A Cached VRAM for 3D Graphics

Michael Deering
Michael Lavelle
Stephen Schlapp
Sun Microsystems Incorporated

Motivation

- **Existing RAMs are too slow for Z-buffered rendering**
  - **DRAM: best bit density**
  - **VRAM: improved bandwidth**
  - **Massive Interleaving: expensive**
  - **SRAM: fast cycle time; expensive**
- **Exotics: optimized for large polygons**

Key Design Concepts

- **Convert read-modify-write cycles into write only operations**
- **Two levels of internal rectangular pixel caches**

Leverage Existing 16Mbit DRAM technology

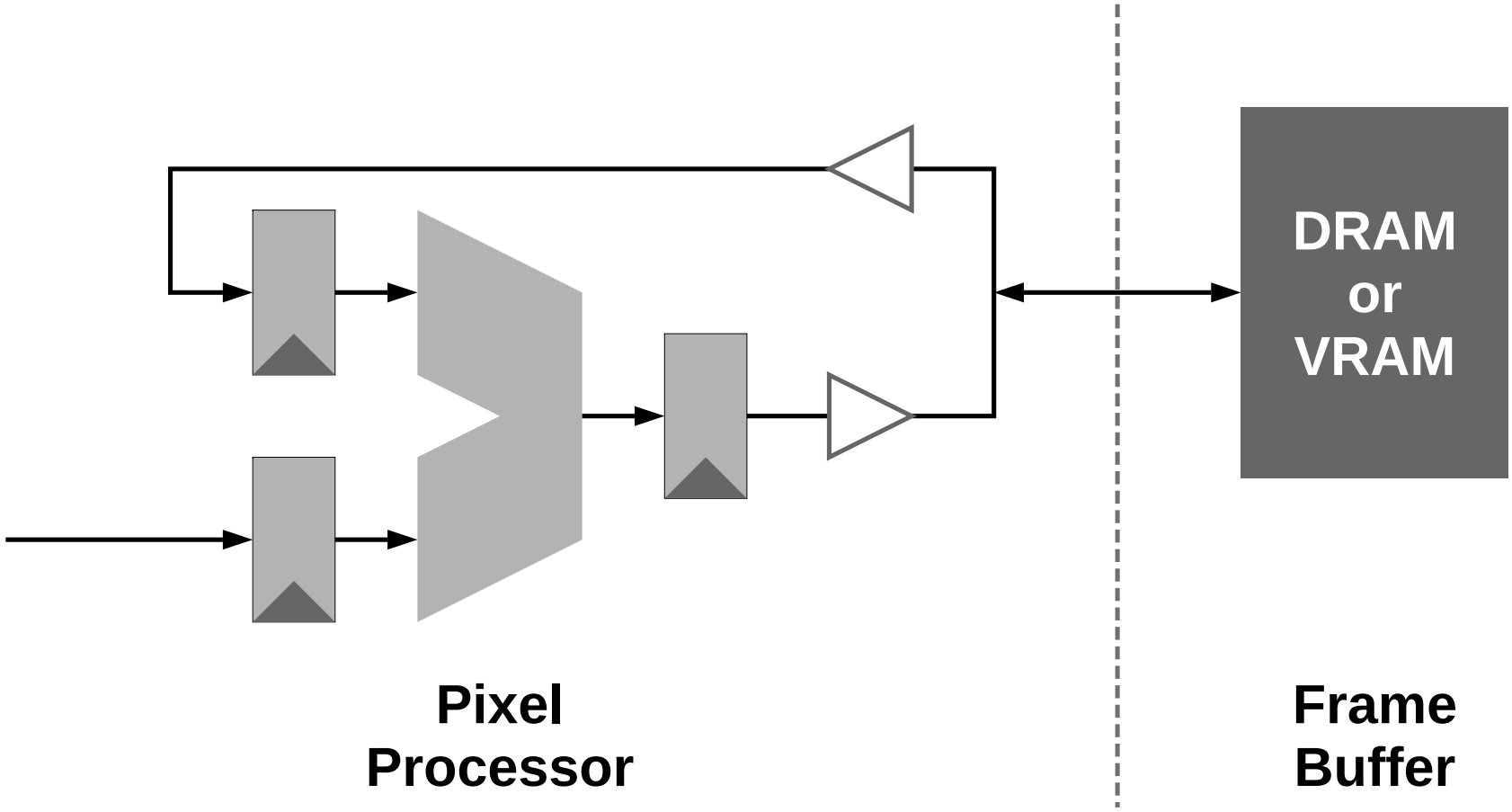
Use standard DRAM process
Use standard circuit designs
Use standard DRAM die size

