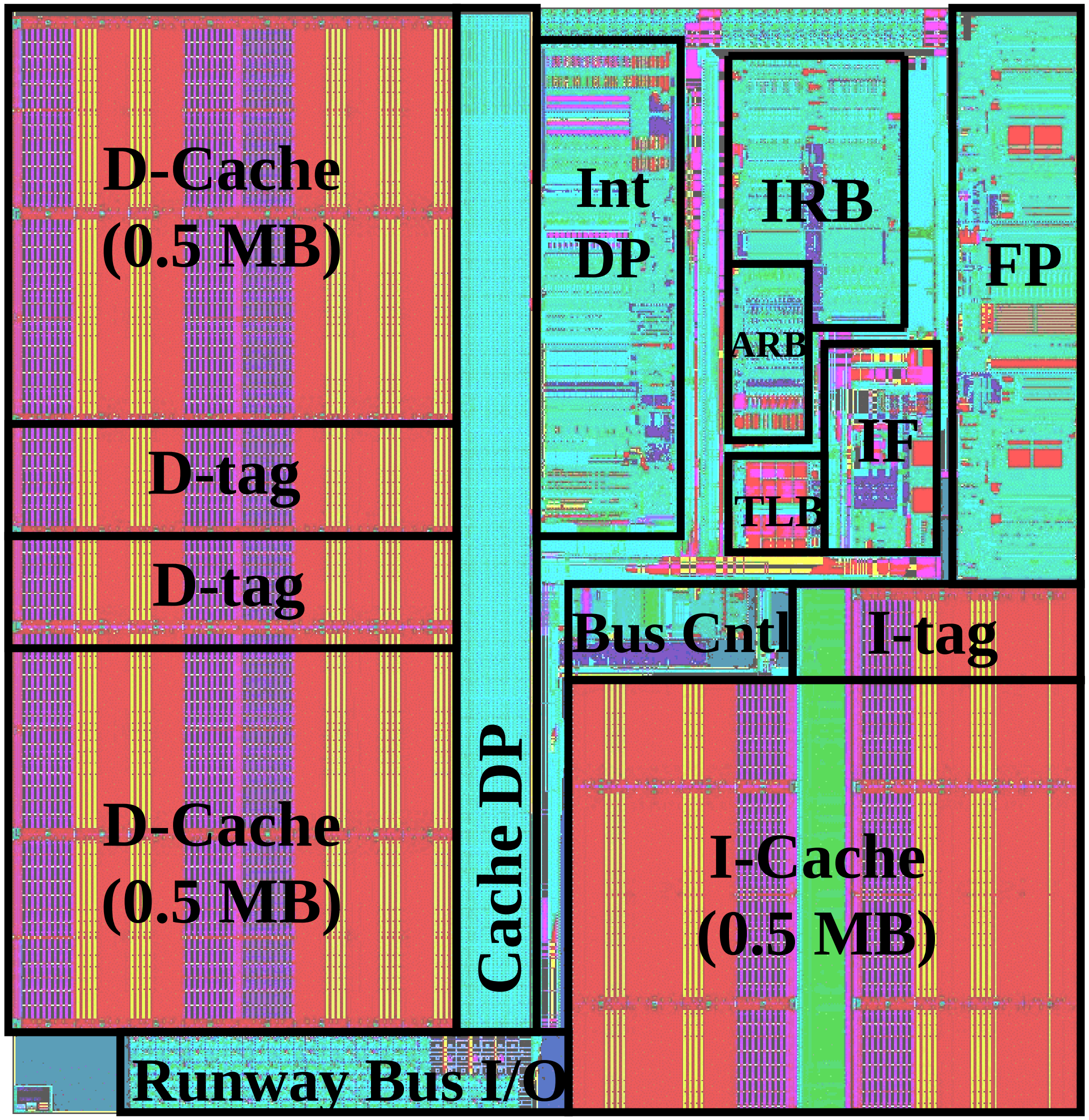
Techniques for Mitigating Memory Latency Effects in the PA-8500 Processor

