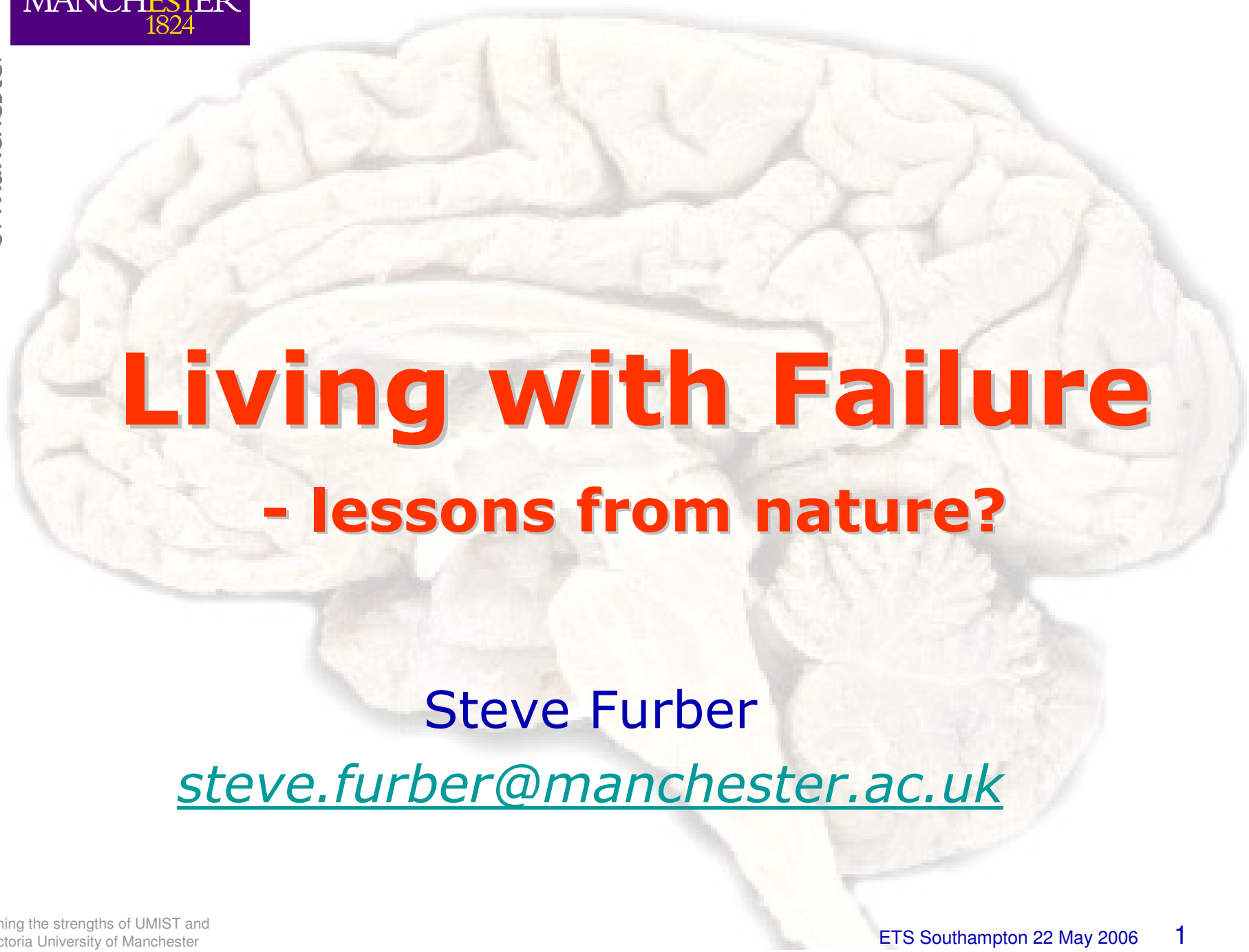


Living with Failure

- lessons from nature?

Steve Furber

steve.furber@manchester.ac.uk

MANCHESTER 1824

The University of Manchester

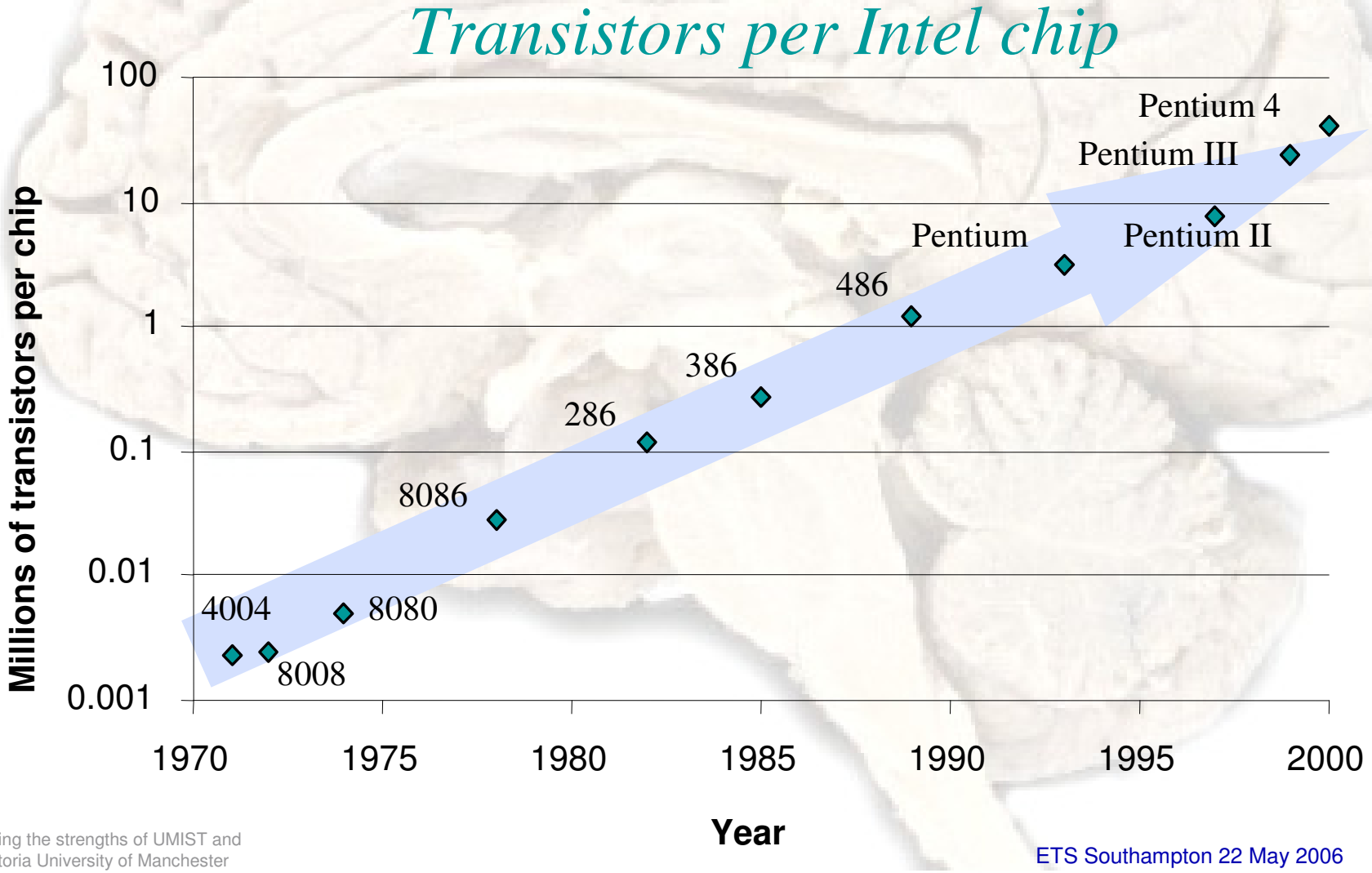
Outline

- Good News/Bad News
- Neurons
- Engineering with Neurons
- A Universal Spiking Neural Network Architecture
- Living with Failure

MANCHESTER 1824

The University of Manchester

The Good News...

