The memory hierarchy: Cache friendly code

Yipeng Huang

Rutgers University

November 25, 2025

Table of contents

Multilevel cache hierarchies

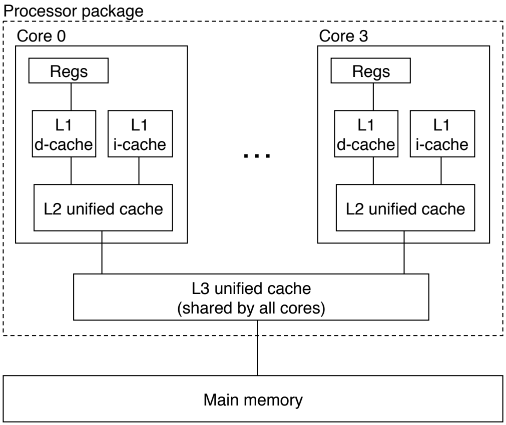


Figure: Intel Core i7 cache hierarchy. Image credit CS:APP

Small fast caches nested inside large slow caches

- L1 data and instruction cache: 32KB, 64 set, 8-way associative, 64B block, 4 cycle latency.
- L2 cache: 256KB, 512 set, 8-way associative, 64B block, 10 cycle latency.
- L3 cache: 8MB, 8192 set, 16-way associative, 64B block, 40-75 cycle latency.

Notice how latency cost increases as *E*-way associativity increases.

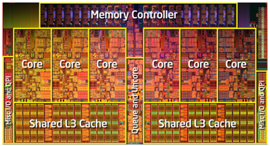


Figure: Intel 2020 Gulftown die shot. Image credit AnandTech

Table of contents

Policies for writes from CPU to memory

How to deal with write-hit?

- **Write-through.** Simple. Writes update both cache and memory. Costly memory bus traffic.
- **Write-back.** Complex. Writes update only cache and set a dirty bit; memory updated only upon eviction. Reduces memory bus traffic. (For multi-core CPUs, motivates complex cache coherence protocols)

How to deal with write-miss?

- **No-write-allocate.** Simple. Write-misses do not load block into cache. But write-misses are not mitigated via cache support.
- **Write-allocate.** Complex. Write-misses will load block into cache.

Typical designs:

- **Simple:** write-through + no-write-allocate.
- **Complex:** write-back + write-allocate.

Table of contents

Cache-friendly code

Algorithms can be written so that they work well with caches

- Maximize hit rate
- Minimize miss rate
- Minimize eviction counts

Do so by:

- Increasing spatial locality.
- Increasing temporal locality.

Advanced optimizing compilers can automatically make such optimizations

- GCC optimizations
- `https://gcc.gnu.org/onlinedocs/gcc/Optimize-Options.html`
- `-floop-interchange`
- `-floop-block`

Loop interchange

Refer to textbook slides on "Rearranging loops to improve spatial locality"

- Loop interchange increases spatial locality.
- In PA5, fourth part "cacheBlocking" you can explore the impact of this on matrix multiplication.
- In practice, programmers do not have to worry about this optimization.
- Optimized automatically in GCC by compiler flag `-floop-interchange` and `-O3`.

Cache blocking

Refer to textbook slides on "Using blocking to improve temporal locality"

- Cache blocking increases temporal locality.
- In PA5, fourth part "cacheBlocking" you can explore the impact of this on matrix multiplication.
- In practice, programmers do not have to worry about this optimization.
- Optimized automatically in GCC by compiler flag `-floop-block`. But it is not part of default optimizations such as `-O3` so you have to remember to set it.

Multilevel cache hierarchies

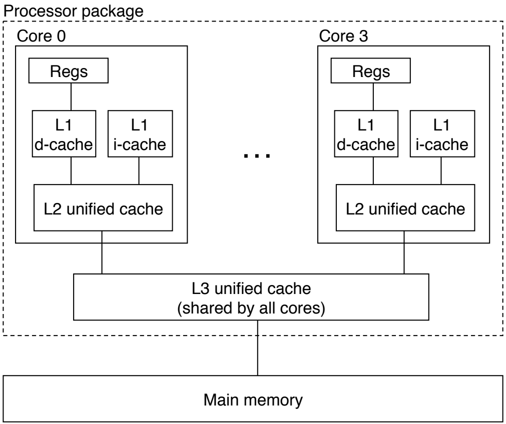


Figure: Intel Core i7 cache hierarchy. Image credit CS:APP

Small fast caches nested inside large slow caches

- L1 data and instruction cache: 32KB, 64 set, 8-way associative, 64B block, 4 cycle latency.
- L2 cache: 256KB, 512 set, 8-way associative, 64B block, 10 cycle latency.
- L3 cache: 8MB, 8192 set, 16-way associative, 64B block, 40-75 cycle latency.

