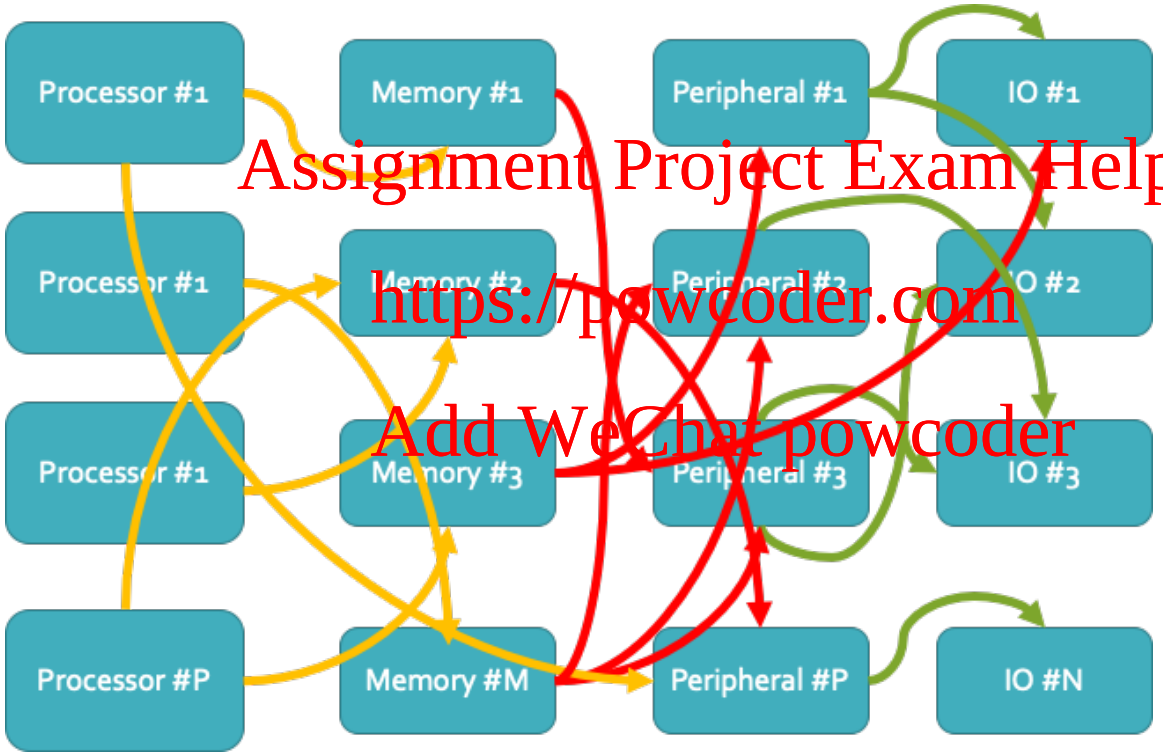
Computer architecture: interconnects

EE3087

Dr Fei Xia and Dr Alex Bystrov

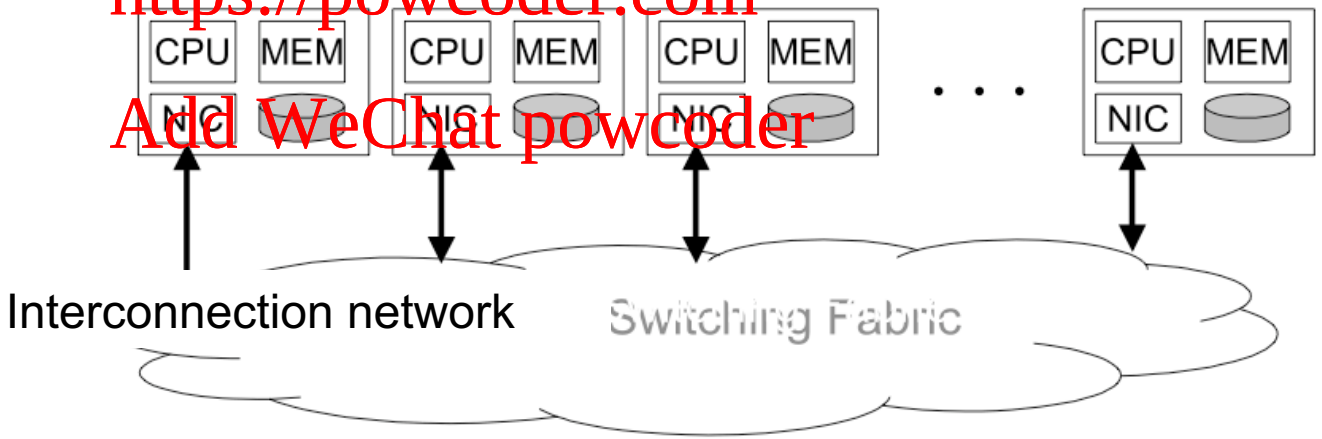
Let's play mash-up



Communication requirements are unpredictable at design time. We need systematic arrangements of interconnects to serve such needs.

