A Performance Comparison of Contemporary DRAM Architectures

Vinodh Cuppu[1], Bruce Jacob[1], Brian Davis[2], and Trevor Mudge[2]

[1] ECE Dept., Univ. Maryland, College Park
[2] EECS Dept., Univ. Michigan, Ann Arbor

OUTLINE:

- **Motivation & Background**
- **Experiments**
- **Results**

UNIVERSITY OF MARYLAND

A COMPARISON OF DRAM ARCHITECTURES

Bruce Jacob

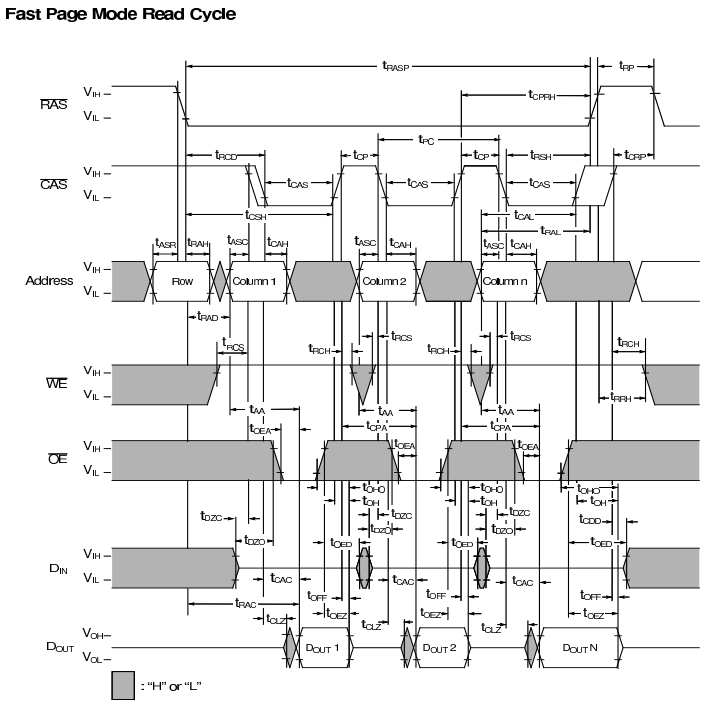
University of Maryland

Dilemma: THIS ...

STATUS QUO in MEMORY-SYSTEM RESEARCH:

```
...

if (memory_instruction(INSTR)) {
   if (L1_cache_miss( data_addr(INSTR) ){
      if (L2_cache_miss( data_addr(INSTR) ){

         cycles += DRAM_LATENCY;

      }
   }
}

...
```

... or THIS



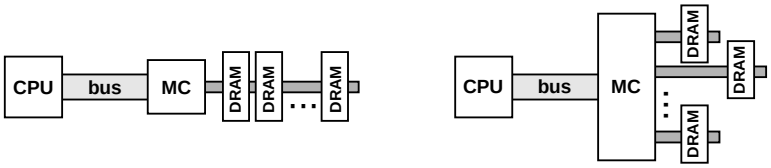
A COMPARISON OF DRAM ARCHITECTURES

Bruce Jacob

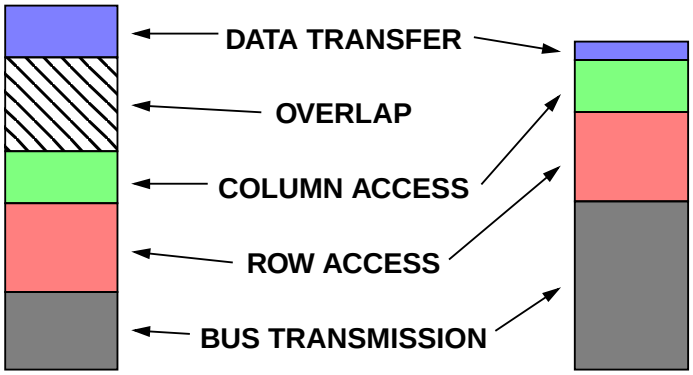
University of Maryland

Motivation

HERE'S WHAT YOU MISS:



DRAM LATENCY:



Goal

PRELIMINARY DRAM STUDY:

- **Bus Transmission**
- **Row Access**
- **Column Access**
- **Data Transfer**
- **Bus Wait/Synch Time**
- **Stalls Due to Refresh**
- **The OVERLAP of These Components (with each other) (with CPU execution)**