David Johnson
Systems Technology Division
Hewlett-Packard Company

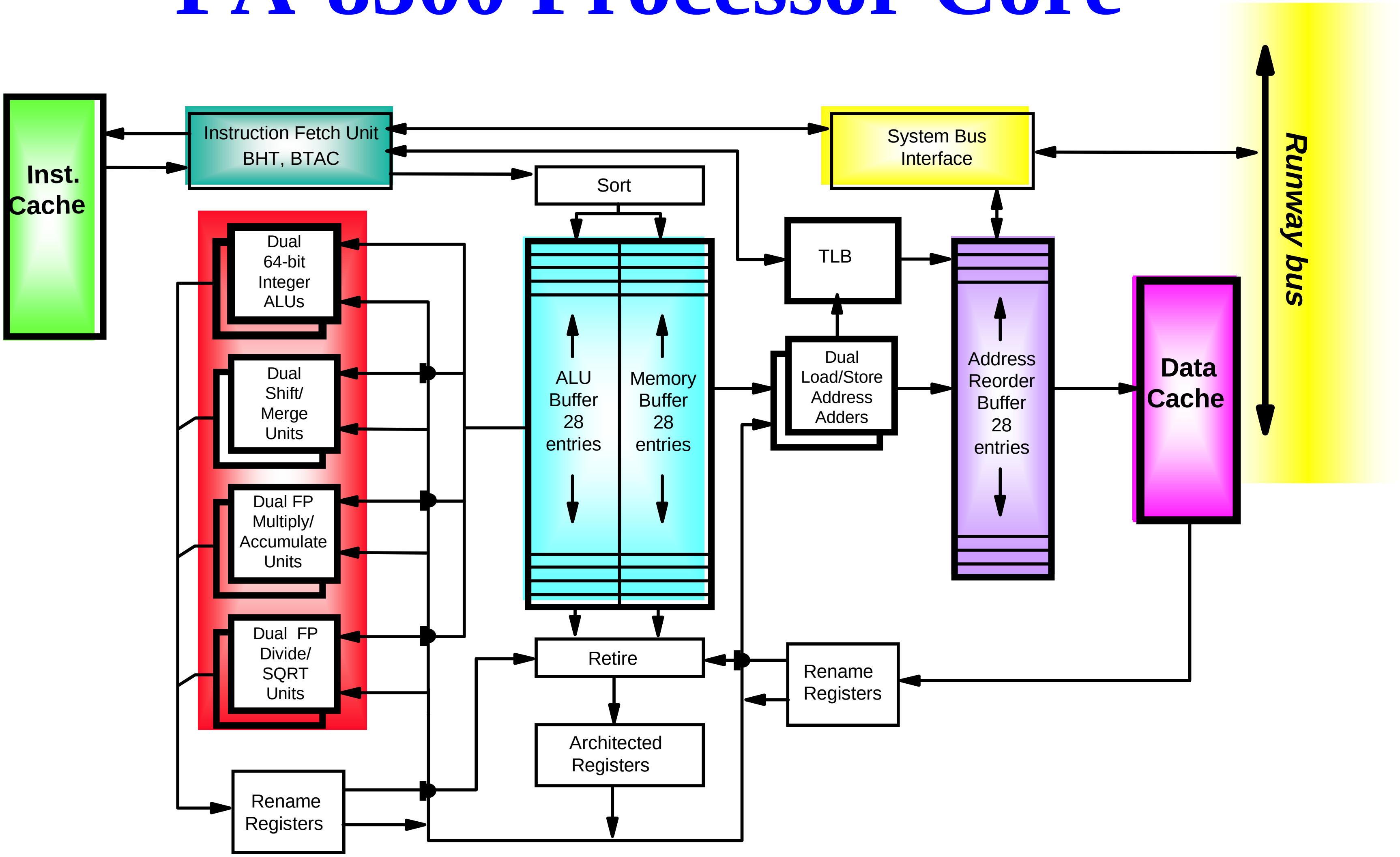
Presentation Overview

- PA-8500 Overview
- Instruction Fetch Capabilities
- Reorder Buffers ("The Queue")
- Data Cache
- System Bus

PA-8500

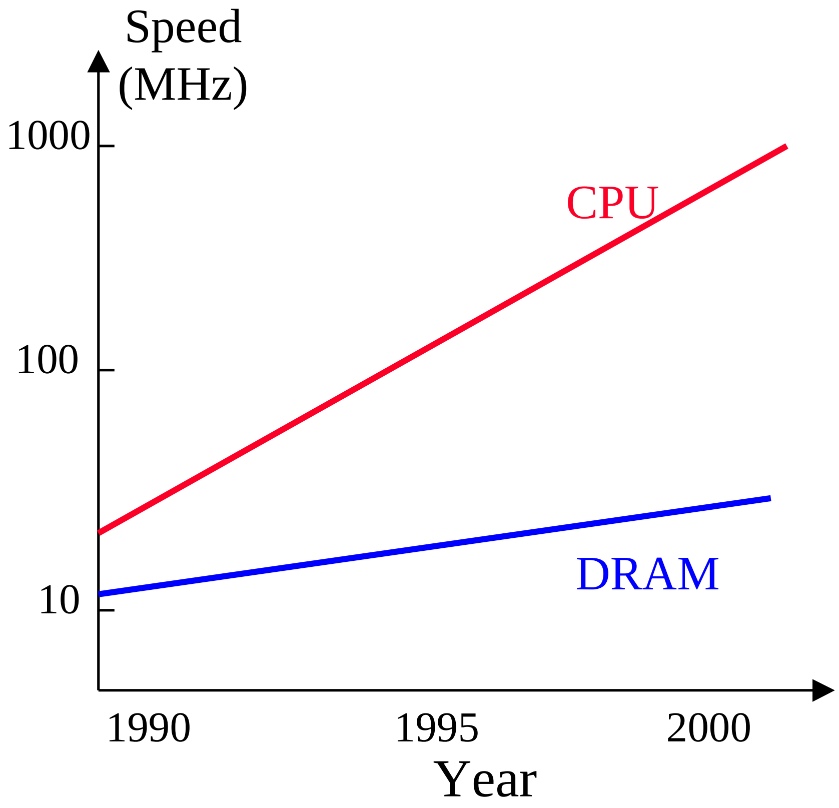

D-Cache
(0.5 MB)
D-tag
D-tag
D-Cache
(0.5 MB)
Cache DP
Int
DP
IRB
ARB
TLB
IF
FP
Bus Cntl
I-tag
I-Cache
(0.5 MB)
Runway Bus I/O