Read-Modify-Write Interface

Used by Z-buffering & RGB Blending

1. **Read old pixel & receive new pixel**
2. **Merge new and old pixels; turn bus**
3. **Write merged pixel**
4. **Turn bus again**

Traditional RMW Frame Buffer Interface



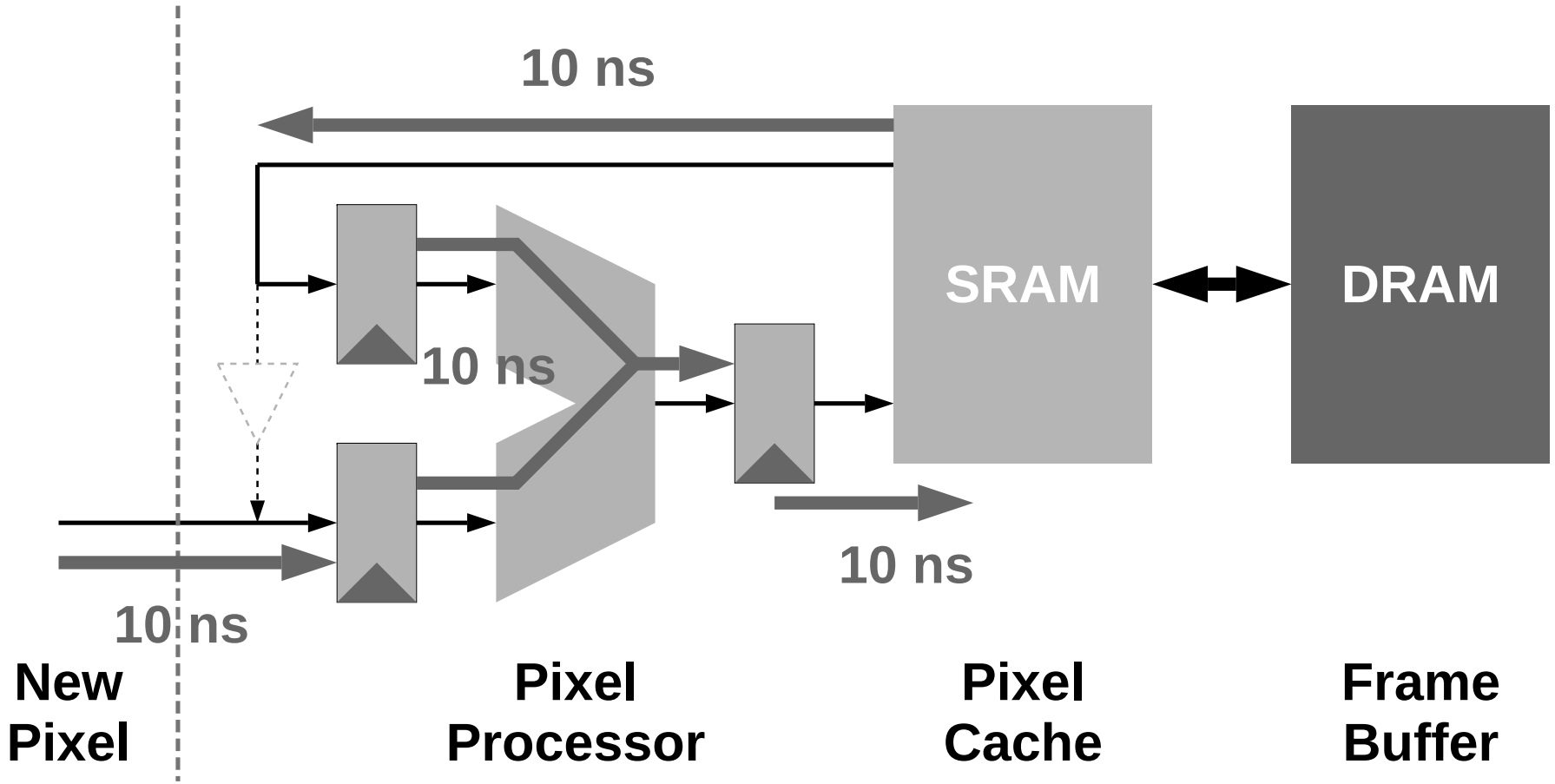
3D RAM Write Only Interface

- **3D RAM removes the bidirectional pin interface bottleneck**
- **Move RMW cycle inside 3D RAM**
- **Z-buffering done inside 3D RAM**
- **Internal ALU merges new and old pixels**
- **External interface is write only**

