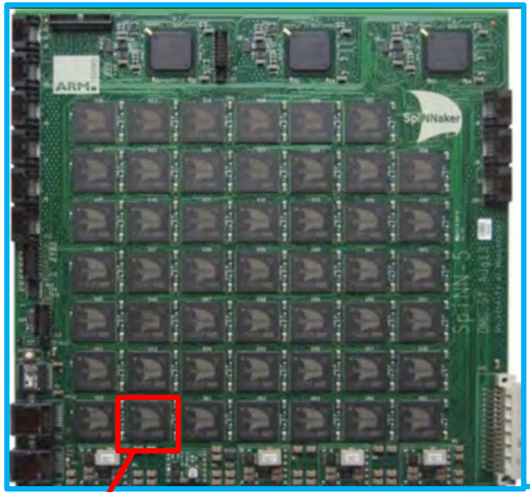
Processor

neuron and synapse
state computations

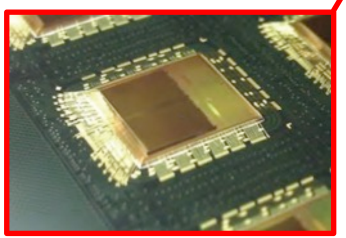
SpiNNaker machines

SpiNNaker board
(864 ARM cores)





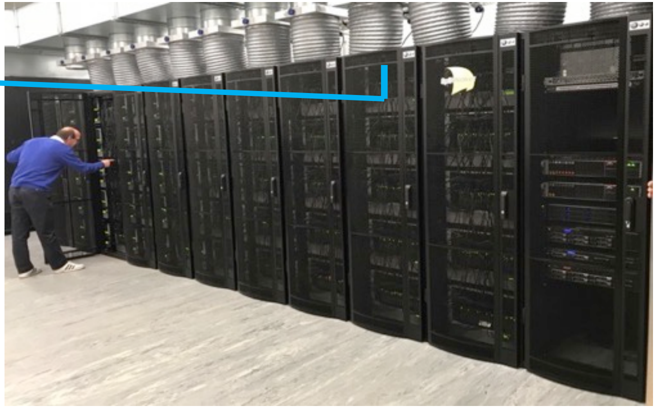
SpiNNaker chip
(18 ARM cores)



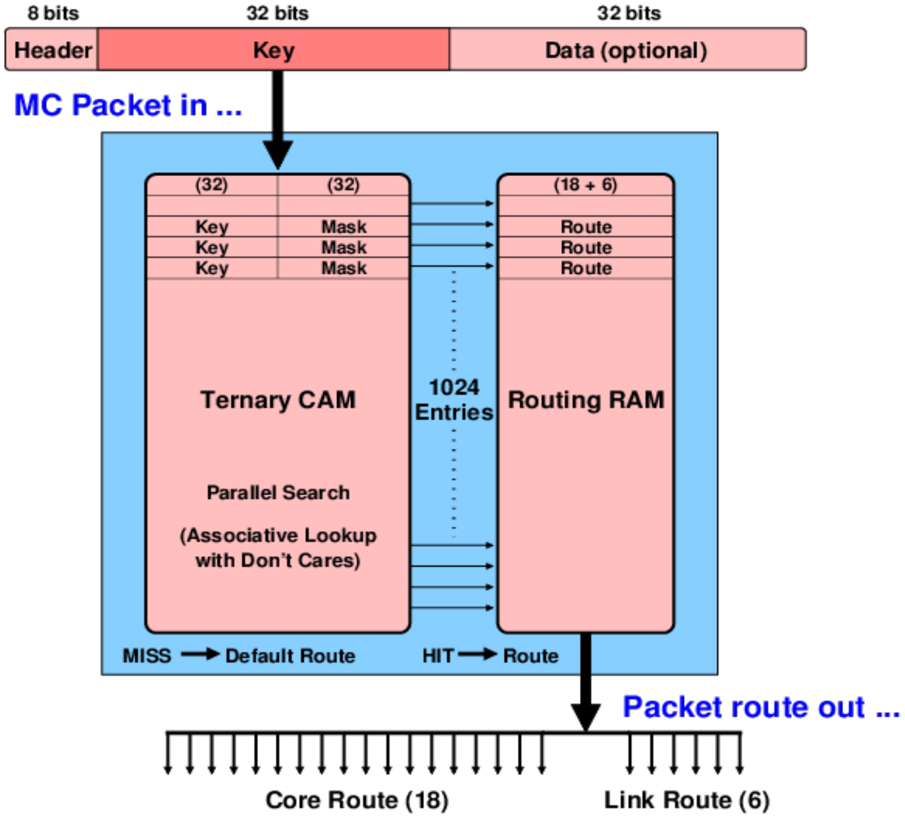
- HBP platform
  - 1M cores
  - 11 cabinets (including server)
- Launch 30 March 2016
  - then 500k cores
  - ~450 remote users
  - 5M SpiNNaker jobs run

