Soar Workshop, June 2013

New Computing Hardware Based On Cognitive Architectures?

Braden Phillips, Michael Liebelt, Brian Ng

Kindly Presented By

John Laird

THE UNIVERSITY *of* ADELAIDE

The Hypothesis

New computing hardware optimised for cognitive architectures will

- better exploit the characteristics of future microelectronics
- accelerate the development of cognitive architectures
- accelerate the development of advanced general agents

- underpin a new generation of computing based on artificial general intelligence

THE UNIVERSITY of ADELAIDE

Do We Need A New Kind Of Computing Machine?

- The von-Neumann architecture has been an incredible success
- 5 decades of performance increases enabled by:
    - increasing number of transistors per IC
        - was doubling every 2 years
        - now doubling every 3 years
        - over next 15 years, expected to double every 3.8 years
    - increasing clock frequency
        - was doubling every 18 months
        - expected to less than double over next 15 years
    - architectural innovations
        - e.g multi-level caches, on-chip networking, multi-core

(figures based on 2011 ITRS roadmap)

Do We Need A New Kind Of Computing Machine?

- Number of transistors per IC still growing
  - 10 billion transistors per IC today
  - expect 100 billion in 2026
- Not seeing a proportional increase in processor performance
  - diminishing returns from ‘more of the same’ (more cores, more cache)
  - power limited to around 150 W per IC
- *How do we achieve higher functionality per transistor per Watt?*

We need a need approach.

- Inspired by the human brain
  - power efficiently achieves complex behaviour
  - using of the order of 100 billion switching devices