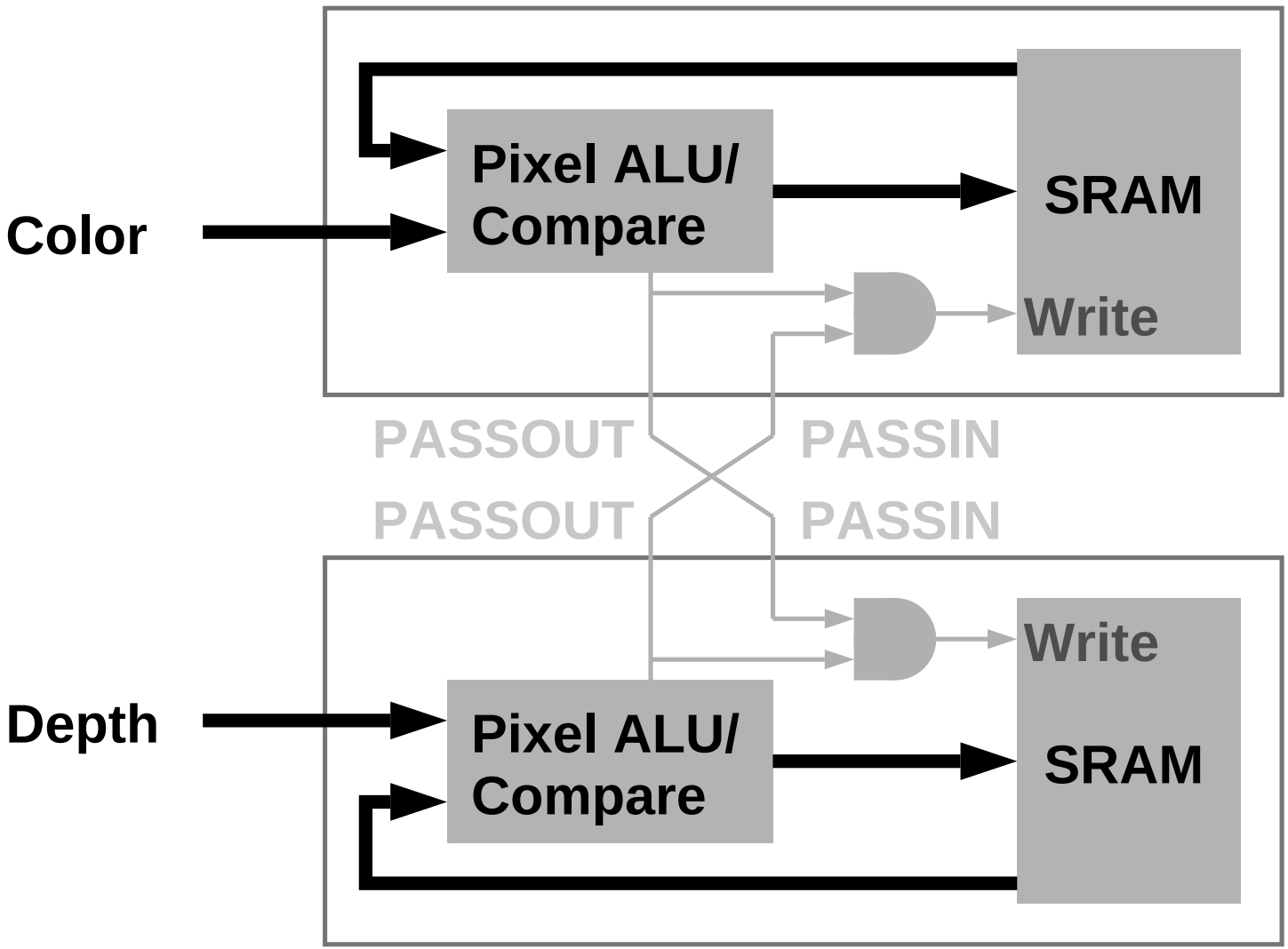
3D RAM Write Only Interface

- **ALU operations are pipelined**
- **All reads, writes, and processing overlapped @ 10ns/pixel rate**
- **No bus turn around required**
- **100 Million pixels/sec per chip!**

