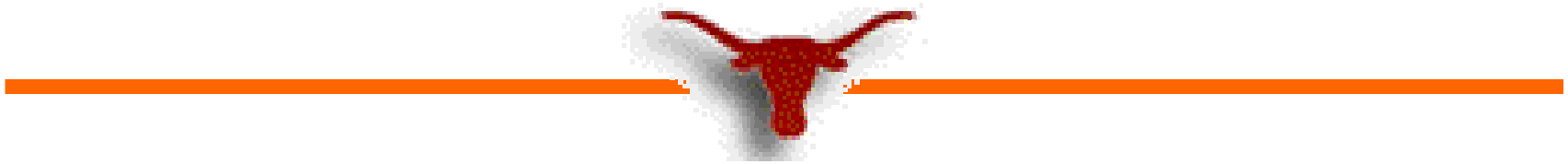


Mechanisms for Streaming Architectures

Stephen W. Keckler

Computer Architecture and Technology Laboratory

Department of Computer Sciences

The University of Texas at Austin

August 22, 2003

Diversity in Streaming Programs

- High variability in attributes of:
  - Computation/memory BW ratio
  - Control structure
  - Memory access: data stream, constants
- Examples
  - DSP/multimedia
    - Subwork parallelism, predictable control flow
  - Scientific
    - Vectorizable with regular or irregular data access
  - Communication (packet processing, encryption, etc.)
    - Regular control flow
    - Irregular control, data dependent control flow
  - Graphics
    - Regular control/data (rasterization)
    - Irregular control/data (shading)

Challenges for Streaming Architectures

- Supporting regular and irregular control structures

RGB to YIQ conversion

```
for(i=0; i<n; i++ {
  read(r[i],g[i],b[i]);
  Y = K1*r+K2*g+K3*b;
  I = K4*r+K5*g+K6*b;
  Q = K7*r+K8*g+K9*b;
  store(Y[i],I[i],Q[i]);
}
```

“Vertex skinning”

```
for(i=0; i<n; i++ {
  read(D[i]);
  for(j=0; j<m[i]; j++)
    D = func(D, table[j]);
  store(D[i]);
}
```

- Different types of storage demands
  - Stream/vector register files (regular access)
  - Scalar constants
  - Indexed tables
- Scaling of kernel processors
  - Large bypass networks, instruction broadcasting
  - Amenable to pipelining – but decreases agility
- Traditional streaming architectures ⇨ regular

DLP Control Characteristics

	Benchmark	Kernel Size (# instr)	Record size (bytes)	ILP	Comp/ Mem	Iterations in inner loop
Multimedia	Convert	15	24	5	2.5	0
	DCT	1728	512	6	27	16
	High pass filter	17	72	3.4	1.7	0
Scientific	FFT	10	48	3.3	1.0	0
	LU	2	16	1	0.7	0
Network	MD5	680	80	1.6	56.7	0
	Blowfish	364	8	2.0	182	16
	Rijndael	650	16	11.8	162.5	10
Graphics	Vertex_simple_light	95	56	4.3	7.3	0
	Fragment_simple_light	64	64	3.0	5.3	0
	Vertex_reflection	94	72	7.1	8.5	0
	Fragment_reflection	98	40	6.2	12.3	0
	Vertex_skinning	112	128	6.8	4.5	Variable

