February 4, 2025

Announcements

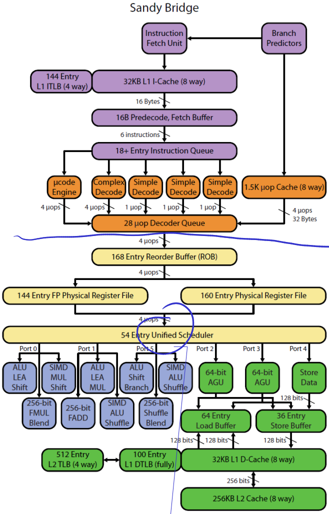
- HW1
- Final talk topic / time assignment: Thu
- First talks: Feb 20.

Goals

- Review
- Memory subsystem

Review

A Glimpse of a More Modern Processor



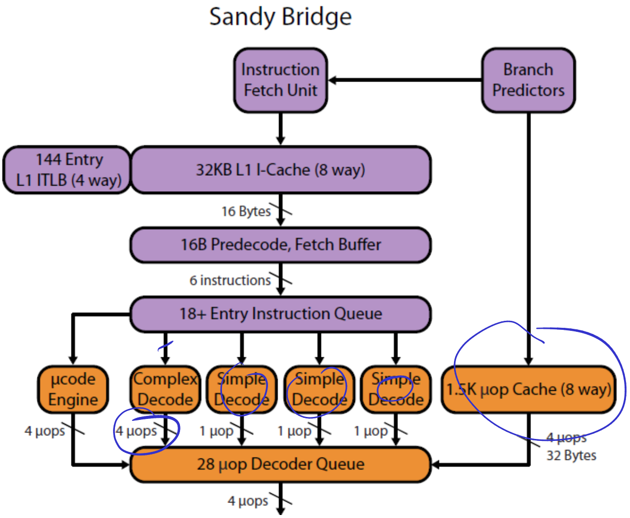
```
mov   $5, %eax
imul  $17
mov   %eax, ...ma

mov   $6, %eax
imul  $17
mov   %eax, ...ma

mov   %ecx, %eax
```

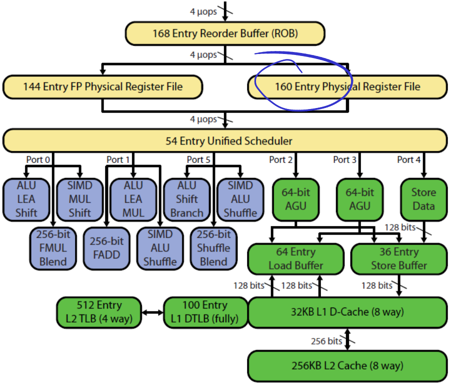
[David Kanter / Realworldtech.com]

A Glimpse of a More Modern Processor: Frontend



[David Kanter / Realworldtech.com]

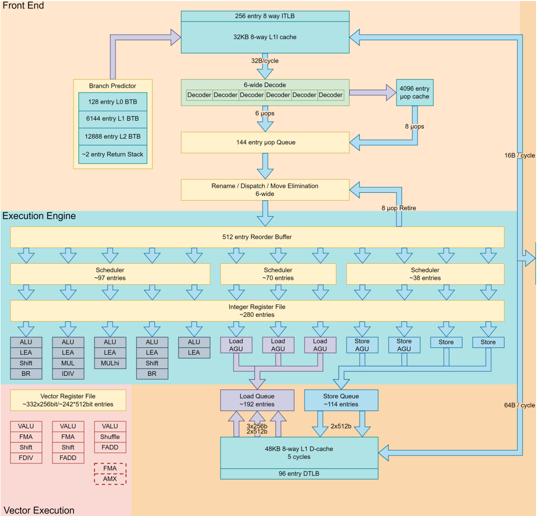
A Glimpse of a More Modern Processor: Backend



- New concept: Instruction-level parallelism ("ILP", "superscalar")
- Where does the IPC number from earlier come from?