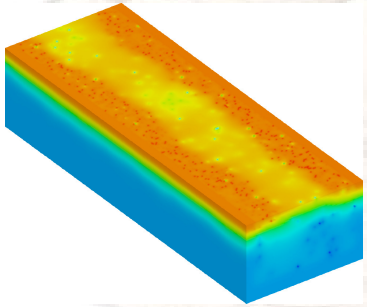


MANCHESTER 1824

The University of Manchester

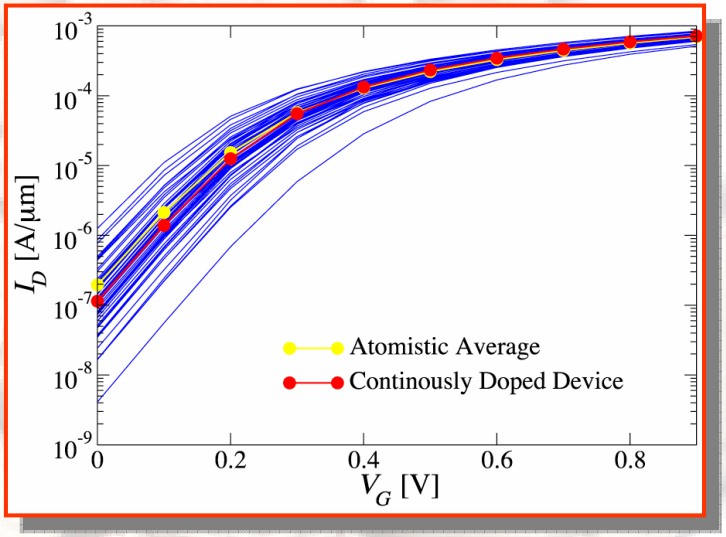
...and the Bad News

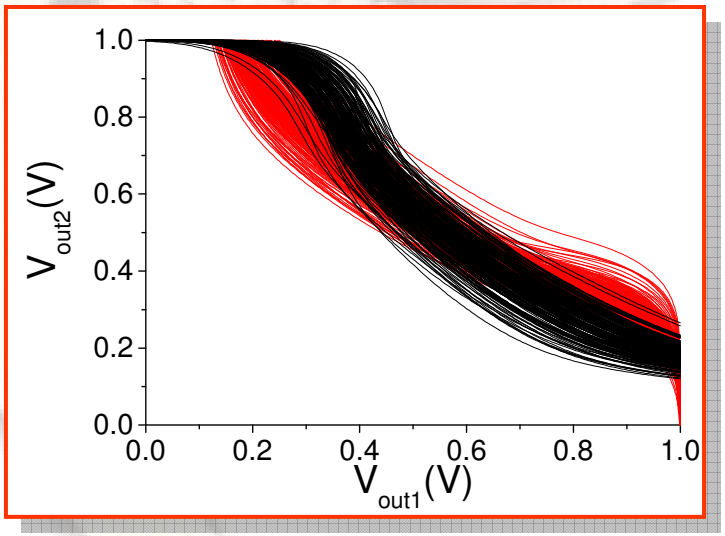
- device variability



&

- component failure

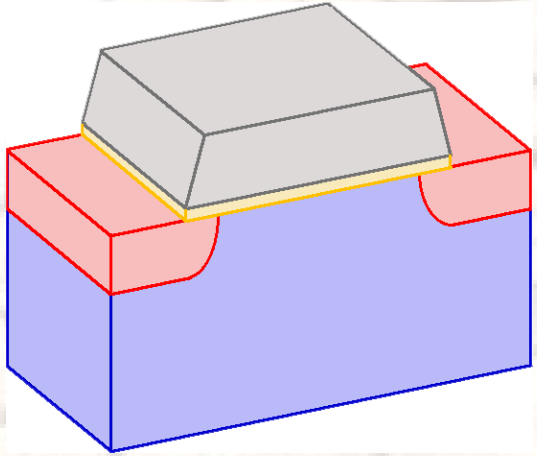




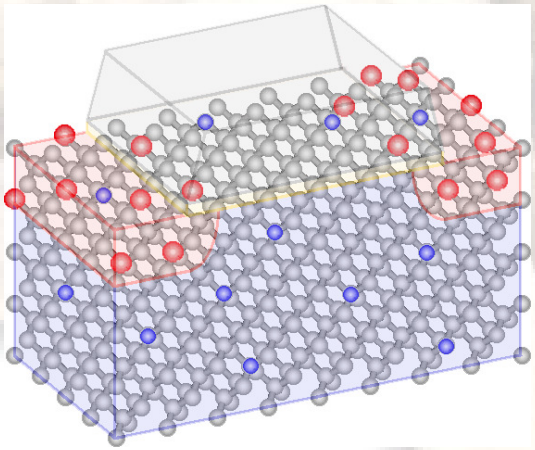
DEVICE MODELLING GROUP

UNIVERSITY of GLASGOW

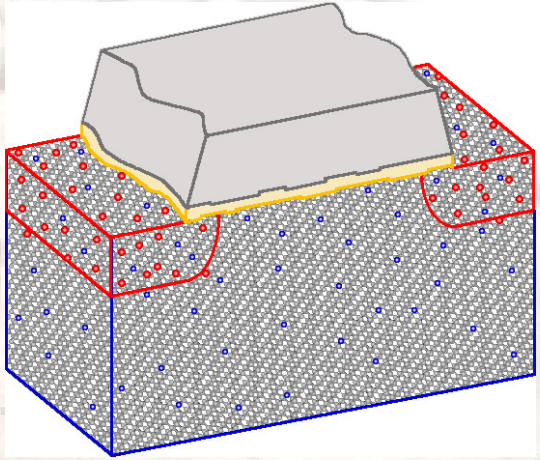
Atomic Scale devices



The simulation
Paradigm now



A 4.2 nm MOSFET
In production 2023



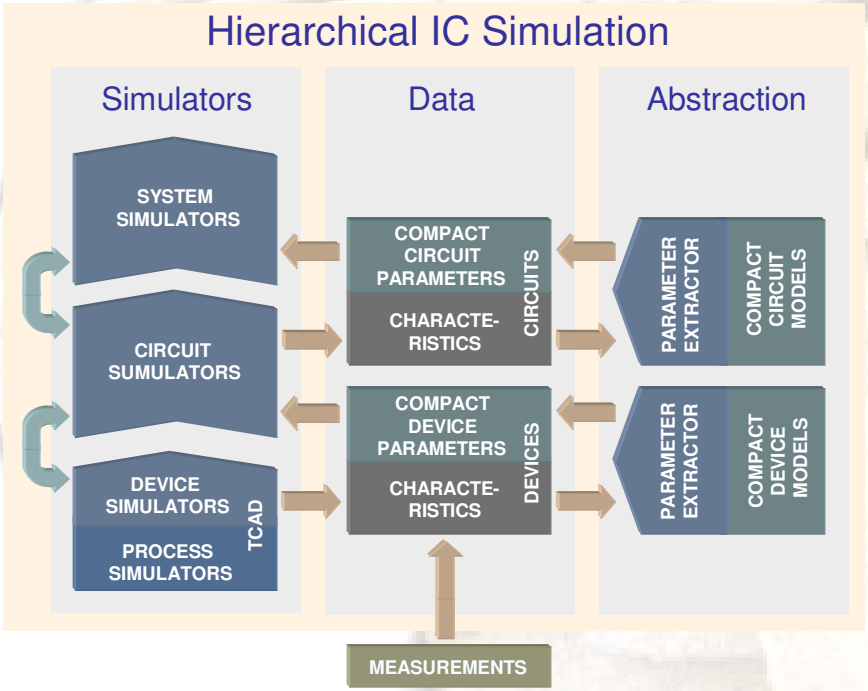
A 22 nm MOSFET
In production 2008

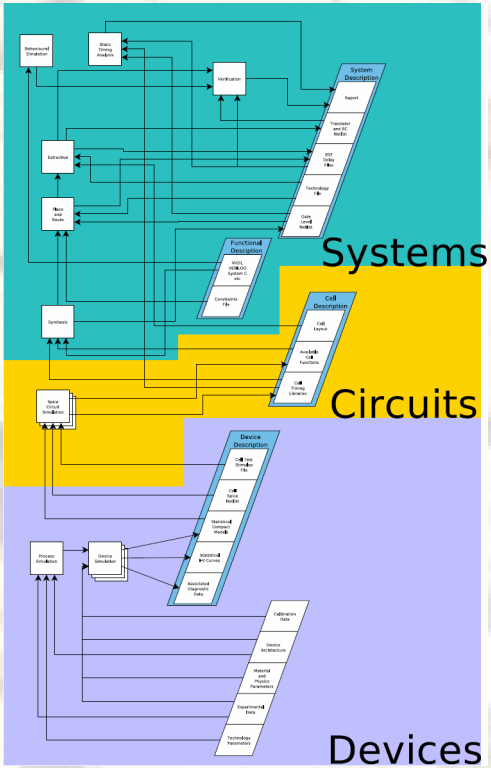
MANCHESTER 1824

The University of Manchester

Meeting the Design Challenges of the Nano CMOS Electronics

A £4.5M EPSRC e-Science project





Simple concept
Integrated Hierarchical
Statistical Design

Complex data and workflows
Data- and Compute-Intensive
Security-Sensitive

MANCHESTER 1824

The University of Manchester

A view from Intel

- the Good News:
  - we will have 100 billion transistor ICs
- the Bad News:
  - billions will fail in manufacture
    - unusable due to parameter variations
  - billions more will fail over the first year of operation