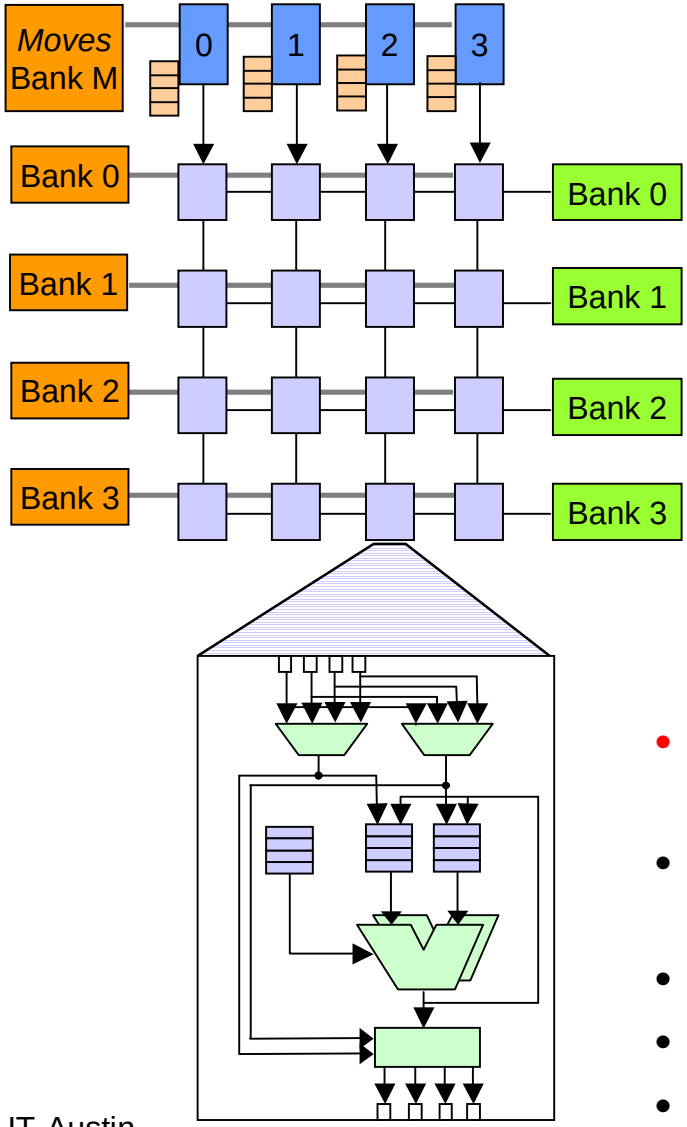
DLP Data Characteristics

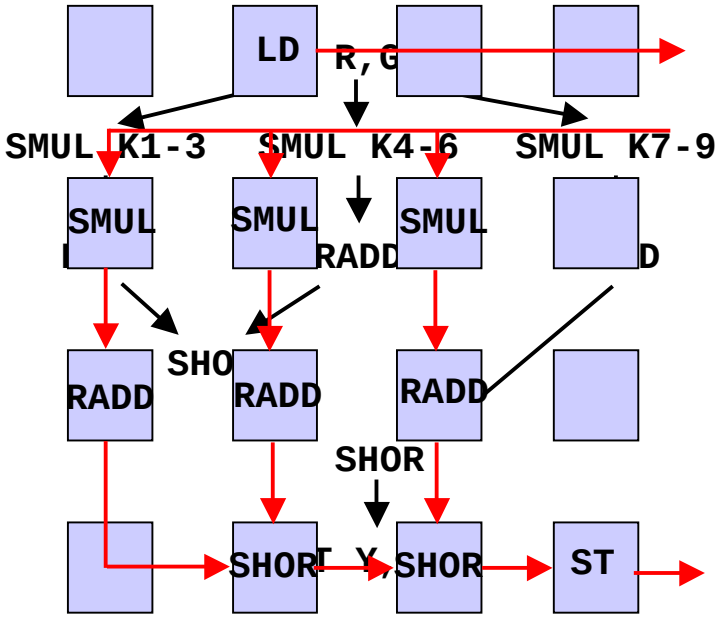
	Benchmark	Read Record size (bytes)	Write Record size (bytes)	# scalar constants	Lookup table size	# of Table Accesses
Multimedia	Convert	24	24	9	0	
Multimedia	DCT	512	512	10	0	
Multimedia	High pass filter	72	8	9	0	
Scientific	FFT	48	32	0	0	
Scientific	LU	16	8	0	0	
Network	MD5	80	16	65	0	
Network	Blowfish	8	8	2	256	82
Network	Rijndael	16	16	18	1024	160
Graphics	Vertex_simple_light	56	48	32	0	
Graphics	Fragment_simple_light	64	32	16	0	
Graphics	Vertex_reflection	72	16	35	0	
Graphics	Fragment_reflection	40	24	7	0	
Graphics	Vertex_skinning	128	72	32	288	36

