Computer Architecture for the Next Millenium

November 1, 1999

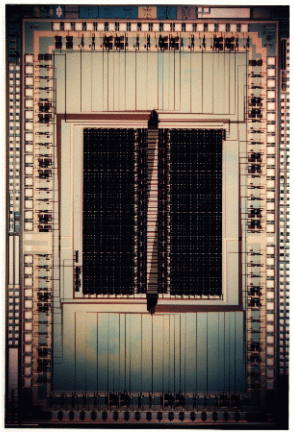
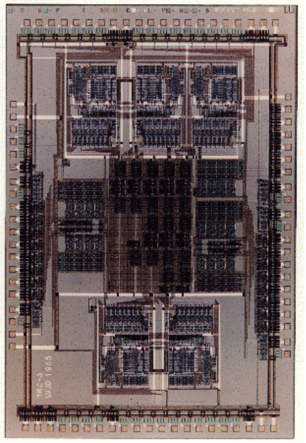
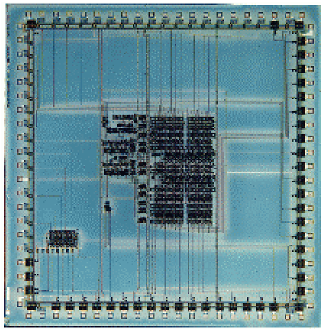
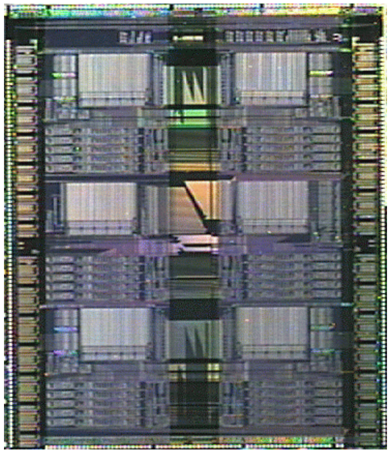
William J. Dally
Computer Systems Laboratory
Stanford University
billd@csl.stanford.edu

Outline

- The Stanford Concurrent VLSI Architecture Group
- Forces acting on computer architecture
  - applications (media)
  - technology (wire-limited)
  - techniques (explicit parallelism)
- Example: register organization
  - distributed register files
- *Imagine* a stream processor
  - 20GFLOPS on a 0.5cm^2 chip
- Tremendous opportunities and challenges for computer architecture in the next millenium
  - its not a *mature* field yet

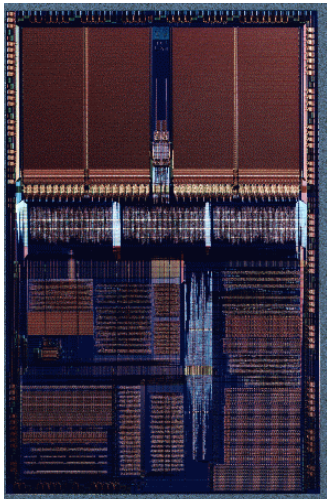
The Concurrent VLSI Architecture Group

- Architecture and design technology for VLSI
- Routing chips
  - Torus Routing Chip, Network Design Frame, Reliable Router
  - Basis for Intel, Cray/SGI, Mercury, Avici network chips

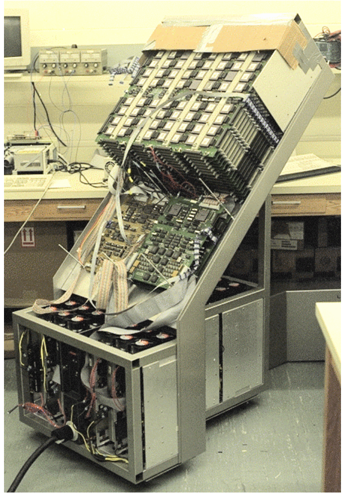





Parallel computer systems

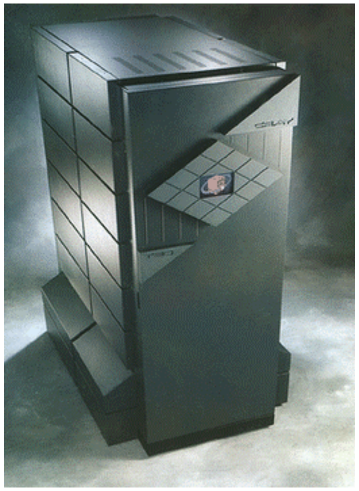
- J-Machine (MDP) led to Cray T3D/T3E
- M-Machine (MAP)
  - Fast messaging, scalable processing nodes, scalable memory architecture



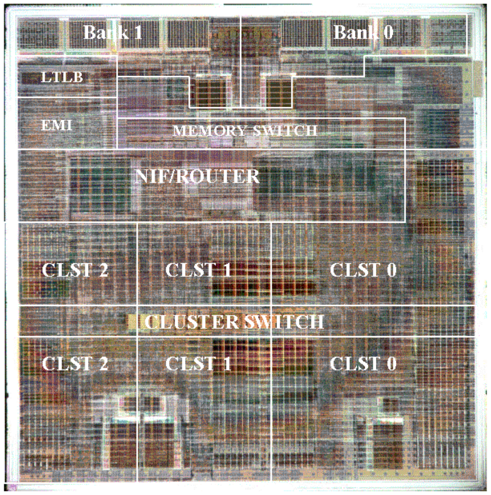
MDP Chip



J-Machine



Cray T3D

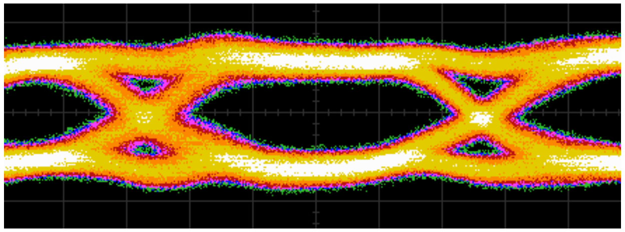


MAP Chip

Design technology

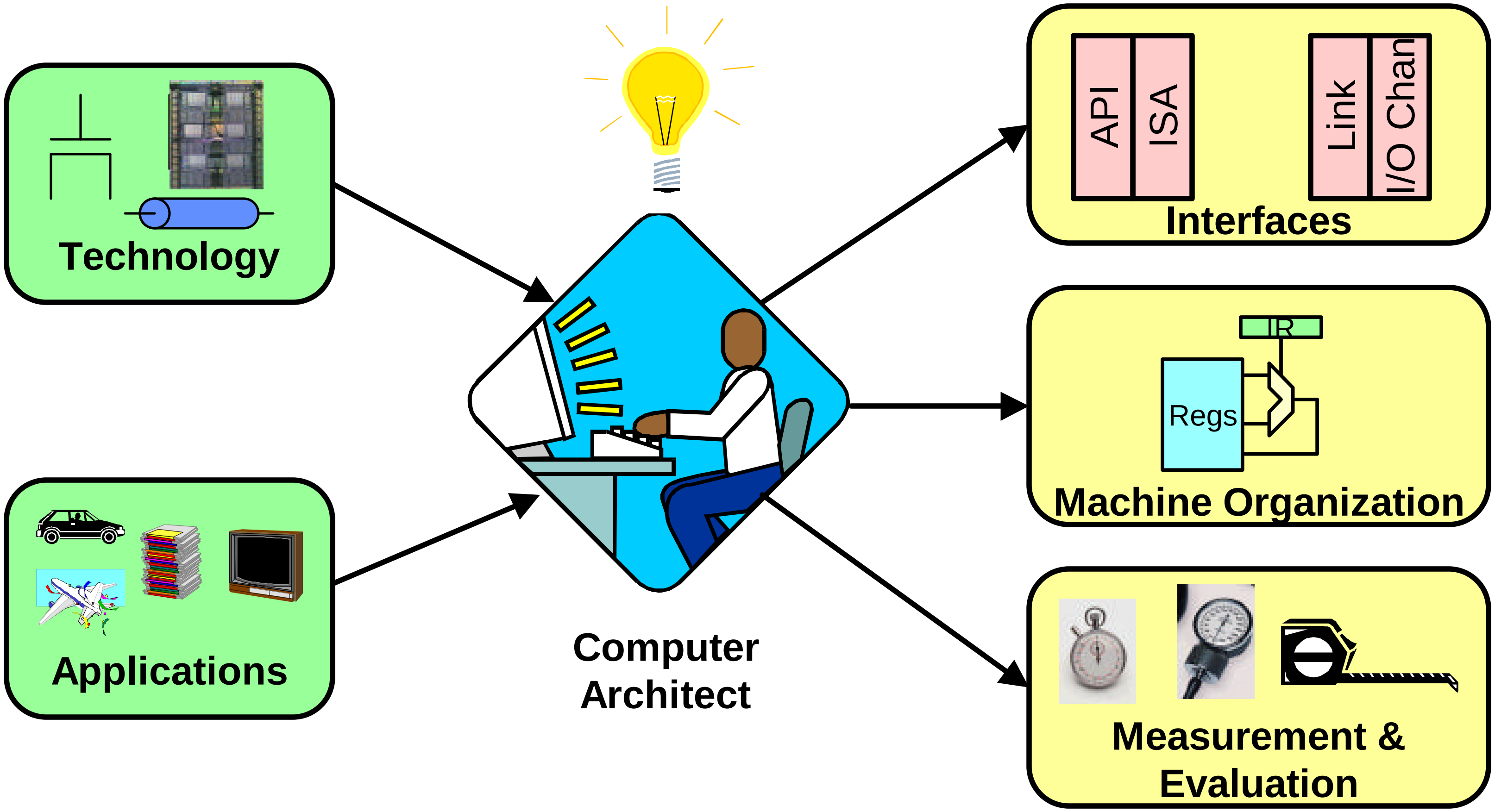
- Off-chip I/O
  - Simultaneous bidirectional signaling, 1989
    - now used by Intel and Hitachi
  - High-speed signalling
    - 4Gb/s in 0.6μm CMOS, Equalization, 1995
- On-Chip Signalling
  - Low-voltage on-chip signalling
  - Low-skew clock distribution
- Synchronization
  - Mesochronous, Plesiochronous
  - Self-Timed Design

250ps/division



4Gb/s CMOS I/O

What is Computer Architecture?