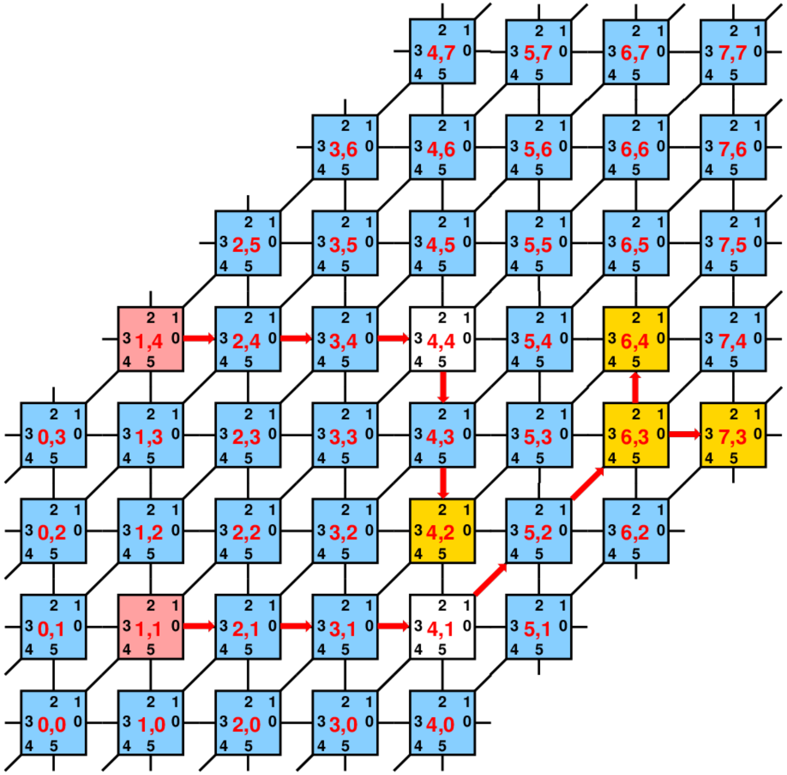
Human Brain Project

SpiNNaker racks
(1M ARM cores)