    - intermittent and permanent faults

(Shekhar Borkar, Intel Fellow)

MANCHESTER 1824

The University of Manchester

A view from Intel

- Conclusions:
  - one-time production test will be out
  - burn-in to catch infant mortality will be impractical
  - test hardware will be an integral part of the design
  - dynamically self-test, detect errors, reconfigure, adapt, ...

(Shekhar Borkar, Intel Fellow)

MANCHESTER 1824

The University of Manchester

Outline

- Good News/Bad News
- Neurons
- Engineering with Neurons
- A Universal Spiking Neural Network Architecture
- Living with Failure

MANCHESTER 1824

The University of Manchester

Biological Neural Systems

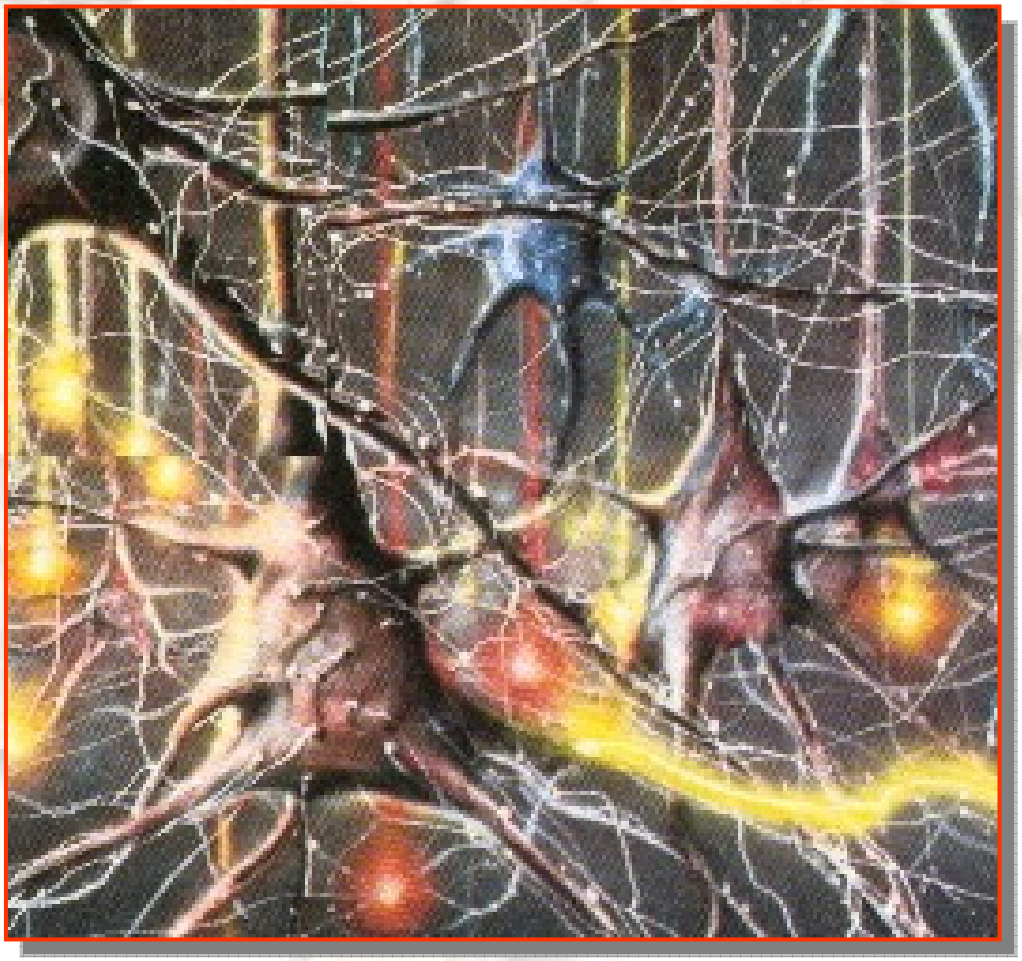
- Brains demonstrate
  - massive parallelism (10^{12} neurons)
  - massive connectivity (10^{15} synapses)
  - excellent power-efficiency (better than CMOS)
  - low-performance components (~ 100 Hz)
  - low-speed communication (~ metres/sec)
  - redundancy (tolerant to component failure)
  - autonomous learning

MANCHESTER 1824

The University of Manchester

Neurons

- A flexible biological control component
  - very simple animals have a handful
  - bees: 850,000
  - humans: 10^{12}