Desirable Attributes of a Stream Processor

- Performs well on DLP programs with different attributes
  - Synchronous core to minimize synchronization overheads on traditional vector/stream applications
  - MIMD-like capabilities for applications with irregular control
  - Support for different types of data structures
- Partitioned and scalable microarchitecture
  - Dataflow instruction execution
  - Limit/eliminate global broadcast of instructions/data
- Decoupled processor core
  - From memory system to enable memory fetch parallelism
  - From other processor cores to enable kernel pipelining
- Efficient instruction distribution and re-use
  - Exploit spatial/temporal locality

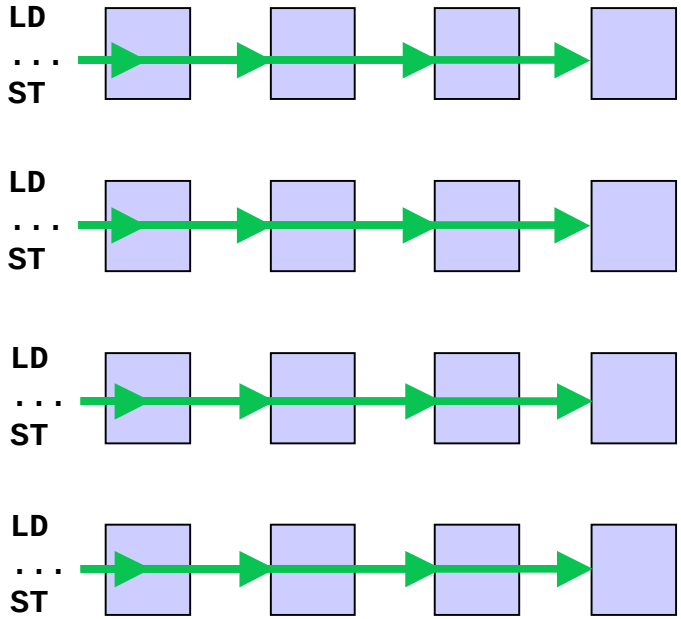
Dataflow Execution in TRIPS Core



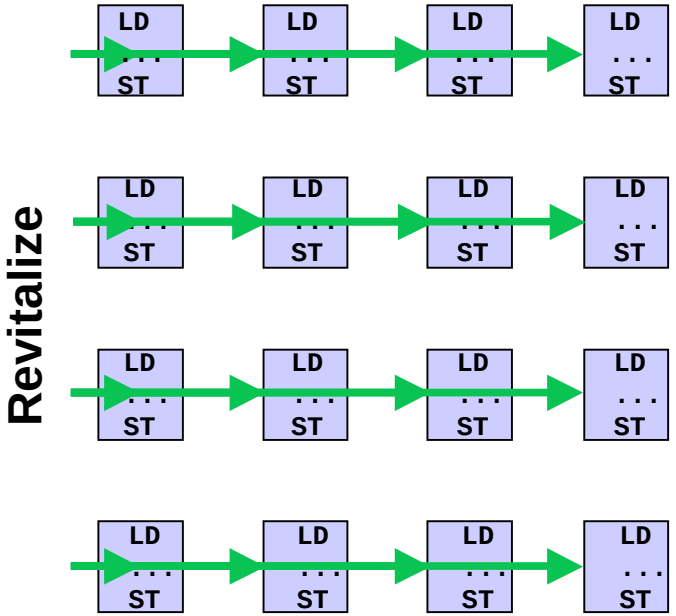


- SPDI: Static Placement, Dynamic Issue
  - Instructions execute in dataflow order
- Instructions stream in from left, data from right
- ALU Chaining
- Pipeline instruction distribution and execution
- Static unrolling for latency tolerance

