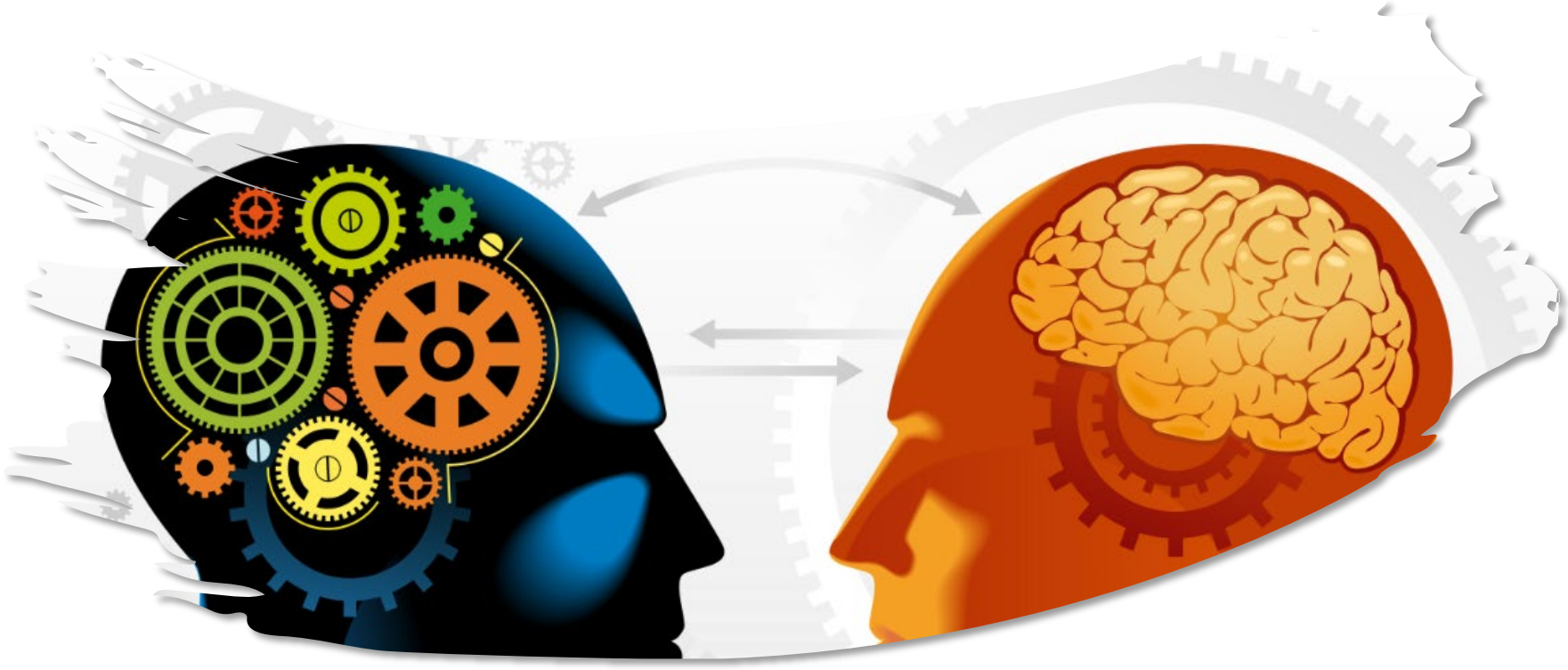


Building the Infinite Brain

COMP 690 (193)

Raghavendra Pradyumna Pothukuchi

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

raghav@cs.unc.edu

Takeaways

What is the challenge with (multicore/distributed) system design and programming?

Implementing and reasoning *correctly* about it

What are two specifications that help the above?

Cache coherence: about writes to the same location across caches
Memory consistency: about ordering of memory accesses

What is cache coherence?

Each write is eventually visible and writes to the same location are serialized

What is memory consistency?

Specification of what values a read can return and when

What are the principles at play in defining and implementing coherence and consistency?