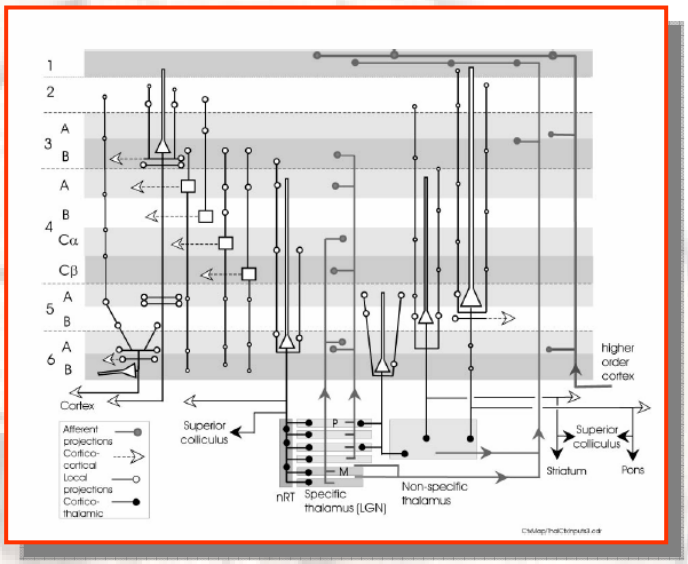
(www.ualberta.ca/ ~kming/homepage.htm)

MANCHESTER 1824

The University of Manchester

Neurons

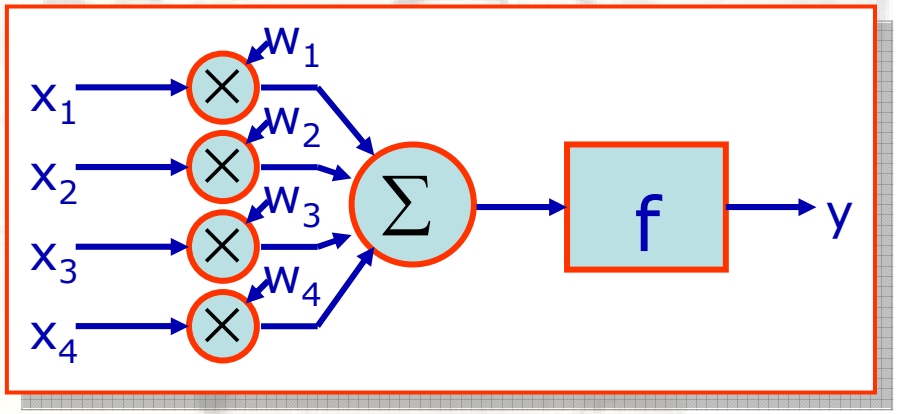
- regular high-level structure
  - e.g. 6-level cortical microachitecture
    - low-level vision, to
    - language, &c
- random low-level structure
  - adapts over time



(faculty.washington.edu/ rhevner/Miscellany.html)

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

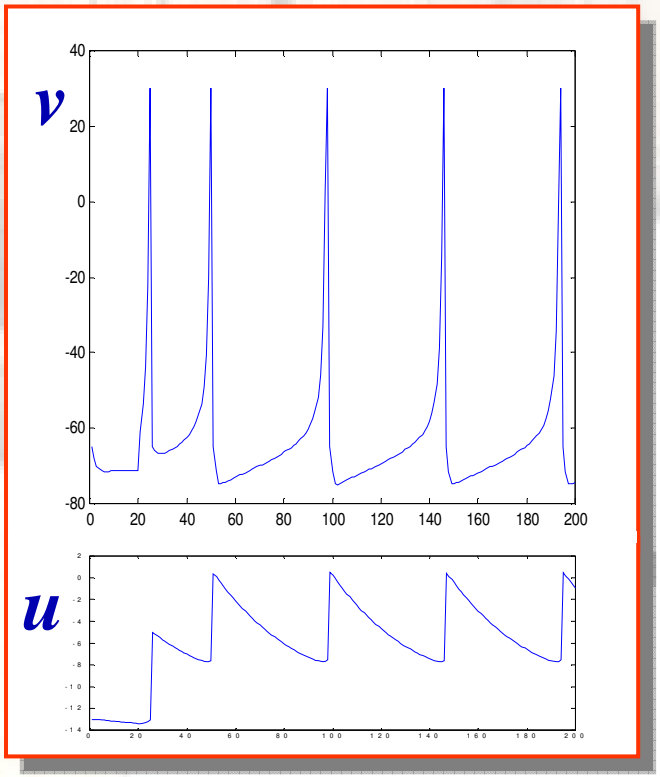
- Izhikevich model (www.izhikevich.com)
  - two variables, one fast, one slow:

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

  - neuron fires when $v > 30$; then:

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

  - a, b, c & d select behaviour, for example:

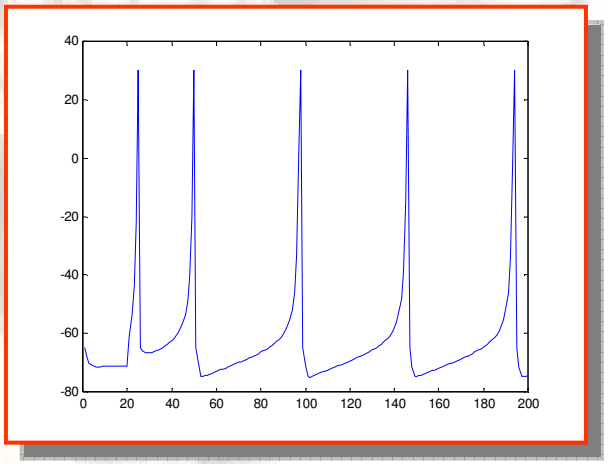


MANCHESTER 1824

The University of Manchester

Communication

- Spikes
  - biological neurons communicate principally via 'spike' events
  - asynchronous
  - information is only:
    - which neuron fires, and
    - when it fires



MANCHESTER 1824

The University of Manchester

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

MANCHESTER 1824

The University of Manchester

Outline

- Good News/Bad News
- Neurons
- Engineering with Neurons
- A Universal Spiking Neural Network Architecture
- Living with Failure

MANCHESTER 1824

The University of Manchester

Neural Engineering

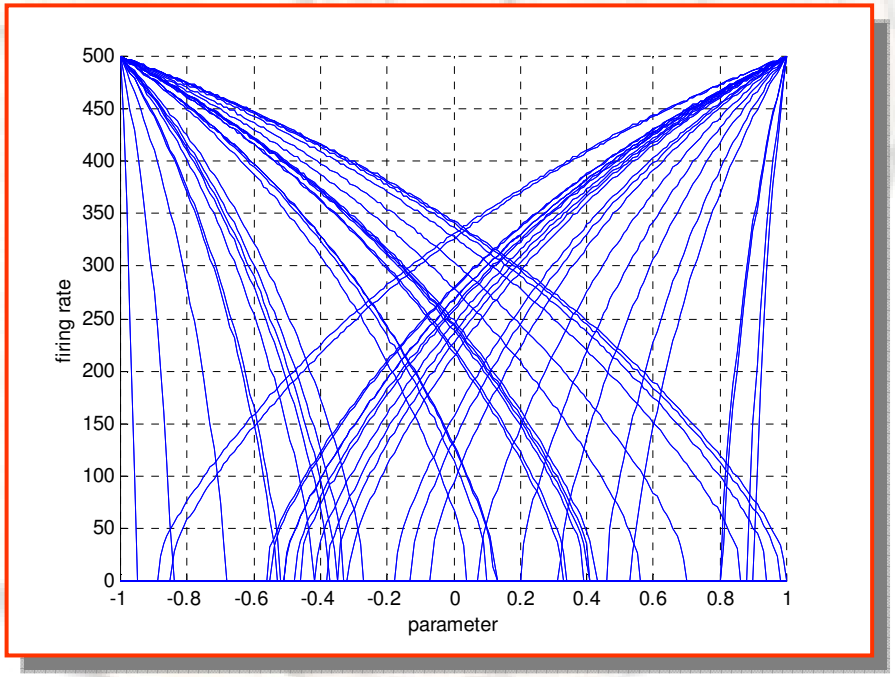
- The Neuron as a component
  - fan-in/-out = ~1,000
  - single output, asynchronous ‘event’
- Capture biology’s fault-tolerance
  - populations of equal peers
    - NB binary codes are out!
  - exploit population diversity
    - variability is now an advantage