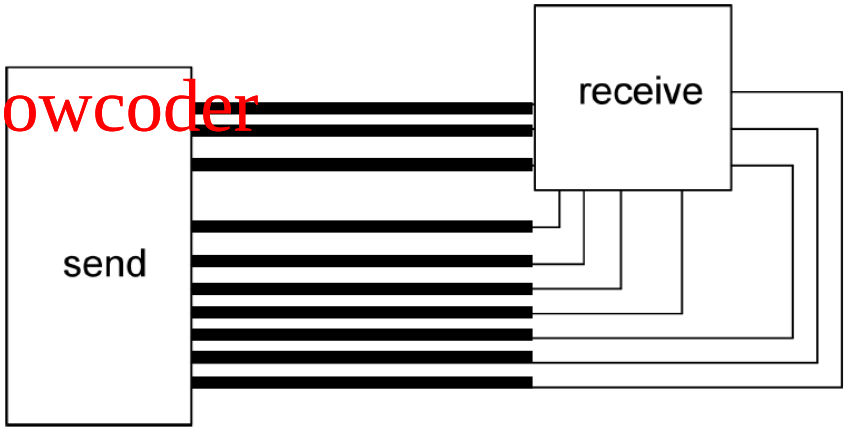
Computer interconnect

- A functional picture
  - Interconnect system is like a sort of memory through which all active units communicate



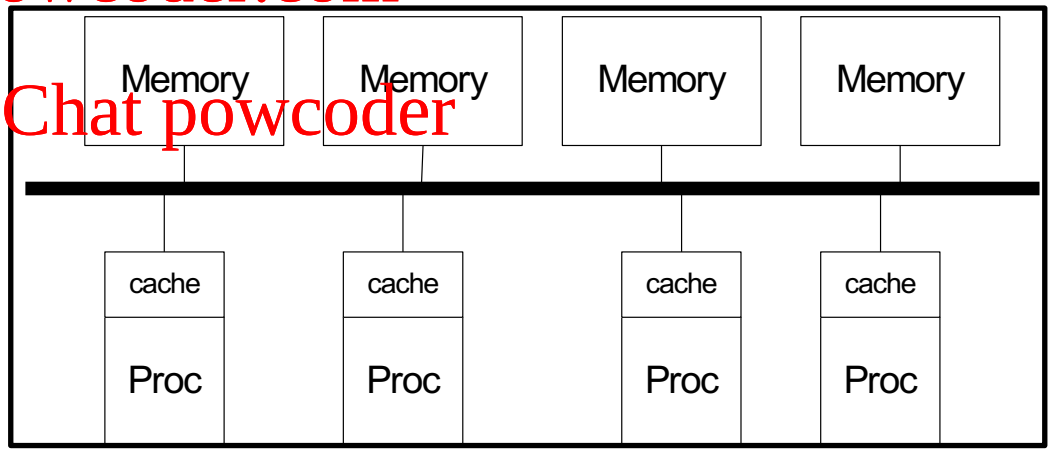
Interconnect design affects real-time operations