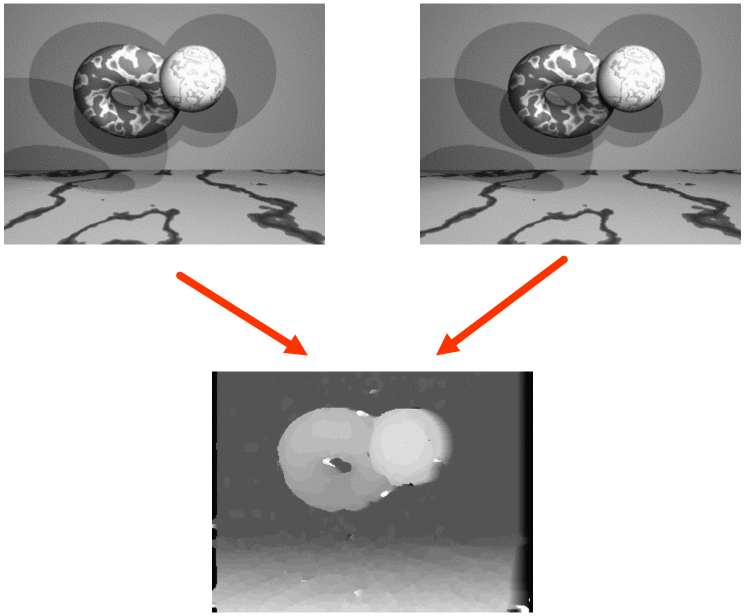


Forces Acting on Architecture

- Applications - shifting towards *media* applications dealing with *streams* of low-precision samples
  - video, graphics, audio, DSL modems, cellular base stations
- Technology - becoming *wire-limited*
  - power and delay dominated by communication, not arithmetic
  - global structures: register files and instruction issue don t scale
- Technique - Micro-architecture - ILP has been *mined out*
  - to the point of diminishing returns on squeezing performance from sequential code
  - explicit parallelism (data parallelism and thread-level parallelism) required to continue scaling performance

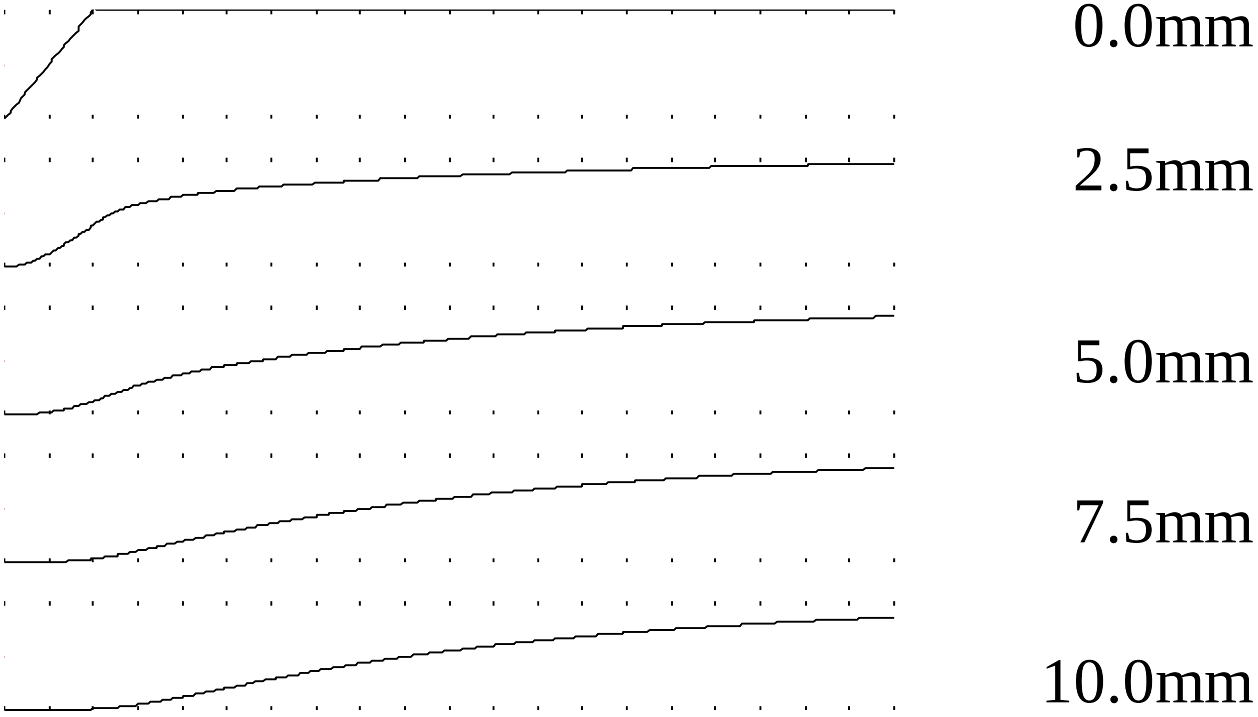
Applications

- Little locality of reference
  - read each pixel once
  - often non-unit stride
  - but there is producer-consumer locality
- Very high arithmetic intensity
  - 100s of arithmetic operations per memory reference
- Dominated by low-precision (16-bit) integer operations

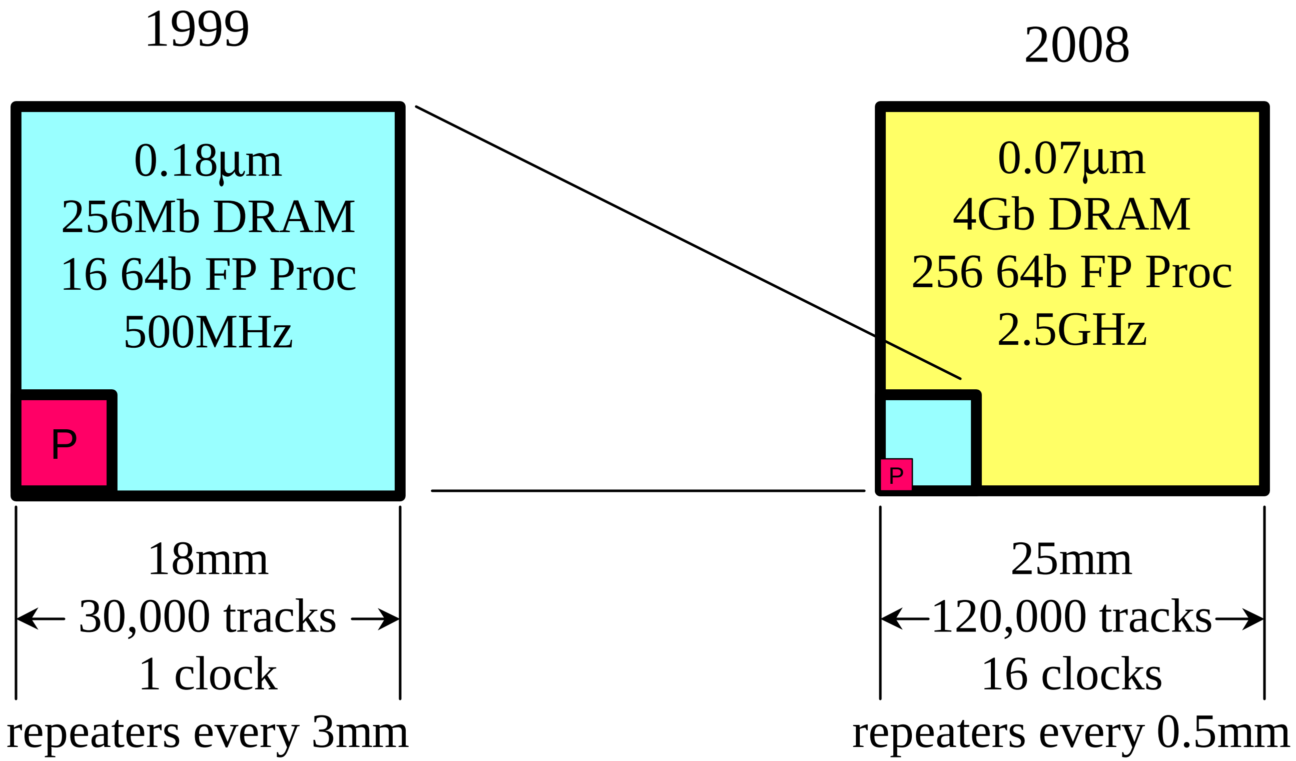


Wires Are Becoming Like Wet Noodles

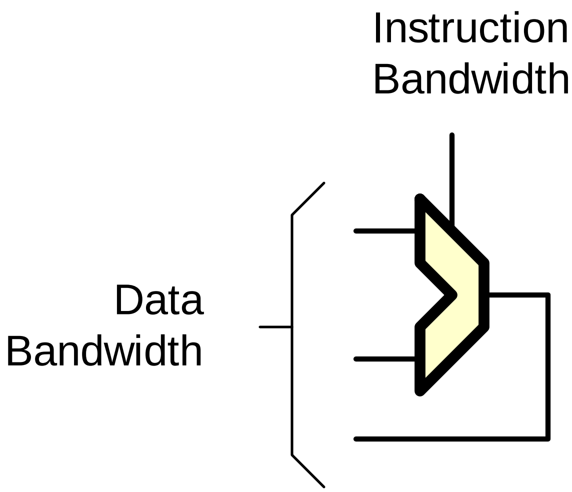
Minimum width wire in an 0.35µm process

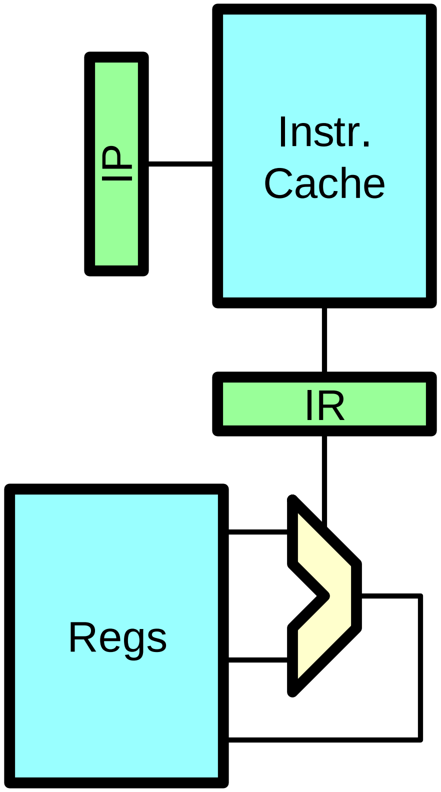


Technology scaling makes communication *the* scarce resource



Care and Feeding of ALUs





‘Feeding’ Structure Dwarfs ALU

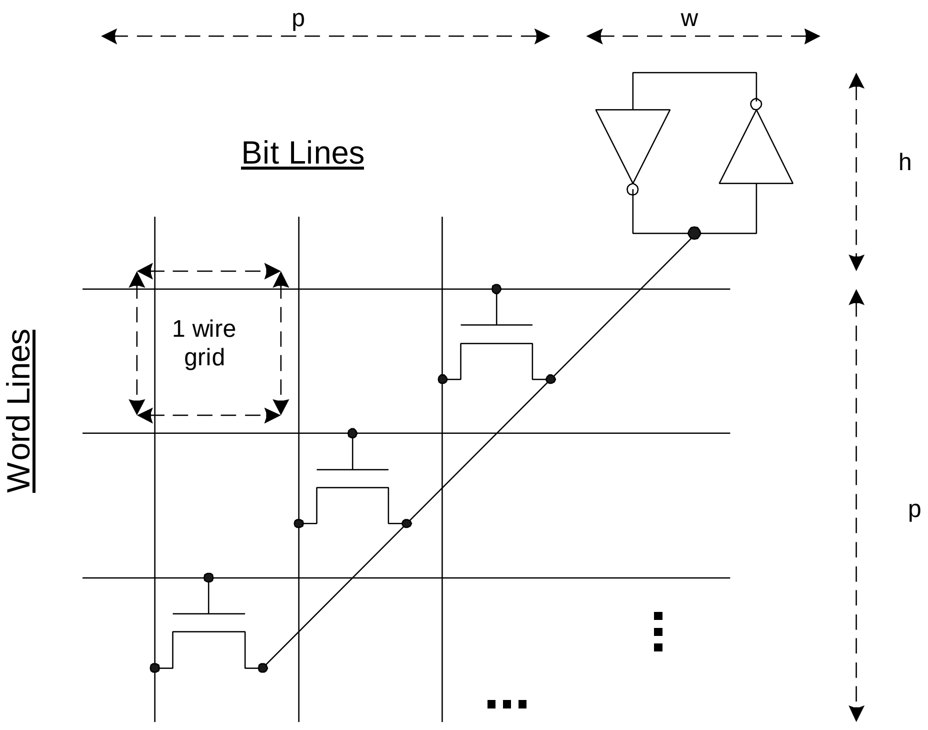
What Does This Say About Architecture?

