Multi-threaded, multi-core embedded systems energy modelling

From instruction set modelling to network & communications modelling

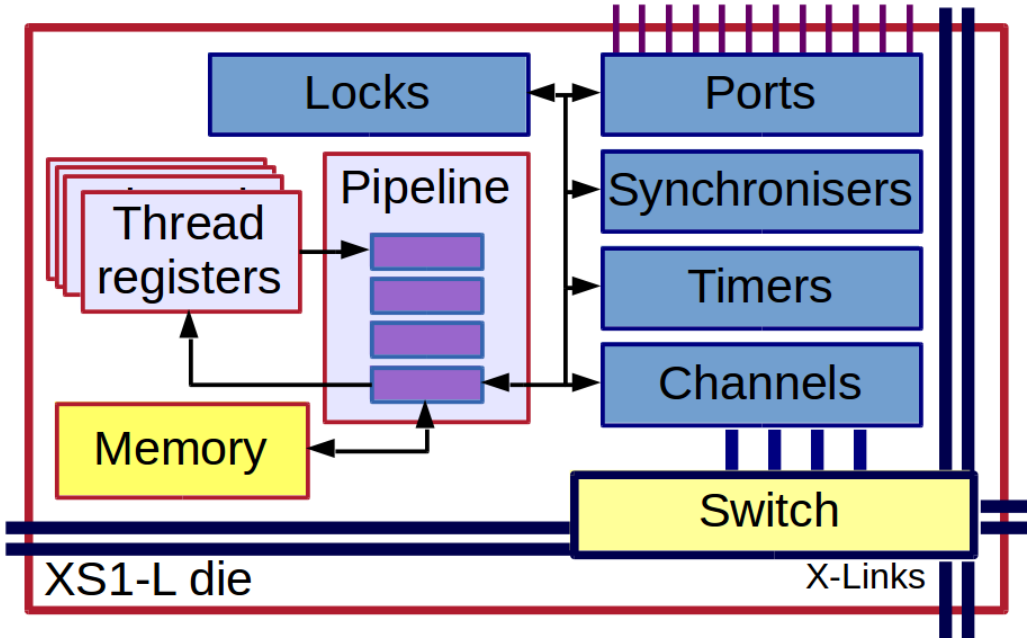

In this talk

- XMOS XS1
  - Multi-threading
  - Modelling
  - Current work
- Multi-core
  - Challenges
  - Approaches
  - Current work