- Multi-bit interconnects can suffer from skew
  - Important for real-time system design
  - Parts of the data arrive earlier than other parts
  - Needs to be carefully considered during the design process



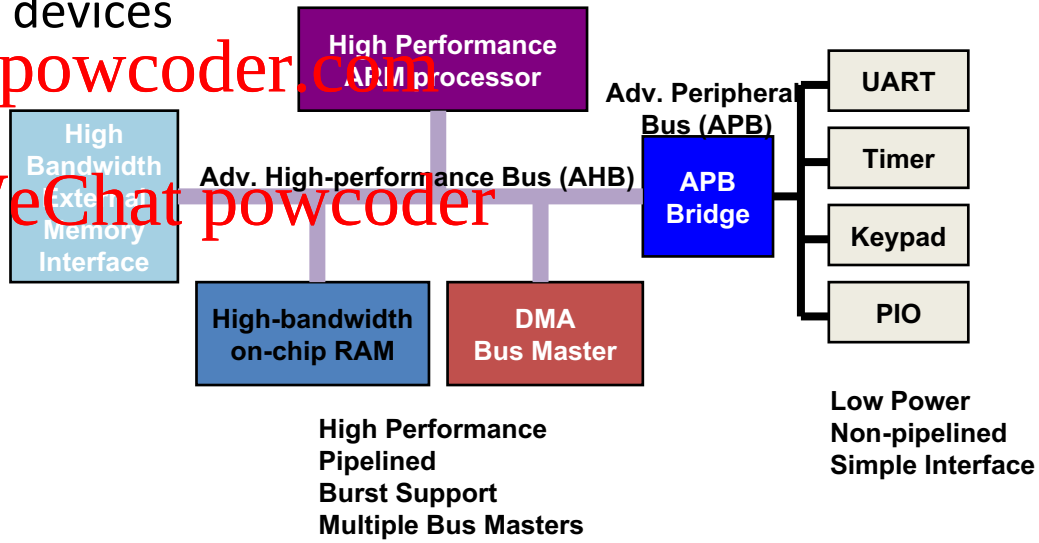
Shared bus

- A number of wires shared between multiple units for data communication
  - At any time, only a pair of the units use these wires
  - Arbitration is used to determine who gets to use the bus when
  - Potential bottleneck
  - Very popular in current systems