PA-8500 Processor Core



HEWLETT PACKARD

Memory Latency



Latency Problems

Instruction Fetches & Loads

Techniques for Hiding Latency

High hit-rate caches

Prefetching

Overlapping cache misses

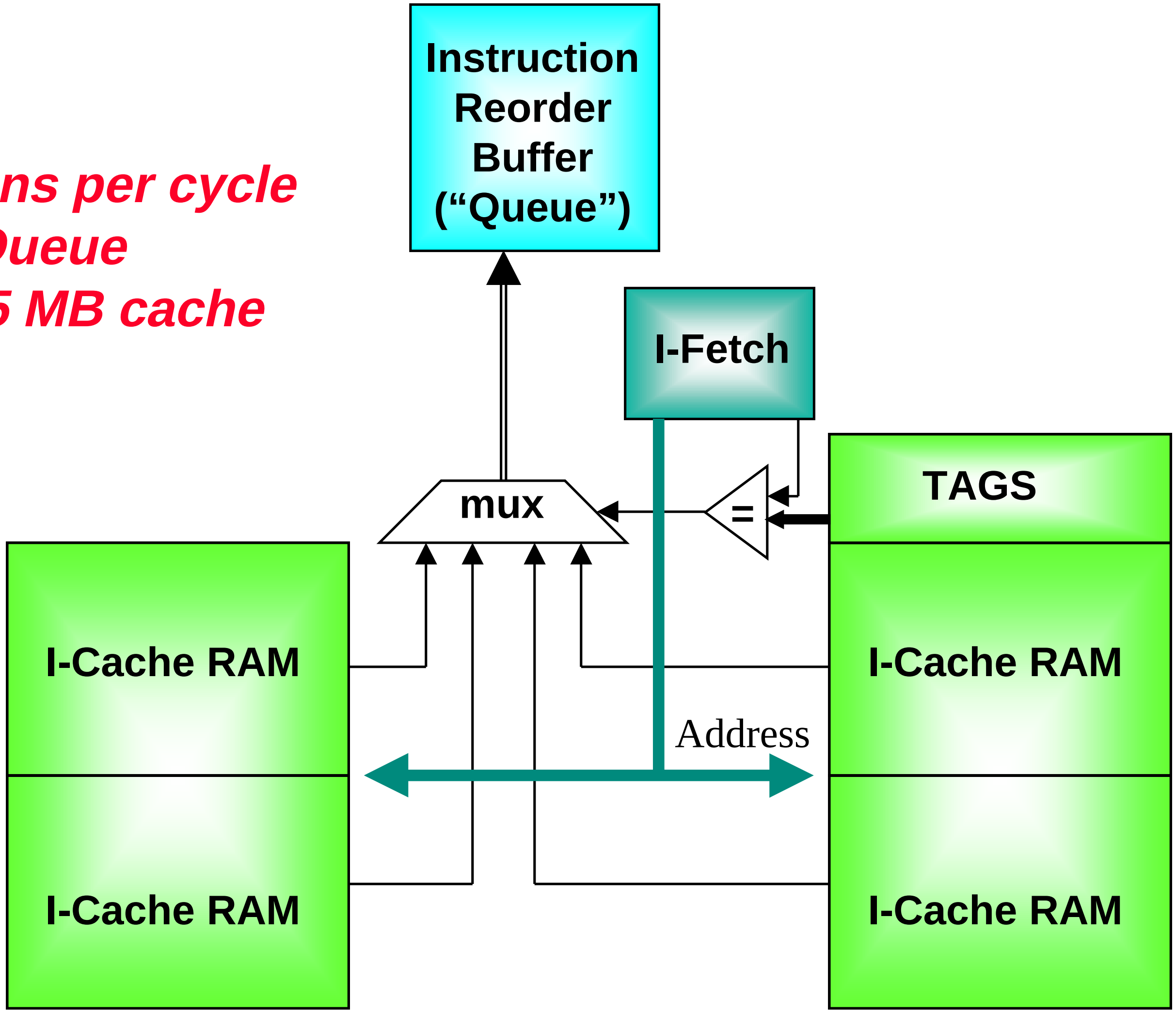
HEWLETT PACKARD

Instruction Fetch Features

- Instruction Cache
  - 0.5 MB on-chip cache
  - 4-way set associative
  - Pipelined 2-cycle access
  - Provides 4 instructions per cycle to CPU core
  - Supports 32-byte and 64-byte line sizes
- Instruction Prefetching