Multi-threaded modelling

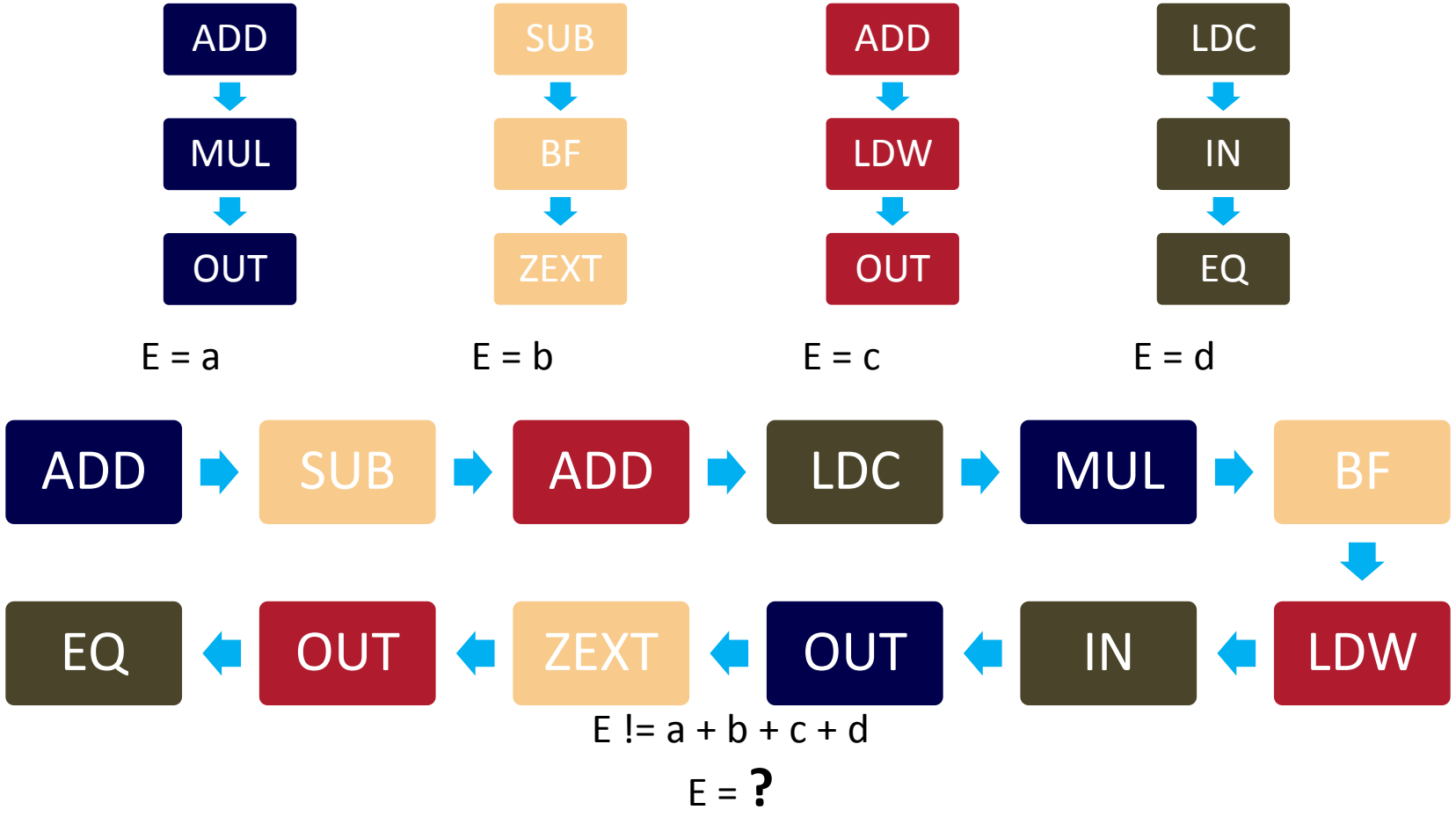
MODELLING THE XS1

XS1 architecture



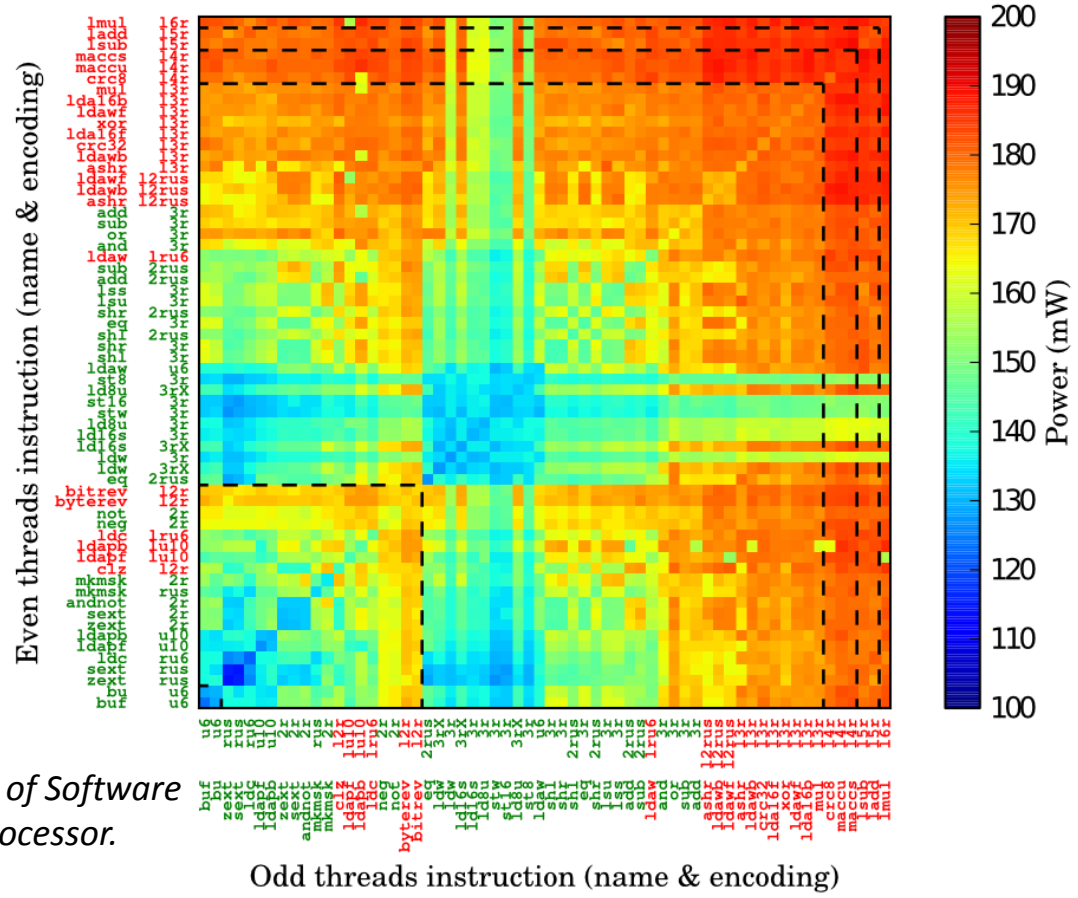
- **Four-stage** pipeline, **round-robin** execution.
  - Time-deterministic.
- **Single-cycle** memory.
  - No caches.
- Dedicated instructions for I/O & **networking**.
  - **Low latency** comms.