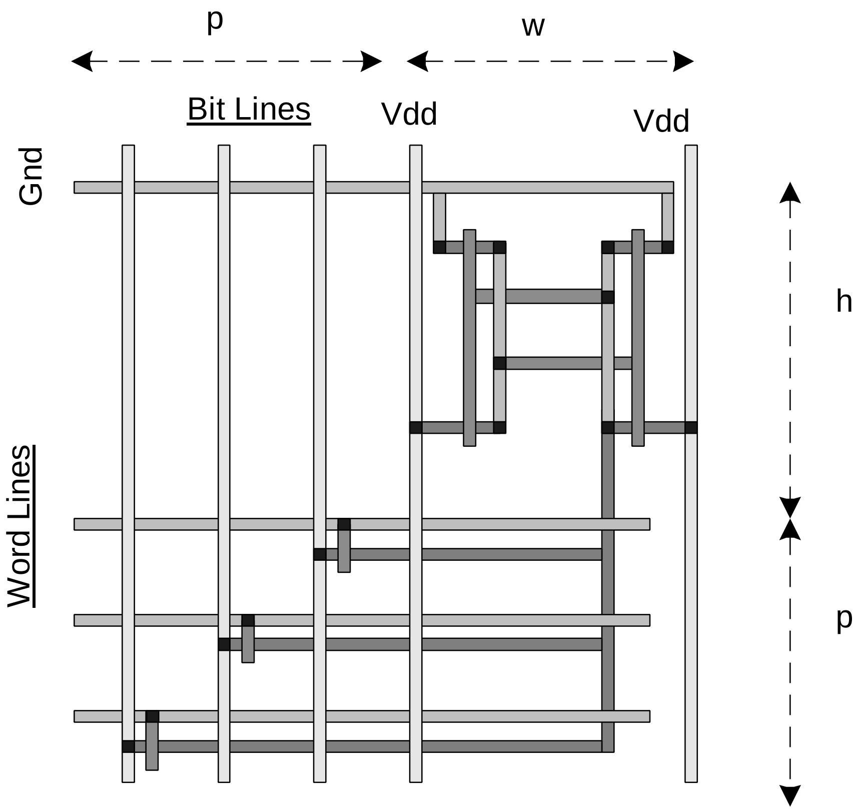
- Tremendous opportunities
  - Media problems have lots of parallelism and locality
  - VLSI technology enables 100s of ALUs per chip (1000s soon)
    - (in 0.18um 0.1mm^2 per integer adder, 0.5mm^2 per FP adder)
- Challenging problems
  - Locality - global structures won t work
  - Explicit parallelism - ILP won t keep 100 ALUs busy
  - Memory - streaming applications don t cache well
- Its time to try some new approaches

Example
Register File Organization

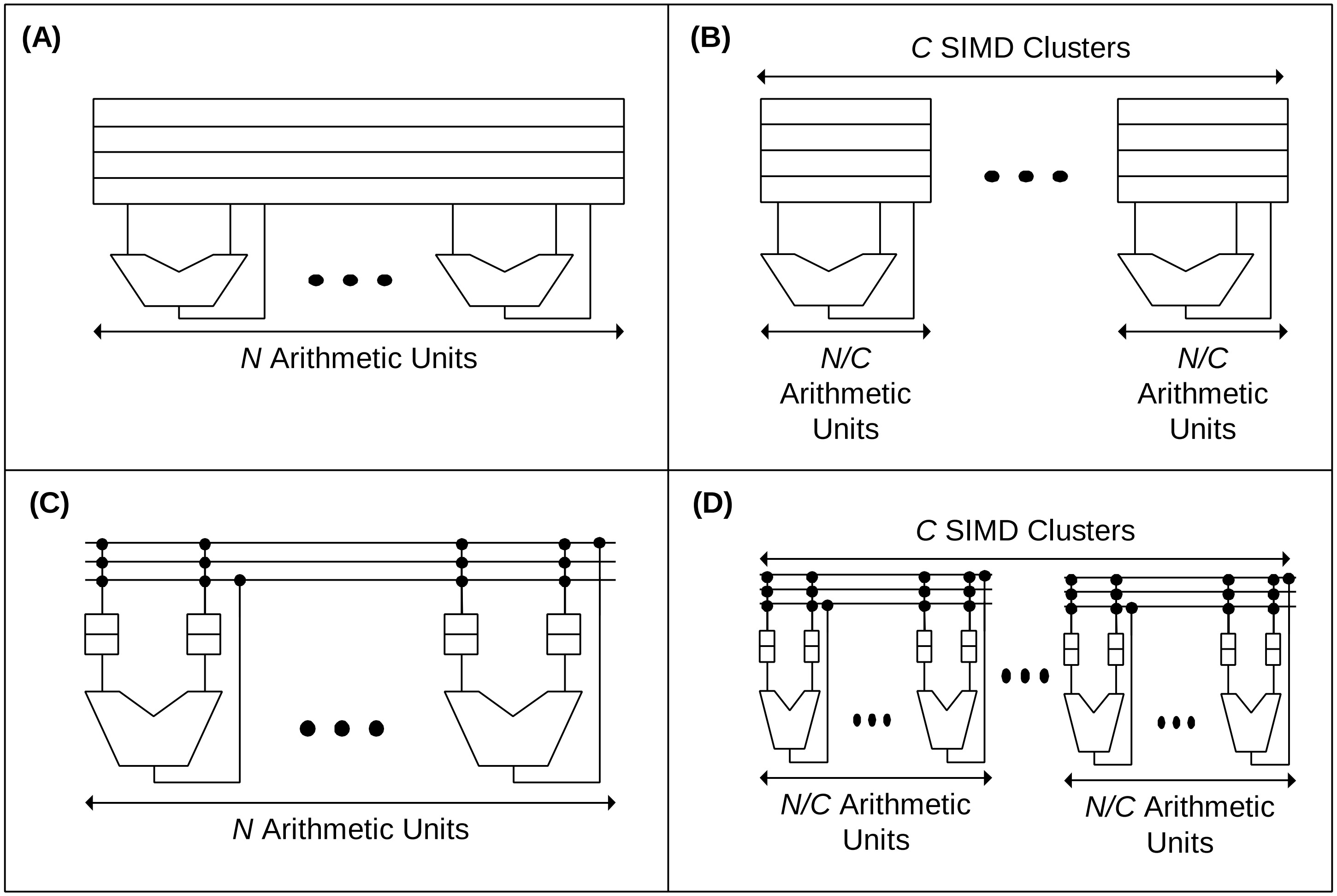
- Register files serve two functions:
  - Short term storage for intermediate results
  - Communication between multiple function units
- Global register files don t scale well as N, number of ALUs increases
  - Need more registers to hold more results (grows with N)
  - Need more ports to connect all of the units (grows with N^2)