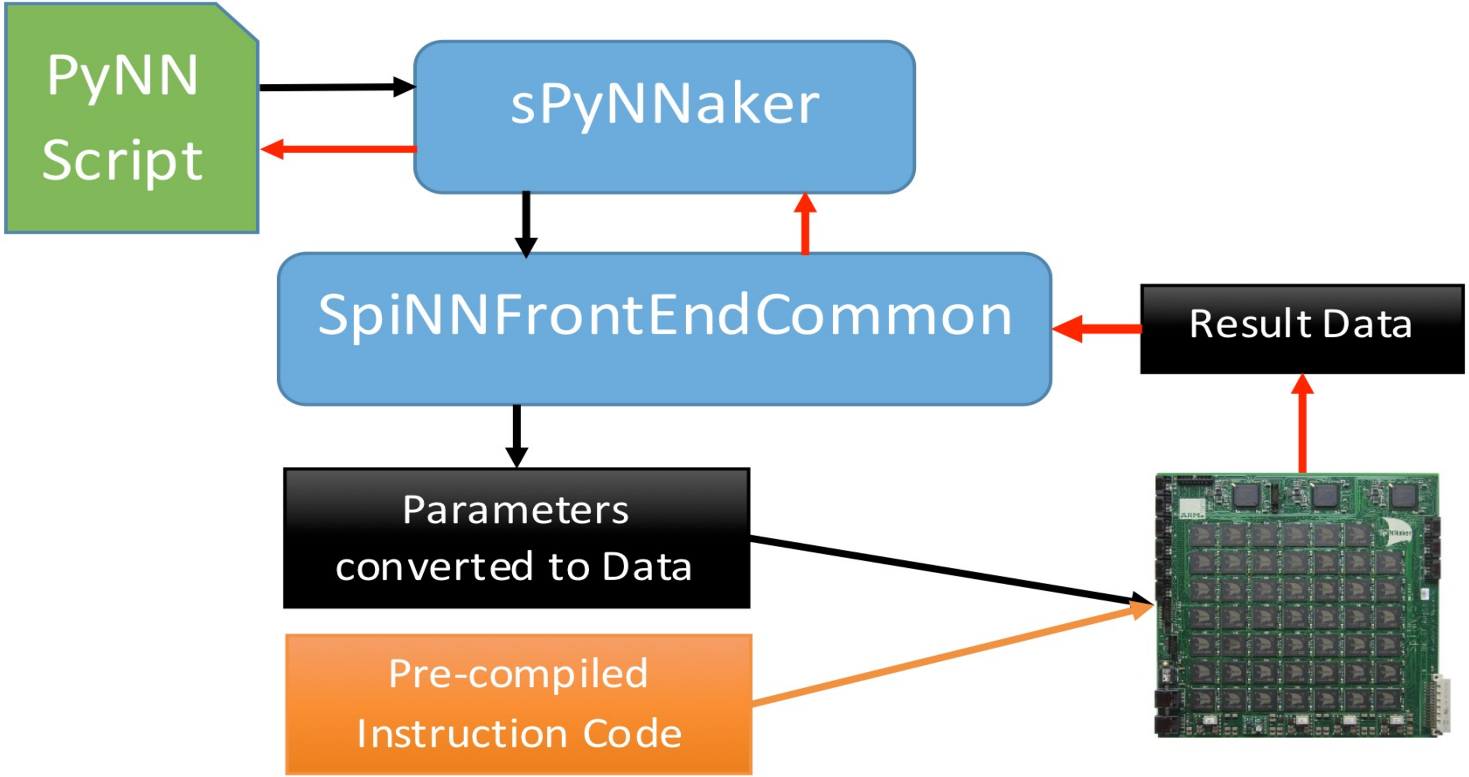


Multicast routing





High-level software flow



PyNN Script

sPyNNaker

SpiNNFrontEndCommon

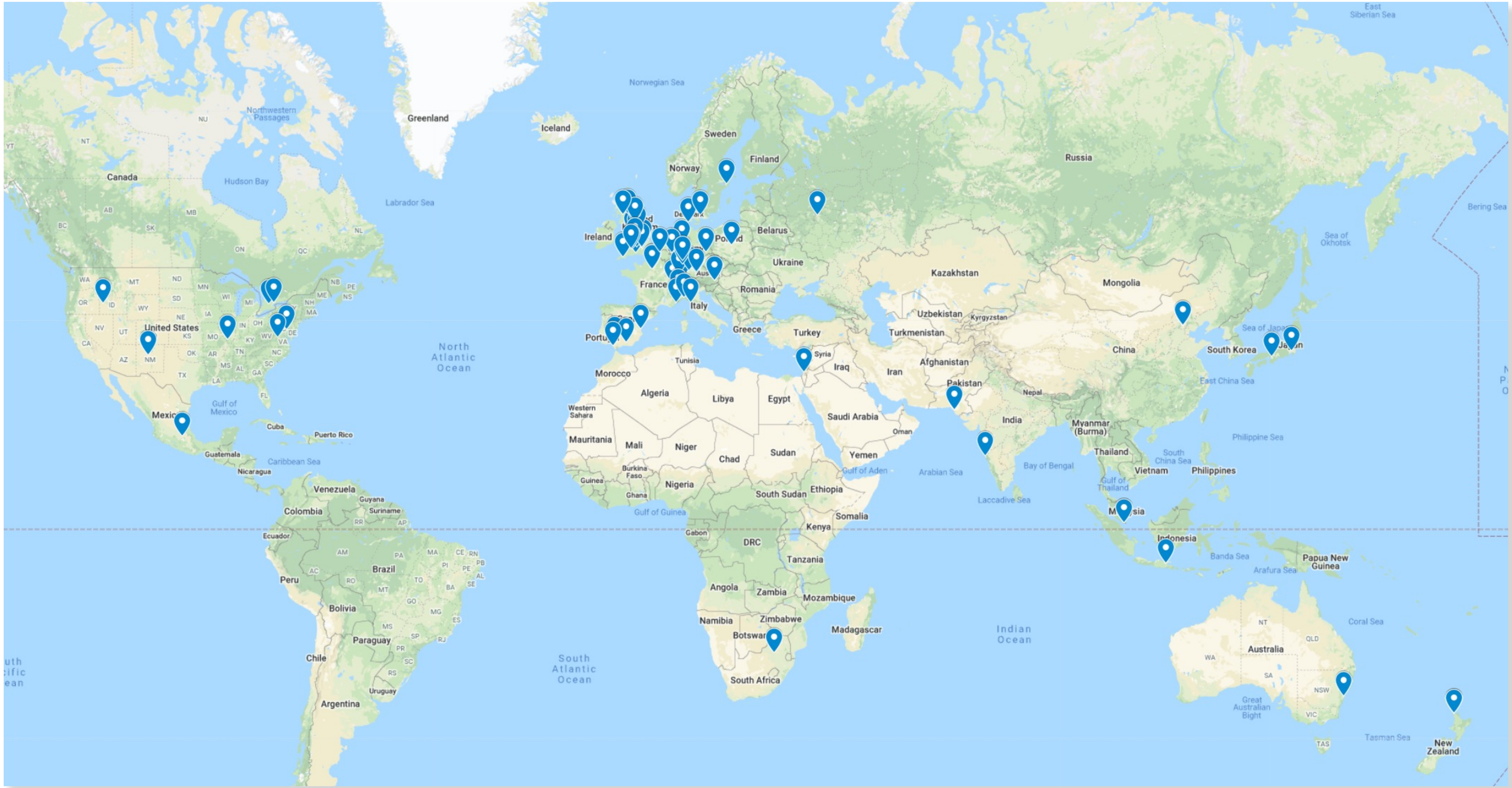
Result Data

Parameters converted to Data

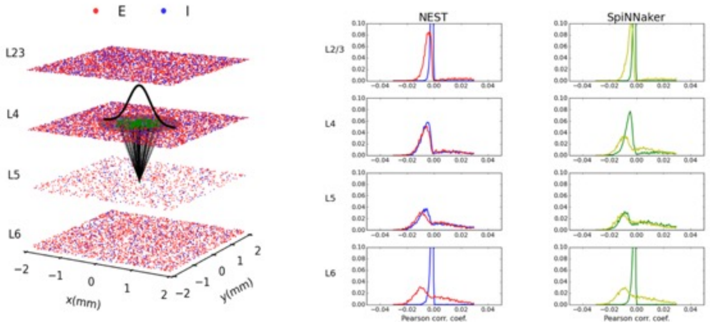
Pre-compiled Instruction Code

SpiNNaker
Biologically
Inspired
Massively
Parallel
Architectures