Modelling a multi-threaded pipeline



ISA Characterization

- Idle power
  ~95mW
- Cheapest instructions
  ~110mW
- Most expensive
  ~200mW
- Worst case
  ~250mW
- Threading level
  1.5x power increase, 4x performance increase.



Steven Kerrison and Kerstin Eder. Energy Modelling of Software for a Hardware Multi-threaded Embedded Microprocessor. Submitted to ACM TECS, under review, Sept. 2013.

Model

$$P = P_{static} + P_{dynamic} \qquad P_{static} = I_{leak} \times V_{core}$$

$$P_{dynamic} = ((C_{idle} + (C_{instr} \times S_N \times O)) \times V_{core}^2 \times F)$$

Base cost

Instruction & thread cost

Over-head

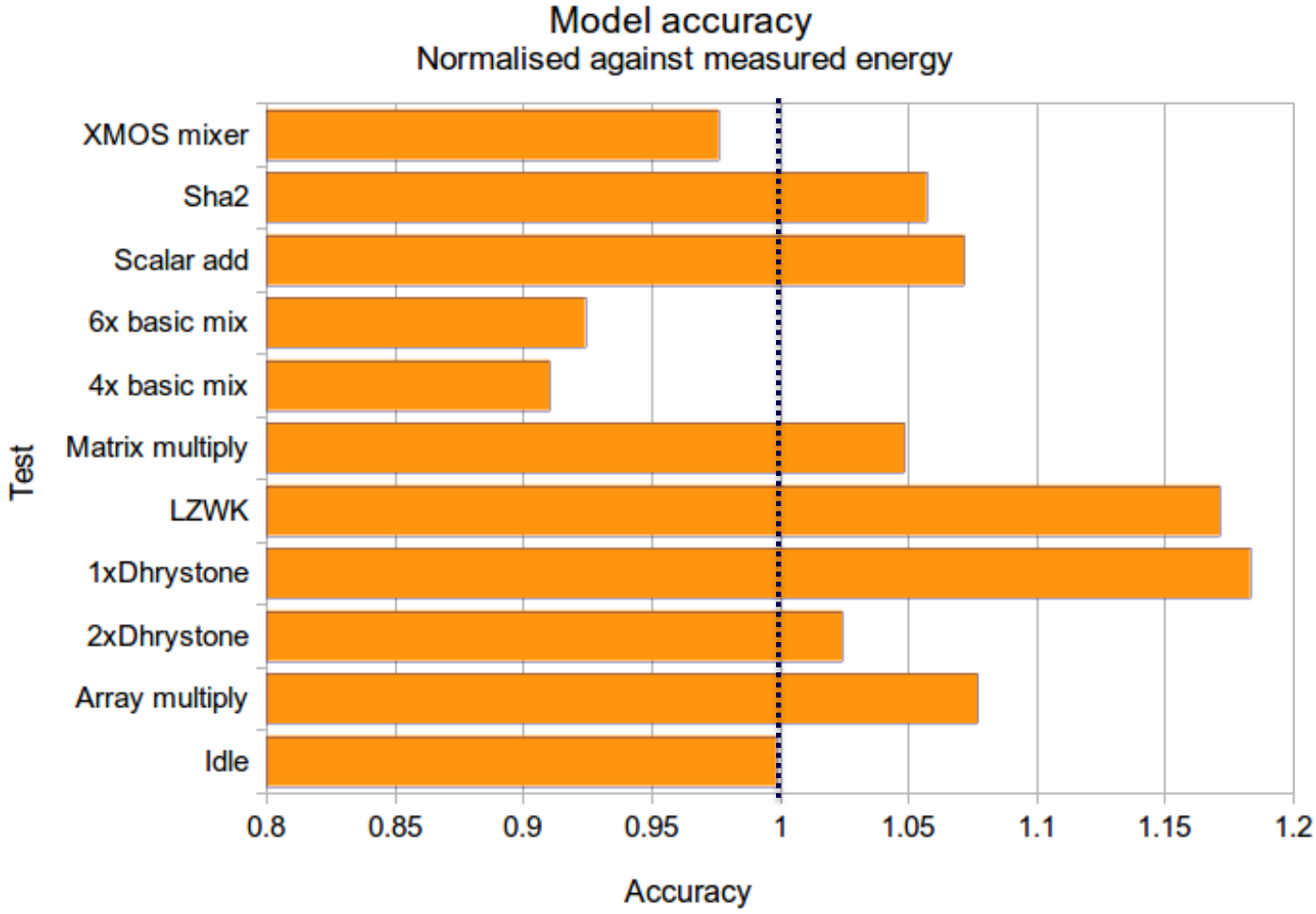
Pipeline fullness

$$N = \min(N_{threads}, 4)$$

$$E = \frac{P \times T}{N}$$

- Express frequency, voltage, thread activity, instruction costs and time consumption.

Preliminary results



What can we do?

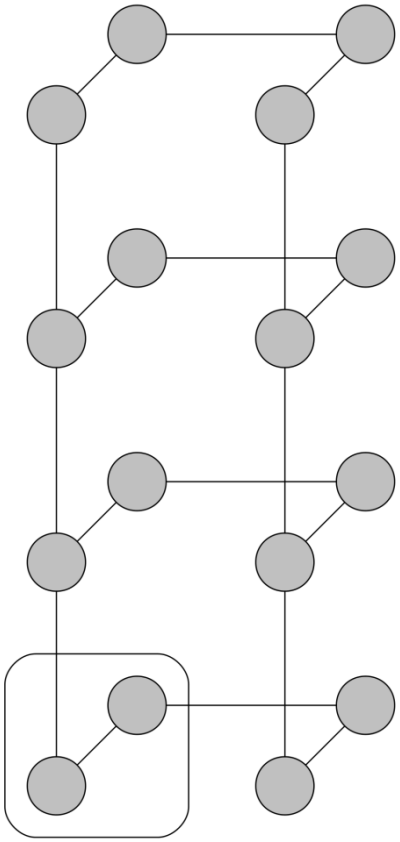
- Modelling of various benchmarks
  - Up to 8 threads (1 core)
- Voltage/frequency parameterised model
- Apply model to different levels:
  - Execution statistics
  - Trace (full or partial)
  - Static analysis
    - ISA
    - LLVM-IR

