

Mamba - Exploring Manycore Architectures

Gregory Chadwick & Simon Moore



UNIVERSITY OF
CAMBRIDGE

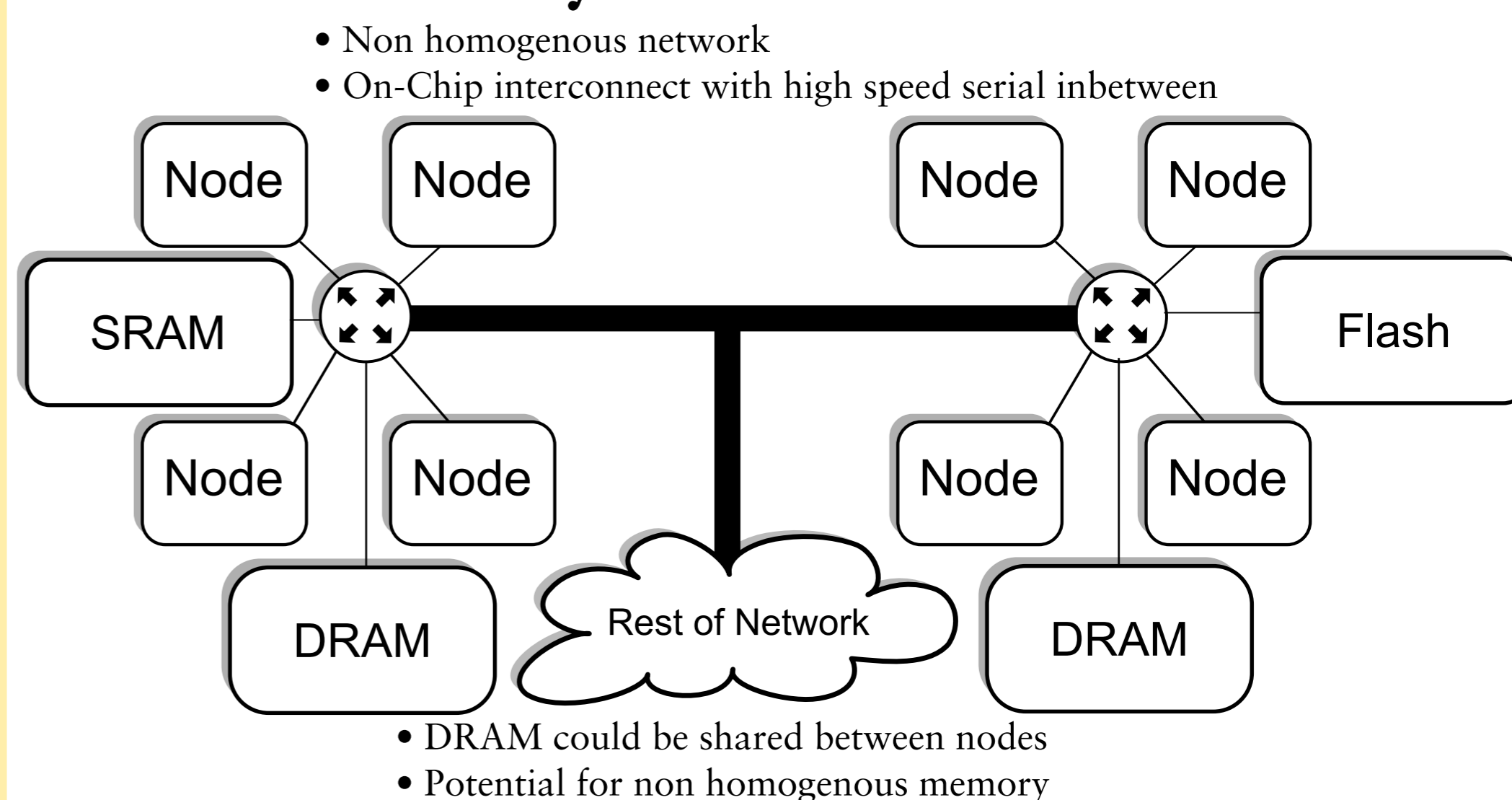
Computer Laboratory

The Challenge

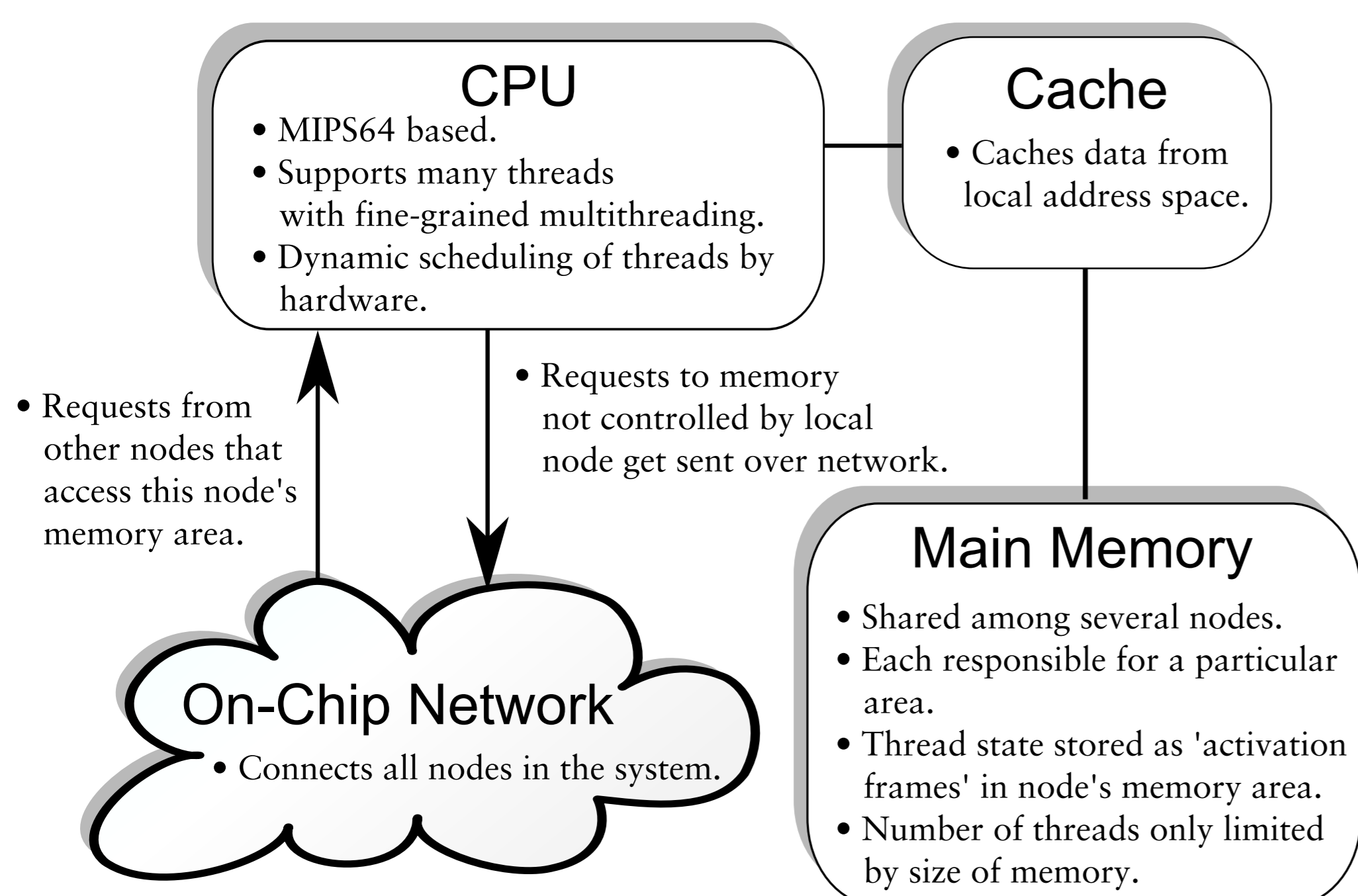
Simple improvements in computational power through technology scaling are no longer available at historical rates. We must turn to explicit parallelism to continue receiving gains.

- How does this affect our programming model?
- What new models are available?
- How can the hardware assist?
- What do we ask of the programmer, the compiler, the hardware and any runtime system?

Nodes Networked Together - Many Possibilities



A Mamba Node



Our Work - Mamba

We are in the process of designing and implementing (Using FPGAs) an architecture we have called Mamba. Mamba is being used to help us explore the multicore design space.