MODEL EXISTING TECHNOLOGY

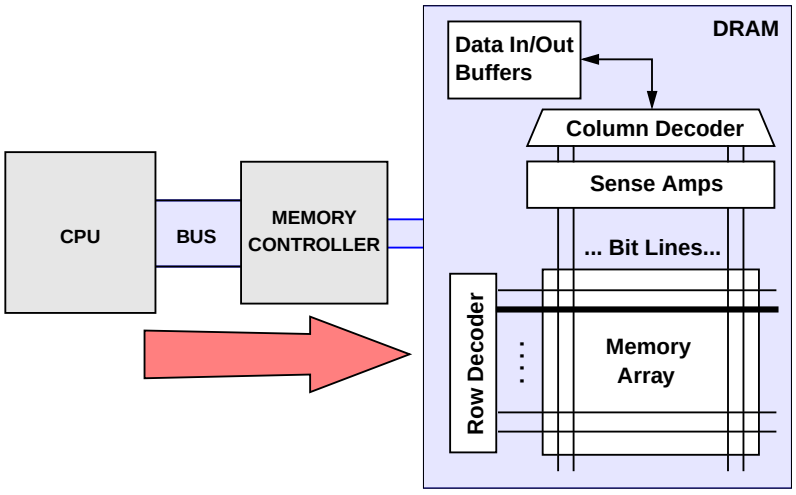
A COMPARISON OF DRAM ARCHITECTURES

Bruce Jacob

University of Maryland

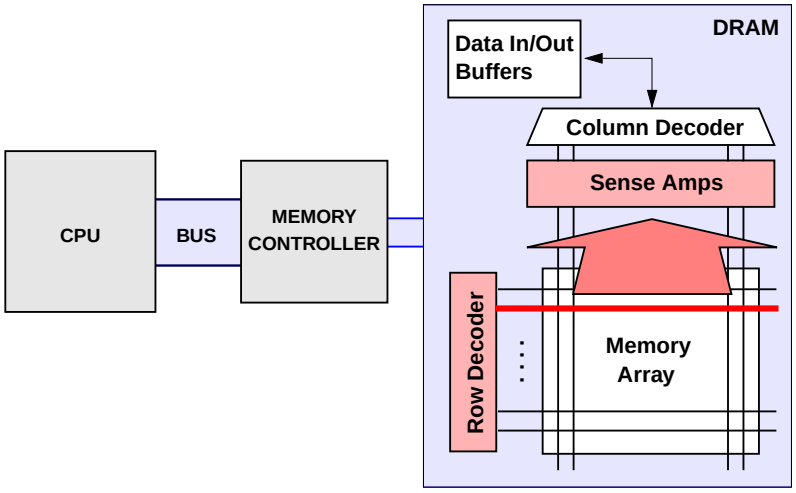
DRAM Primer

BUS TRANSMISSION



DRAM Primer

ROW ACCESS



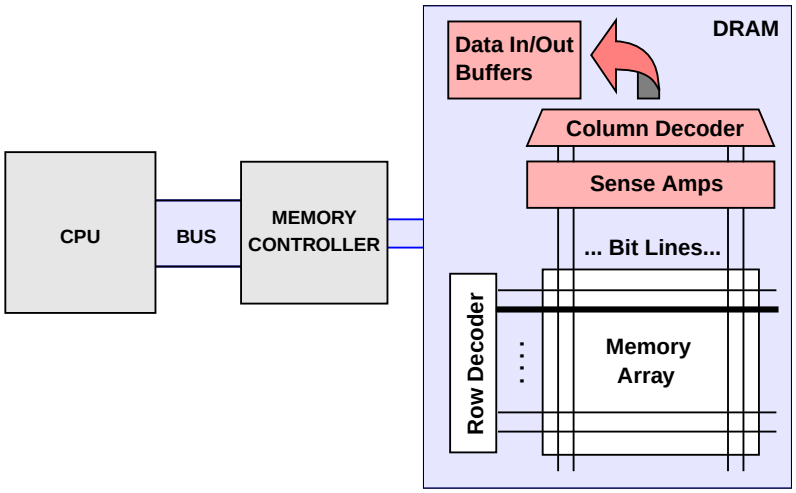
A COMPARISON OF DRAM ARCHITECTURES

Bruce Jacob

University of Maryland

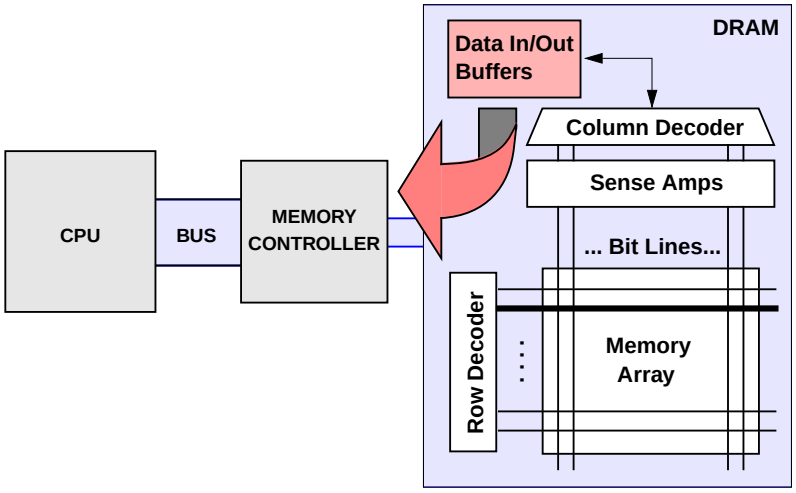
DRAM Primer

COLUMN ACCESS



DRAM Primer

DATA TRANSFER



note: page mode enables overlap with COL