Modelling and analysis considerations

MULTI-THREADED & MULTI-CORE

Considerations

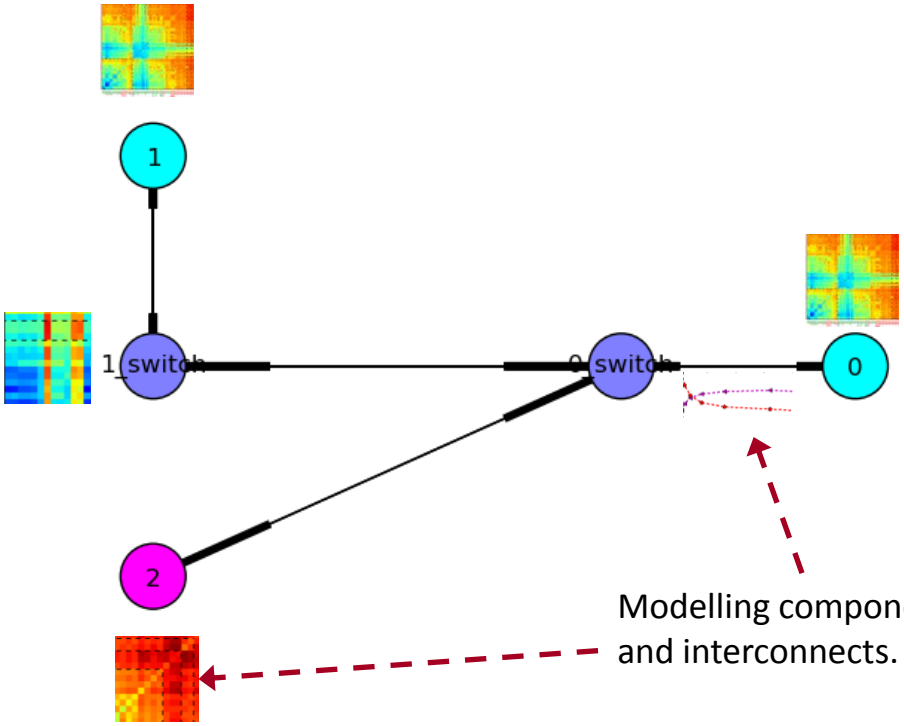
- Core-local multi-threading
  - Very fast (<10 clocks)
  - Lots of bandwidth
- Chip-local multi-core
  - Quite fast (10s of clocks)
  - Multiple lanes
- Multi-chip multi-core
  - Fairly fast (10s-100s clocks)
  - More bandwidth contention



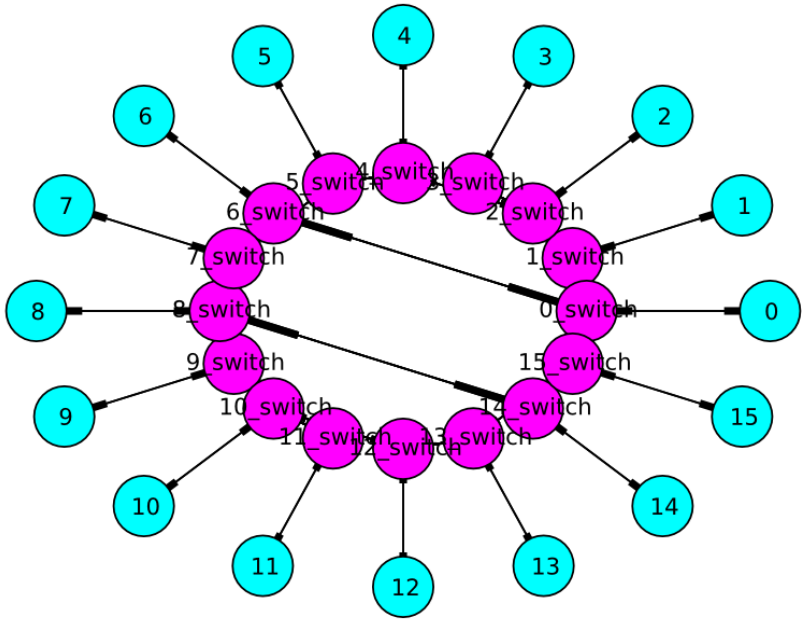
Swallow, 16x XS1 processors per board

Modelling

- 2-tile XMOS network with USB

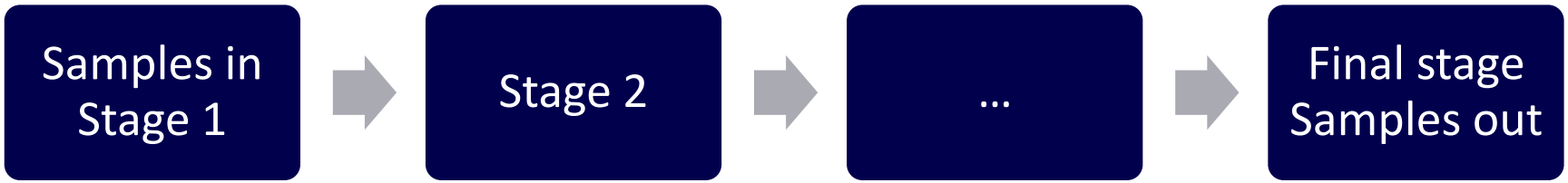


- 16-tile XMOS board (UoB project Swallow)