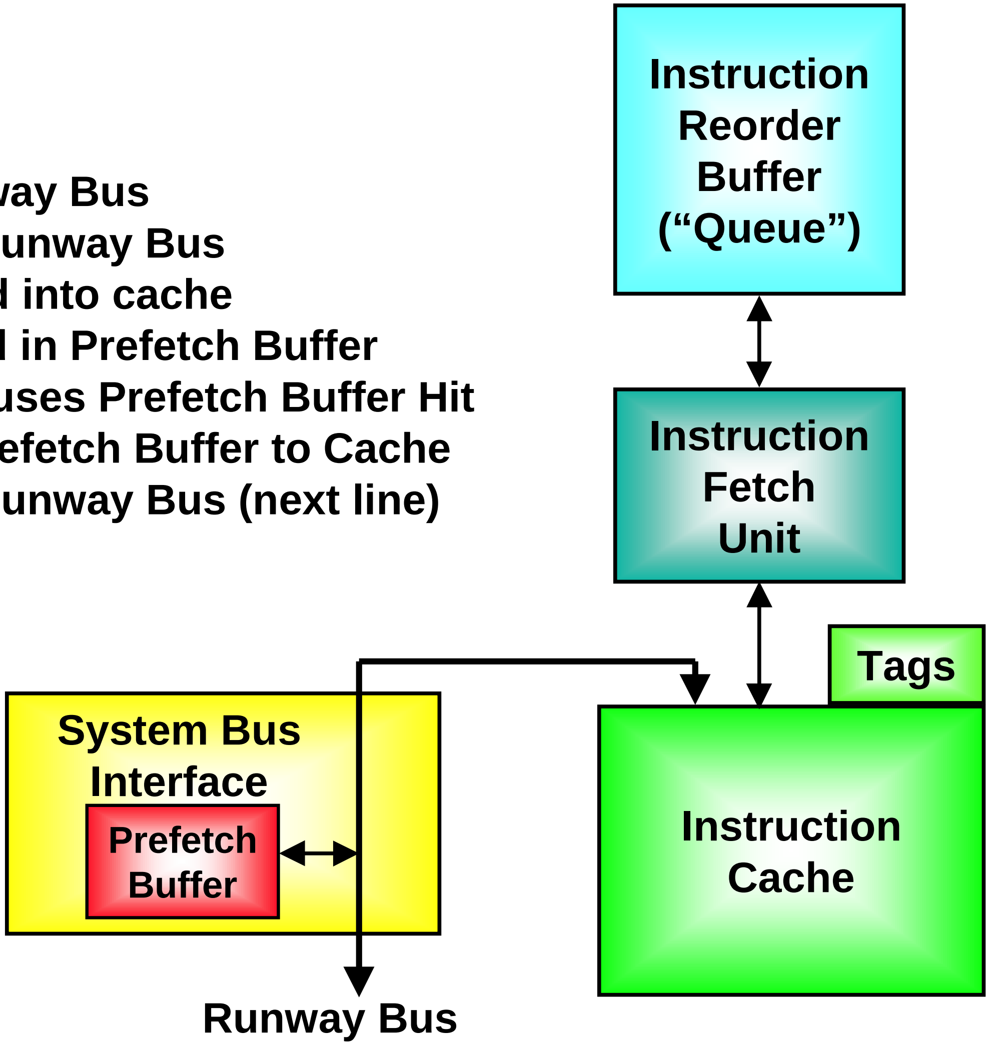
PA-8500 I-Cache Composition

4 Instructions per cycle to Queue from a 0.5 MB cache

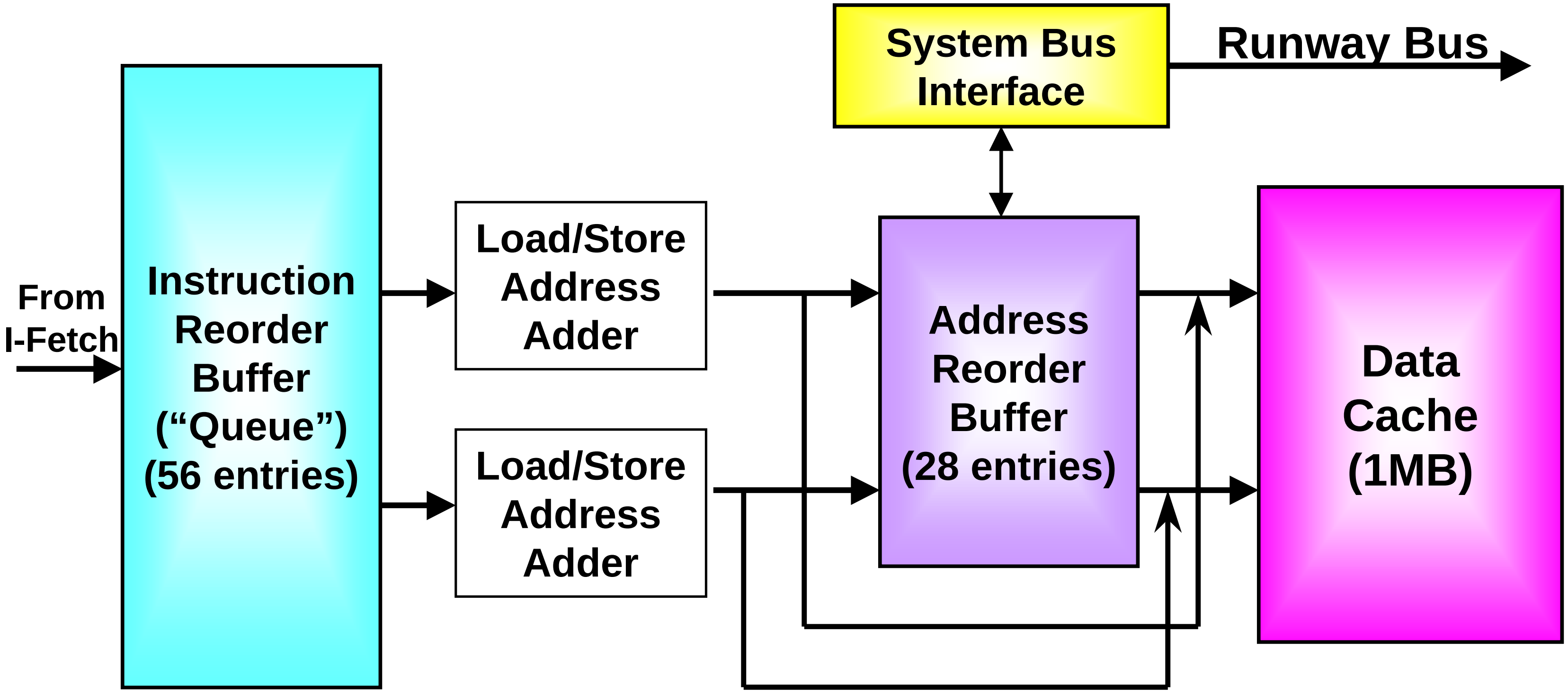


PA-8500 Instruction Prefetching

1. I-Miss from cache
2. I-Miss issued to Runway Bus
3. I-Prefetch issued to Runway Bus
4. I-Miss Return inserted into cache
5. I-Prefetch Return held in Prefetch Buffer
6. I-Miss from Cache causes Prefetch Buffer Hit
7. I-Miss moved from Prefetch Buffer to Cache
8. I-Prefetch issued to Runway Bus (next line)