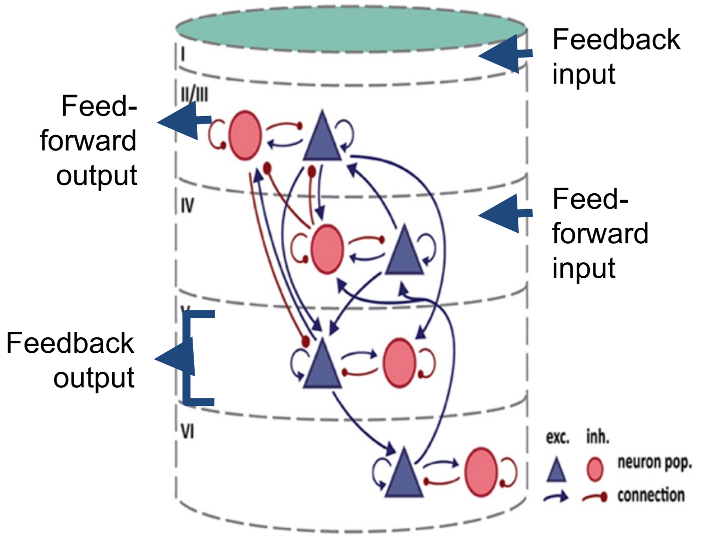
SpiNNaker machines



Cortical microcircuit



- ***Realtime execution of cortical model***
  - 1mm^2 cortex
    - 77k neurons
    - 285M synapses
    - 0.1 ms time-step
- ***Best previous versions of this model***
  - HPC: 3x slow-down
  - GPU: 2x slow-down
- ***Will scale to 100mm^2 without slow-down***
  - on current machine, simply by using more boards