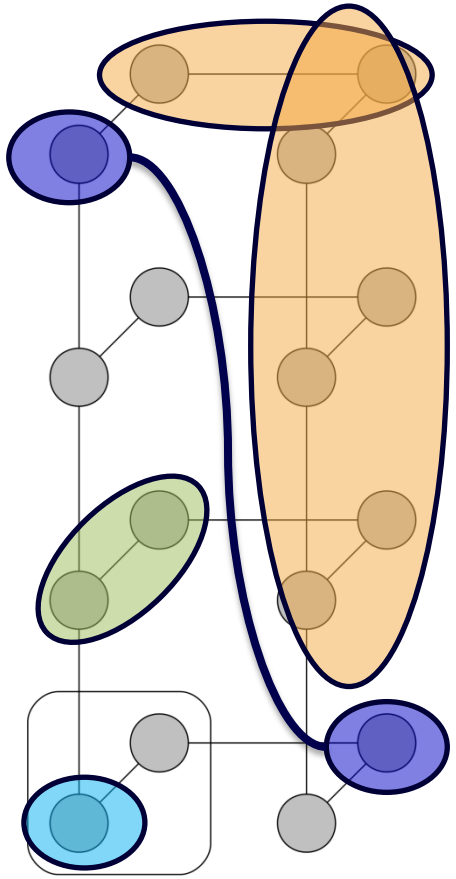
Biquad filter

- N-stage biquad filter
- Each stage is a thread
  - Channel communication
- Pipeline construction



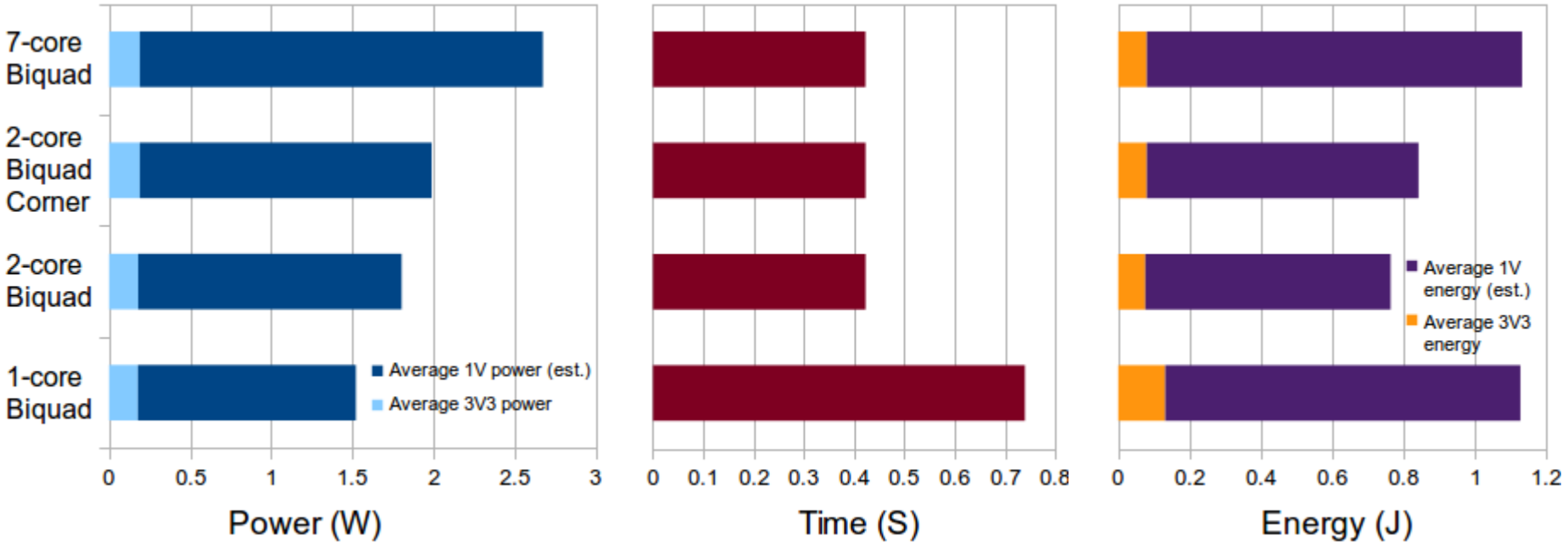
Biquad filter

- Implemented in various configurations on Swallow.
- Each implementation's placement is colour coded.