Register Cells are Mostly Switch

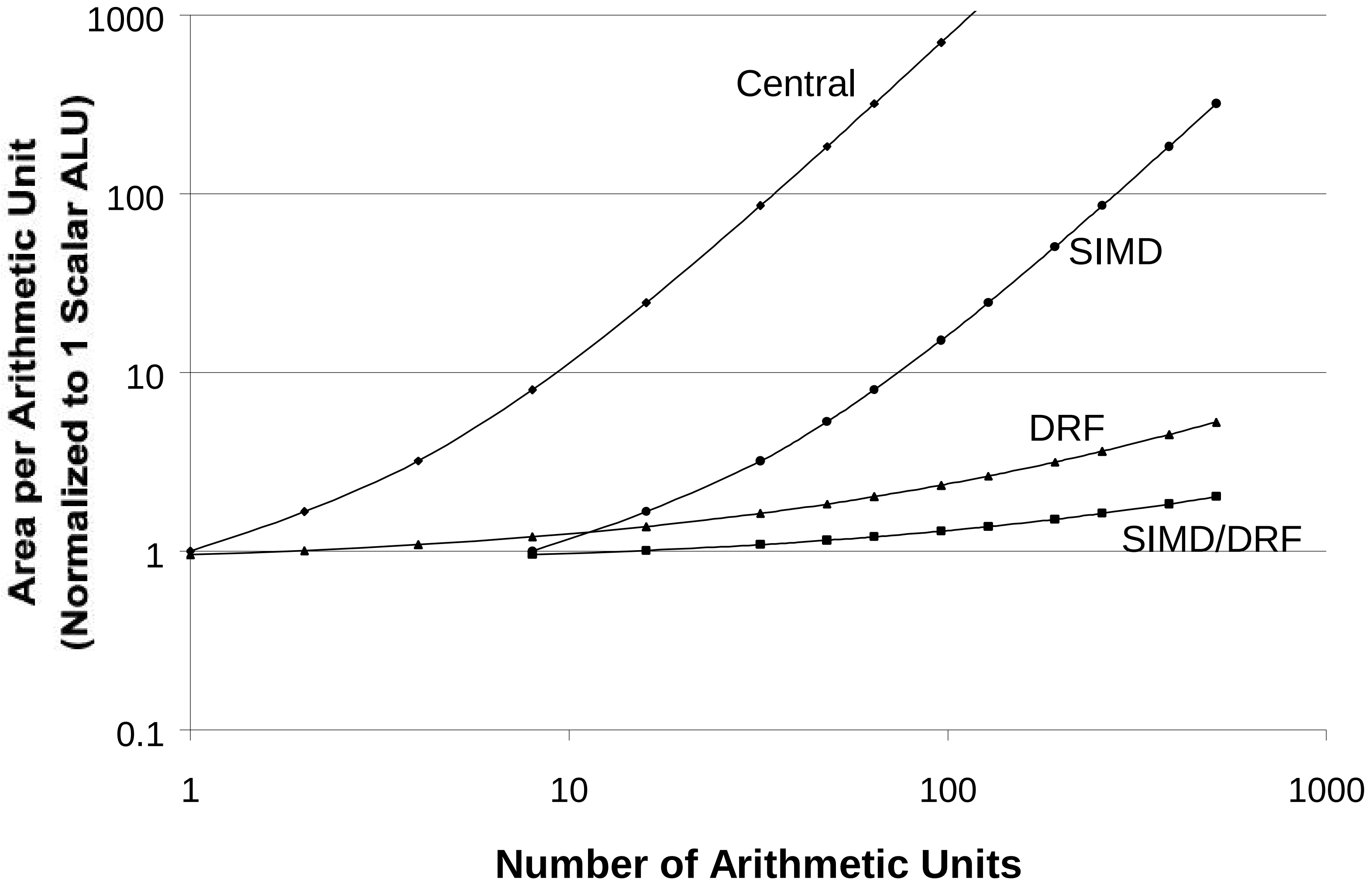




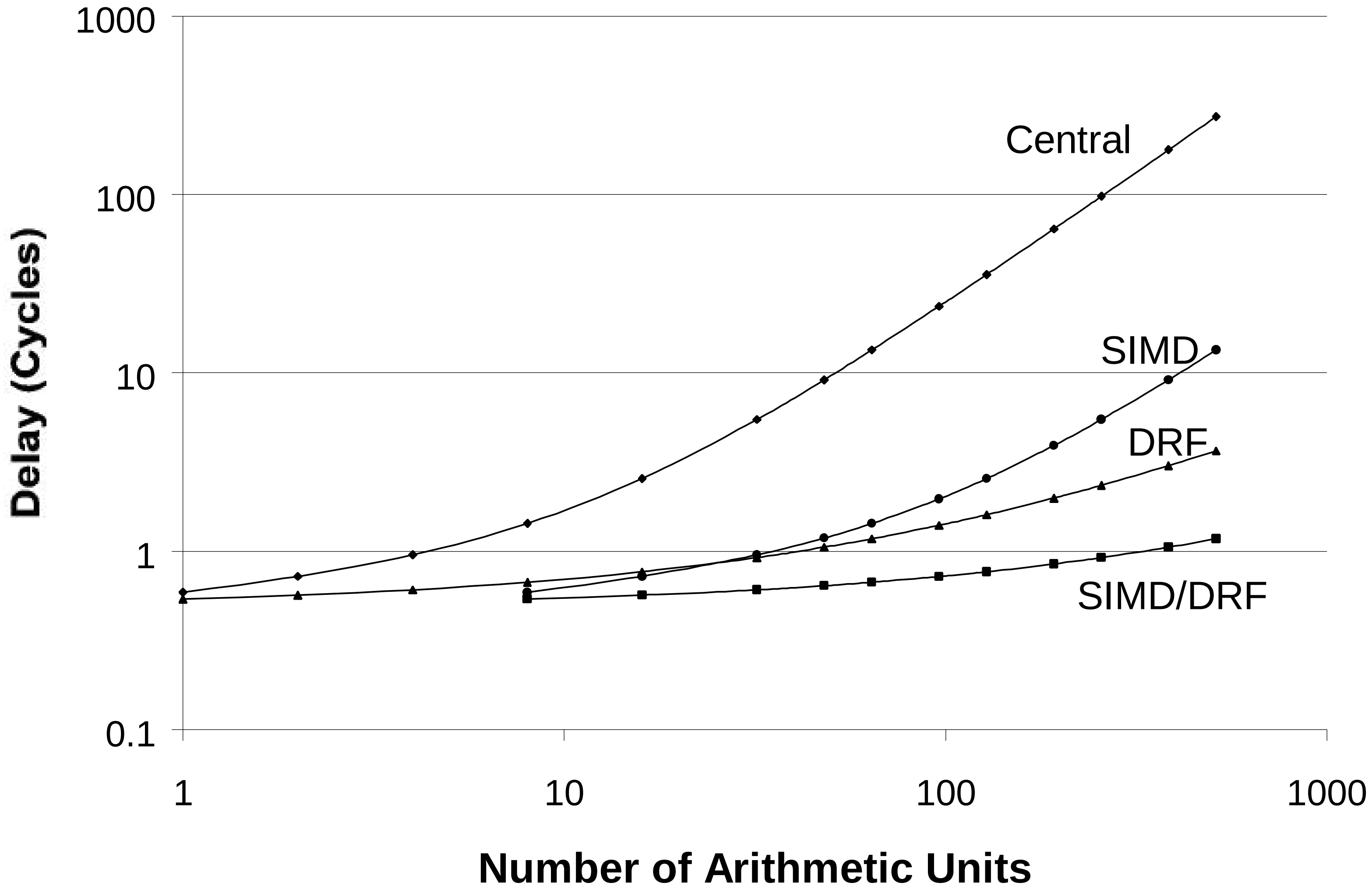
Register Architecture for ‘wide ’Processors