Feedback input

Feed-forward output

Feed-forward input

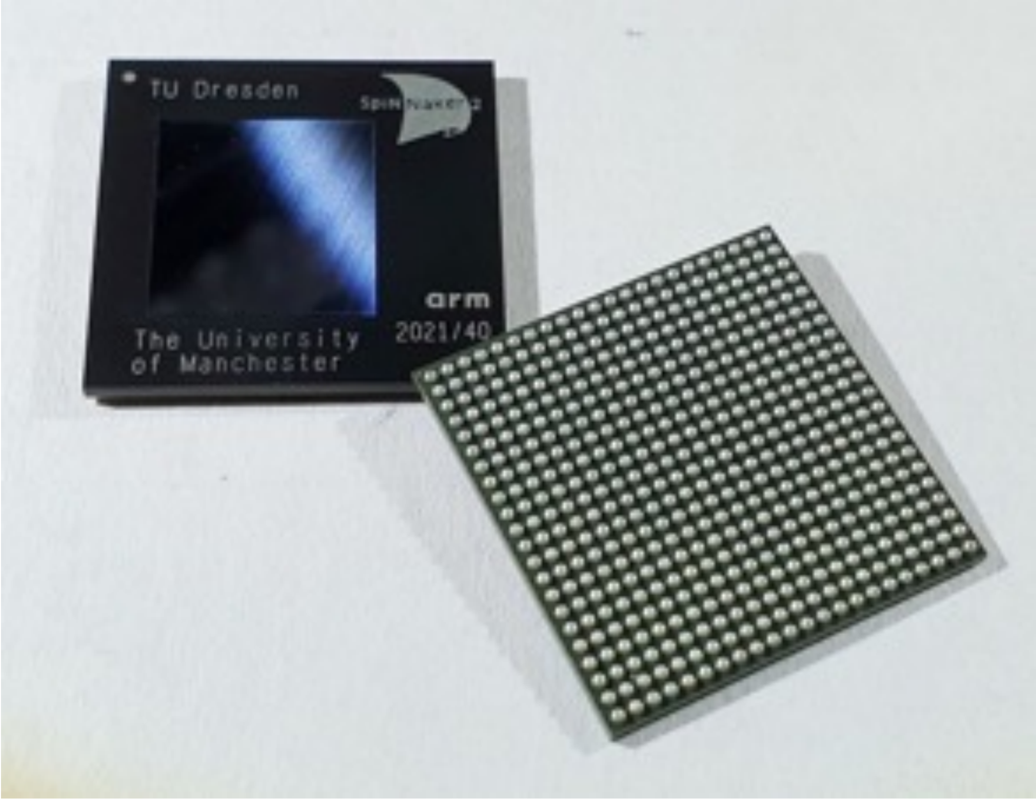
Feedback output

S.J. van Albada, 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 2018.

Oliver Rhodes, Luca Peres, Andrew G. Rowley, Andrew Gait, Luis A. Plana, Christian Brenninkmeijer & Steve.B. Furber, "*Real-time cortical simulation on neuromorphic hardware*", Phil Trans Roy Soc A, December 2019.

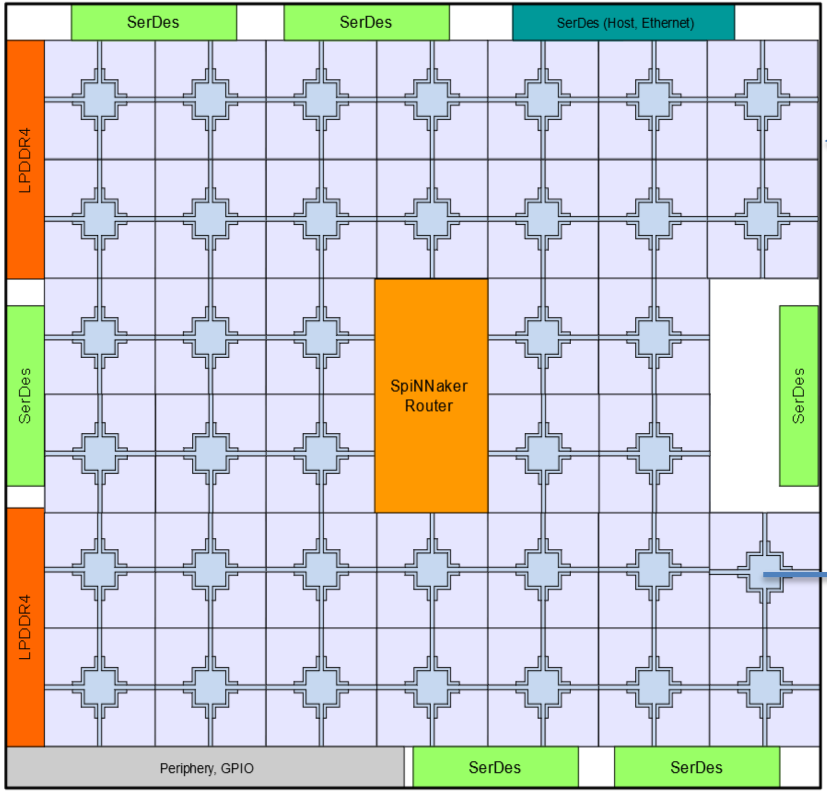
SpiNNaker2

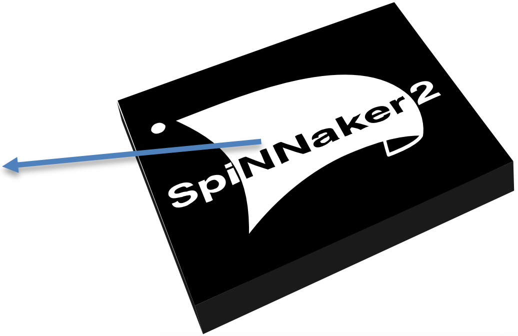
- 152 ARM-based processing elements
- 4 GByte DRAM
- 7 energy-efficient chip-to-chip links
- 10x ***SpiNNaker1***
- co-developed with TU Dresden