Comms example: Biquad filter

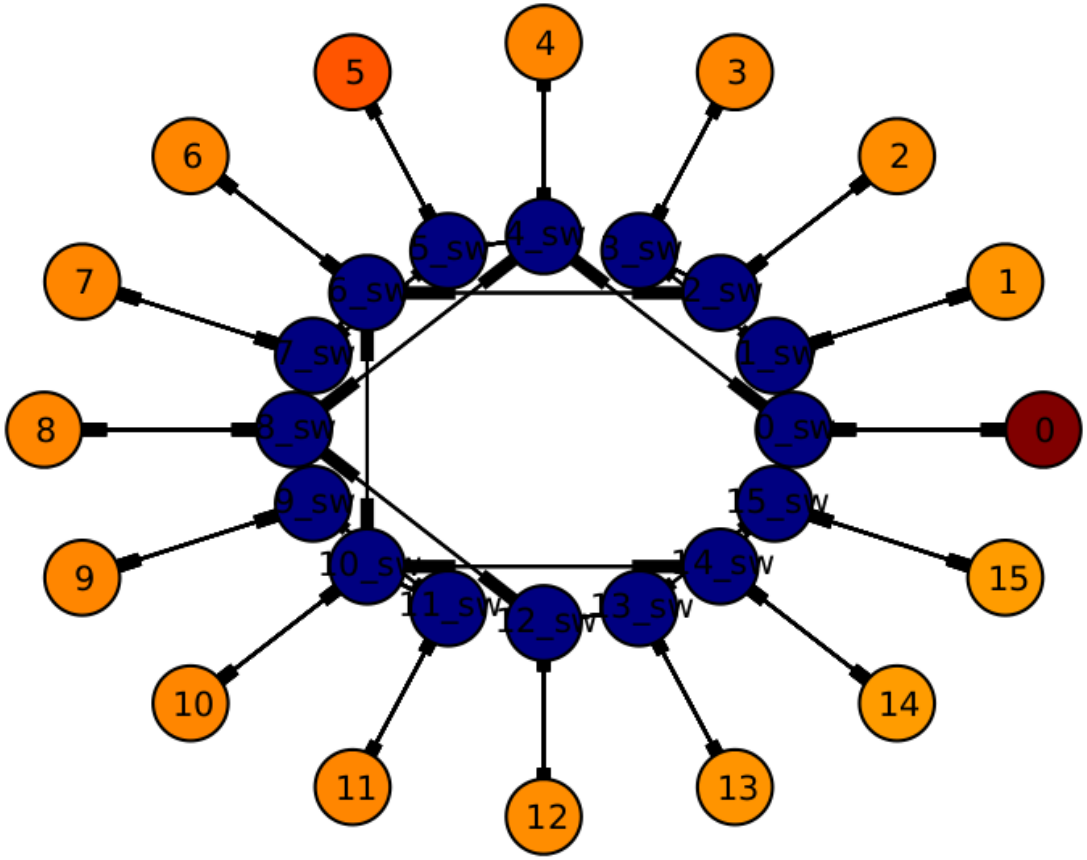
- Active cores, latency, contention and under/over-allocation all affect total energy.
- Power, time & energy a valuable triple.



Biggest impacts

- Task placement in a multi-core system
  - Determines V/F of cores, speed of comms, switching costs.
- Latency between communications
  - Having idle, powered on cores is bad
  - Excessive-synchronisation is bad
- Core execution is time-deterministic only in-between I/O and events!

Visualising MTMC energy consumption



Current work

- Evaluating benchmarks & test cases for a multi-core network model against real hardware.
- Working towards utilising this information in static analysis.