Reorder Buffers

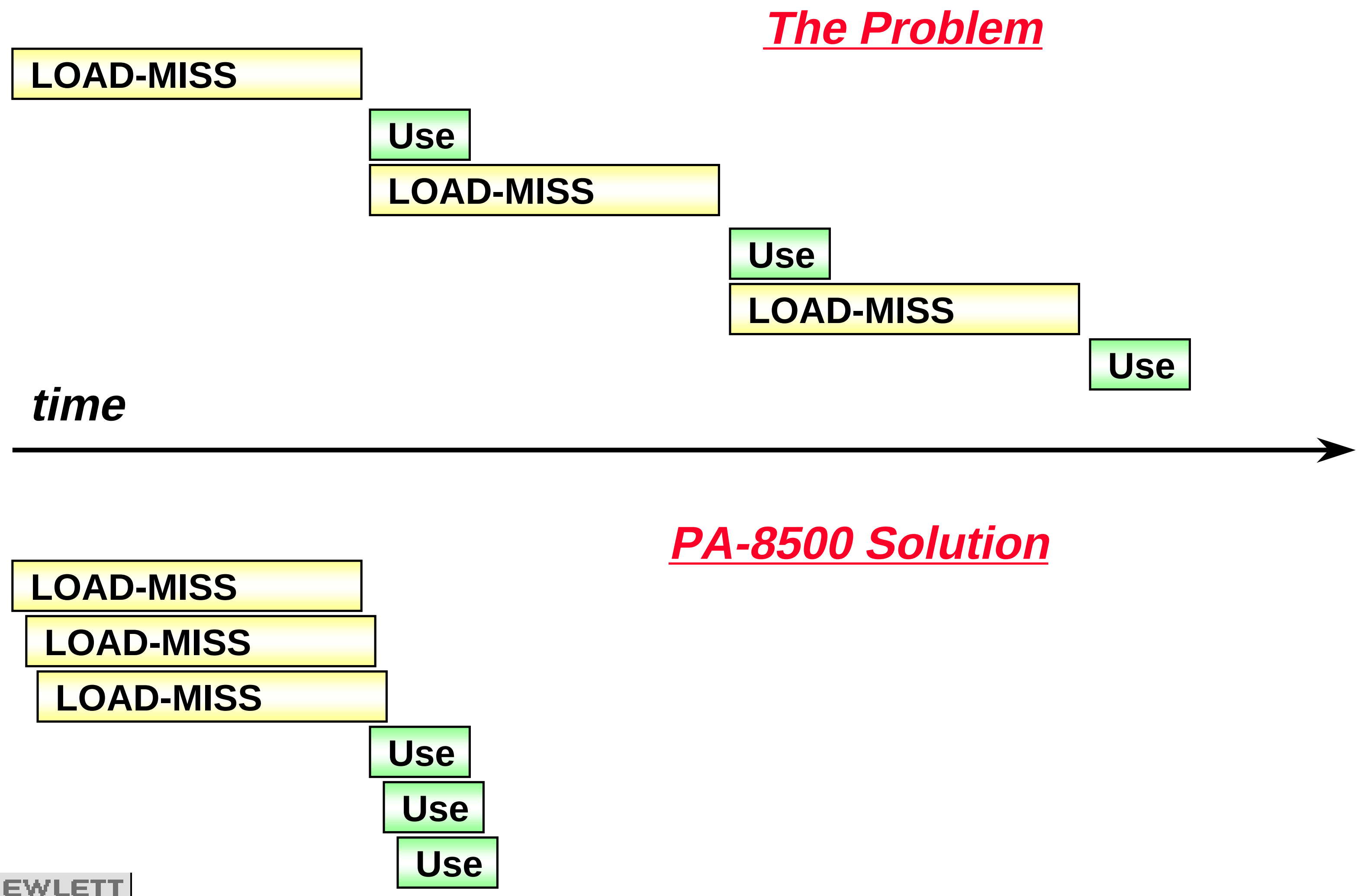


Cycle by cycle progression of a load instruction

Insert	Launch	Address	Cache	Cache	RR	Retire

HEWLETT PACKARD

LOAD-MISS Overlapping



The Problem

LOAD-MISS

Use

LOAD-MISS

Use

LOAD-MISS

Use

time

PA-8500 Solution

LOAD-MISS

LOAD-MISS

LOAD-MISS

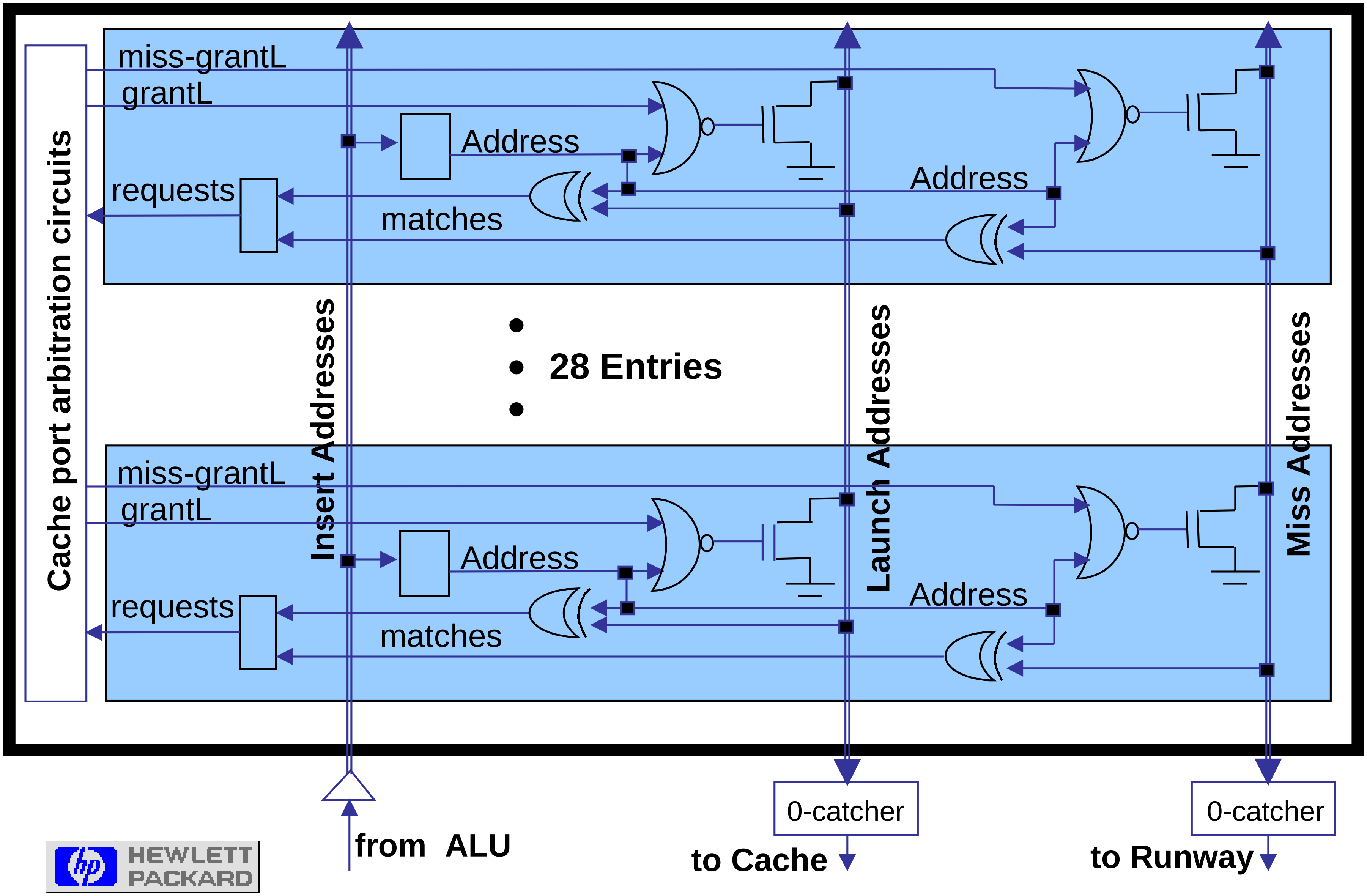
Use

Use