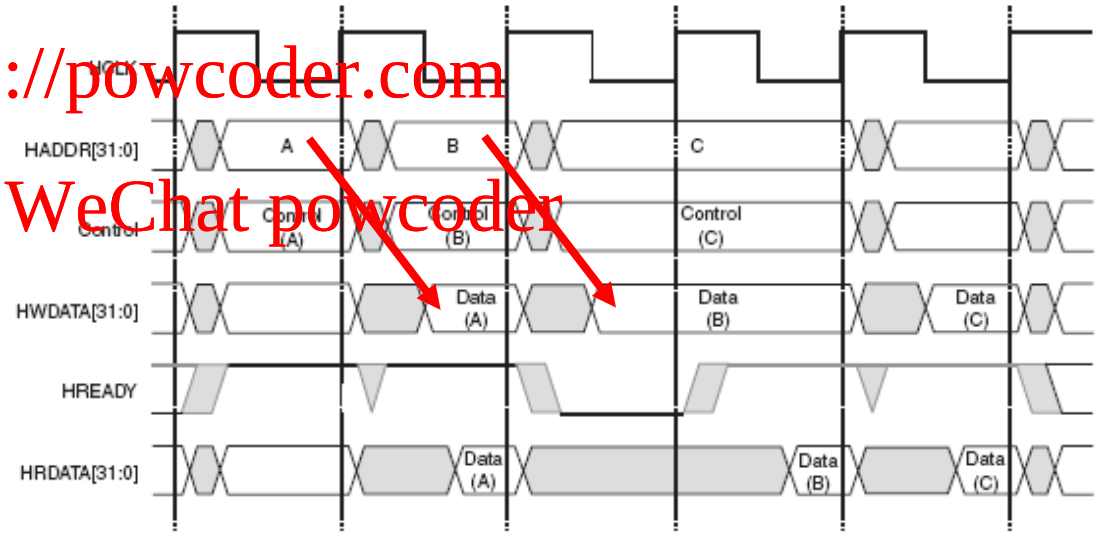
Bus example: ARM AMBA

- Advanced microprocessor bus architecture
  - Divided into two parts
  - For two different types of devices



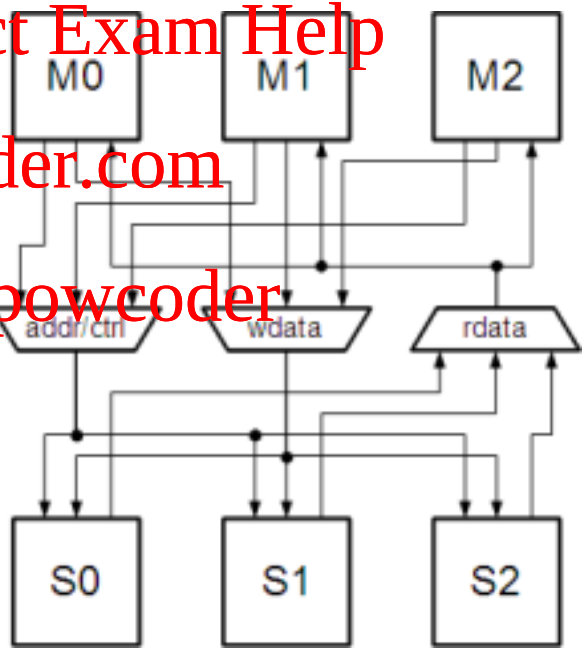
AHB pipelining

- AHB provides transaction pipelining
  - Increases bus bandwidth
  - Separate address and data buses allows the address and data phases of write/read to be pipelined
  - When one transaction is in the address phase, the previous one may be moving data



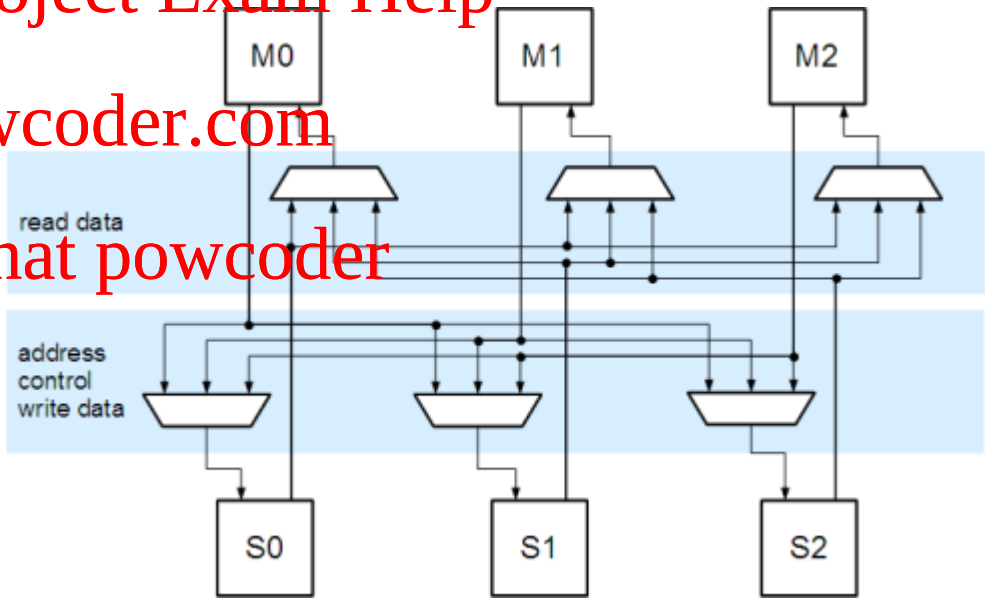
Parallelism?