Notice how latency cost increases as *E*-way associativity increases.

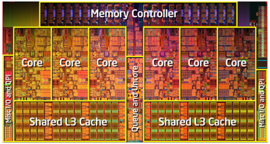


Figure: Intel 2020 Gulftown die shot. Image credit AnandTech

Cache oblivious algorithms

The challenge in writing code / compiling programs to take advantage of caches:

- Programmers do not easily have information about target machine.
- Compiling binaries for every envisioned target machine is costly.

Matrix transpose baseline algorithm: iteration

$$\mathbf{A} = \begin{bmatrix} a_{0,0} & a_{0,1} & a_{0,2} & a_{0,3} \\ a_{1,0} & a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,0} & a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,0} & a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix}$$

$$\mathbf{B} = \mathbf{A}^{\intercal} = \begin{bmatrix} a_{0,0} & a_{1,0} & a_{2,0} & a_{3,0} \\ a_{0,1} & a_{1,1} & a_{2,1} & a_{3,1} \\ a_{0,2} & a_{1,2} & a_{2,2} & a_{3,2} \\ a_{0,3} & a_{1,3} & a_{2,3} & a_{3,3} \end{bmatrix}$$

```
for ( size_t i=0; i<n; i++ ) {
  for ( size_t j=0; j<n; j++ ) {
    B[ j*n + i ] = A[ i*n + j ];
  }
}
```

Matrix transpose via recursion

$$\mathbf{A} = \begin{bmatrix} A_{0,0} & A_{0,1} \\ A_{1,0} & A_{1,1} \end{bmatrix} = \left[\begin{array}{cc|cc} a_{0,0} & a_{0,1} & a_{0,2} & a_{0,3} \\ a_{1,0} & a_{1,1} & a_{1,2} & a_{1,3} \\ \hline a_{2,0} & a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,0} & a_{3,1} & a_{3,2} & a_{3,3} \end{array}\right]$$

$$\mathbf{B} = \mathbf{A}^\intercal = \begin{bmatrix} A_{0,0}^\intercal & A_{1,0}^\intercal \\ A_{0,1}^\intercal & A_{1,1}^\intercal \end{bmatrix} = \left[\begin{array}{cc|cc} a_{0,0} & a_{1,0} & a_{2,0} & a_{3,0} \\ a_{0,1} & a_{1,1} & a_{2,1} & a_{3,1} \\ \hline a_{0,2} & a_{1,2} & a_{2,2} & a_{3,2} \\ a_{0,3} & a_{1,3} & a_{2,3} & a_{3,3} \end{array}\right]$$

Strategy:

- Divide and conquer large matrix to transpose into smaller transpositions.
- After some recursion, problem will fit well inside cache capacity.
- Once enough locality exists withing subroutine, switch to plain iterative approach.

Advantages:

- No need to know about cache capacity and parameters beforehand.
- Works well with deep multilevel cache hierarchies: different amounts of locality for each cache level.

Memory hierarchy implications for software-hardware abstraction

It is not entirely true the architecture can hide details of microarchitecture

Even less true going forward. What to do?

Application level recommendations

- Use industrial strength, optimized libraries compiled for target machine.
- Lapack, Linpack, Matlab, Python SciPy, NumPy...
- `https://people.inf.ethz.ch/markusp/teaching/263-2300-ETH-spring11/slides/class08.pdf`

Algorithm level recommendations

Deploy cache-oblivious algorithm implementations.

Compiler level recommendations

- Enable compiler optimizations—*e.g.*, `-O3`, `-floop-interchange`, `-floop-block`.
- `https://gcc.gnu.org/onlinedocs/gcc/Optimize-Options.html`

Table of contents

Computer organization

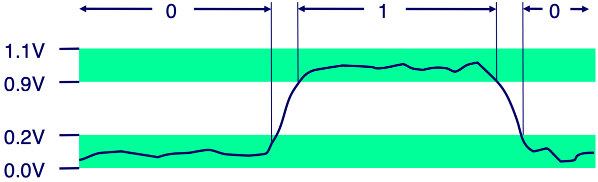
Layer cake

- Society
- Human beings
- Applications
- Algorithms
- High-level programming languages
- Interpreters
- Low-level programming languages
- Compilers
- Architectures
- Microarchitectures
- Sequential/combinational logic
- Transistors
- Semiconductors
- Materials science