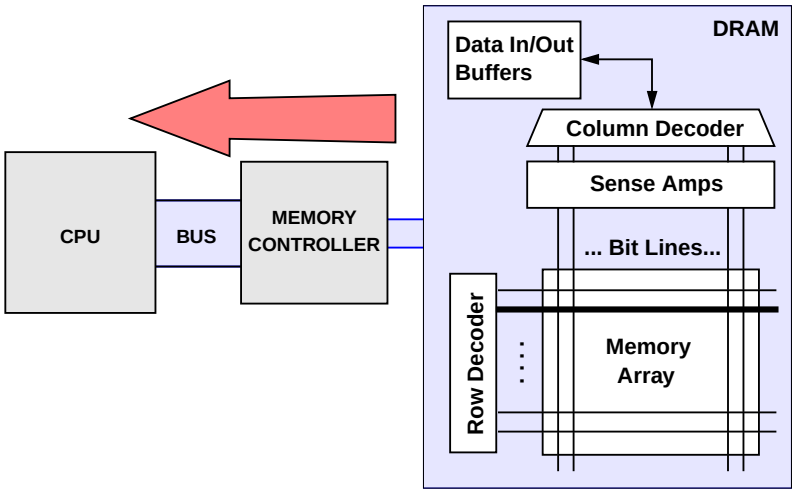
A COMPARISON OF DRAM ARCHITECTURES

Bruce Jacob

University of Maryland

DRAM Primer

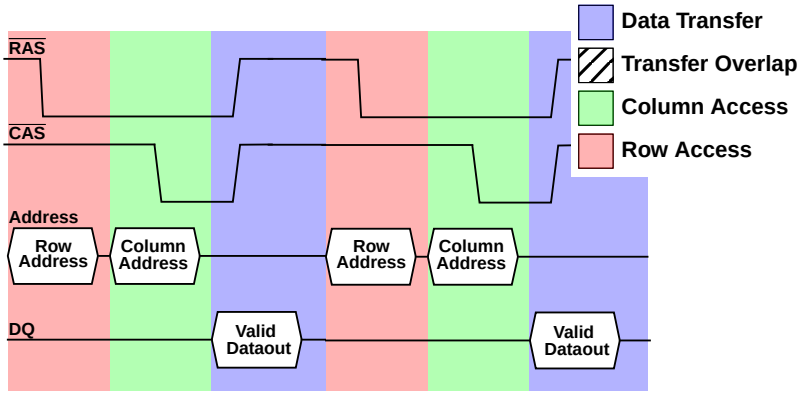
BUS TRANSMISSION



note: overlapped component not shown

DRAM Primer

Read Timing for Conventional DRAM



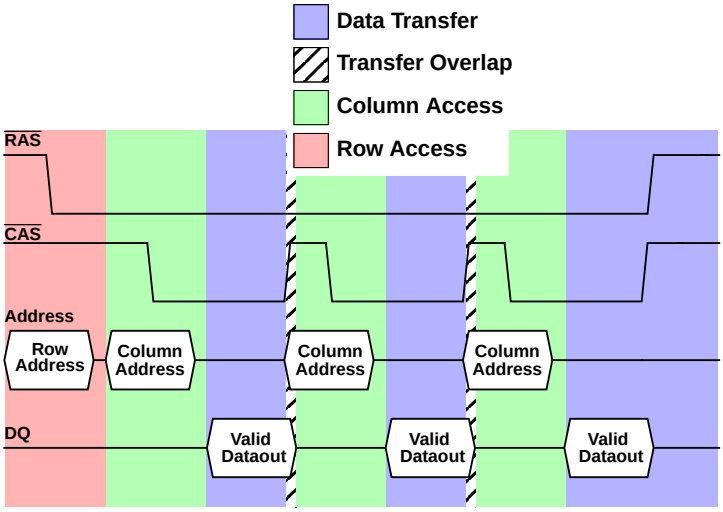
A COMPARISON OF DRAM ARCHITECTURES

Bruce Jacob

University of Maryland

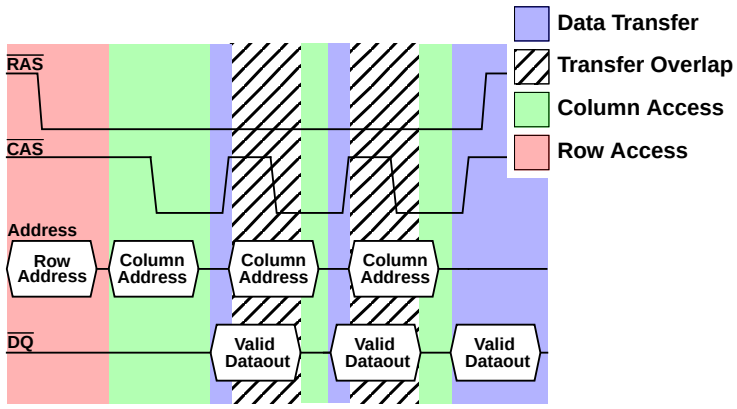
DRAM Primer

Read Timing for Fast Page Mode DRAM



DRAM Primer

Read Timing for Extended Data Out DRAM