MANCHESTER 1824

The University of Manchester

Neural Engineering

- Firing rate population codes
  - N neurons
  - diverse tuning
  - collective coding of a physical parameter
  - accuracy $\propto \sqrt{N}$
  - robust to neuron failure



(Neural Engineering, Eliasmith & Anderson 2003)

MANCHESTER 1824

The University of Manchester

Neural Engineering

- Single spike/neuron codes
  - choose N to fire from a population of M
  - order of firing may or may not matter

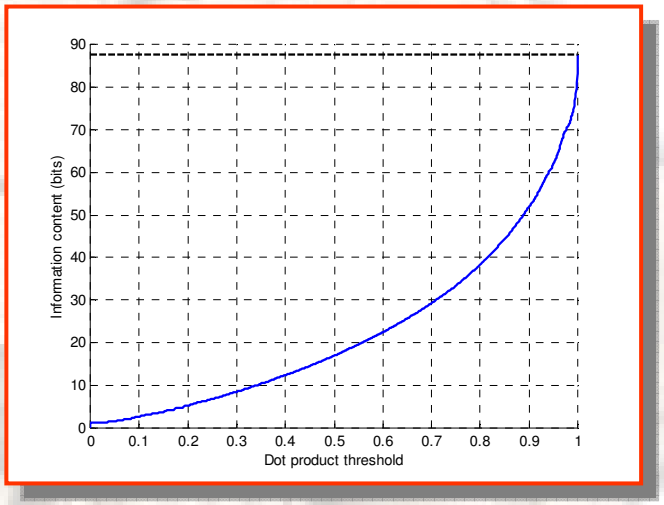
	Unordered N-of-M	Ordered N-of-M	M-bit binary
Number of codes	C_N^M	M!/(M-N)!	2^M
e.g. M=100, N=20	10^{21}	10^{39}	10^{30}
e.g. M=1000, N=200	10^{216}	10^{591}	10^{301}

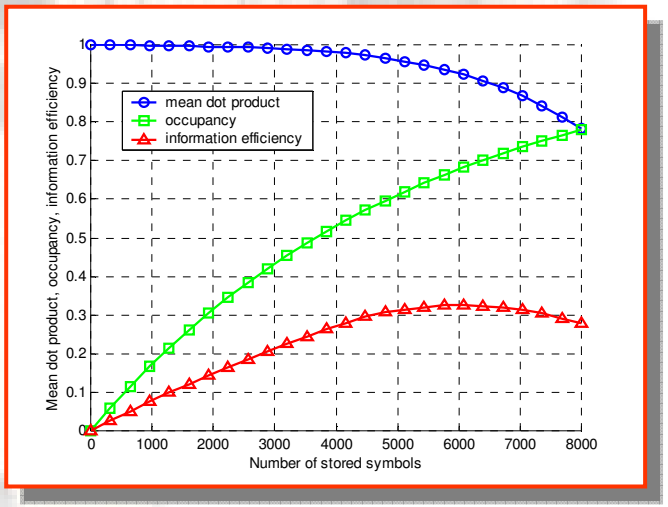
MANCHESTER 1824

The University of Manchester

Neural Engineering

- Rank-order codes
  - information content of ordered 11-of-256 code vs fault-tolerance:
  - associative memory performance against load:



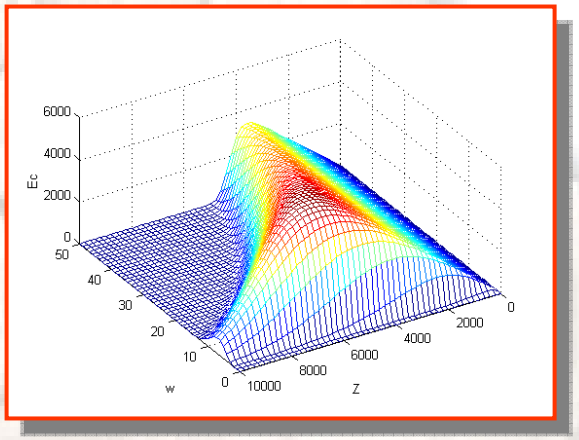


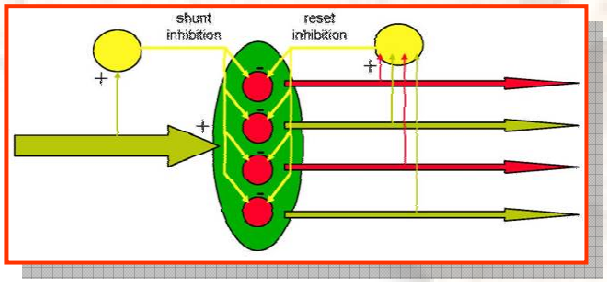
MANCHESTER 1824

The University of Manchester

Neural Engineering

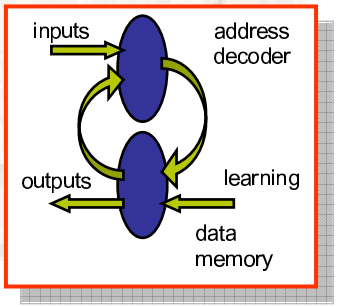
- *N*-of-*M* sparse distributed memory (SDM)
  - suits spiking neuron implementation
  - scalable, inexact associative memory





- Rank-order coded SDM
  - static memory stores dynamic neural codes

- Neural asynchronous FSM
  - learns and recalls time sequence data



(work by Joy Bose)

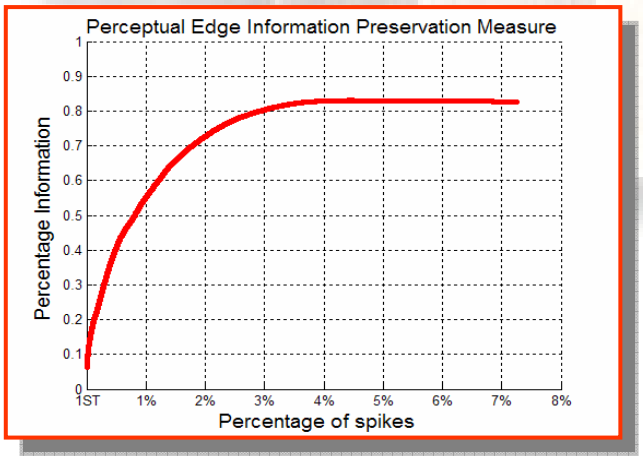
MANCHESTER 1824

The University of Manchester

Neural Engineering

- Rank-order coded images
  - temporal dynamics of information conveyance to brain through optic nerve

Original 0.5% 1% 10% 50%



(work by Basabdatta Sen, using image information measure from Petrovic and Xydeas, 'Objective evaluation of signal-level image fusion performance')

MANCHESTER 1824

The University of Manchester

Neural Engineering

- Testing
  - how do you test a fault-tolerant system of equal-peer neural components?
    - check each neuron works
      - can you scan-test a neuron?
    - statistical analysis
      - e.g. memory performance against predicted
    - measure system-level functional accuracy?
      - get it to sit GCSEs?