3DRAM Write Only Interface



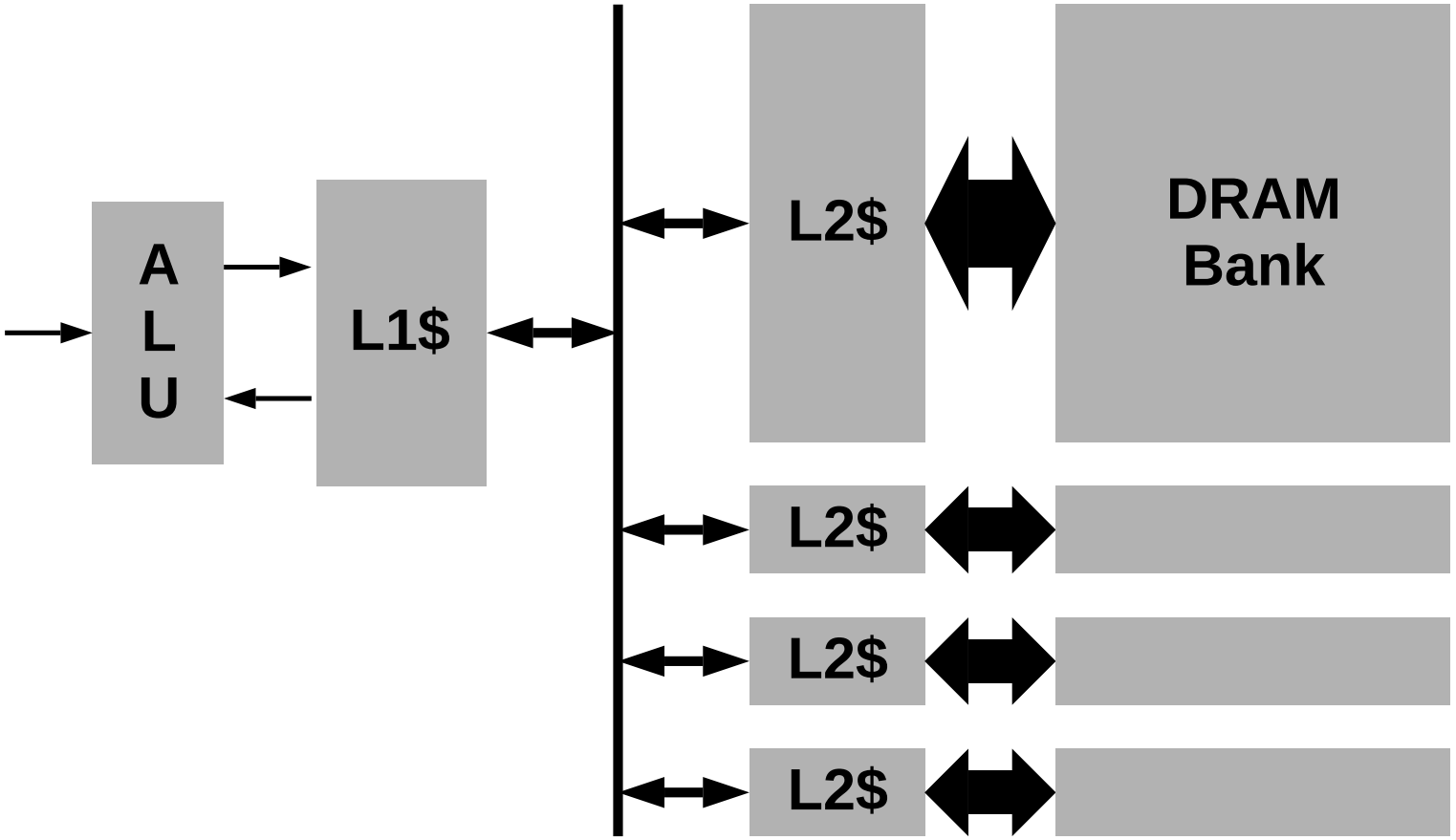
Conditional Write



On-chip Caching System

- **Rectangular blocks for graphics**
  - **Near-isotropic performance**
- **Two levels of cache**
  - **L1$ small, fast, multiported SRAM**
  - **L2$ big, slow, multiported buffer**
- **Multiple DRAM banks**
  - **Hide DRAM access & precharge**