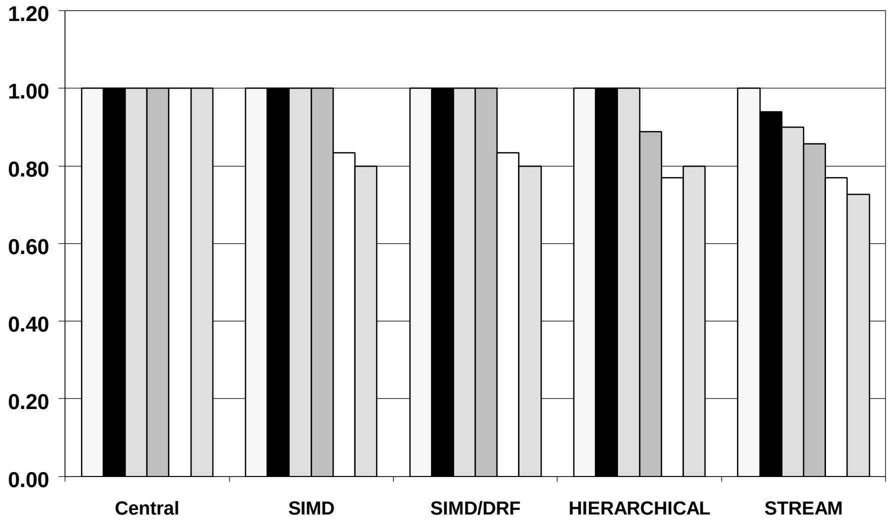
Area of Register Organizations



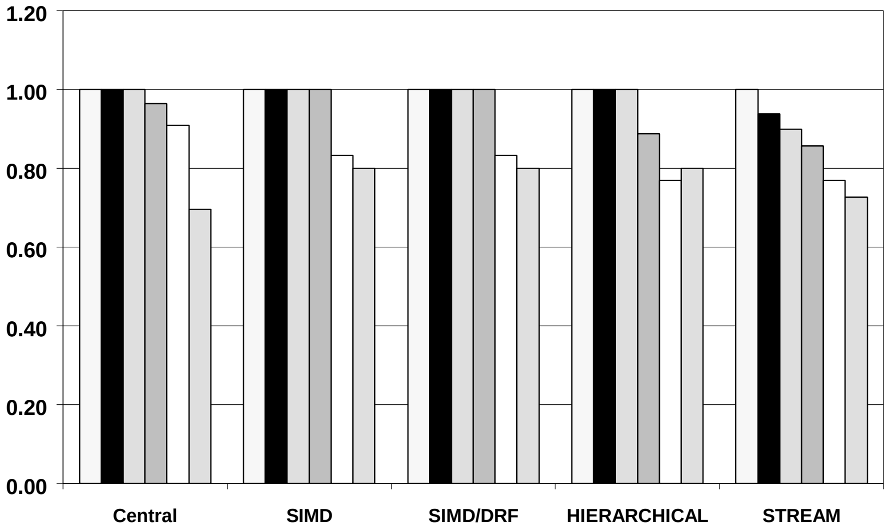
Delay of Register Organizations



Performance of Register Organizations

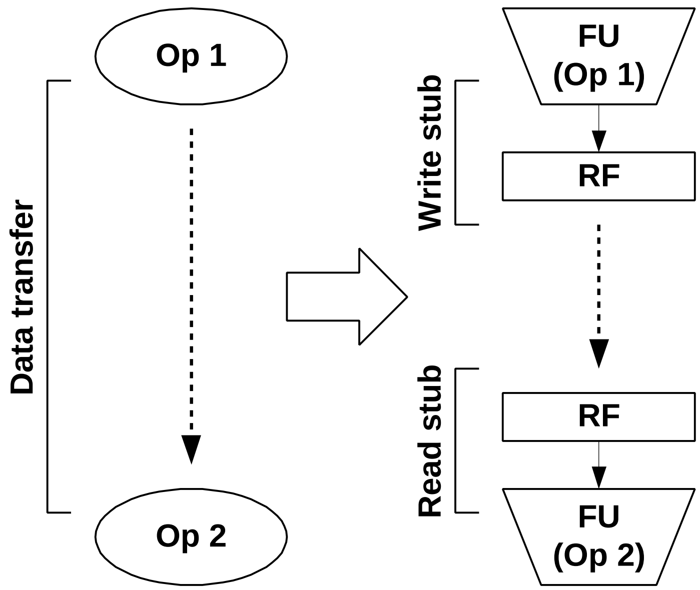


(A) Raw Performance

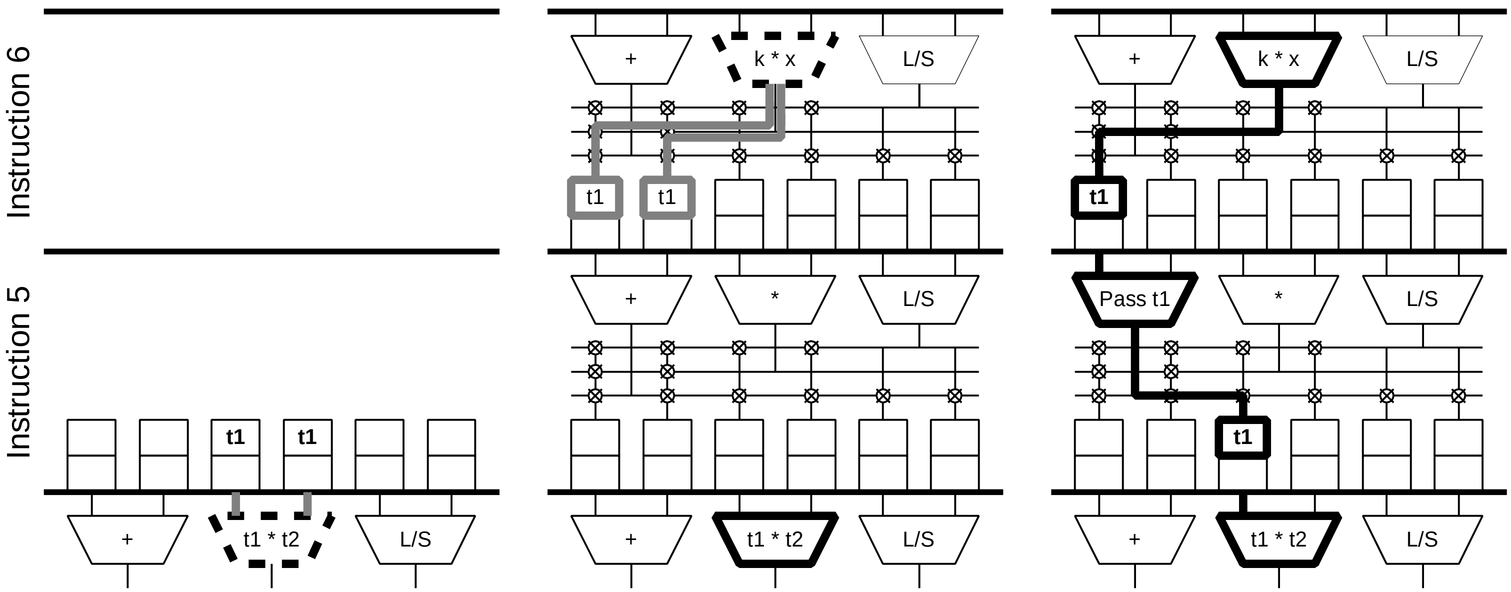


(B) Performance with Latency

Stubs Abstract the Communication Between Operations



A Communication Example

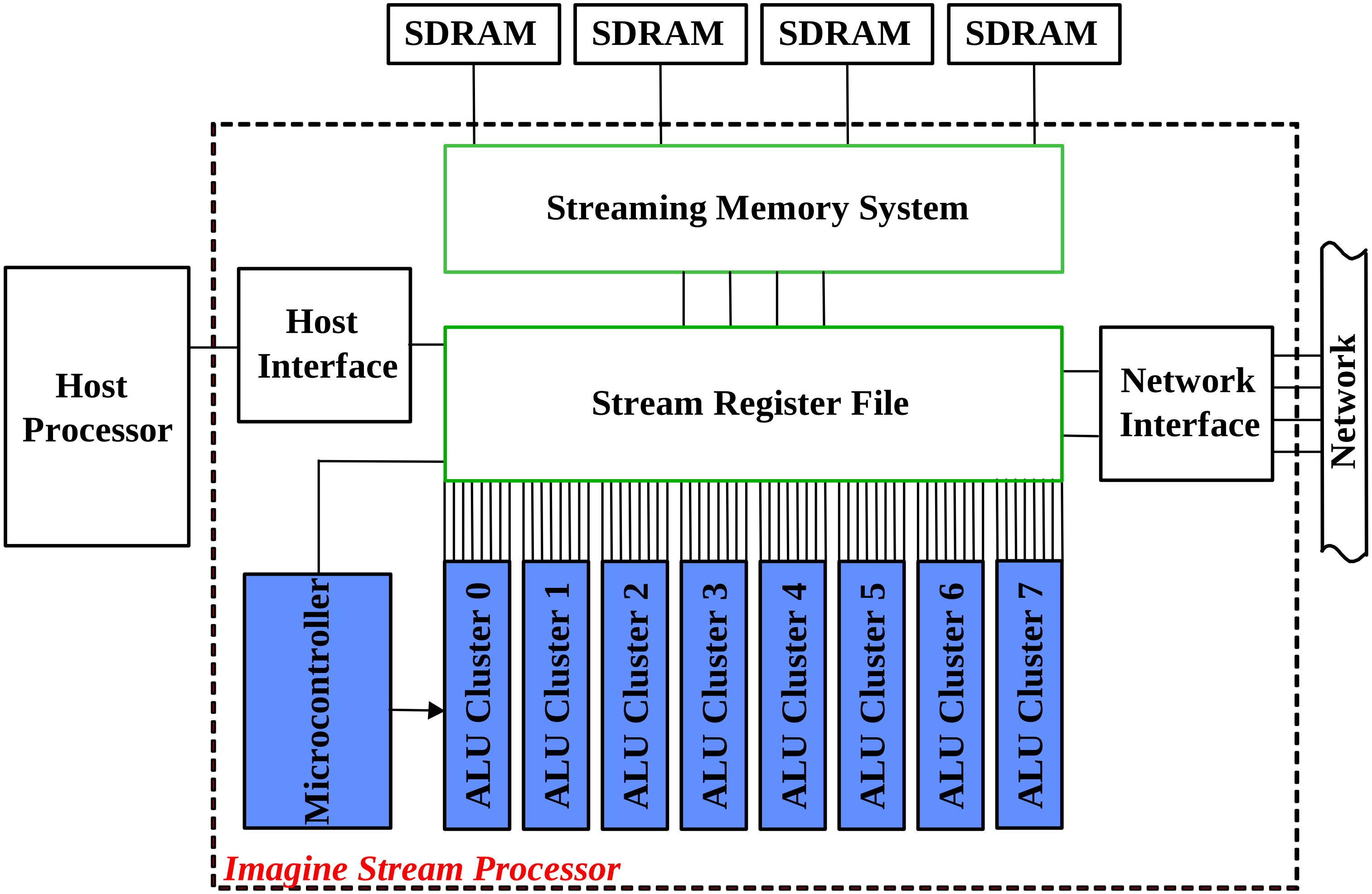


(a)

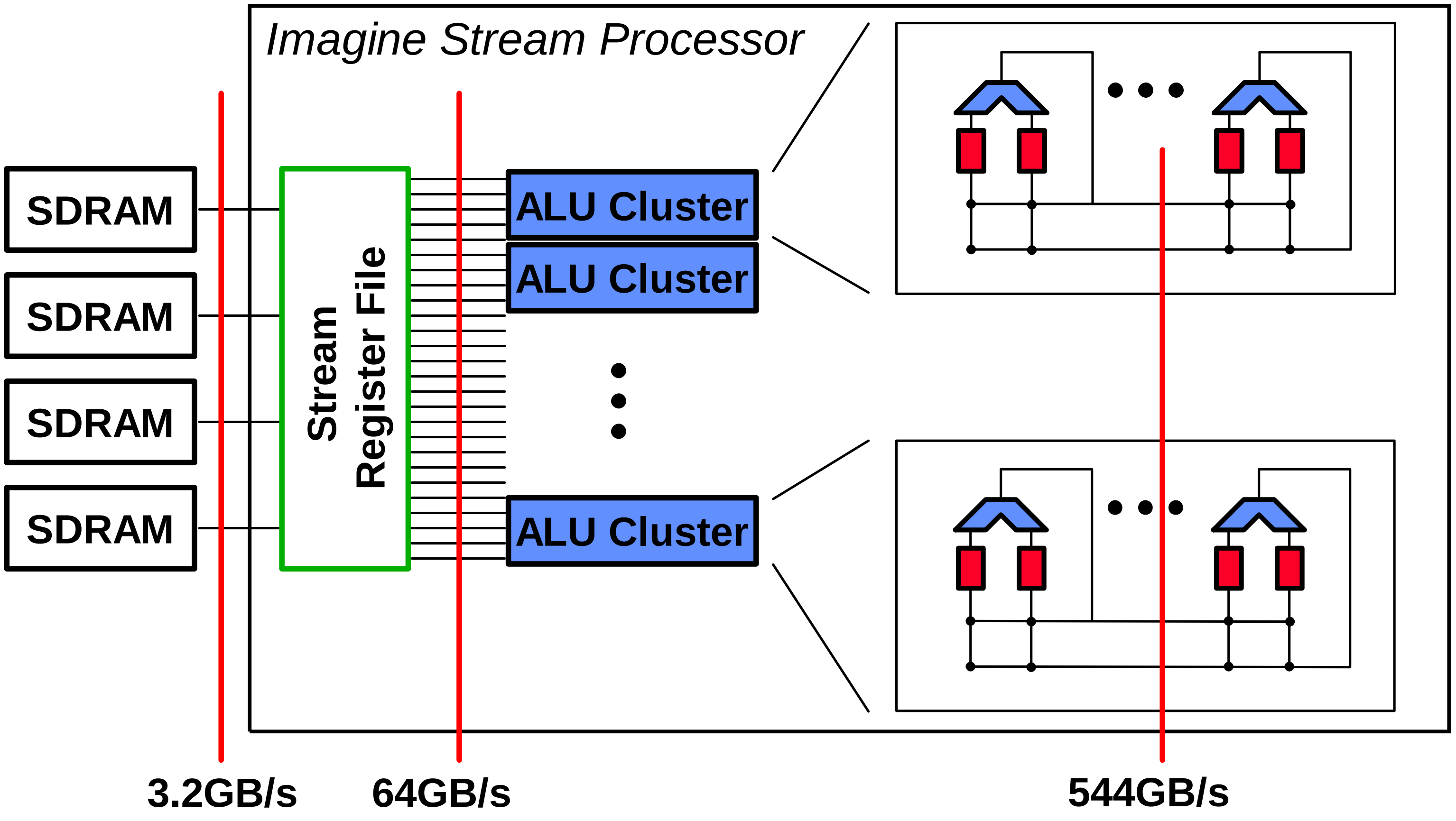
(b)

(c)