Instruction Fetch Efficiency

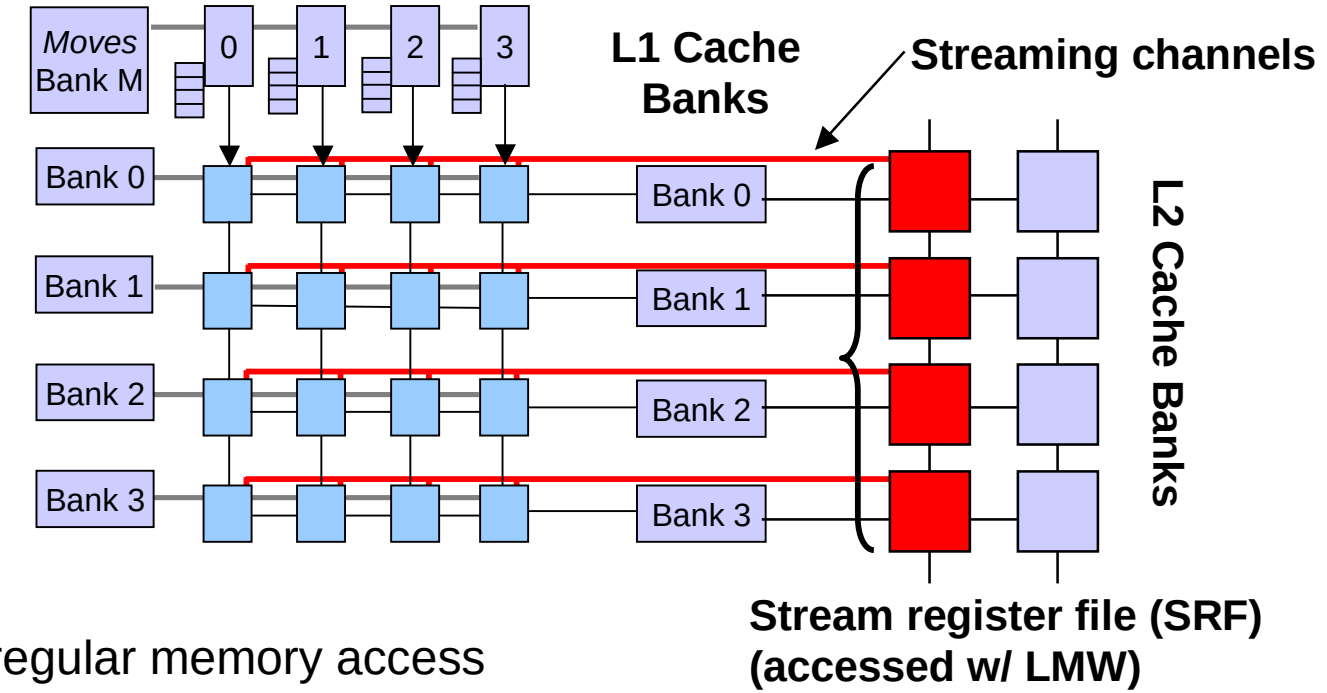


- Spatial loop unrolling
  - Copy same instruction sequence across multiple execution units



- Mapping reuse
  - Reset previously mapped instructions
  - Re-execution without refetch