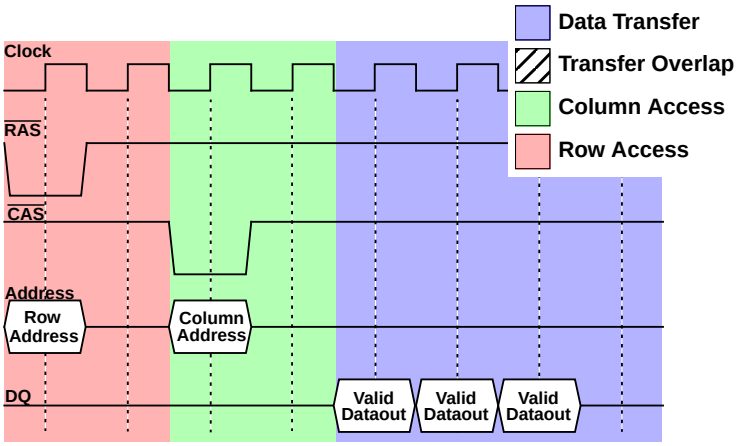
A COMPARISON OF DRAM ARCHITECTURES

Bruce Jacob

University of Maryland

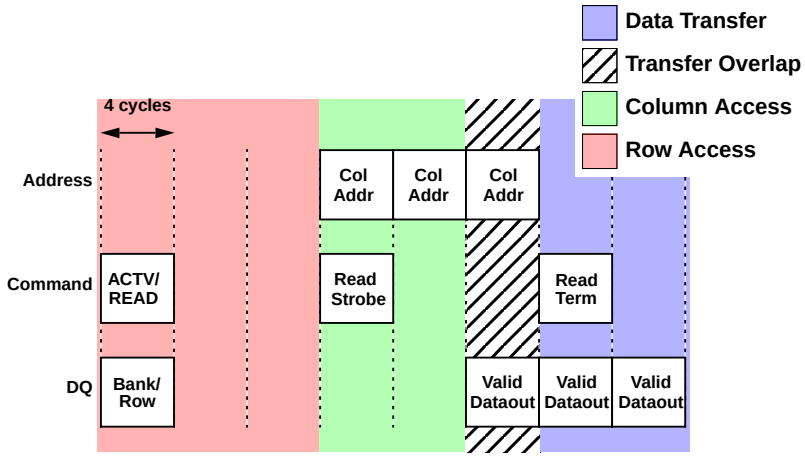
DRAM Primer

Read Timing for Synchronous DRAM



DRAM Primer

Read Timing for Rambus DRAM



A COMPARISON OF DRAM ARCHITECTURES

Bruce Jacob

University of Maryland

Simulator Overview

CPU: SimpleScalar v3.0a

- **8-way out-of-order**
- **L1 cache: split 64K/64K, lockup free x32**
- **L2 cache: unified 1MB, lockup free x1**
- **L2 blocksize: 128 bytes**