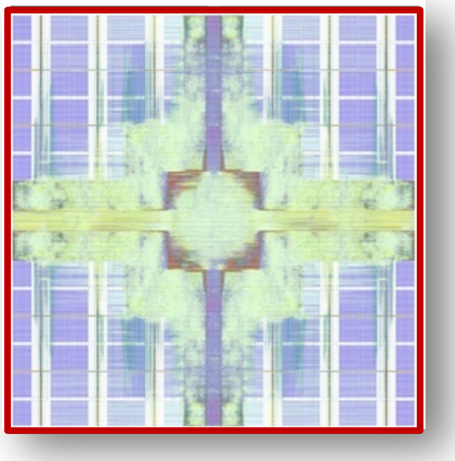


SpiNNaker2 chip overview

- 152 ARM-based processing elements (PEs)
- 4 GByte LPDDR4 DRAM
- 7 energy efficient chip-to-chip links







SpiNNaker2 Processing Element

Dynamic Power Management for enhanced energy efficiency

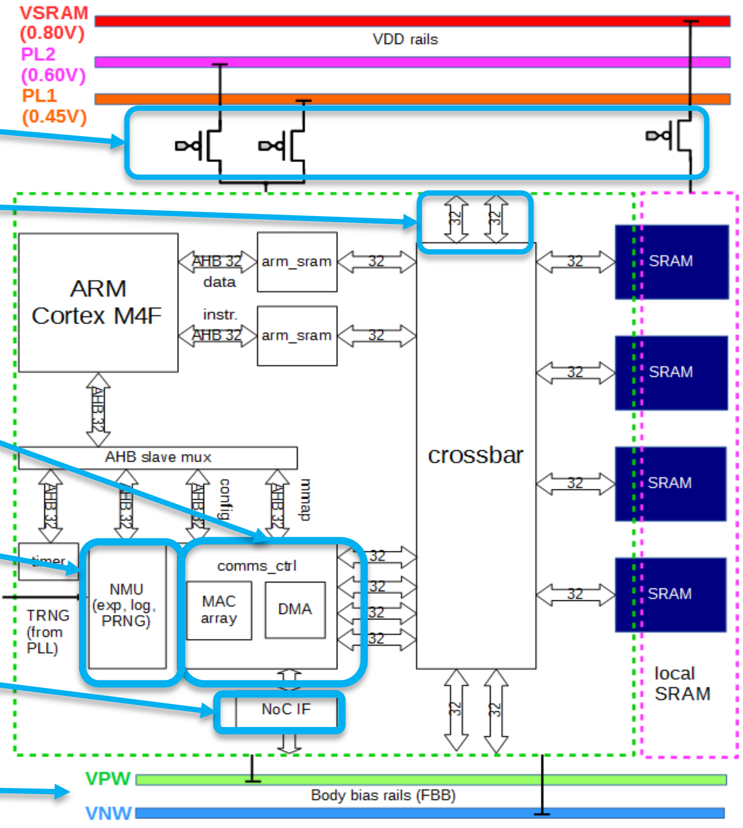
Memory sharing for flexible code, state and weight storage

Multiply-Accumulate accelerator for machine learning

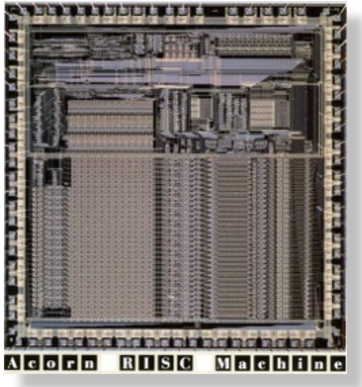
Neuromorphic accelerators and random generators for synapse and neuron computation

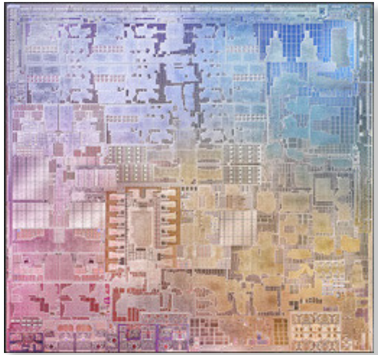
Network-on-Chip for efficient spike communication