Use

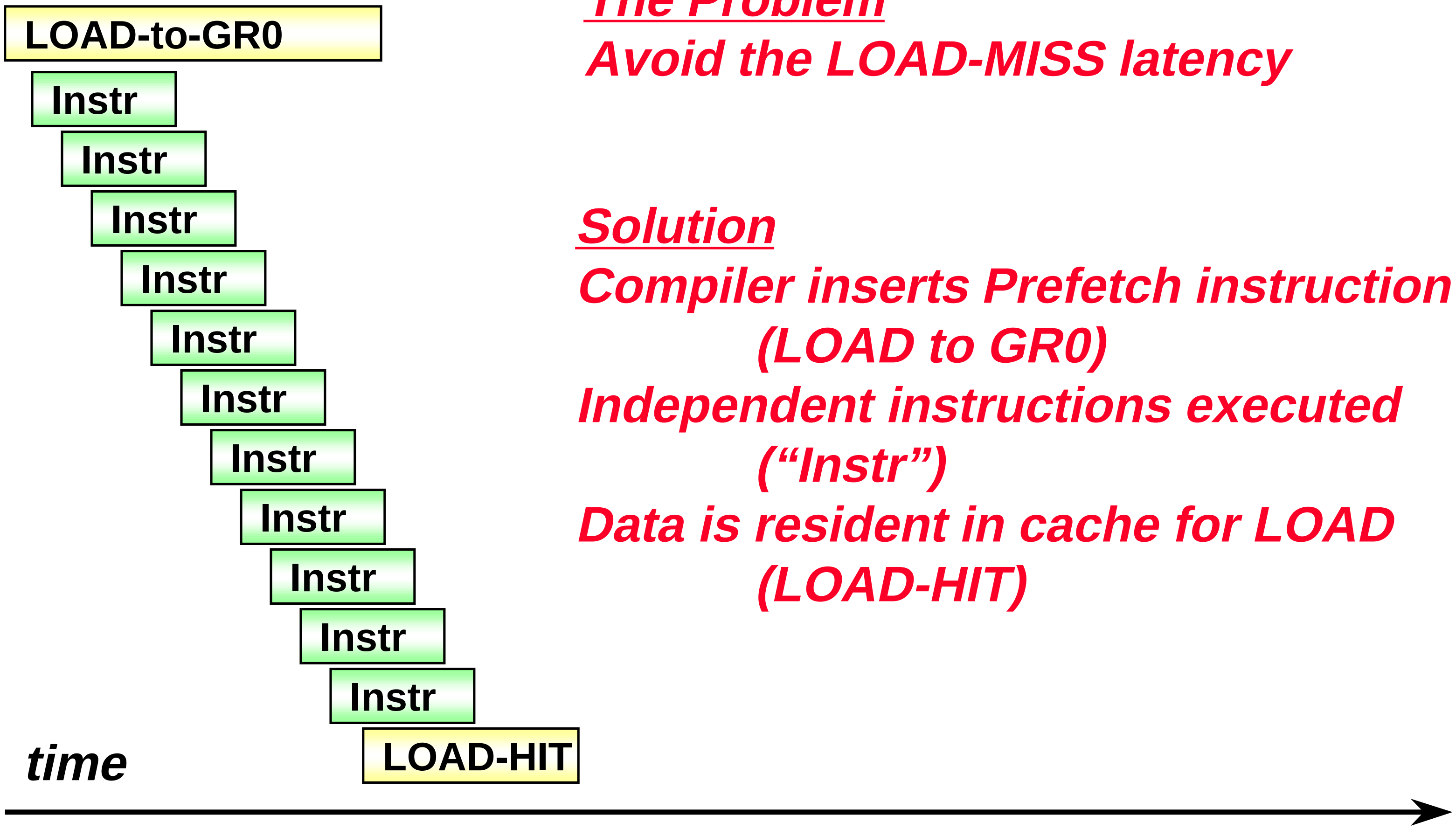
HEWLETT
PACKARD

Address Reorder Buffer: High-Speed Custom Circuitry



HEWLETT PACKARD

Data Prefetching



LOAD-to-GR0

Instr

Instr

Instr

Instr

Instr

Instr

Instr

Instr

Instr

Instr

Instr

LOAD-HIT

time

The Problem

Avoid the LOAD-MISS latency

Solution

Compiler inserts Prefetch instruction (LOAD to GR0)