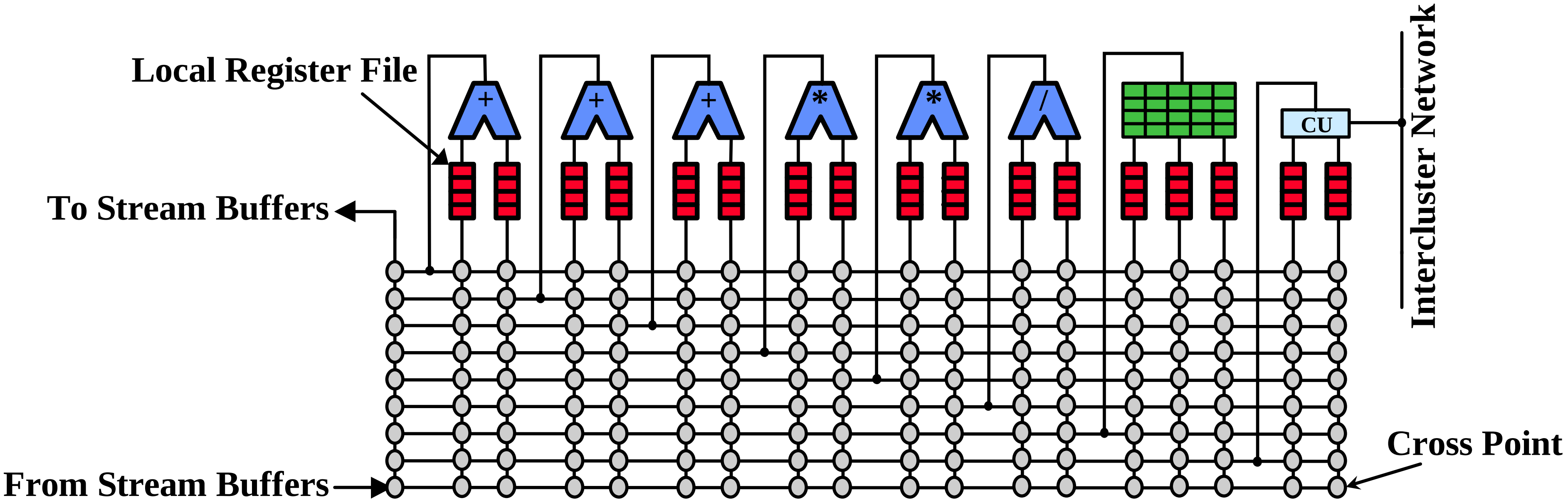
The Imagine Stream Processor



Data Bandwidth Hierarchy



Cluster Architecture

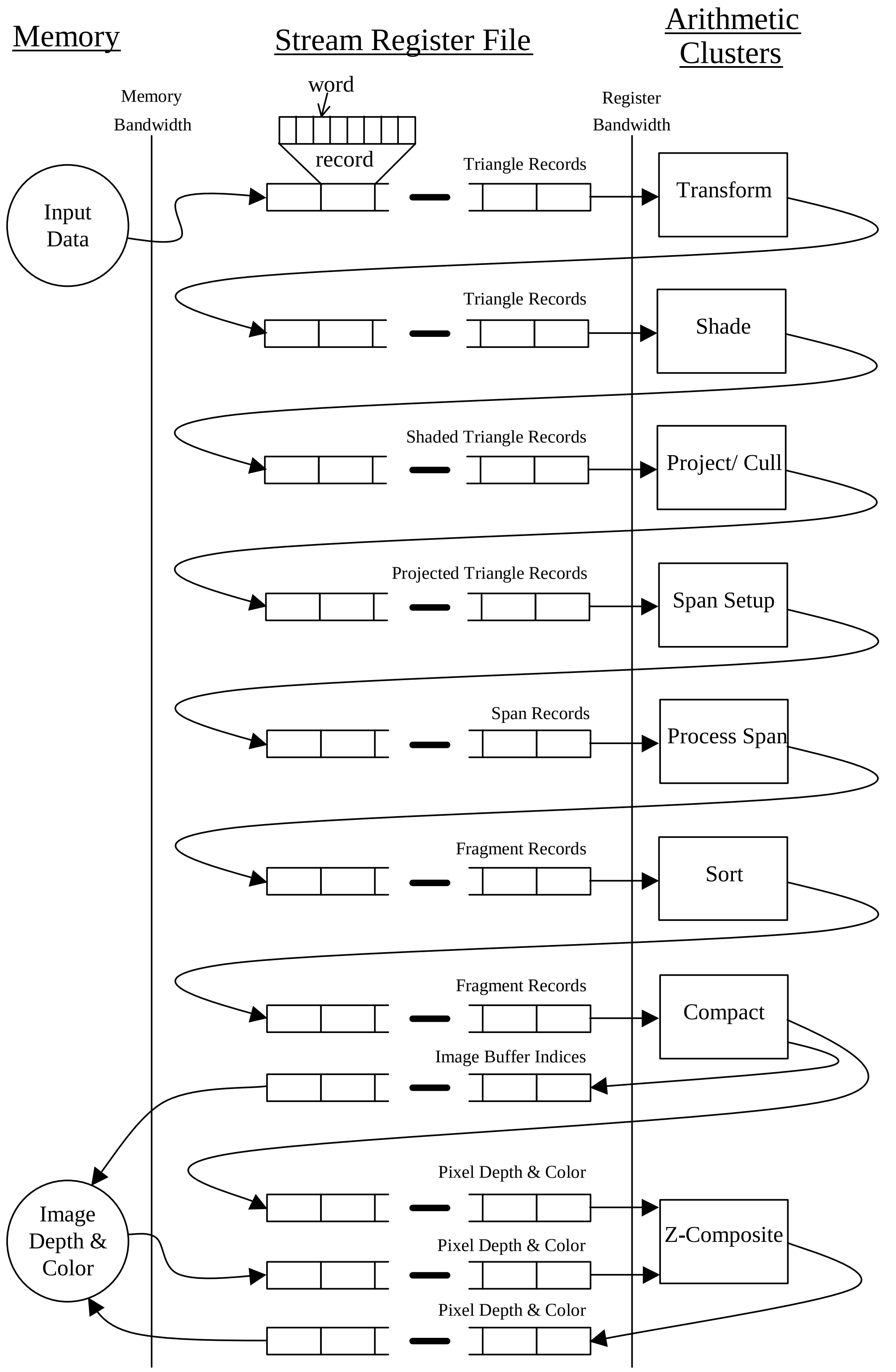


- VLIW organization with shared control
- Local register files provide high data bandwidth

Imagine is a *Stream Processor*

- Instructions are Load, Store, and Operate
  - operands are streams
  - also Send and Receive for multiple-imagine systems
- Operate performs a *compound stream operation*
  - read elements from input streams
  - perform a local computation
  - append elements to output streams
  - repeat until input stream is consumed
  - (e.g., triangle transform)
- Order of magnitude less global register bandwidth than a vector processor

Triangle Rendering



Bandwidth Demands

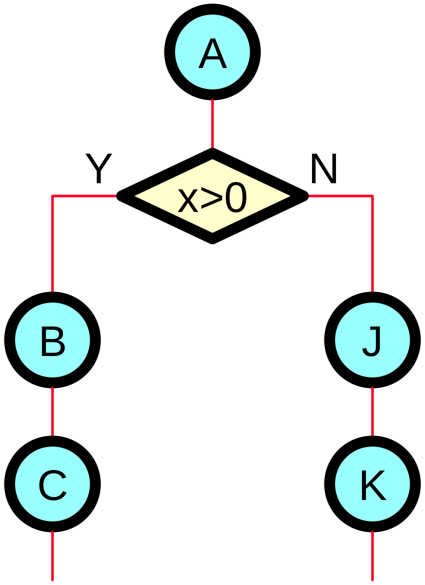
Transform Kernel

References (per Δ)	Stream	Scalar		Vector	
Memory	5.5	117	(21.3)	48	(8.7)
Global RF	48	624	(13.0)	261	(5.4)
Local RF	372	N/A		N/A	

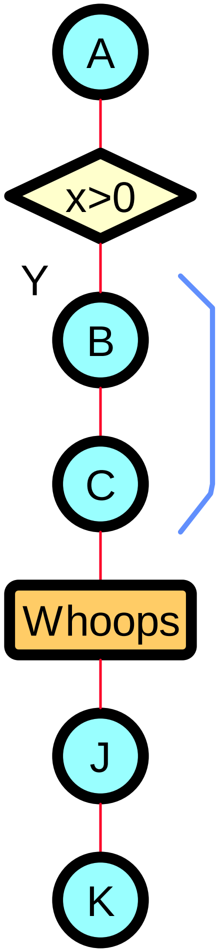
Data Parallelism is easier than ILP

Kernel	1 to 8 Cluster Speedup
FFT (1024)	6.4
DCT (8x8)	7.8
Blockwarp (8x8)	7.2
Transform (Δ)	8.0
Harmonic Mean	7.3

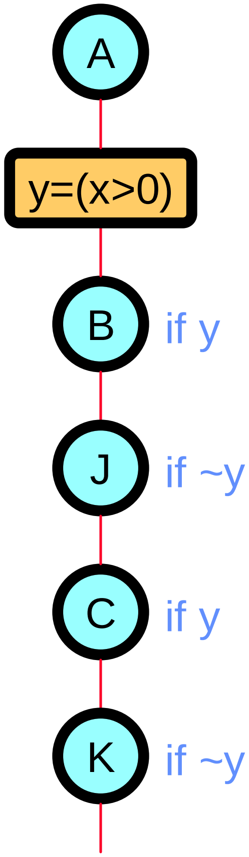
Conventional Approaches to Data-Dependent Conditional Execution