Adaptive Body Biasing for energy efficient low voltage operation



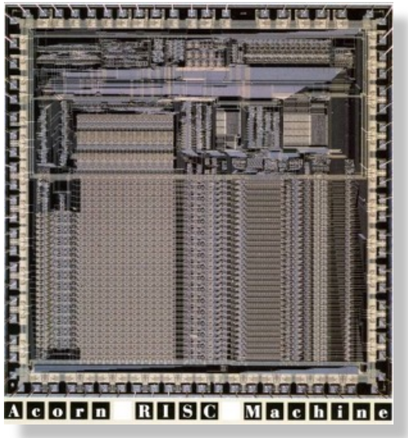
Outline



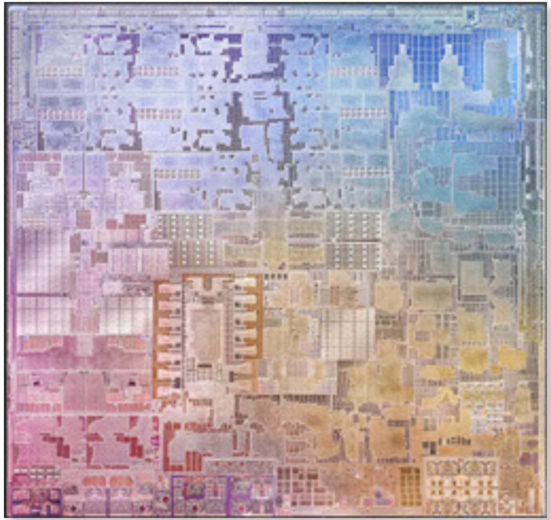


- from little Acorns…
- building brains
- 40 years of Moore’s Law
- how it started… how it’s going

Apple Silicon M1 (2020)



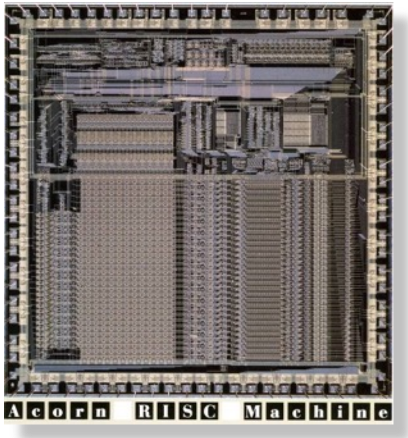
(die images roughly to scale)



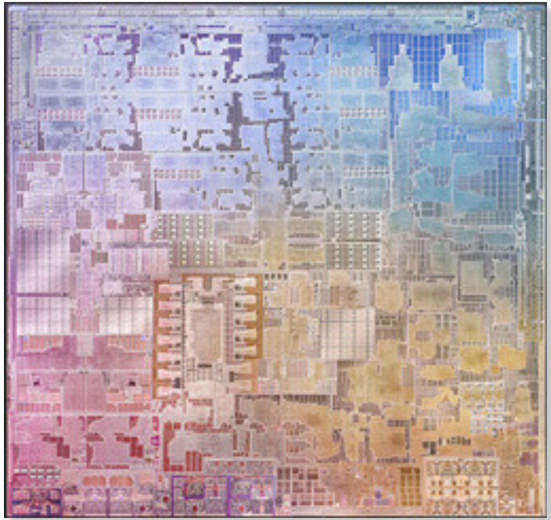
Moore's Law (1965):

- The number of transistors on a chip doubles every ~2 years
  - X 1,000 every 20 years
  - X 1,000,000 in 40 years
- ARM1: 25,000 transistors
- M1: 16 billion transistors

~40 years of Moore's Law



(die images roughly to scale)

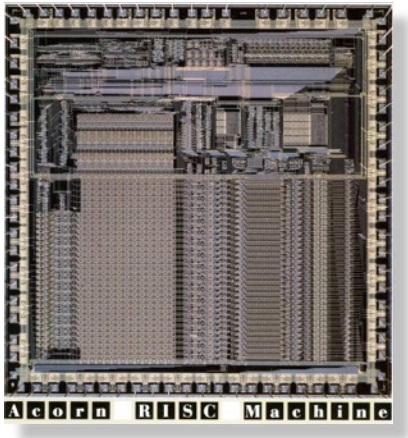


Feature size:

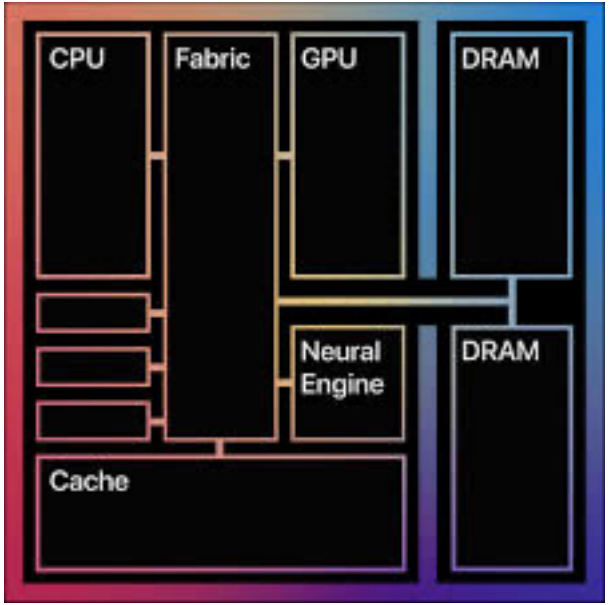
- ARM1: 3μm
- M1: 5nm
  - ~1,000x smaller
  - ~1,000,000x denser