Abstraction, Efficiency, “Yummy”, Concurrency, Approximate



For Today

- Quick review
- **Parallel architectures**



Defining a Parallel Architecture

> "A Survey of Parallel Computer Architectures", Ralph Duncan (CDC) 1990

Low-level parallelism

Pipelining, superscalar, CPU+I/O separation

Parallel architecture

Multiple processors executing concurrently to solve a problem, with an explicit high-level framework

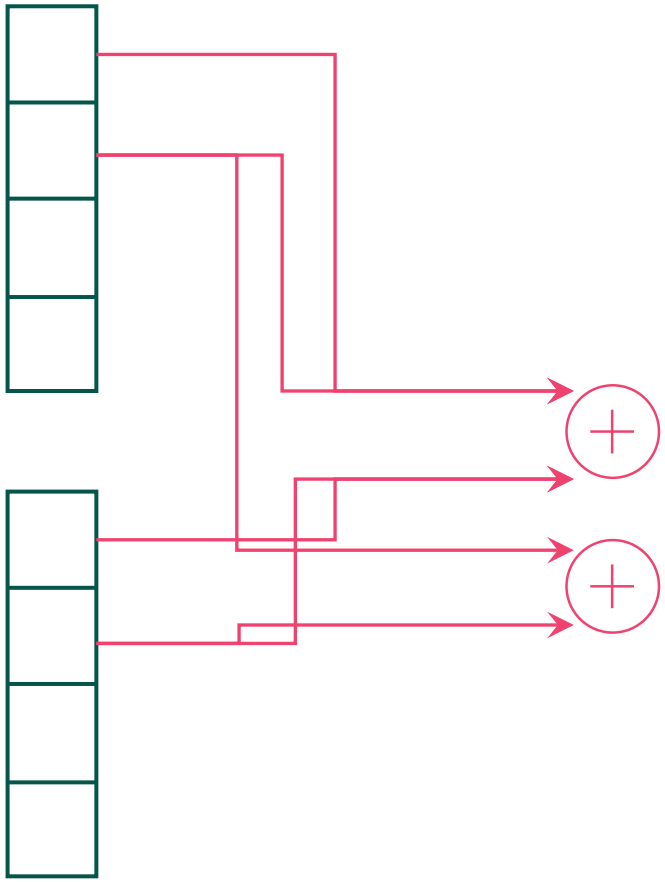
Various classifications and taxonomies

Flynn: Single/Multiple Instruction and Single/Multiple Data

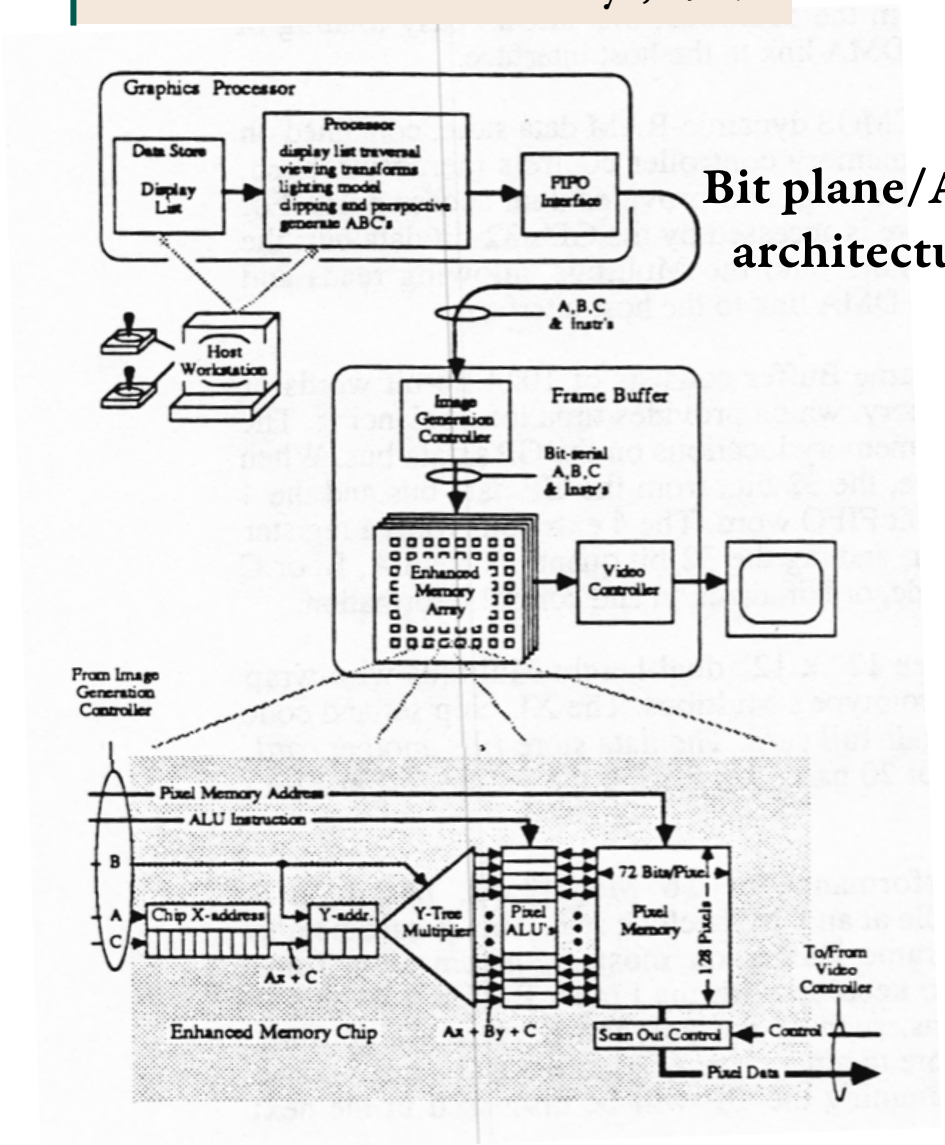


Synchronous Architectures

Vector processors



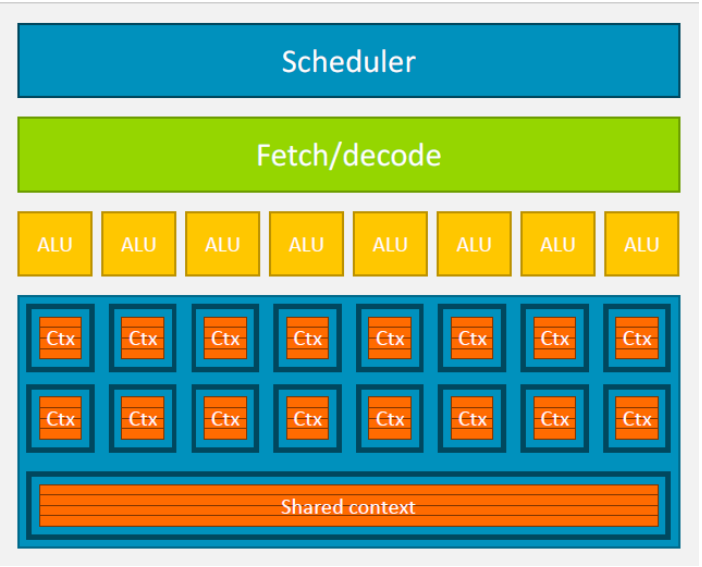
"Pixel-Planes 4: A summary", 1987



Bit plane/Array architectures

Single Instruction Multiple Data

GPUs (Multiple Threads)