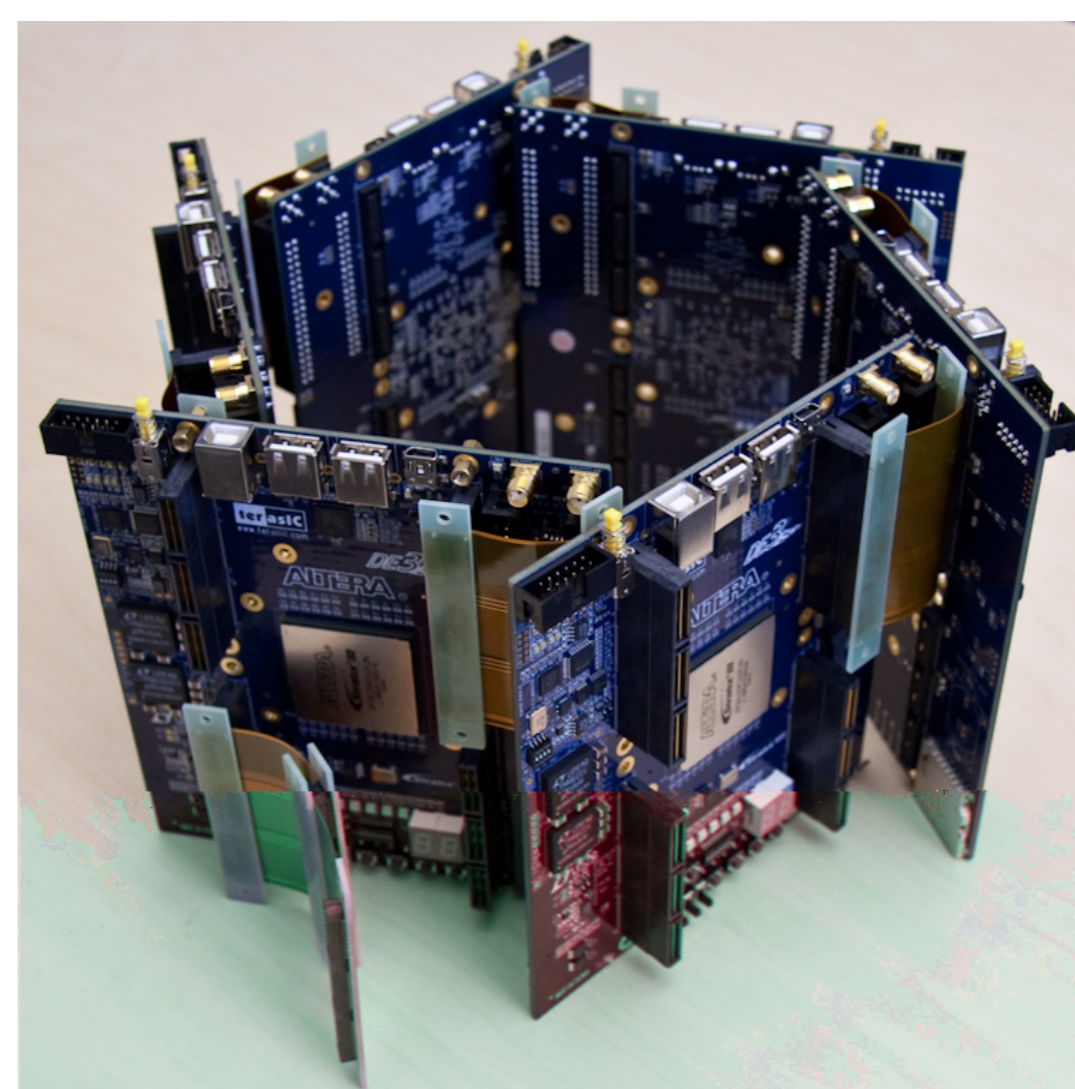
- Mamba consists of a network of nodes, each with its own CPU, cache and area of memory.
- Mamba supports many threads, which are very lightweight.
 - Encourage programmers to parallelise as much as possible to gain scalability
- Each word of memory has a presence bit, which is used as a synchronisation mechanism.
 - Used for fine grained synchronisation, to make it easier to use many threads
- First implementation completed using a Stratix III on a DE3 board.

Computation and Communication

Technology scaling favours transistors over wires, so communication becomes relatively more expensive than computation.

- We need to minimise the communication within a system.
- We may get gains from replacing communication with computation.
- We need to make the communication within a system more explicit to a programmer so they can better optimise for it.

DE3 Boards - Used to implement initial version of Mamba



- FPGAs allow us to run test programs at a decent speed, without having to do a full, expensive and inflexible silicon implementation.

Gregory.Chadwick@cl.cam.ac.uk
Simon.Moore@cl.cam.ac.uk

Computer Architecture Group

<http://www.cl.cam.ac.uk/research/comparch/>