NB: ~4 Silicon atoms per nm

~40 years of Moore's Law



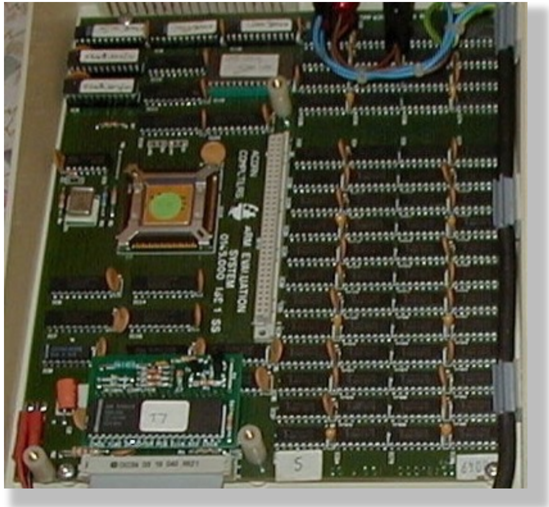
(die images roughly to scale)

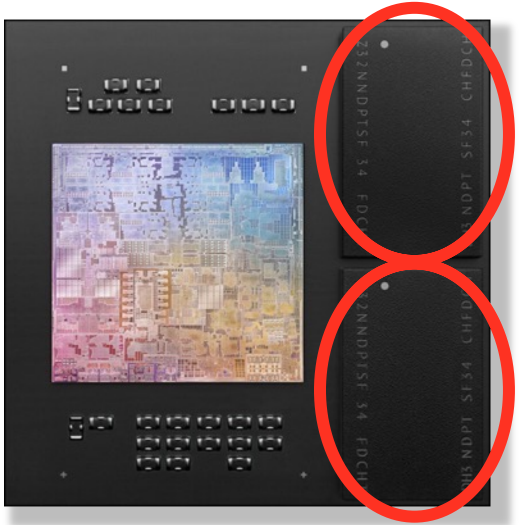


Performance:

- ARM1: 6 MHz, 32-bit
- M1: 3.2 GHz, 64-bit
  - ~1,000x faster
  - 8 ARM cores
    - 4 performance
    - 4 efficiency
- ~10,000x throughput

~40 years of Moore's Law

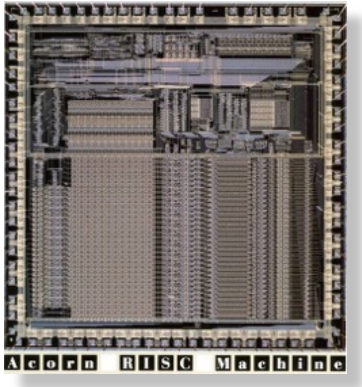


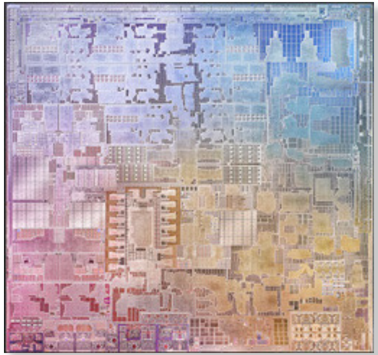


Memory:

- ARM1: 4 Mbytes
  - in 64 packages
- M1: 8/16 Gbytes
  - In 2 packages
- ~1,000x memory size
- ~100,000x density

Outline

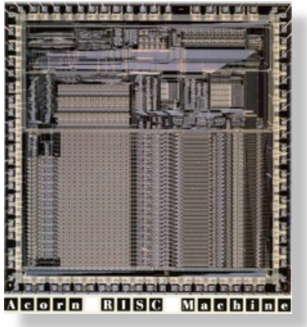




- from little Acorns…
- building brains
- 40 years of Moore’s Law
- how it started… how it’s going

The Call to ARMs

How it started…



How it's going…

nowopen

SpiNNaker book

- Open Access in PDF form (i.e. free!)
  - $90 on paper

20 years in conception and 15 in construction, the SpiNNaker project has delivered the world's largest neuromorphic computing platform incorporating over a million ARM mobile phone processors and capable of modelling spiking neural networks of the scale of a mouse brain in biological real time…

http://dx.doi.org/10.1561/9781680836523