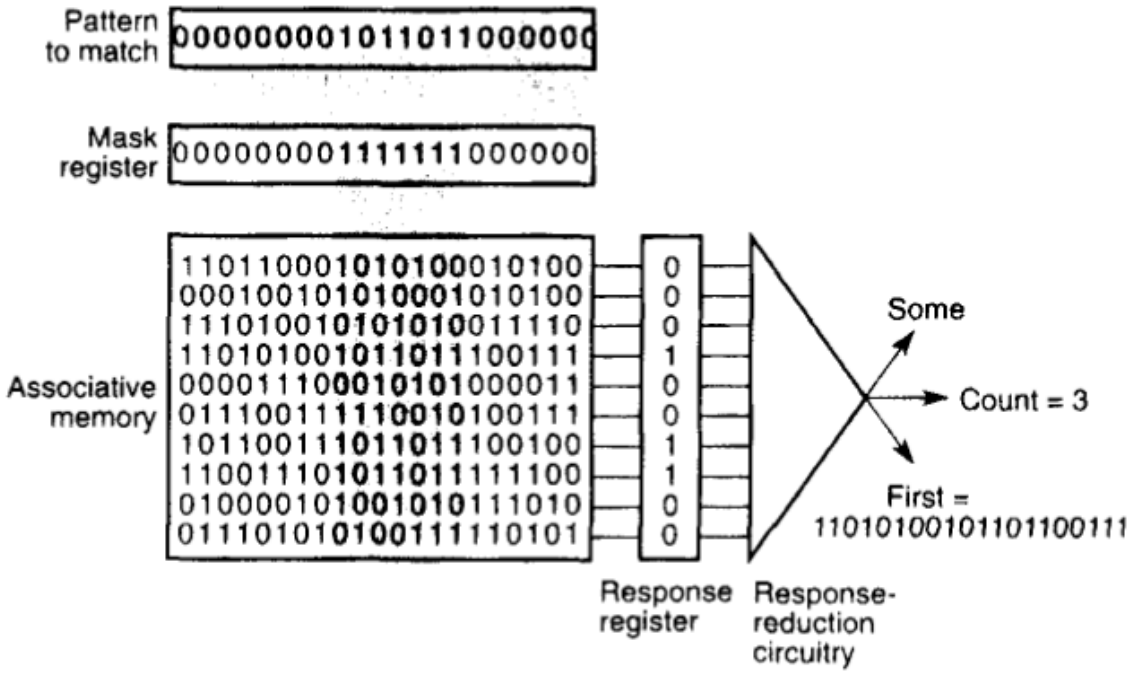


Synchronous Architectures

Associative processors

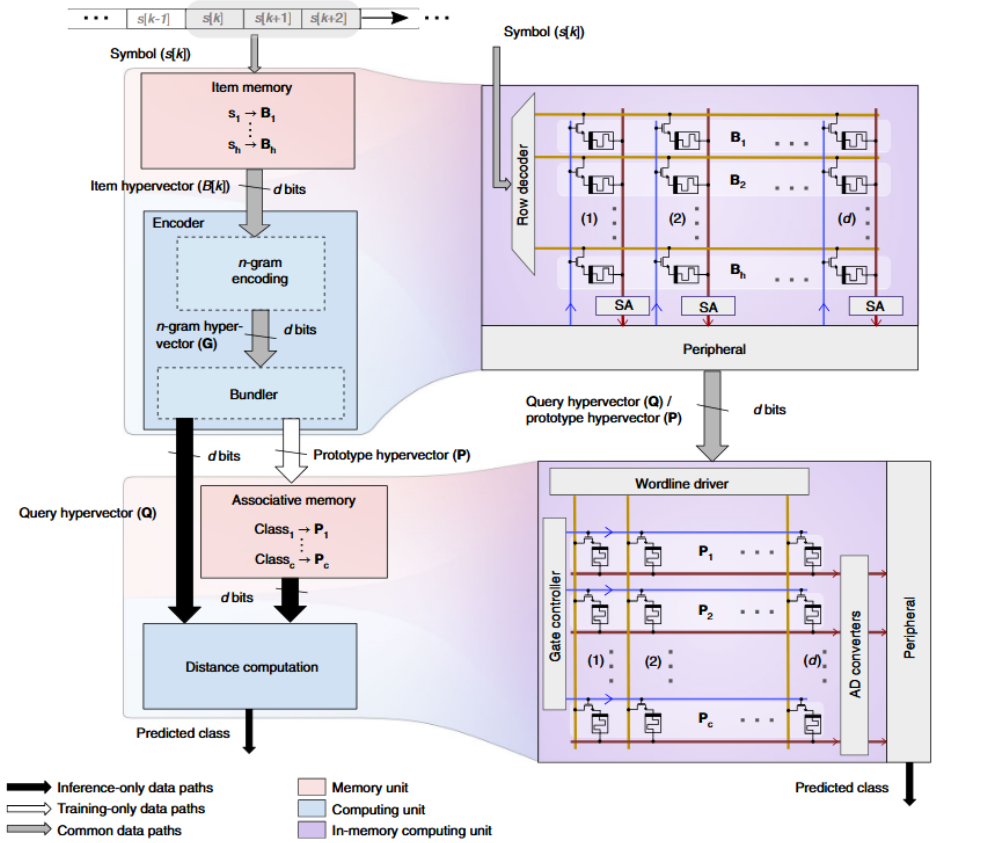
Single Instruction Multiple Data

"Associative Processing and Processors", 1994



Memex!