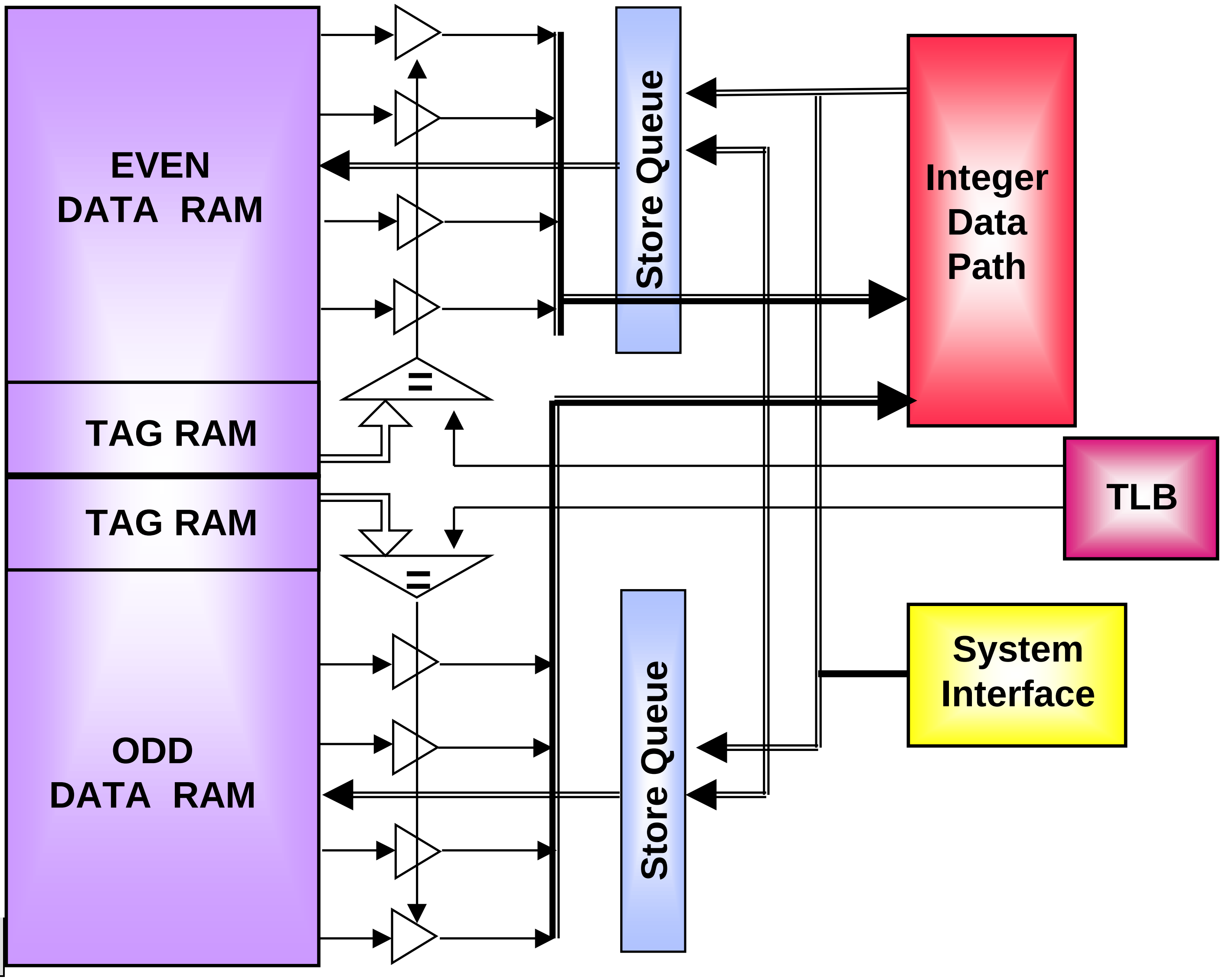
Independent instructions executed ("Instr")

Data is resident in cache for LOAD (LOAD-HIT)