- No parallelism between multiple cores because they share the same multiplexes



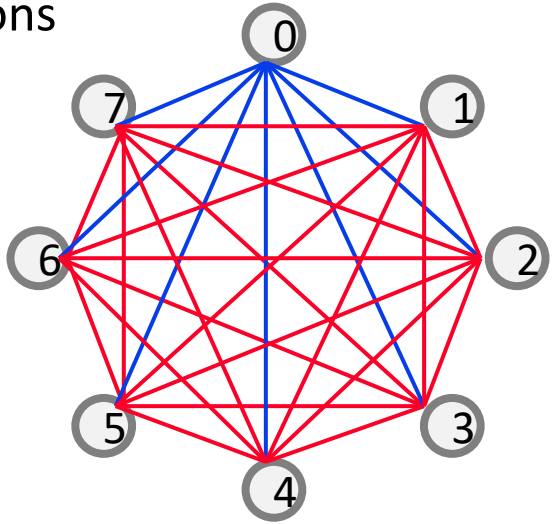
Improved version

- AXI4 AMBA standard
  - Multi-layering allows the write and read/address multiplexes to be in different layers
  - Can have one active mux per layer at any one time
  - Can have one reading at the same time as another writing between two different masters and two different slaves



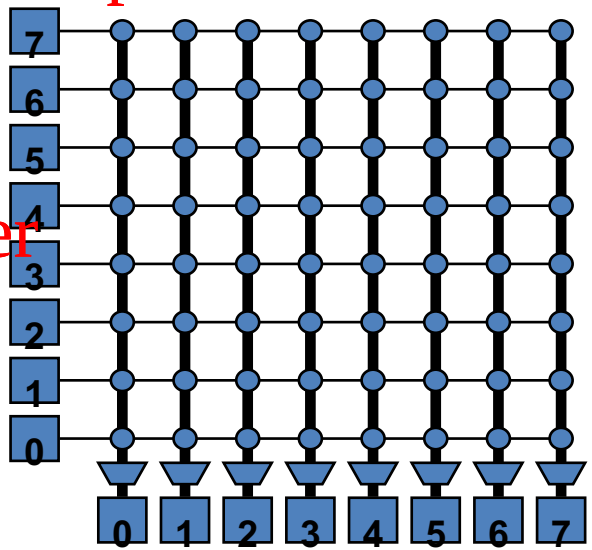
Point to point

- Ideal situation
  - Every interconnect system tries to functionally do this
  - Actually doing this is not realistic in most situations
  - Very high cost (O(N2) number of direct links)
  - Not feasible for 2D chip layout