"In-memory hyperdimensional computing", 2020

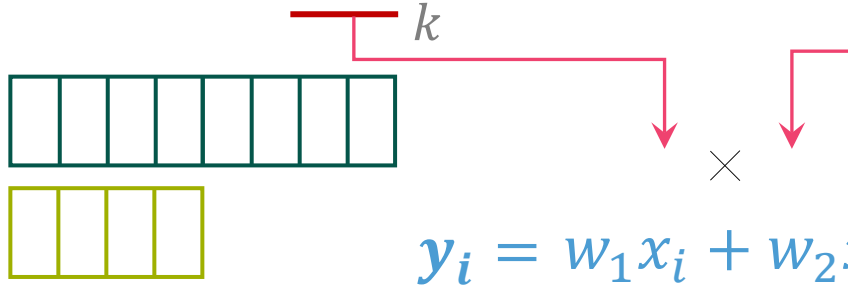


Synchronous Architectures

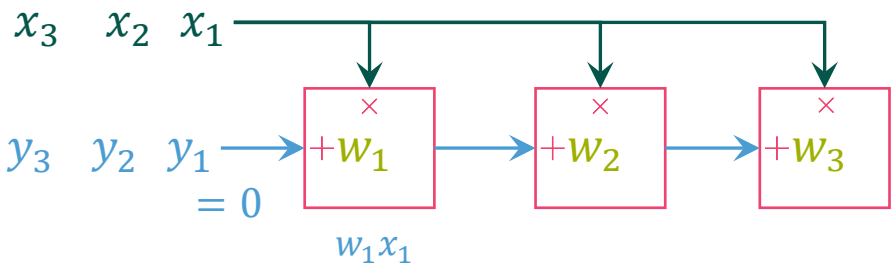
Systolic arrays

"Systolic Arrays for (VLSI)",
H. T. Kung and C. E. Leiserson, 1978

$$\boldsymbol{x} = \{x_1, x_2, \ldots x_i \ldots x_n\} \qquad \boldsymbol{w} = \{w_1, w_2, \ldots w_k\}$$