On-chip Caching System



L1 Cache

- 3 port SRAM
  - 10 ns 32 bit read port to Pixel ALU
  - 10 ns 32 bit write port from Pixel ALU
  - 20 ns 256 bit bidirectional to global bus
- Eight 256 bit blocks
- Writeback
- N-way associativity

Global Bus

- **Connects L1$ to L2$**
- **One L1$ block wide**
- **Transfers 256 bits in 20 ns**
- **Bidirectional L1$⇔L2$**
- **L1$⇒L2$ transfers are plane and byte maskable**

L2 Cache

- **Four 10,240 bit page buffers**
- **Each buffer has 3 ports**

  20 ns 256 bit to global bus
  40 ns 640 bit to video buffer
  120 ns 10,240 bit to DRAM
- **Write through**
- **Direct mapped**

DRAM Banks

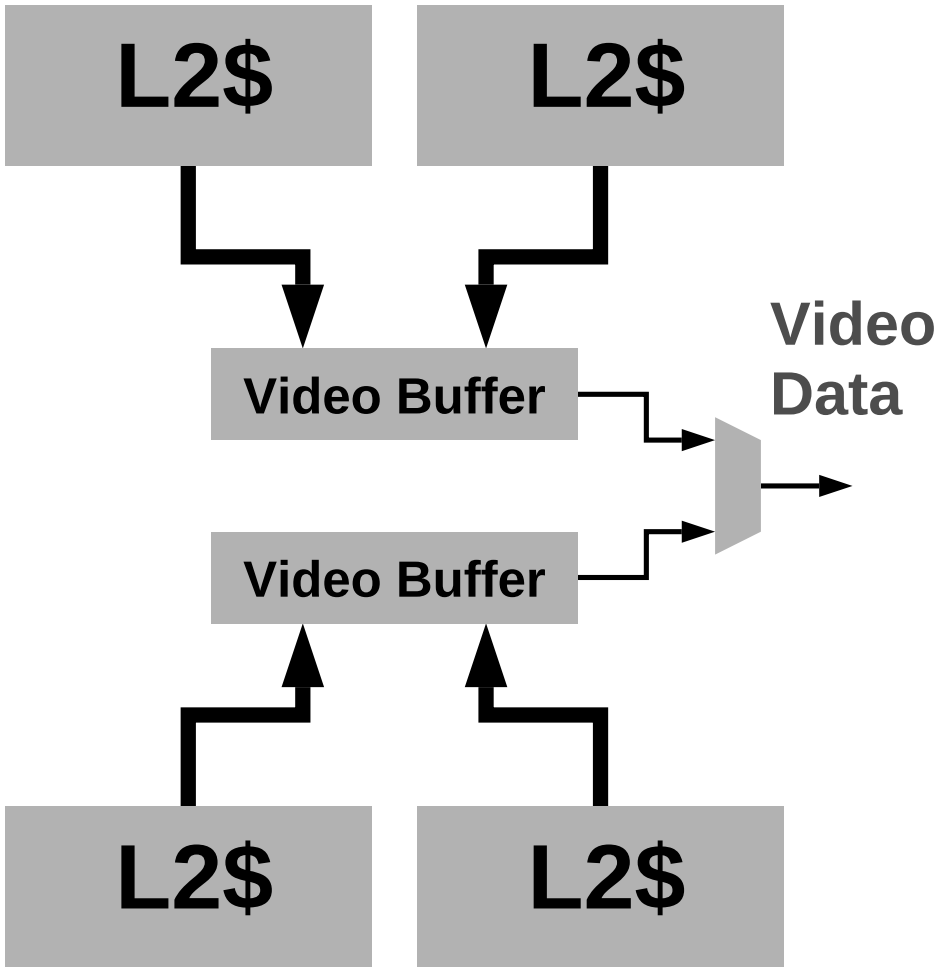
- 4 independent banks
- 256 pages per bank
- 10,240 bits per page
- 10 Mb per chip