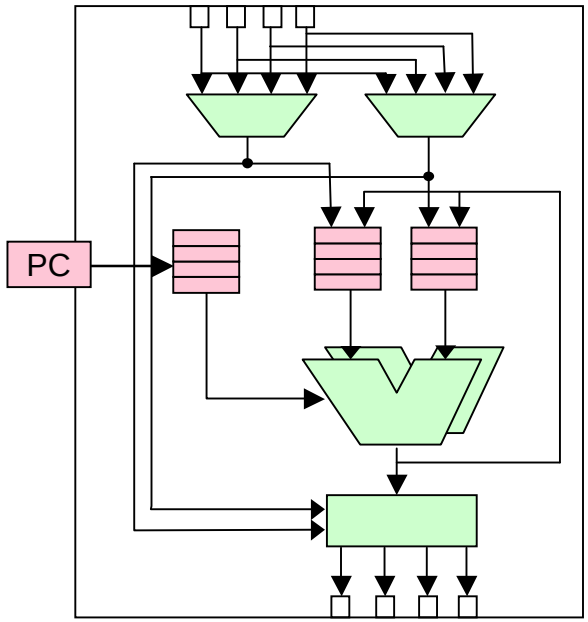
Memory Accesses



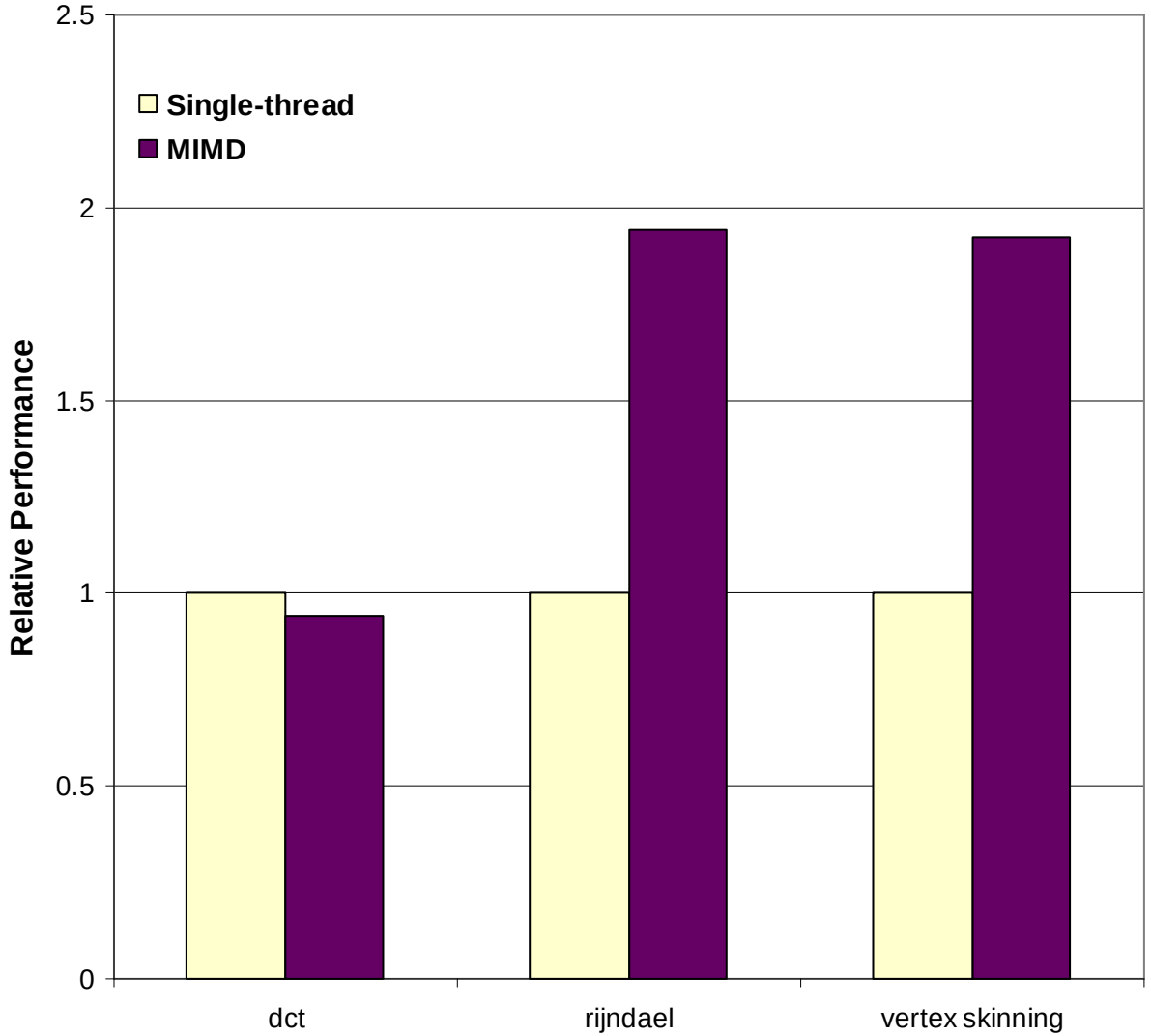
- Irregular memory access
  - Map to hardware cache hierachy
- Regular data accesses
  - Subset of L2 cache banks configured as Stream Register File (SRF)
  - High bandwidth data channels to SRF, reduced address BW
  - DMA engines transfer between SRF and DRAM or other SRFs
- Constants saved in reservation stations with corresponding instructions