$$y_i = w_1 x_i + w_2 x_{i+1} + \cdots w_k x_{i+k-1}$$



w_1x_1

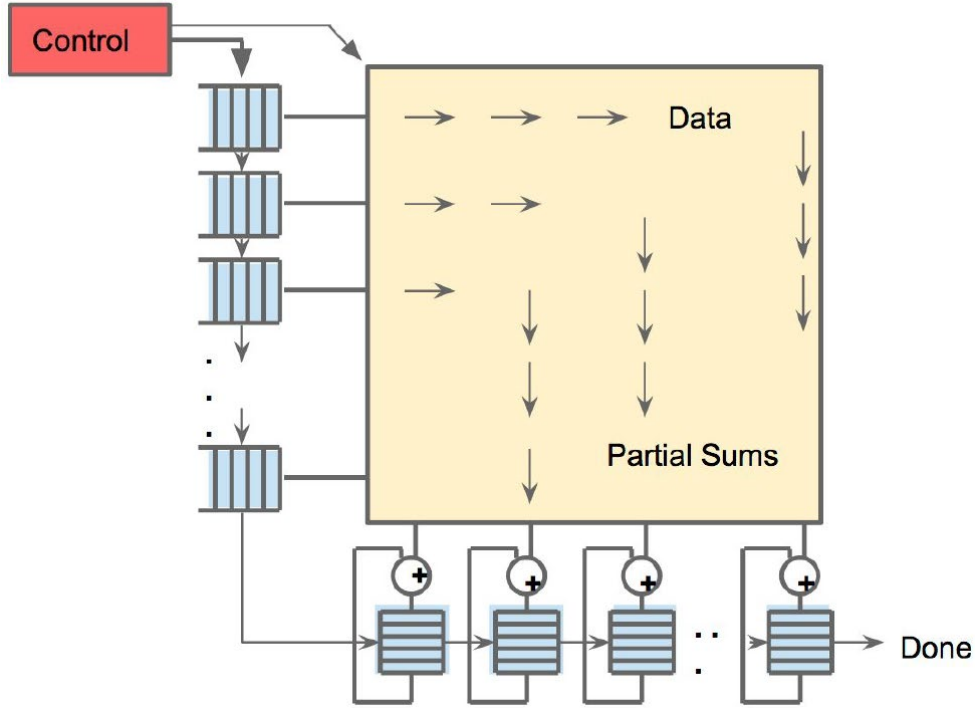
w_1x_2 $\quad w_1x_1 + w_2x_2$

w_1x_3 $\quad w_1x_2 + w_2x_3$ $\quad w_1x_1 + w_2x_2 + w_3x_3$

Algorithmic pipelining

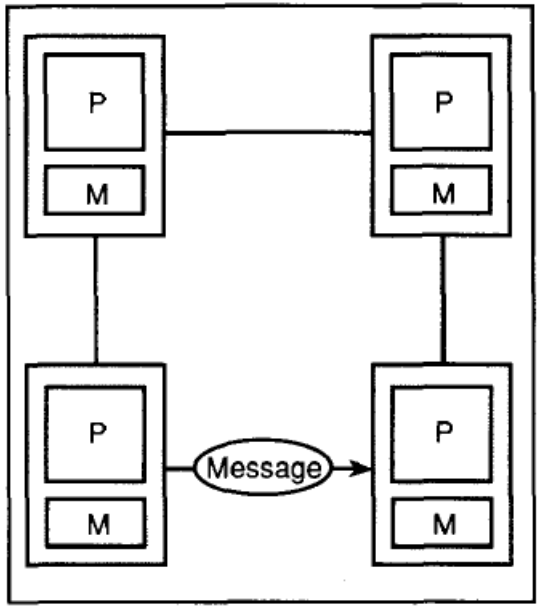
TPUs

"In-Datacenter Performance Analysis of a Tensor Processing Unit",
Google, ISCA 2017

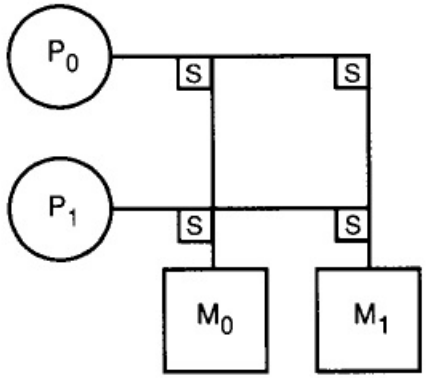


Multiple Input Multiple Data Architectures

Distributed memory



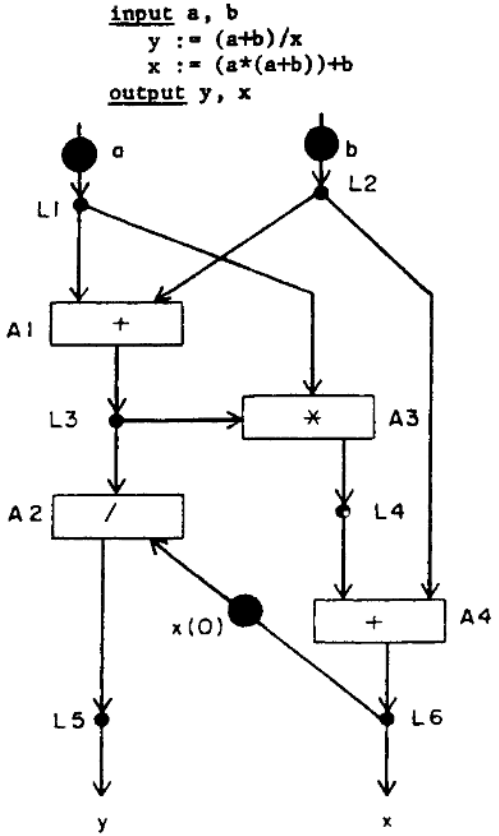
Shared memory

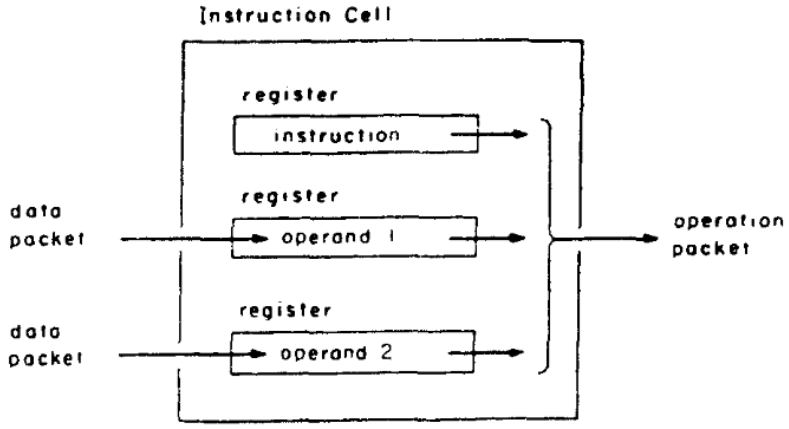


Dataflow Architectures

> "Properties of a model for parallel computations: determinacy, termination, queueing", R. M. Karp and R. E. Miller, 1966