Fast Fills

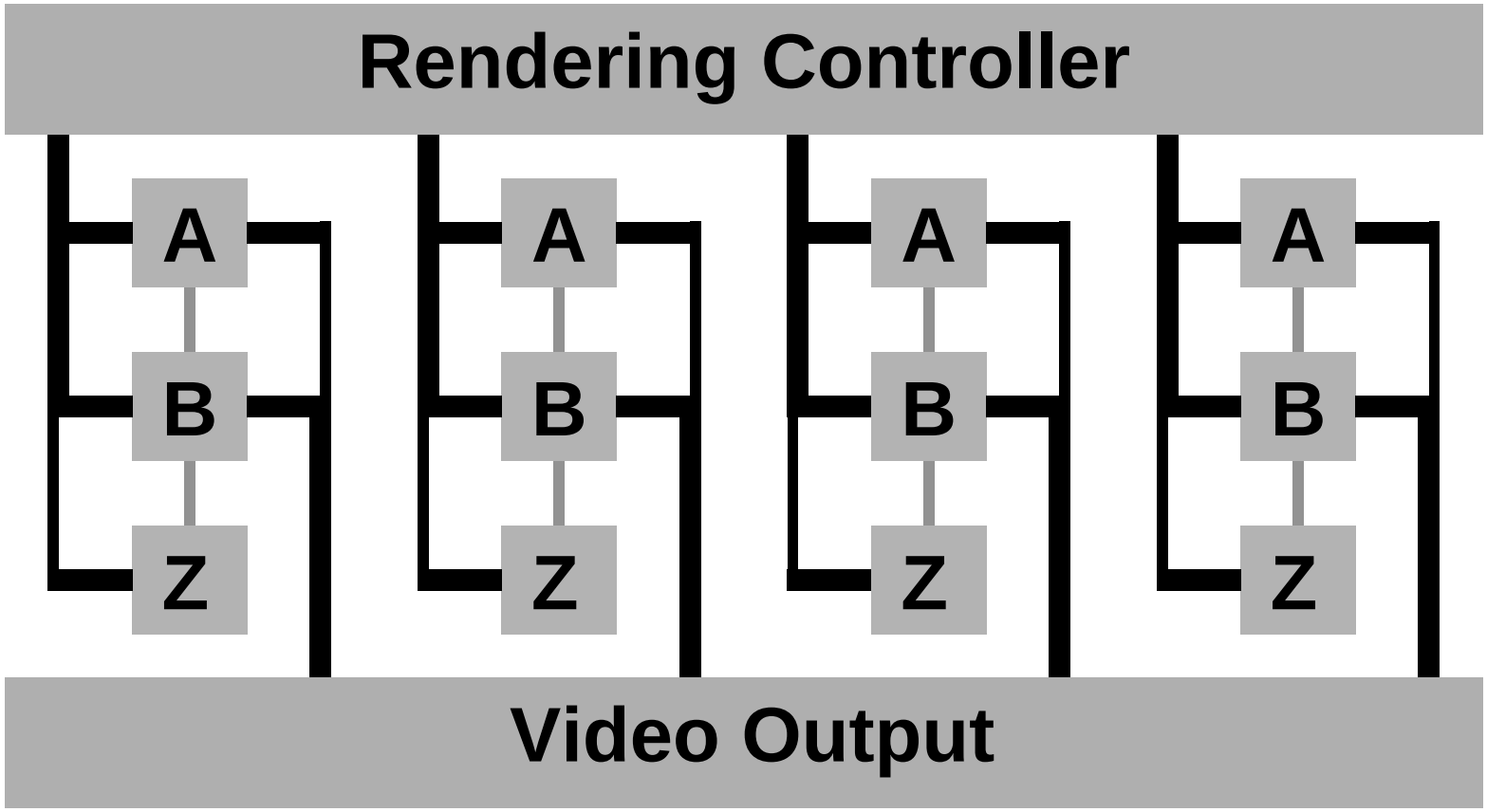
- **Must clear the screen before rendering each frame**
- **Fast constant color rectangle fill**
- **Worth accelerating in hardware**
- **Uses existing busses**

Video Port

- Frees random access port for rendering
- Two 20 pixel shift registers
  - Inexpensive; small die area
  - ping pong: seamless video
- Frequent reloads (~600 ns)



Typical Frame Buffer Organization



12 chip system; 1280x1024; double buffer + Z

Performance

Triangles	50 pixel	2-4M /sec
	100 pixel	1.5-1.8M /sec
Vectors	10 pixel	3.5-7.5M /sec
Fast Clear	Block Fill	1.2-1.6G pixel/sec
	Page Fill	24-32G pixel/sec
Image Copy	Write	264-400M pixel/sec
	Read	132-200M pixel/sec

12 chip system; 1280x1024; double buffer + Z

Silicon Summary

- **Mitsubishi 16M DRAM process**
- **Test silicon works**
- **Samples available later this year**
- **Area**
  - **~10% Video Output**
  - **~12% Pixel ALU**
  - **~8% L1$**
  - **The rest is DRAM and L2$**