[David Kanter / Realworldtech.com]

A Glimpse of a More Modern Processor: Golden Cove

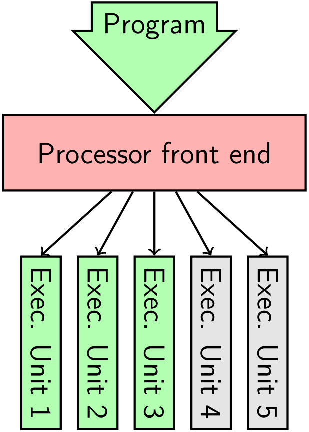


[Wikipedia 🅭]

Demo

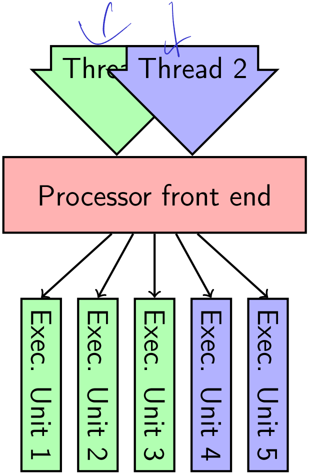
Demo: intro/More Pipeline Mysteries

SMT/“Hyperthreading”



Q: Potential issues?

SMT/"Hyperthreading"



Q: Potential issues?

- Contention (not just cache?)

- "for losers"

- Power?
  Efficiency?

Outline

Introduction

- Notes
- Notes (unfilled, with empty boxes)
- Notes (source code on Github)
- About This Class
- Why Bother with Parallel Computers?
- Lowest Accessible Abstraction: Assembly
- Architecture of an Execution Pipeline
- Architecture of a Memory System
- Shared-Memory Multiprocessors

Machine Abstractions

Performance: Expectation, Experiment, Observation

Performance-Oriented Languages and Abstractions

More Bad News from Dennard

Parameter	Factor
Dimension	$1/\kappa$
Line Resistance	κ
Voltage drop	κ
Response time	1
Current density	κ

[Dennard et al. '74, via Bohr '07]

- The above scaling law is for on-chip interconnects.
- Current ∼ Power *vs.* response time

Getting information from

- processor to memory
- one computer to the next

is

- slow (in *latency*)
- power-hungry

Somewhere Behind the Interconnect: Memory

Performance characteristics of memory:

- Bandwidth
- Latency

Flops are cheap
Bandwidth is money
Latency is physics — M. Hoemmen

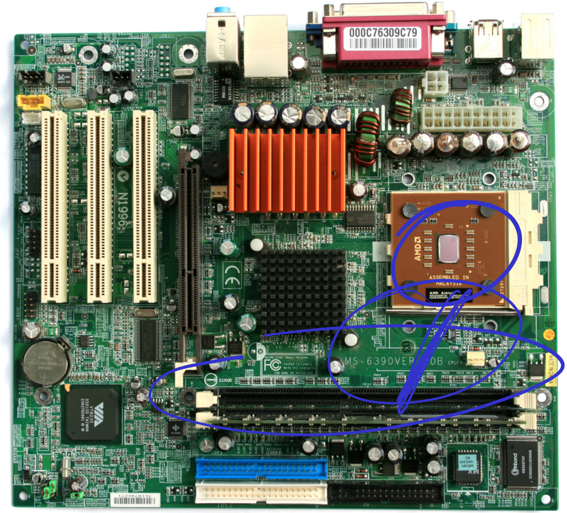
Minor addition (but important for us)?

SRAM

DRAM

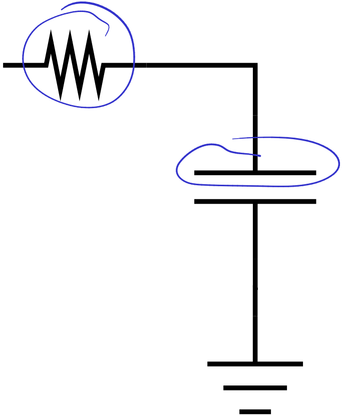
Bandwidth is money and code structure.

Latency is Physics: Distance

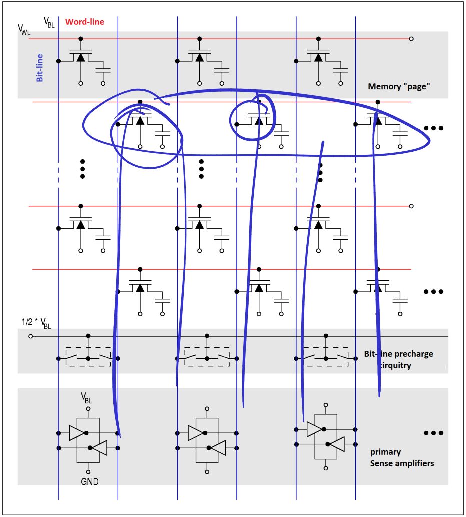


[Wikipedia 🅭]

Latency is Physics: Electrical Model



Latency is Physics: DRAM



[Wikipedia ⓒ]

Latency is Physics: Performance Impact?