THE UNIVERSITY of ADELAIDE

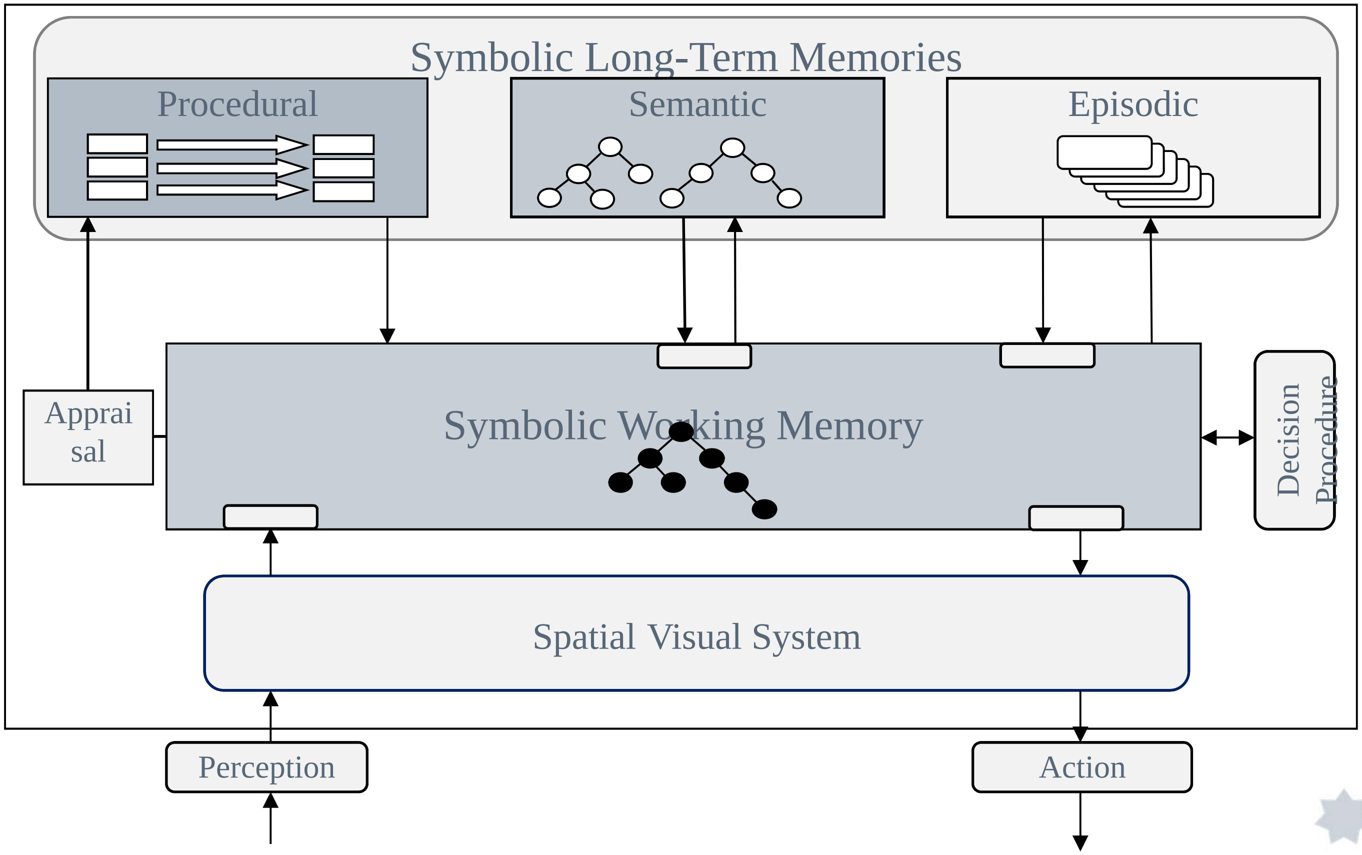
Future Directions In Microelectronics

- 100 billion switching elements
  - = fine grained parallelism
  - = distributed/interleaved storage and computation
  - – power limitation ⇒ cannot all be switching all the time
- Clock frequencies in the 10s GHz
  - = faster than switching in the brain
  - – much less interconnected than the brain
  - – suggests may need to packet switch data (on-chip networks) or have longer wires at lower switching frequency
- Less reliable circuit operation
  - – increase in transient and permanent errors expected
  - – may need fault-tolerant architectures

Cognitive Architectures As The Basis Of New Computing Machines?

- Cognitive architectures are at the right level of abstraction i.e. the same as computer architecture
- Fixed processing blocks map directly to hardware units
- Naturally present opportunities for fine-grained parallelism and for distributed memory and computation (within blocks) e.g.
  - production rule matching
  - spreading activation
  - memory searches
- Also parallelism at a coarser scale (between blocks) e.g.
  - learning
  - episodic memory

Soar 9 Structure

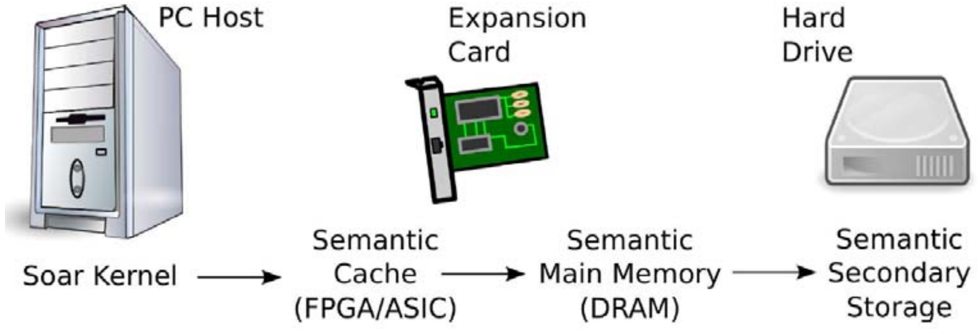


A Note On Hardware Development

- Incremental development is possible
  - Can realise hardware blocks as separate modules e.g. a plug-in episodic memory box
  - Some blocks in Soar have changed very slowly over time e.g. the Rete
- Hardware can be flexible
  - Modern hardware development cycle is not unlike software
  - Designed in a hardware description language (e.g. Verilog)
  - Automatic synthesis to reprogrammable logic (e.g. FPGA) or to application specific IC.

A Hardware Semantic Memory For Soar

- Aim: a plug-in semantic memory system
  - scalable to very large knowledge bases
  - low latency (e.g. < 5 µs queue-based retrieval[2])
  - base-level and spreading activation
- Approach:
  - semantic memory hierarchy
  - virtual memory


PC Host
Expansion Card
Hard Drive
Soar Kernel
Semantic Cache (FPGA/ASIC)
Semantic Main Memory (DRAM)
Semantic Secondary Storage

[2] the main limitation is PCIe bus latency.

A Hardware Semantic Memory For Soar

- We are beginning with a level 1 semantic cache block.
- Why start here?
  - fine-grained parallelism for search and spreading activation
  - lots of transistors with limited switching activity
  - a stepping-stone to episodic memory

Some Items To Ponder

- What aspects of Soar would most benefit from hardware acceleration?
  - We have started with semantic memory…

    …and we plan to move on to episodic memory.
  - We have also started looking at the Rete. Looks promising:
    - Lots of fine-grained parallelism.
    - Not much communication between processing nodes.
  - What else?
- How has the evolution of Soar been influenced by considerations of the underlying processor architecture?
  - What might be possible if we changed the assumptions?