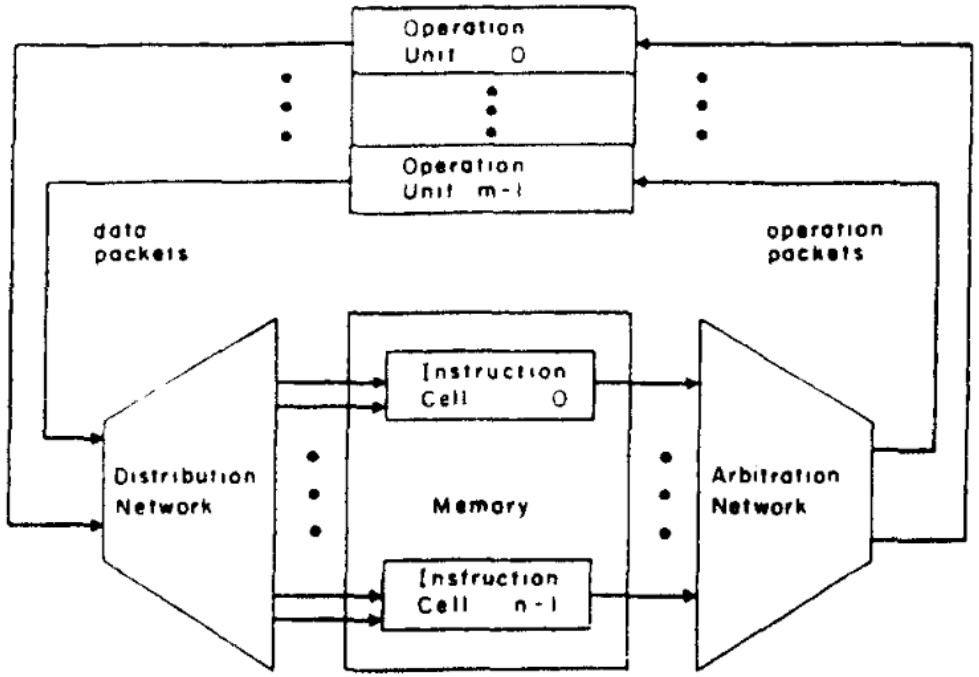
> "A Preliminary Architecture for a Basic Dataflow Processor", J. B. Dennis and D.P. Misunas, 1974