MIMD Extensions

- Extensions to ALU nodes
  - Local control (PC)
    - Independent loops
    - Conditionals
  - Treat frame space as local instruction and data memory
- Tightly coupled MIMD
  - Central sequencer
    - Distributes kernel code to ALUs
    - Determines when to proceed to next kernel
  - Inter-ALU communication under investigation
- Issue: limited reservation station storage
  - Add more local instruction storage
  - Aggregate multiple ALUs into larger logical ALU

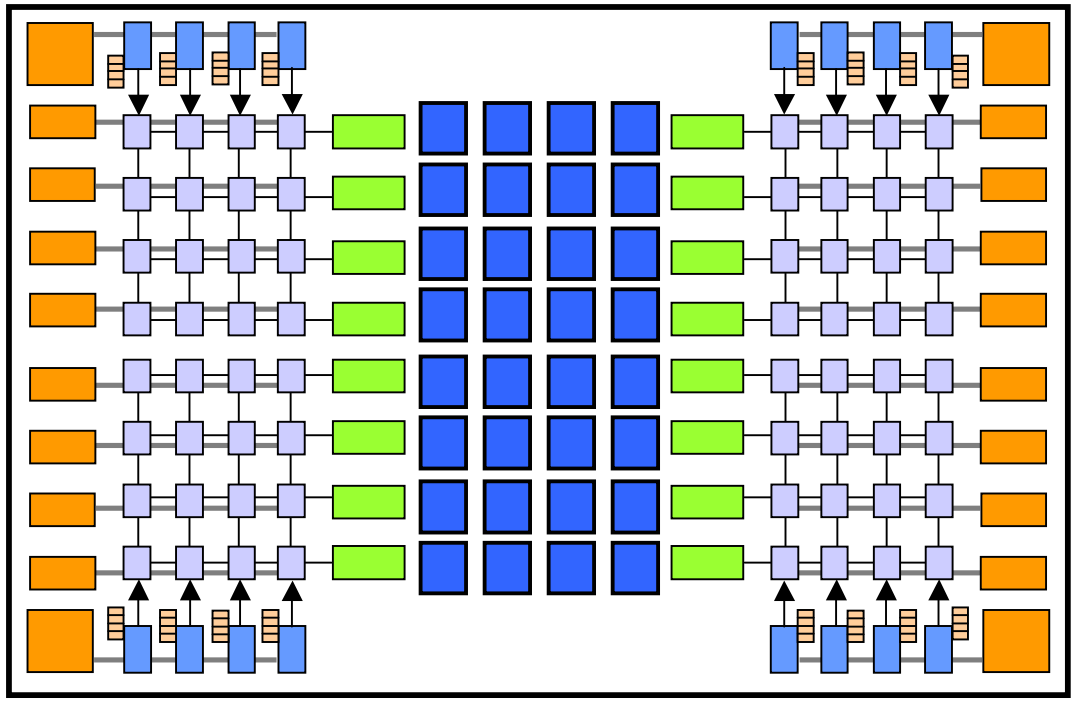


Performance Comparisons



TRIPS Chip

- 4 cores (with streaming support)
- L2 cache and SRF memory banks
- Pipelining across kernels mapped to different cores
  - Extend to system through off-chip channels



Summary

- Streaming applications: irregular and regular
  - Irregular control and data access
  - Irregularity increasing (particularly in graphics domain)
- Limitations of many stream/vector kernel processors
  - Execution model demands regular control flow
    - Clever extensions such as conditional streams
  - Poor support for irregular table access
    - Scatter/gather for sparse/irregular vectors
- Hybridization can extend range of applications
  - Efficient caching support
  - Flexible instruction execution
    - Decouple execution engines and memory
    - Single instruction stream for regular control flow, loop-carried dependencies
    - Tightly coupled MIMD for irregular control flow
    - Less synchronization means better scalability