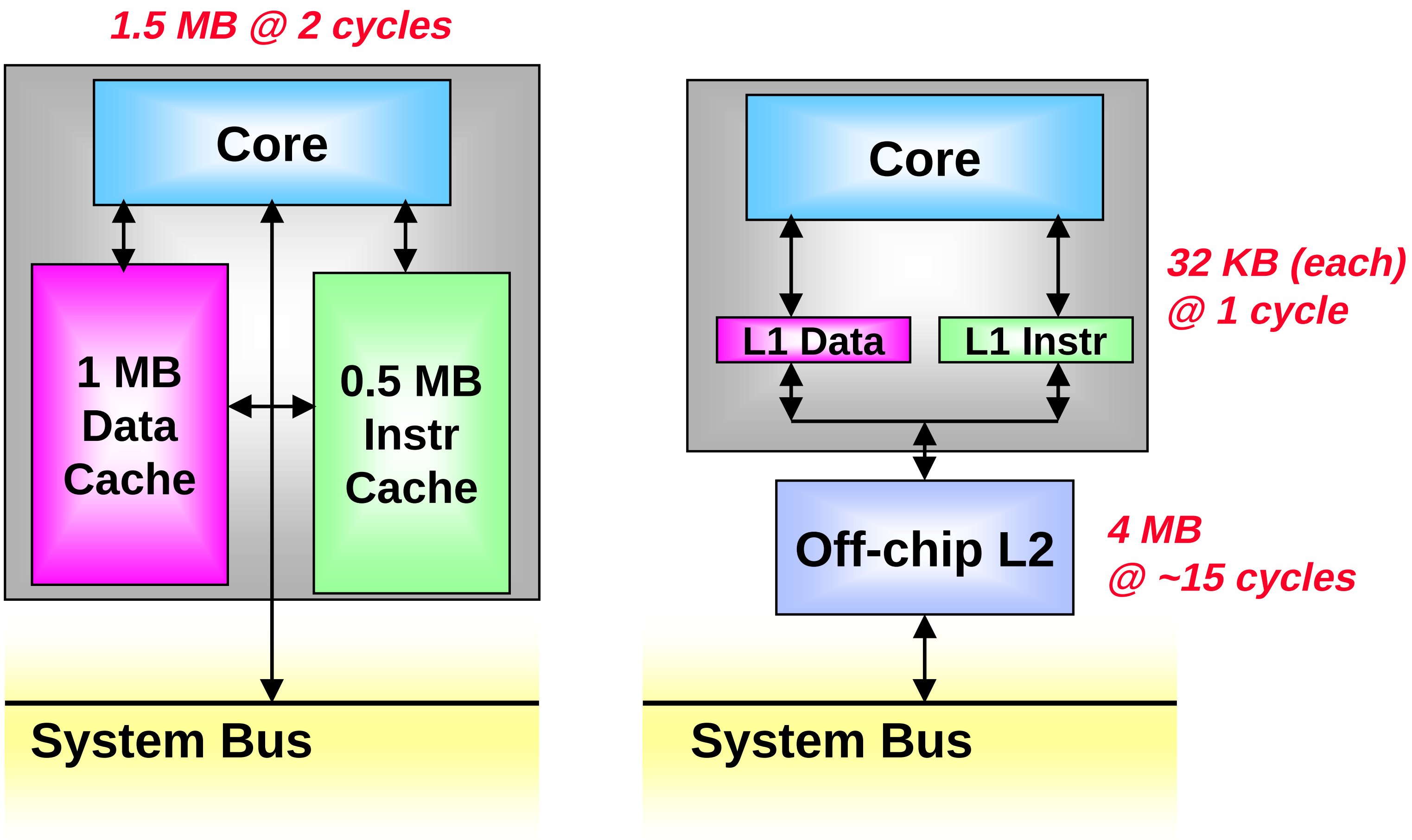
Data Cache Features

- 1.0 MB on-chip cache
- 4-way set associative
- 2-cycle pipelined access
- Two accesses per cycle
- Supports 32-byte and 64-byte line sizes
- Sophisticated Store Queue

Data Cache


EVEN
DATA RAM
TAG RAM
TAG RAM
ODD
DATA RAM
=
=
Store Queue
Store Queue
Integer
Data
Path
TLB
System
Interface

Single-Level vs. Multi-Level Cache Designs



1.5 MB @ 2 cycles

Core

1 MB Data Cache

0.5 MB Instr Cache

System Bus

Core

L1 Data

L1 Instr

32 KB (each) @ 1 cycle

Off-chip L2

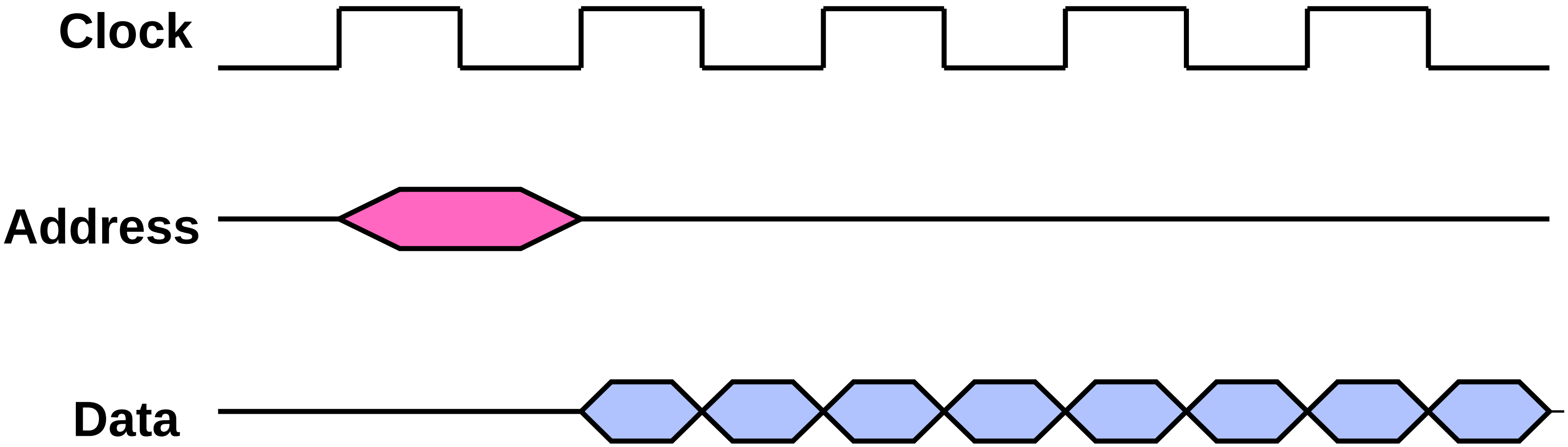
4 MB @ ~15 cycles

System Bus

System Bus Interface

- Split-transaction bus with out-of-order returns
- Multiple transactions in flight simultaneously
- Priority given to latency-sensitive transactions
- Asynchronous Interface
- Turbo Mode

Turbo Mode



High-Speed Data Transfer between Memory and CPU

Mitigating Memory Latency Effects

- Large Caches
- Out-of-Order Queue
- Flexible System Interface
- Custom Circuit Design

The PA-8500 Achieves Superb Performance !

HEWLETT PACKARD