Data-Dependent Branch



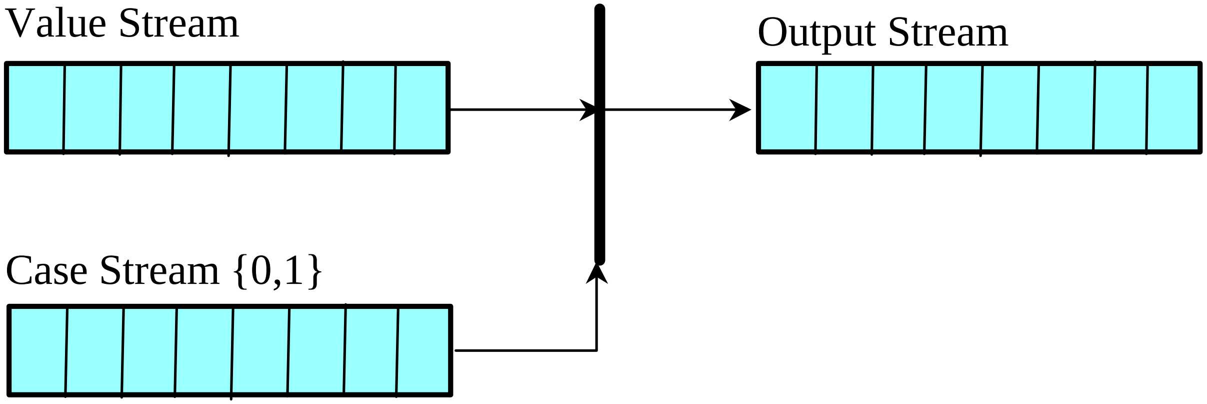
Speculative Loss D x W ~100s



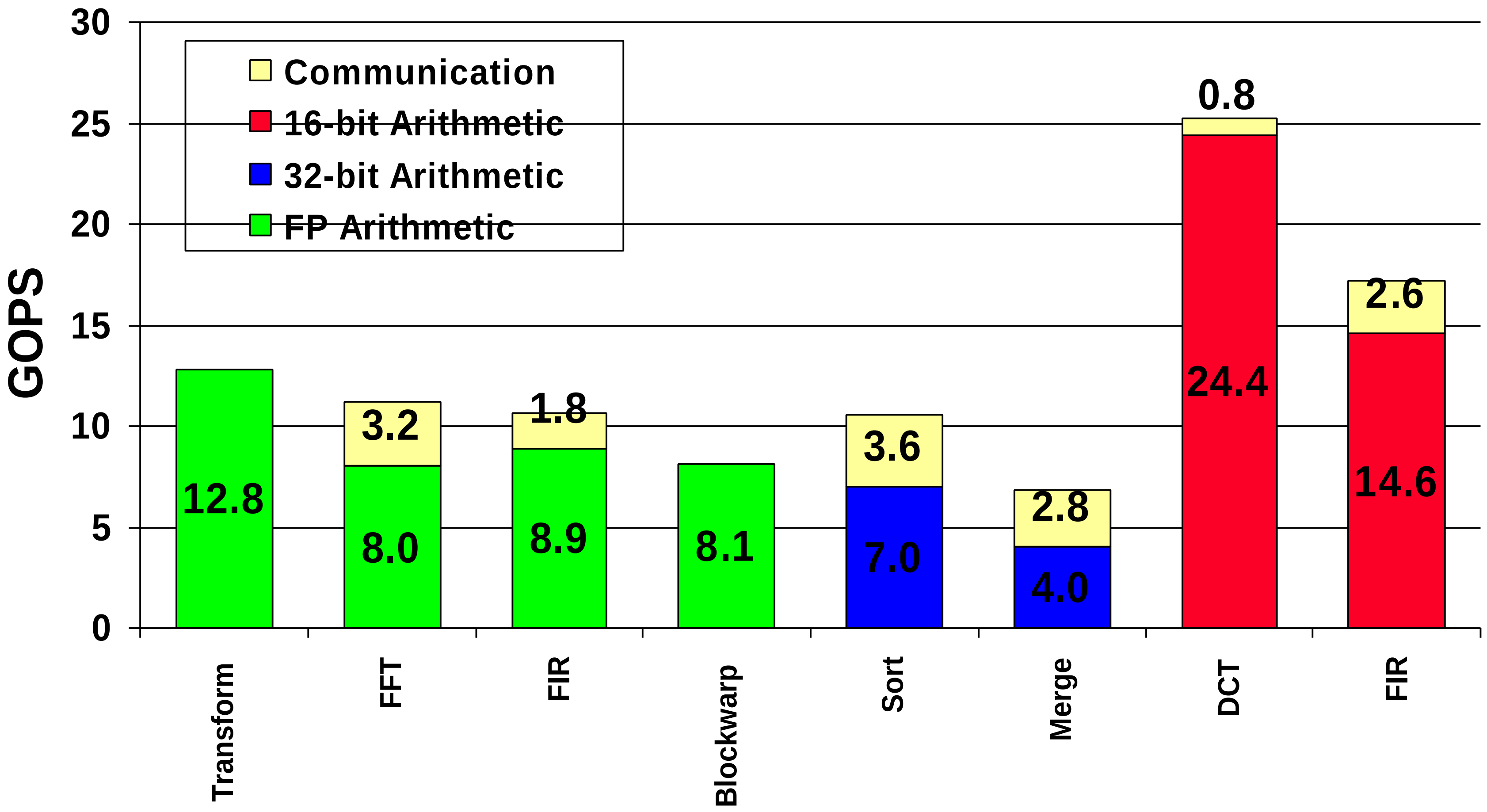
Exponentially Decreasing Duty Factor

Zero-Cost Conditionals

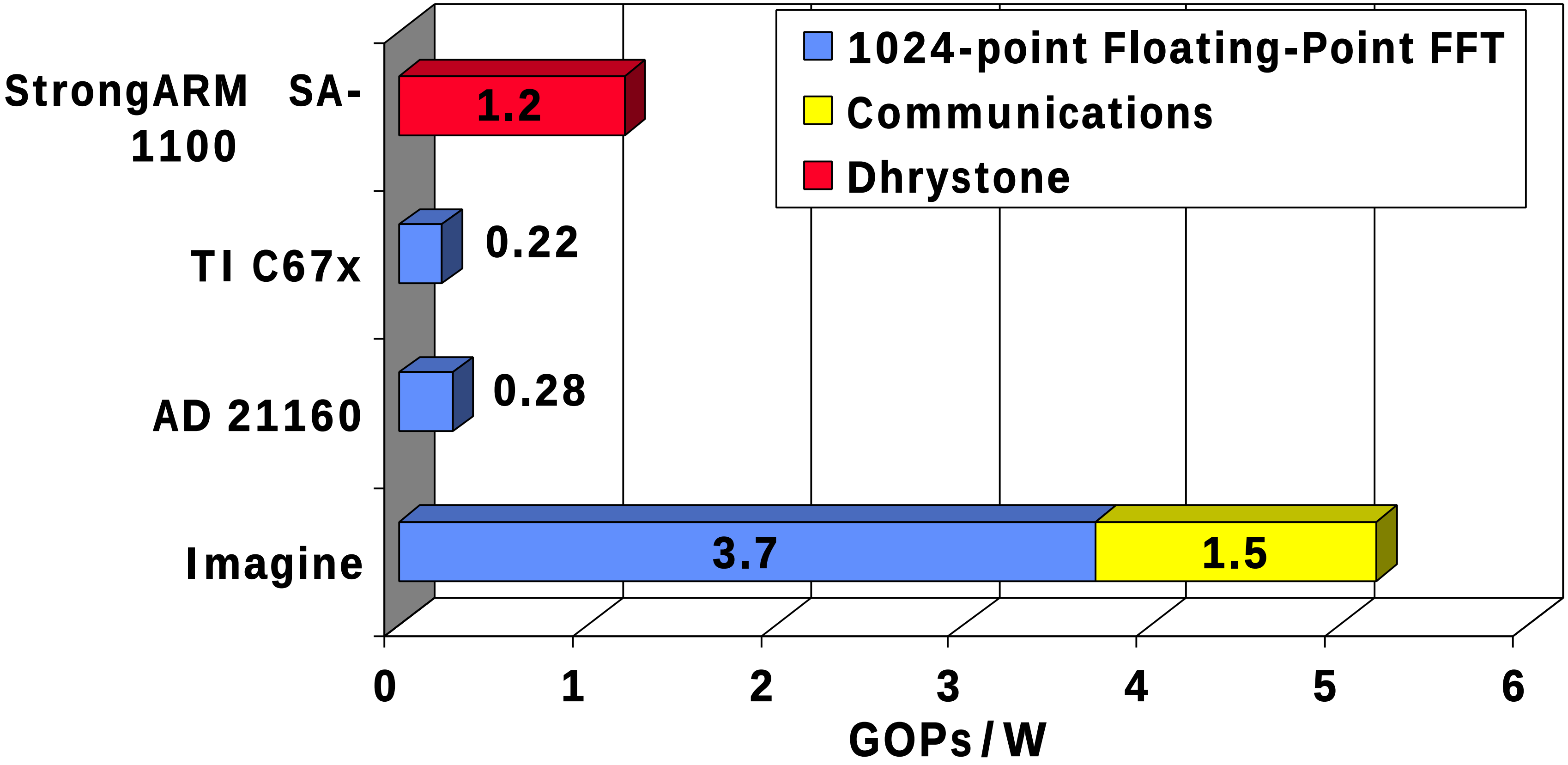
- Most Approaches to Conditional Operations are Costly
  - Branching control flow - dead issue slots on mispredicted branches
  - Predication (SIMD select, masked vectors) - large fraction of execution ‘opportunities ’go idle.
- Conditional Streams
  - append an *element* to an output stream depending on a *case* variable.



Sustainable Performance

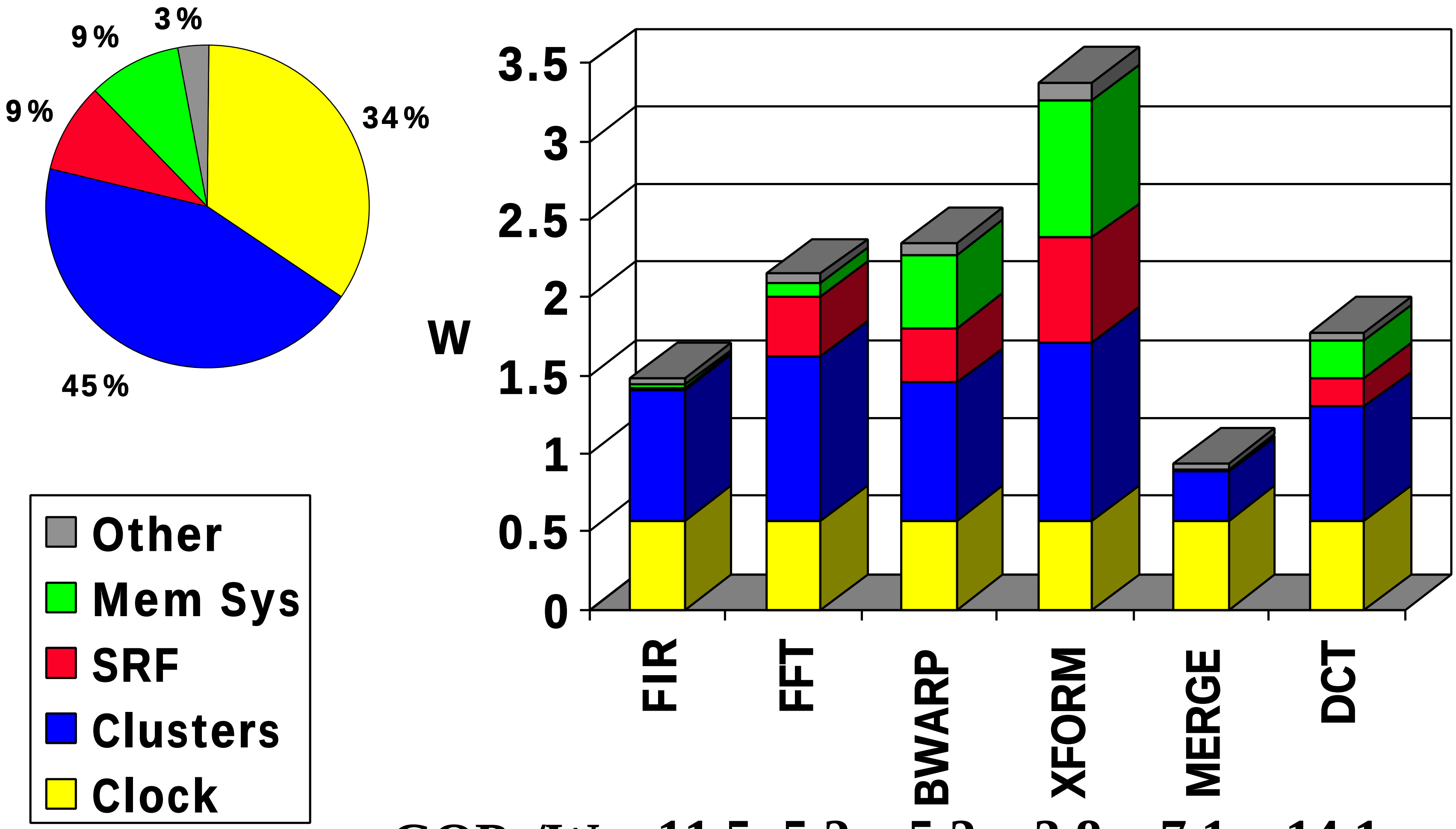


Power Comparison



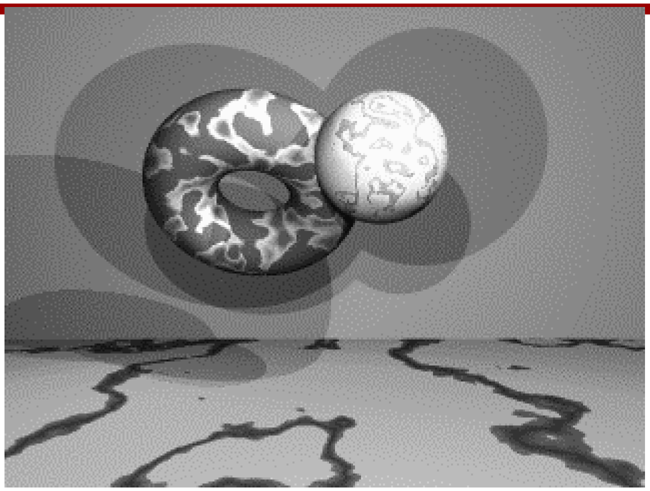
-Source: Web Pages of Intel, TI, and Analog Devices