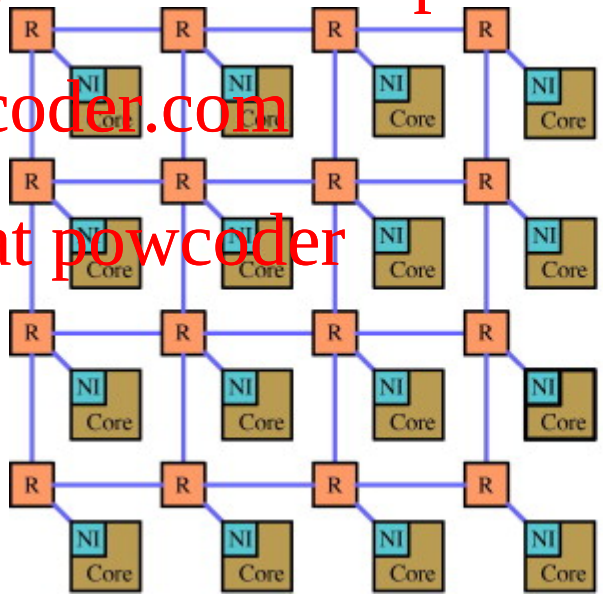
Crossbar

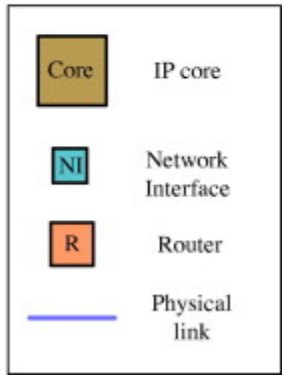
- Switches at the intersecting points programmably link pairs of communicating units
- Parallelism among non-conflicting pairs of units (those who do not share an intersection)
- High cost $O(N^2)$
  - Not scalable to large N
- Used in real examples (core-to-cache networks)



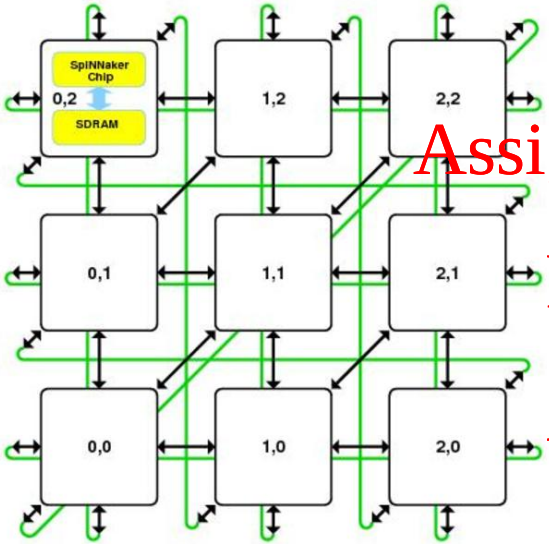
Networks on chip (NoC)

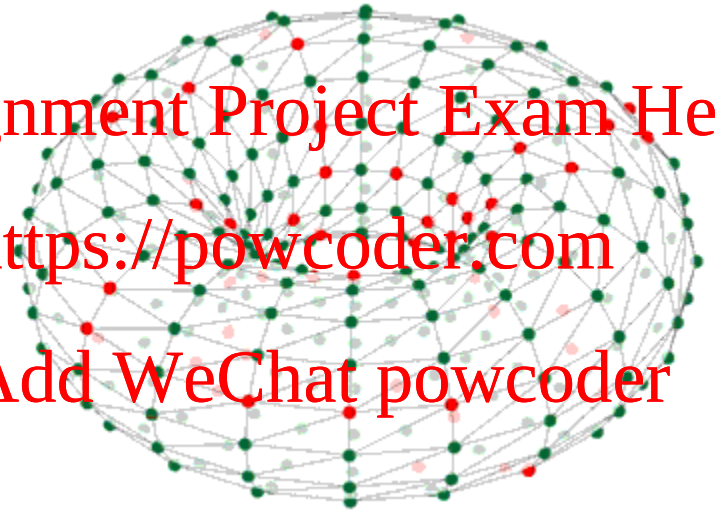
- Imagine running a computer network on a single chip
  - Packet-based data transfer
  - Network protocols
  - Widely used
  - Scalable O(N)
  - Complicated system for routing and arbitration, although simpler routing protocols than off-chip networks
  - FIFO buffers everywhere





Example: SpiNNaker





Neuralmorphic computing based on the spiking neural network model

Brain simulation is the most prominent application

Million-core supercomputer based on a NoC interconnect fabric

Quiz

- You are asked to design the interconnect for a 16-core system. The cores are organized in four groups (with 4 cores each), each with a shared L3 cache and main memory. What type of interconnect would suit such a system? Explain with appropriate reasoning.
  - There is no one ‘best’ answer, you have to consider your use case scenarios.
  - When faced with such a question, your answer needs to be convincingly argued with internal consistency