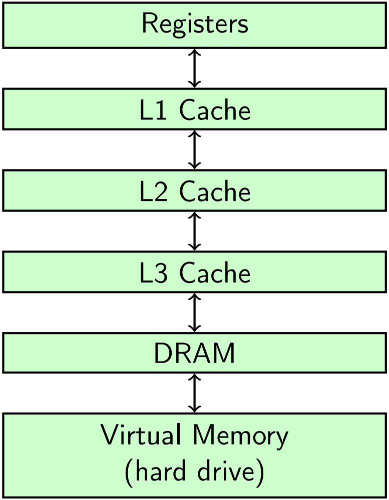
What is the performance impact of high memory latency?

Stall.

Idea:

- Put a look-up table of recently-used data onto the chip.
- Cache

Memory Hierarchy

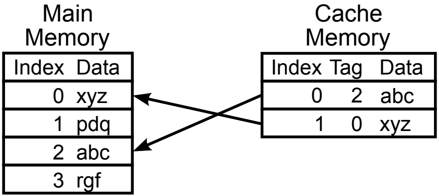


Registers	1 kB, 1 cycle
L1 Cache	10 kB, 10 cycles
L2 Cache	100 kB, 10 cycles
L3 Cache	10 MB, 100 cycles
DRAM	1 GB, 1000 cycles
Virtual Memory (hard drive)	1 TB, 1 M cycles

A Basic Cache

Demands on cache implementation:

- Fast, small, cheap, low power
- Fine-grained
- High “hit”-rate (few “misses”)



Main Memory

Index	Data
0	xyz
1	pdq
2	abc
3	rgf

Cache Memory

Index	Tag	Data
0	2	abc
1	0	xyz

Design Goals: at odds with each other. Why?

lookup logic is not free

[Wikipedia 🅭]

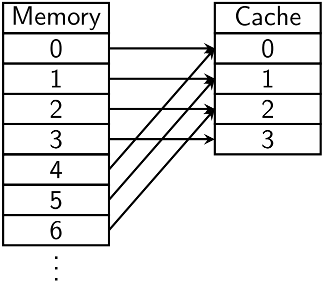
Caches: Engineering Trade-Offs

Engineering Decisions:

- More data per unit of access matching logic
  → Larger "Cache Lines"
- Simpler/less access matching logic
  → Less than full "Associativity"
- Eviction strategy
- Size

Associativity

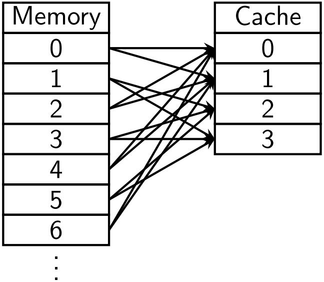
Direct Mapped:



Memory
0
1
2
3
4
5
6
⋮

Cache
0
1
2
3

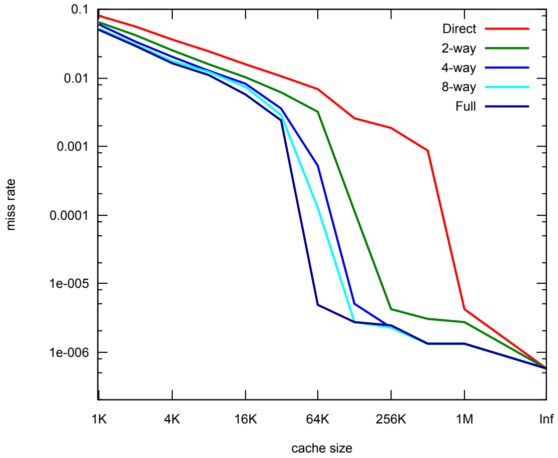
2-way set associative:



Memory
0
1
2
3
4
5
6
⋮

Cache
0
1
2
3

Size/Associativity vs Hit Rate



Miss rate versus cache size on the Integer portion of SPEC CPU2000 [Cantin, Hill 2003]