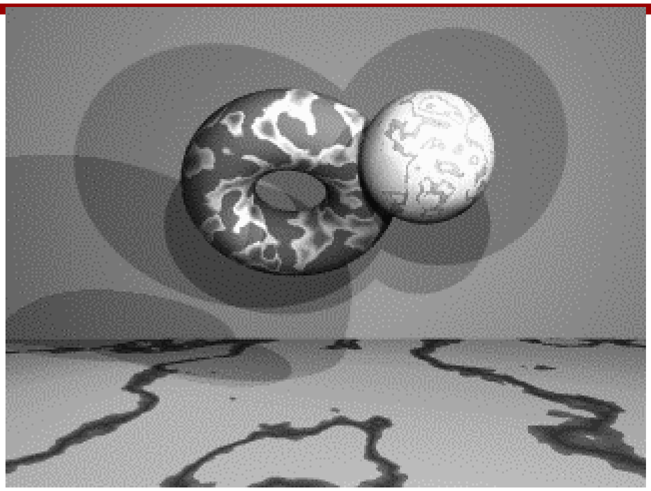
Power and Performance

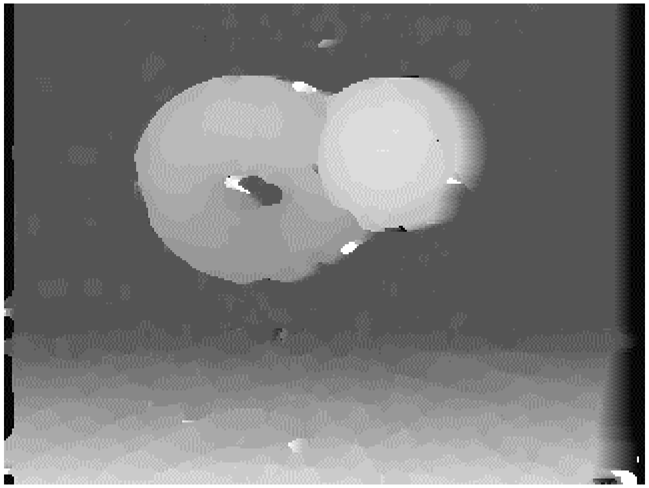


GOPs/W: 11.5 5.2 5.2 3.8 7.1 14.1

A Look Inside an Application
Stereo Depth Extraction



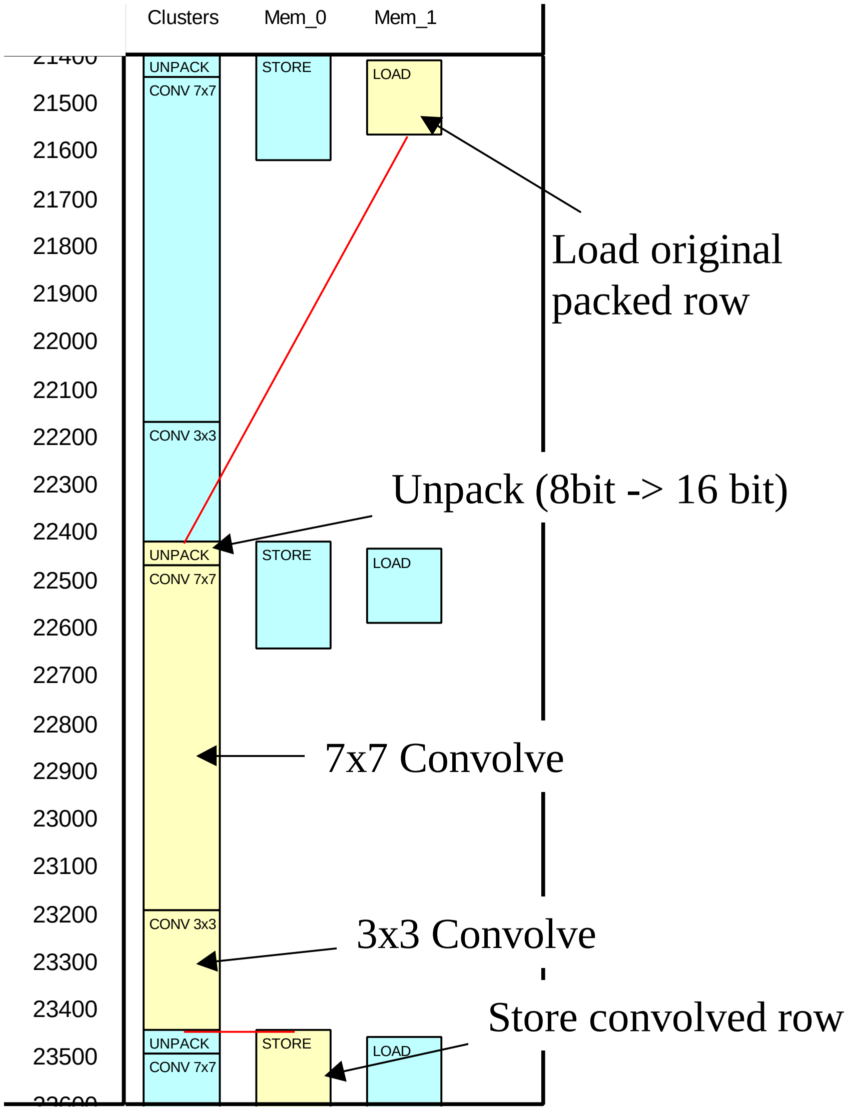




- 320x240 8-bit grayscale images
- 30 disparity search
- 220 frames/second
- 12.7 GOPS
- 5.7 GOPS/W

Stereo Depth Extractor

Convolutions



Load original packed row

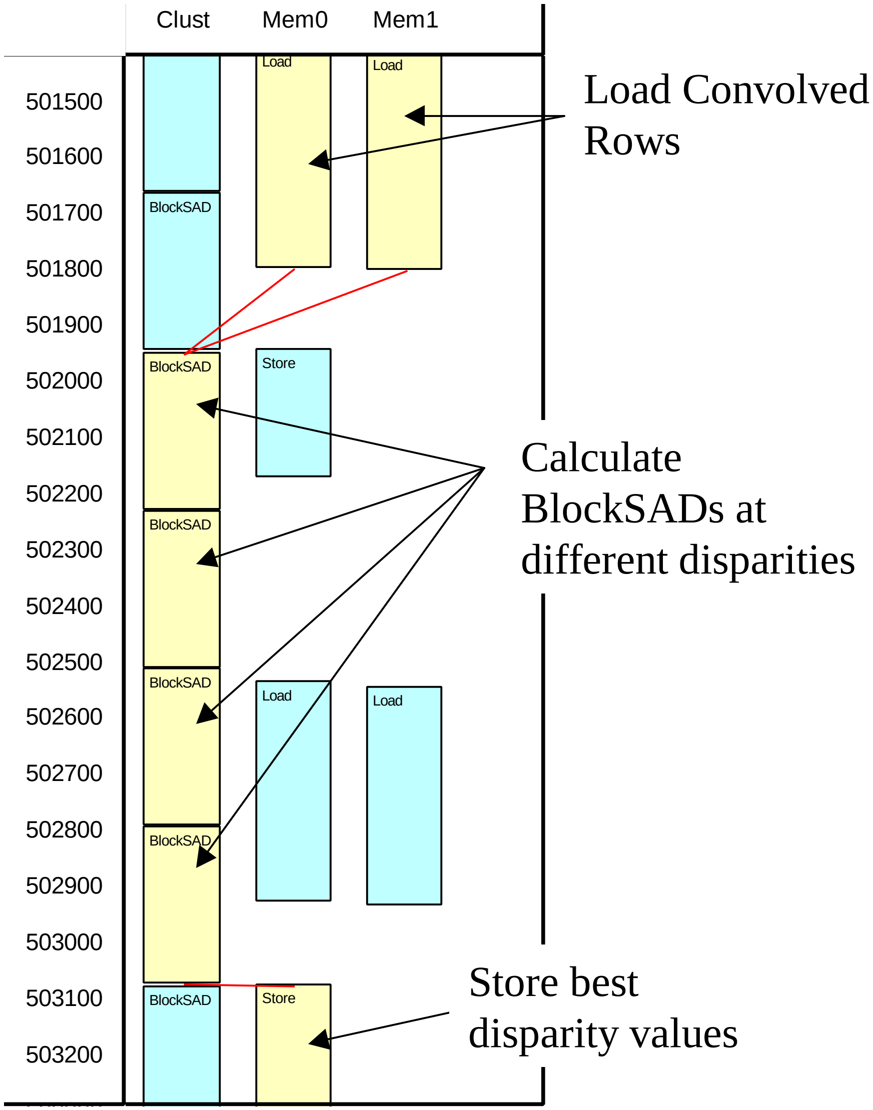
Unpack (8bit -> 16 bit)

7x7 Convolve

3x3 Convolve

Store convolved row

Disparity Search

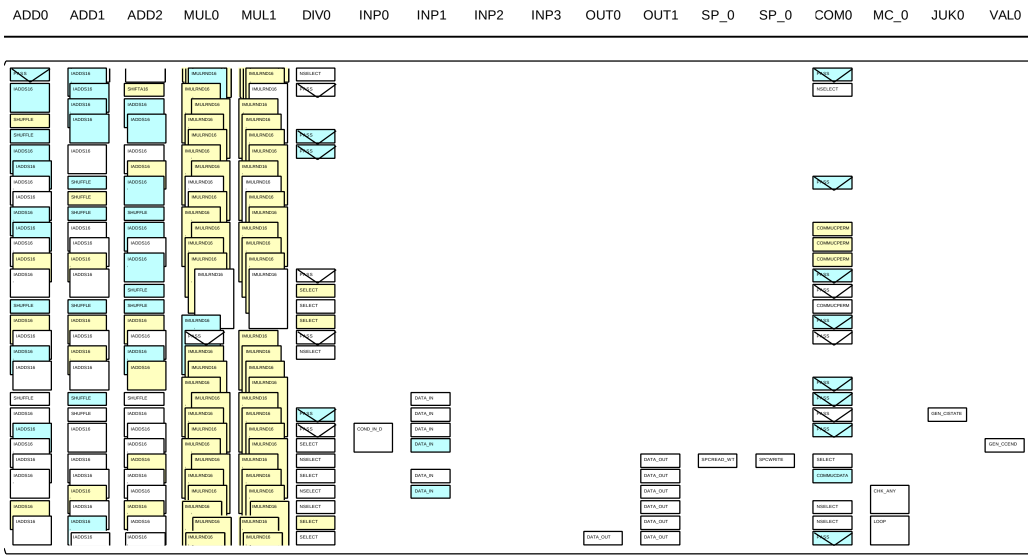


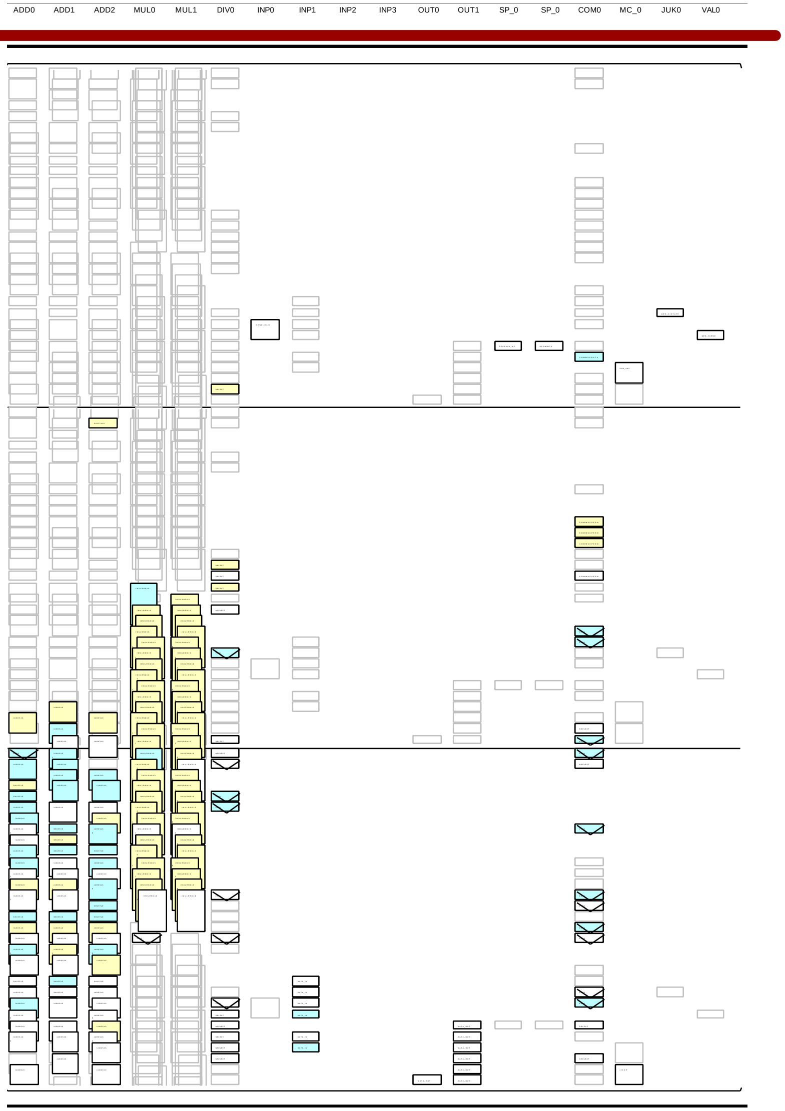
Load Convolved Rows

Calculate BlockSADs at different disparities

Store best disparity values

7x7 Convolve Kernel





Imagine Summary

- Imagine operates on *streams* of records
  - simplifies programming
  - exposes locality and concurrency
- Compound stream operations
  - perform a subroutine on each stream element
  - reduces global register bandwidth
- Bandwidth hierarchy
  - use bandwidth where its inexpensive
  - distributed and hierarchical register organization
- Conditional stream operations
  - sort elements into homogeneous streams
  - avoid predication or speculation

Computer Architecture for the Next Millenium

- Applications and technology are changing
  - *media* applications process streams of low-precision samples
  - wires dominate gates
- ILP is at the point of diminishing returns
- Tremendous opportunities for new architectures
  - new applications have *lots* of parallelism and locality
  - modern technology can build chips with 100s of ALUs (32b FP) 1000s in the near future
- The challenge is to develop architectures
  - that can harness this potential performance
  - in a way that can be easily programmed
- Stream processing is one approach, there are many others. We need to start exploring them