MANCHESTER 1824

The University of Manchester

Outline

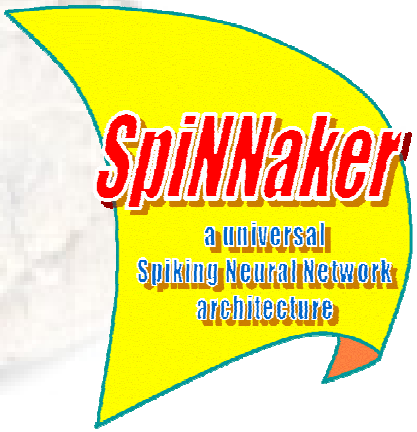
- Good News/Bad News
- Neurons
- Engineering with Neurons
- A Universal Spiking Neural Network Architecture
- Living with Failure

MANCHESTER 1824

The University of Manchester

SpiNNaker project

- 20 ARM968 processors per chip
  - 100 million transistors per chip
- Scalable up to systems with 10,000s of chips
  - 100,000s of processors
  - ~10^{14} MIPS total
- Power ~ 25μw/neuron



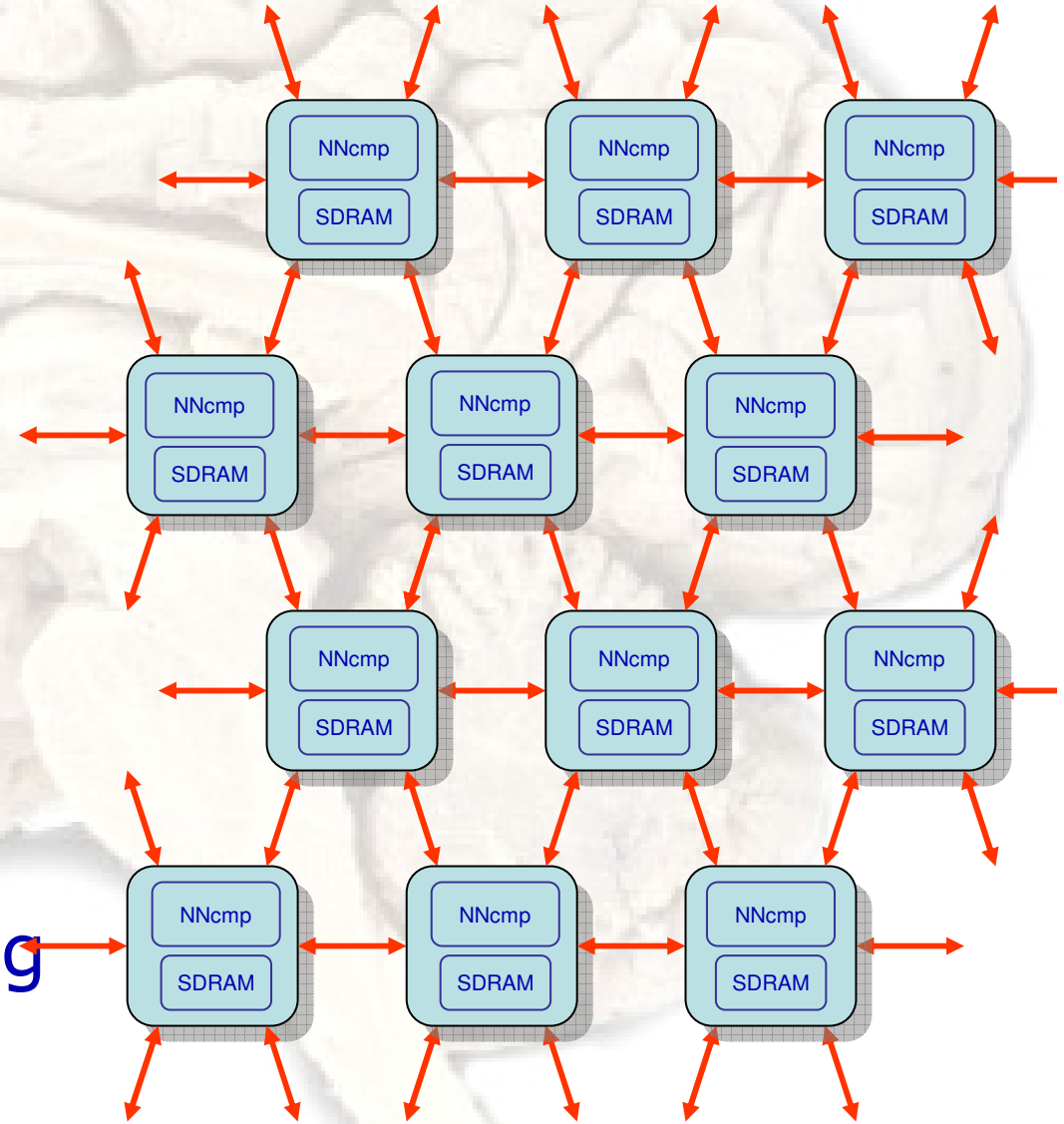
EPSRC University of Southampton MANCHESTER 1824 ARM Silistix

MANCHESTER 1824

The University of Manchester

2D Multi-Chip System

- 2-chip nodes
  - CMP + SDRAM
- Point-to-point links
  - self-timed
  - NoC protocol
- Async packets
  - source key routing
  - key is payload