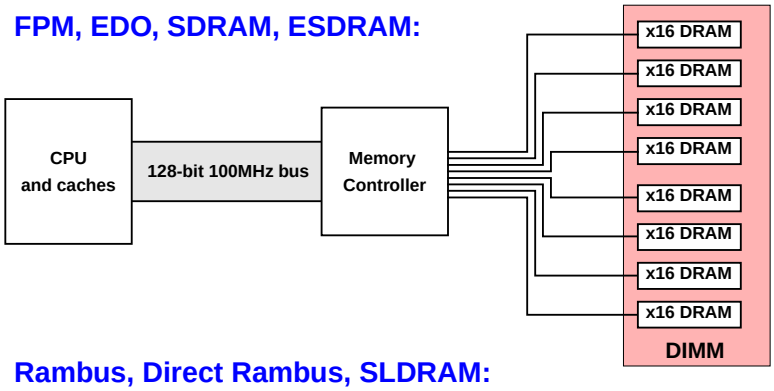
Main Memory: 8 64Mb DRAMs

- **100MHz/128-bit memory bus**
- **Optimistic *open-page* policy (*close-immediately* can be calculated)**

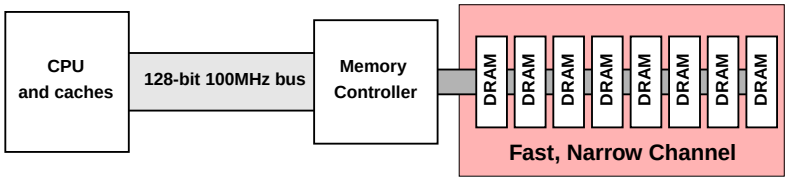
Represents a “typical” workstation

DRAM Configurations

FPM, EDO, SDRAM, ESDRAM:



Rambus, Direct Rambus, SLDRAM:



Note: TRANSFER WIDTH of Direct Rambus Channel

- **equals that of ganged FPM, EDO, etc.**
- **is 2x that of Rambus & SLDRAM**

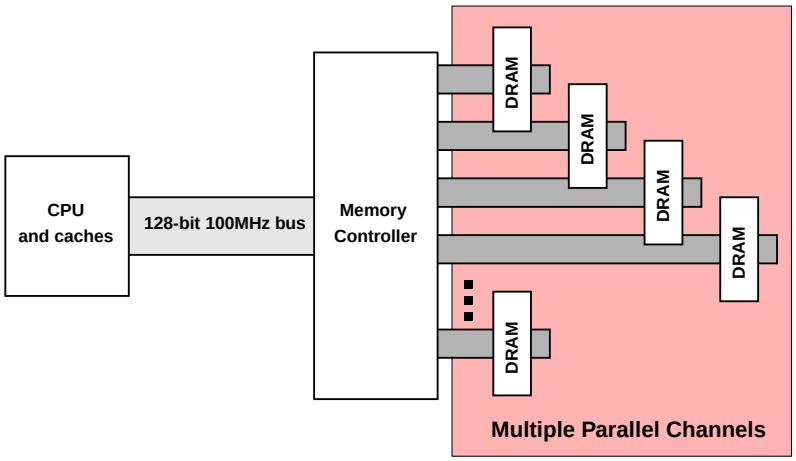
A COMPARISON OF DRAM ARCHITECTURES

Bruce Jacob

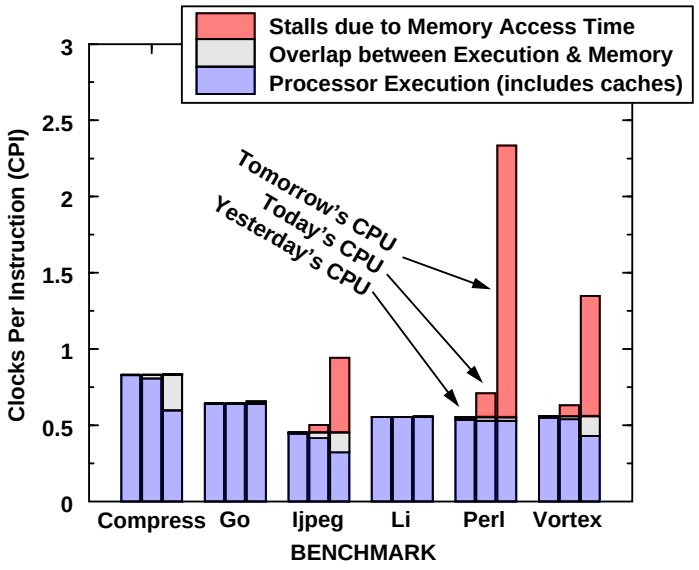
University of Maryland

DRAM Configurations

Strawman: Rambus, etc.



Overhead: Memory vs. CPU



Variable: speed of processor & caches

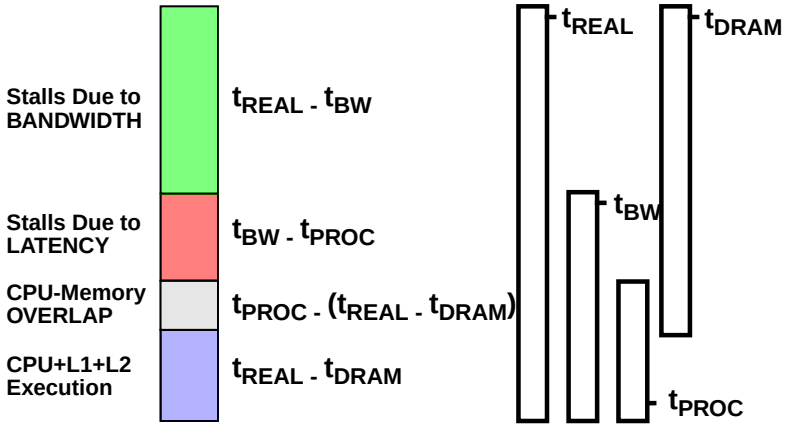
A COMPARISON OF DRAM ARCHITECTURES

Bruce Jacob

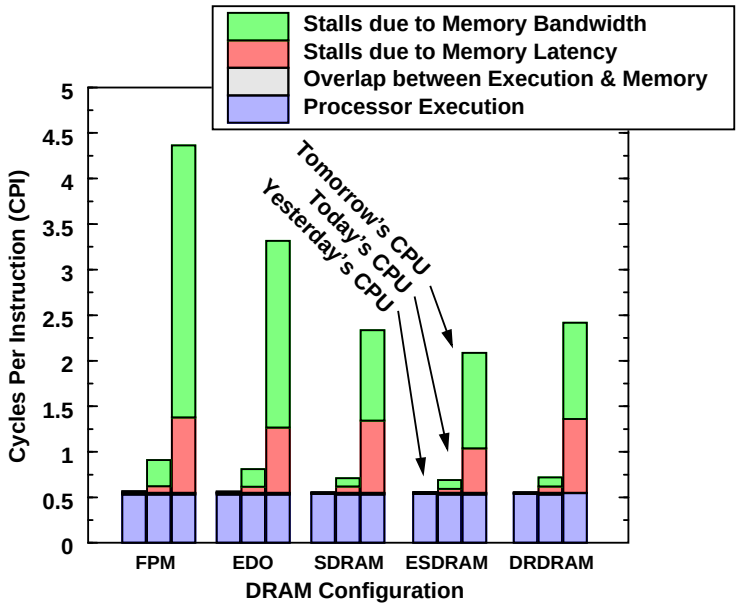
University of Maryland

Definitions (var. on Burger, et al)

- t_{PROC} — processor with perfect memory
- t_{REAL} — realistic configuration
- t_{BW} — CPU with wide memory paths
- t_{DRAM} — time seen by DRAM system