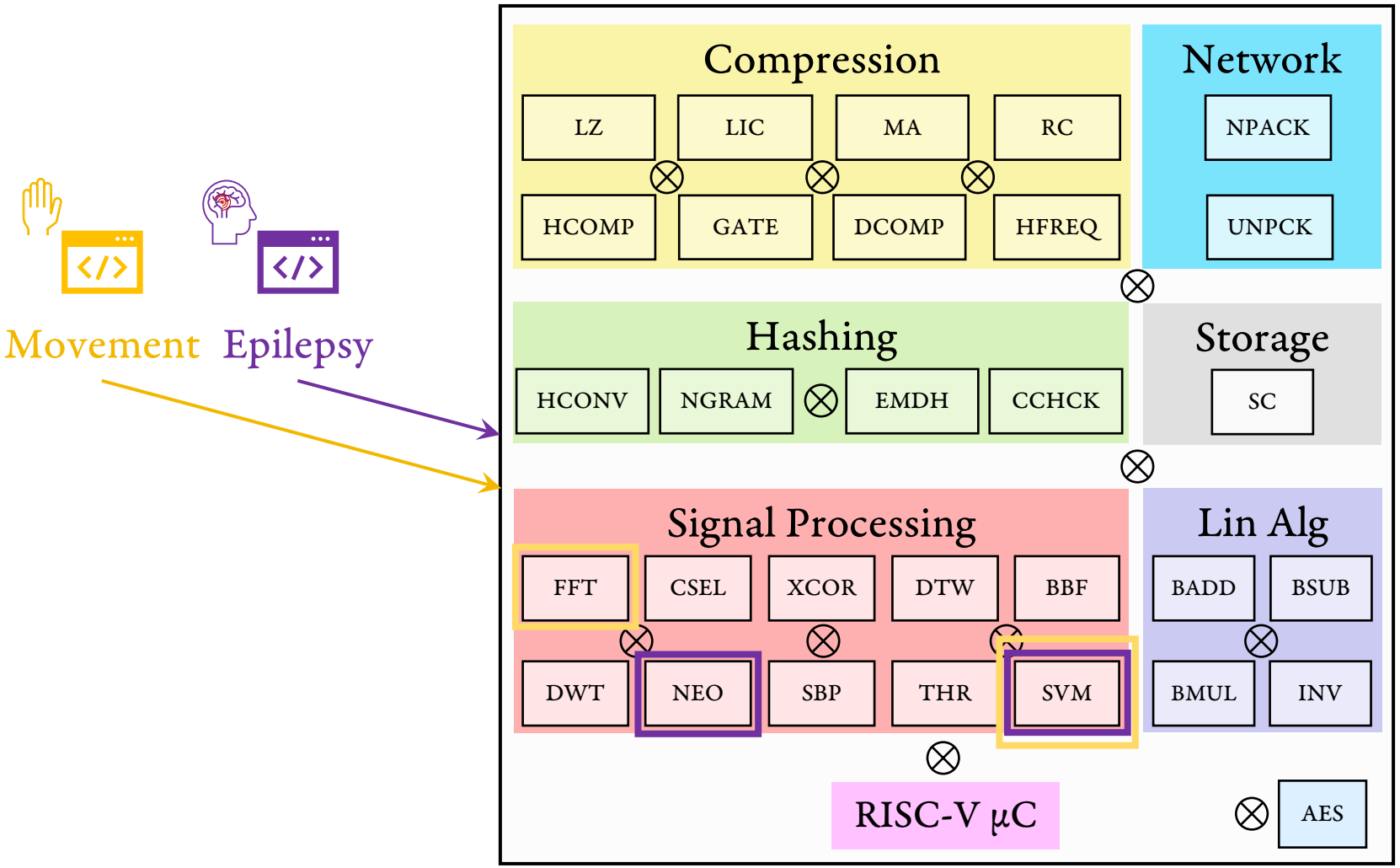
Out-of-order execution is dataflow

Programming?



Example from a BCI Processor (SCALO)

Streaming neural data processing through dataflow



Neural data is already processed in dataflow pipelines

Granularity of processing is bigger than arithmetic units

Supports Multiple Instruction Single Data processing

Exposes accelerators to the software

Image Credits (Educational, Fair Use)

- Title image: VLADGRIN, https://www.istockphoto.com/vector/human_-machine-gm147409511-16840728 (Educational fair use)
- Infinite brain: Science wonder stories, May 1930, Illustrator: Frank R Paul, Editor: Hugo Gernsback
- Brain color, ICs, cloud server, black rat: No attribution required (Hiclipart)
- Hand with spoon: public domain freepng
- Signals: https://www.nature.com/articles/nrn3724
- Thought clouds: F. Willett et al./*Nature* 2021/Erika Woodrum, https://med.stanford.edu/neurosurgery/news/2022/bci-award. https://www.the-scientist.com/news-opinion/brain-computer-interface-user-types-90-characters-per-minute-with-mind-68762
- Picture of scientists: https://www.cs.auckland.ac.nz/~brian/rutherford8.html (original: Pierre de Latil), Bush (Carnegie Science), Others (Wikipedia, National Academies, IEEE, and university profile images)
- Flowchart: Pause08 – flaticon.com; Digital brain: Smashicons – flaticon.com; Quantum processor icons created by Paul J. - Flaticon
- Server rack: upklyak – freepik.com
- Arm, Lotus: Adobe stock
- Quantum processor: Rigetti computing
- Images of implanted users: Top: Case Western Reserve University (https://thedaily.case.edu/man-quadriplegia-employs-injury-bridging-technologies-move-just-thinking/), Bottom: Jan Scheuermann (University of Pittsburgh/UPMC; https://www.upmc.com/media/news/bci-press-release-chocolate)
- Images of wearable BCIs: Cognixion, NextMind
- Types of BCIs: "Brain–computer interfaces for communication and rehabilitation,
- Illustrative BCI: Neuralink
- Electrodes: "Electrochemical and electrophysiological considerations for clinical high channel count neural interfaces", Vatsyayan et al.
- Form factors: Neuropace, Medtronic, Bloomberg, "Fully Implanted Brain–Computer Interface in a Locked-In Patient with ALS" by Vansteensel et al., Blackrock Neurotech
- Jose Delgado's video: Online, various sources (CNN, Youtube)
- Video of Kennedy and Ramsey: Online, various sources (Youtube, Neural signals)
- Code snippet inspiration: ECE 252 slides at Duke (Dan Sorin et al.)
- Apple processor pipeline: https://dougallj.github.io/applecpu/firestorm.html

Logos, trademarks are all properties of respective owners
Not to be shared outside the course