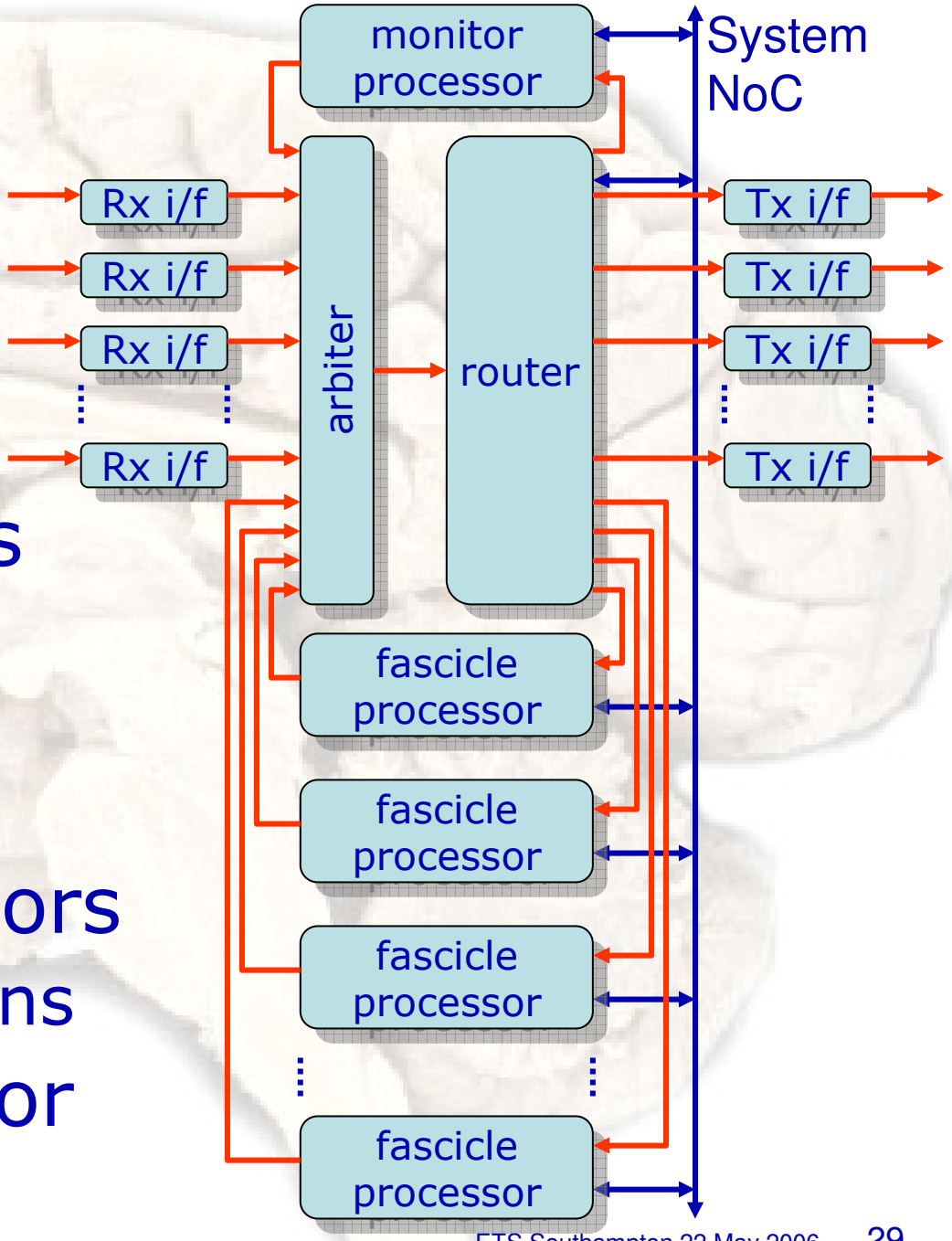


MANCHESTER 1824

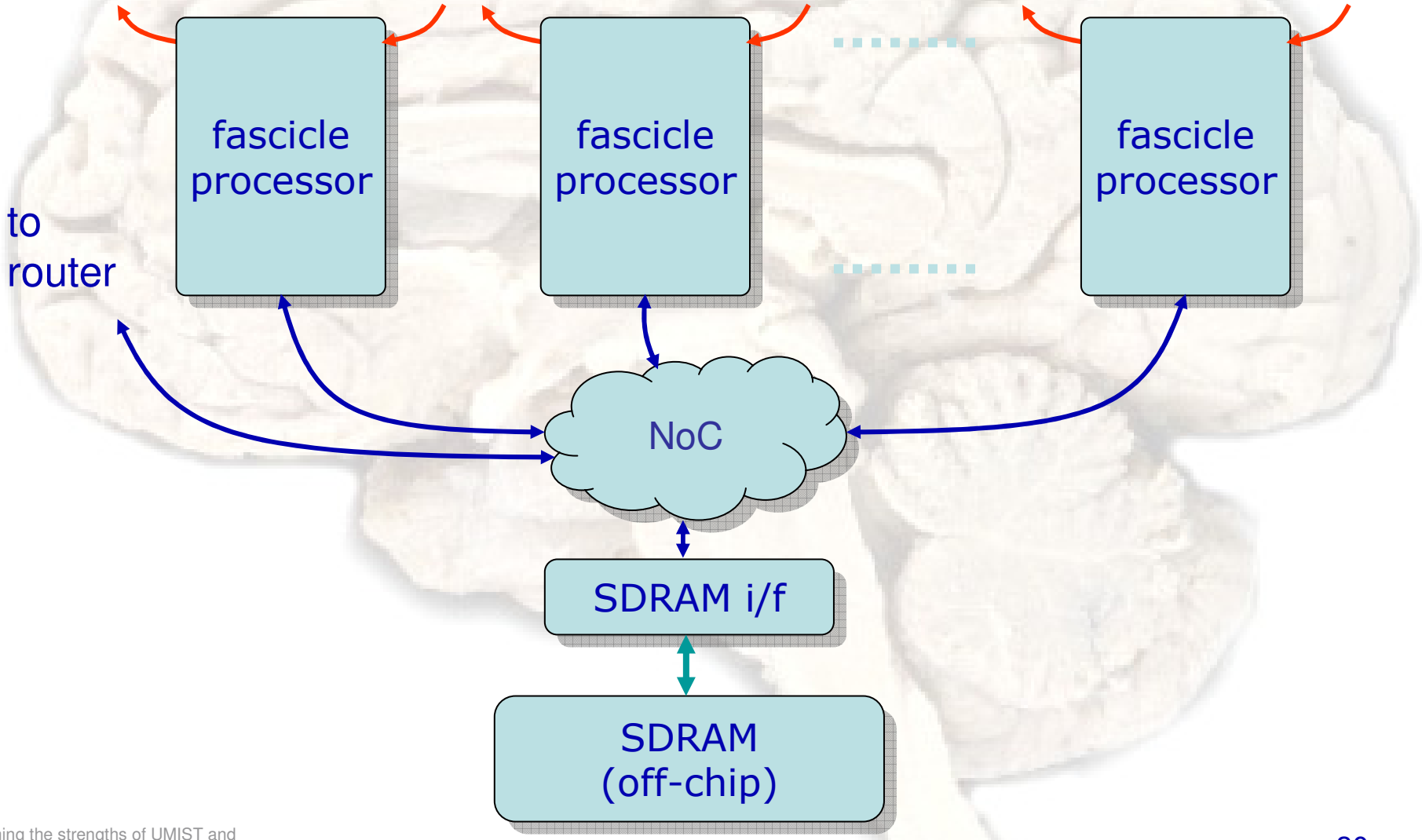
The University of Manchester

Comms NoC

- Communications router
  - associative
  - multicast
- Fascicle processors
  - each ~1k neurons
- Monitor processor