Memory & CPU — PERL

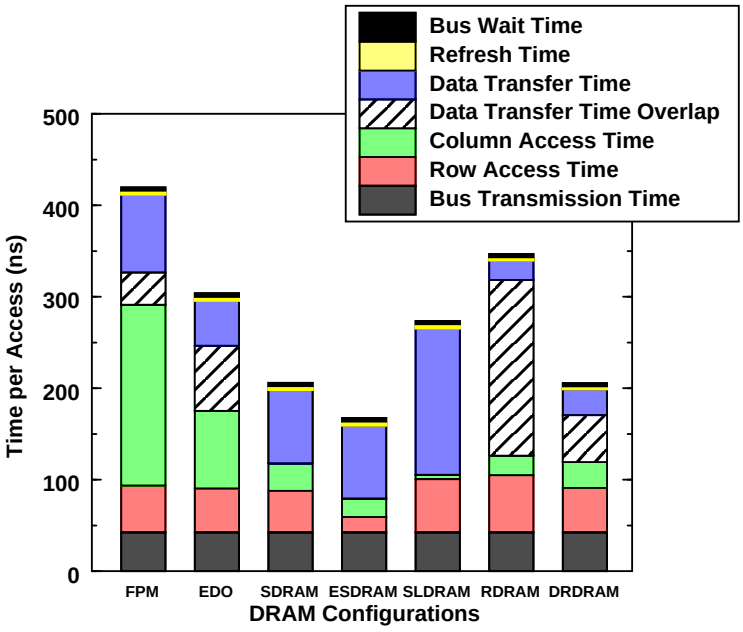


A COMPARISON OF DRAM ARCHITECTURES

Bruce Jacob

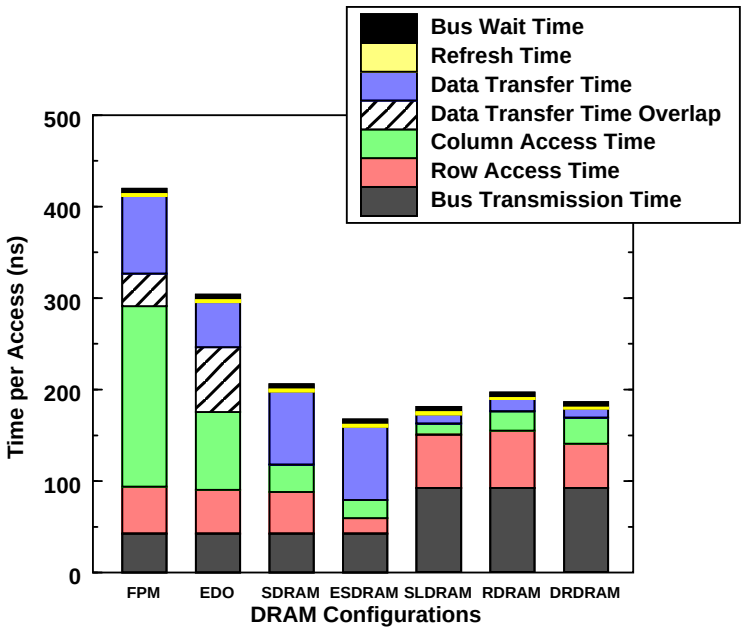
University of Maryland

Average Latency of DRAMs



note: SLDRAM & RDRAM 2x data transfers

Ganged Rambus Channels



A COMPARISON OF DRAM ARCHITECTURES

Bruce Jacob

University of Maryland

Cost-Performance

FPM, EDO, SDRAM, ESDRAM:

- **Lower Latency => Wide/Fast Bus**
- **Increase Capacity => Decrease Latency**
- **Low System Cost**

Rambus, Direct Rambus, SLDRAM:

- **Lower Latency => Multiple Channels**
- **Increase Capacity => Increase Capacity**
- **High System Cost**

1 DRDRAM = Multiple SDRAM

Conclusions

100MHz/128-bit Bus is Current Bottleneck

- **Solution: Fast Bus/es & MC on CPU (*e.g.* Compaq Alpha, Sony Emotion, ...)**

Current DRAMs Solving Bandwidth Problem (but not Latency Problem)

There is Locality in DRAM Accesses (but how important is this?)

SPECint '95 Fits in 1MB Cache

A COMPARISON OF DRAM ARCHITECTURES

Bruce Jacob

University of Maryland

Future Work

Improve Model:

- **DDR, DDR-II, MoSys, VCDRAM, etc.**
- **More realistic bus (scheduling, turnaround, etc.)**
- **Memory controller overhead**
- **Dual-bus latency vs. single-bus (include memory controller on CPU)**

Exploit DRAM Concurrency

Large Systems (bandwidth or latency?)

Small Systems: DSP + IRAM = D-IRAM