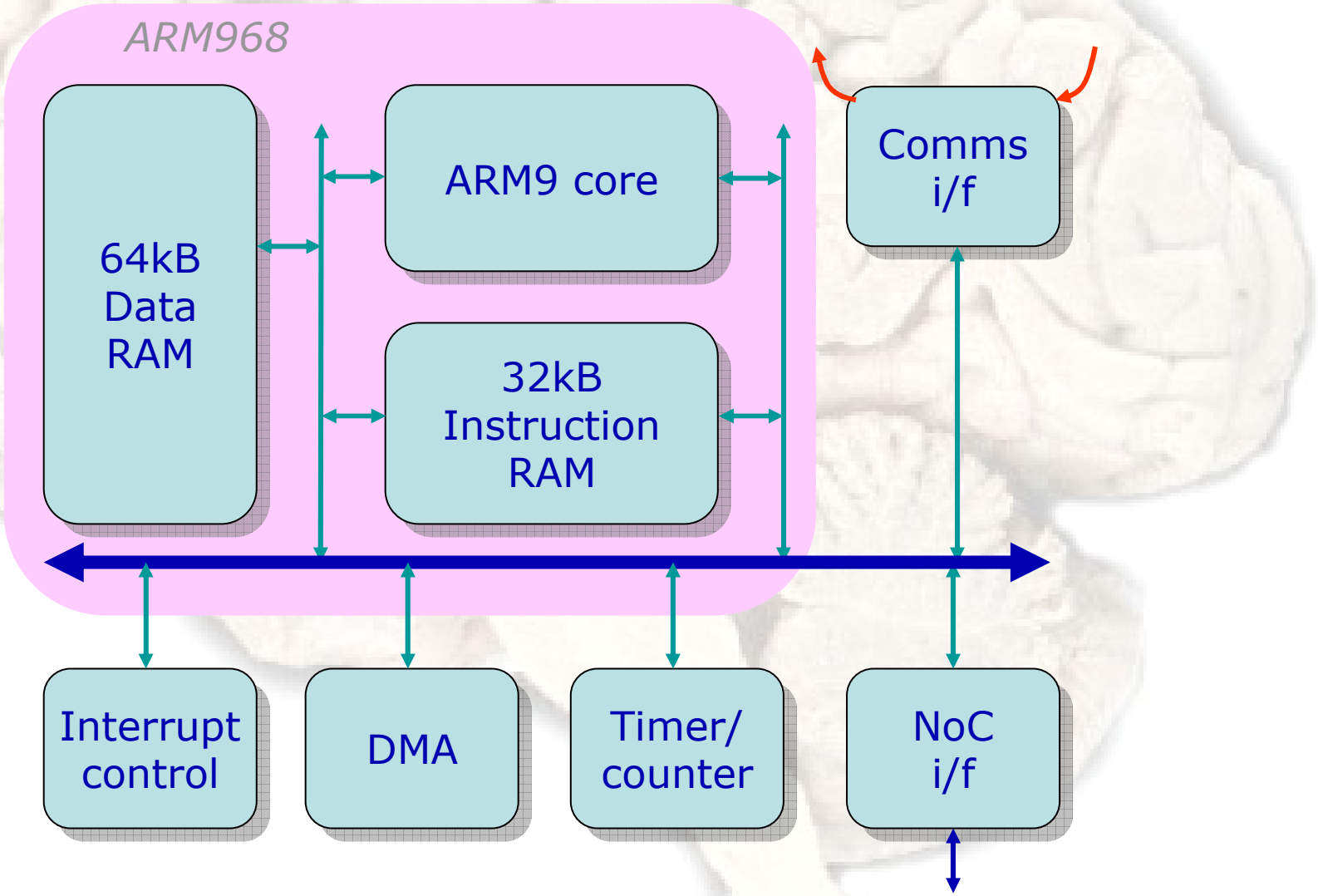
System NoC


fascicle processor
fascicle processor
fascicle processor
to router
NoC
SDRAM i/f
SDRAM (off-chip)

MANCHESTER 1824

The University of Manchester

Fascicle processor



MANCHESTER 1824

The University of Manchester

Fault-tolerance

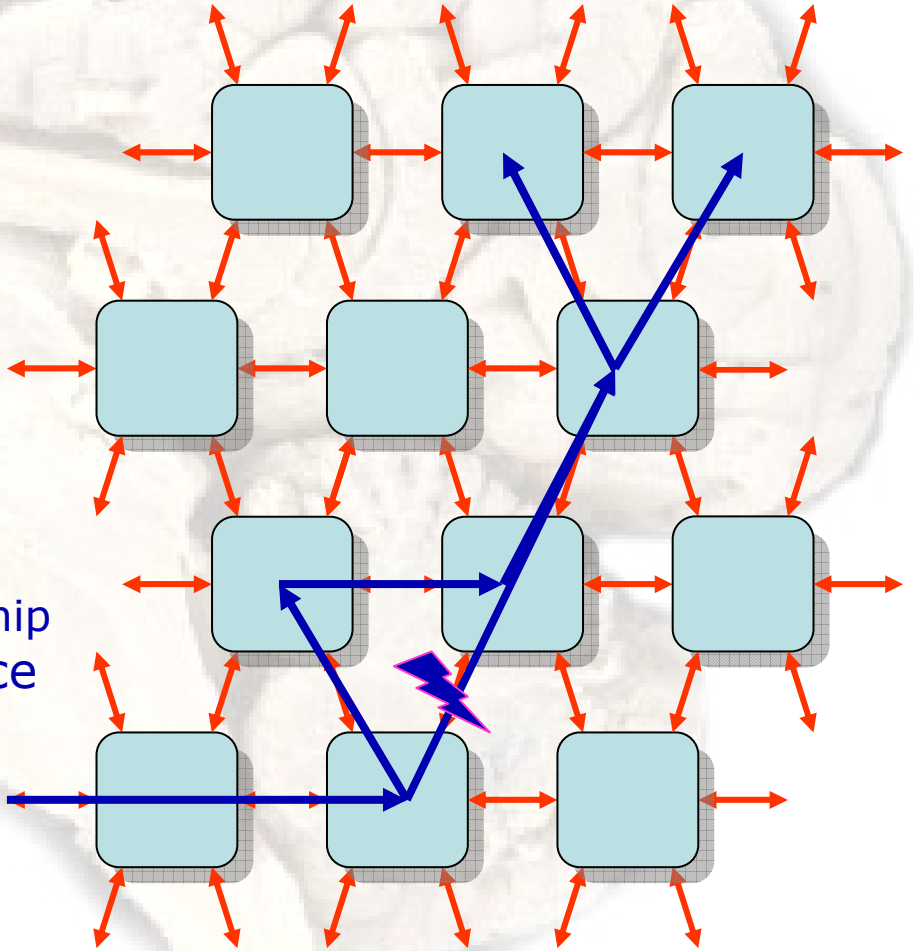
- System of 10,000 ICs will suffer component failure
- Objective:
  - real-time performance
    - so check-point & wind-back inapplicable
  - failures cause transient performance dip
    - system recovers as gracefully as possible
- Biological system is fault-tolerant
  - adult humans lose 1 neuron/sec
- But electronic model concentrates functionality
  - many more local 'single points of failure'
  - more akin to a stroke than losing a neuron

MANCHESTER 1824

The University of Manchester

Fault-tolerance

- Approach
  - high-speed hardware 'emergency' services
  - distributed software monitoring layer
    - detect failures & take appropriate action
  - run-time functional migration
    - move work to another processor
      - on same or different chip
    - reroute packets to reduce congestion
  - still some transient loss
    - dropped packets
    - neural state in failed processors



MANCHESTER 1824

The University of Manchester

Production Test

- We can tolerate:
  - a couple of dead ARM968 subsystems
  - localised memory faults
  - broken Router entries
- There are single points of failure
  - SDRAM i/f, Router function, power shorts, ...
- What is the production acceptance level?
  - allow for more early-life component failures
- Is scan useful?
  - needs full start-up and run-time self-test
- Focus on:
  - embedded self-test, run-time error detection
  - use the same approach for production test?

MANCHESTER 1824

The University of Manchester

Outline

- Good News/Bad News
- Neurons
- Engineering with Neurons
- A Universal Spiking Neural Network Architecture
- Living with Failure

MANCHESTER 1824

The University of Manchester

Living with Failure

- Future IC technologies promise
  - high device variability
  - high component failure rates
- Designing with ~30% failures
  - is beyond current knowledge
  - but biology manages it
- Reliable systems on unreliable platforms
  - a Grand Challenge for electronic design?
  - requiring a rethink of test methodologies?