Why binary

Everything is bits

- **Each bit is 0 or 1**
- **By encoding/interpreting sets of bits in various ways**
  - Computers determine what to do (instructions)
  - … and represent and manipulate numbers, sets, strings, etc…
- **Why bits? Electronic Implementation**
  - Easy to store with bistable elements
  - Reliably transmitted on noisy and inaccurate wires

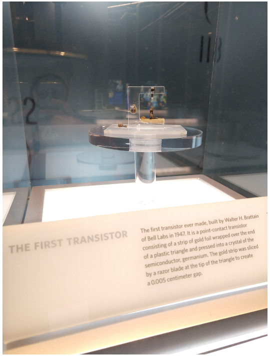


To build logic, we need switches

Vacuum tubes a.k.a. valves

Figure: Source: By Stefan Riepl (Quark48) - Self-photographed, CC BY-SA 2.0 `https://commons.wikimedia.org/w/index.php?curid=14682022`

Transistors



- The first transistor. Developed at Bell Labs, Murray Hill, New Jeresy
- `https://www.bell-labs.com/about/locations/`

MOSFETs

MOS: Metal-oxide-semiconductor

- A sandwich of conductor-insulator-semiconductor.

FET: Field-effect transistor

- Gate exerts electric field that changes conductivity of semiconductor.

NMOS, PMOS, CMOS

PMOS: P-type MOS

- positive gate voltage, acts as open circuit (insulator)
- negative gate voltage, acts as short circuit (conductor)

NMOS: N-type MOS

- positive gate voltage, acts as short circuit (conductor)
- negative gate voltage, acts as open circuit (insulator)

CMOS: Complementary MOS

- A combination of NMOS and PMOS to build logical gates such as NOT, AND, OR.
- We'll go to slides posted in supplementary material to see how they work.

Combinational vs. sequential logic

Combinational logic

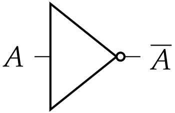
- No internal state nor memory
- Output depends entirely on input
- Examples: NOT, AND, NAND, OR, NOR, XOR, XNOR gates, decoders, multiplexers.

Sequential logic

- Has internal state (memory)
- Output depends on the inputs and also internal state
- Examples: latches, flip-flops, Mealy and Moore machines, registers, pipelines, SRAMs.

Table of contents

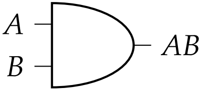
NOT gate



A	$\overline{A}$
0	1
1	0

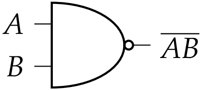
Table: Truth table for NOT gate

AND gate, NAND gate



A	B	AB
0	0	0
0	1	0
1	0	0
1	1	1

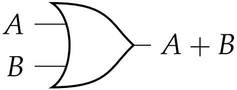
Table: Truth table for AND gate



A	B	$\overline{AB}$
0	0	1
0	1	1
1	0	1
1	1	0

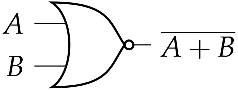
Table: Truth table for NAND gate

OR gate, NOR gate



A	B	$A + B$
0	0	0
0	1	1
1	0	1
1	1	1

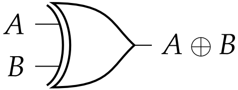
Table: Truth table for OR gate



A	B	$\overline{A + B}$
0	0	1
0	1	0
1	0	0
1	1	0

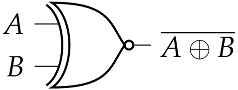
Table: Truth table for NOR gate

XOR gate, XNOR gate



A	B	$A \oplus B$
0	0	0
0	1	1
1	0	1
1	1	0

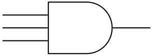
Table: Truth table for XOR gate

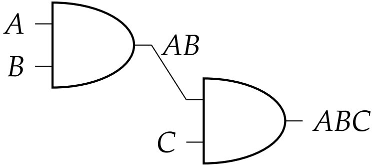


A	B	$\overline{A \oplus B}$
0	0	1
0	1	0
1	0	0
1	1	1

Table: Truth table for XNOR gate

More-than-2-input AND gate

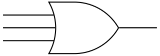


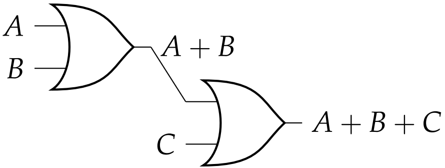


A	B	C	ABC
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	0
1	1	0	0
1	1	1	1

Table: Truth table for three-input AND gate

More-than-2-input OR gate





A	B	C	$A + B + C$
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	1
1	0	0	1
1	0	1	1
1	1	0	1
1	1	1	1

Table: Truth table for three-input OR gate

Table of contents

The set of logic gates {NOT, AND, OR} is universal

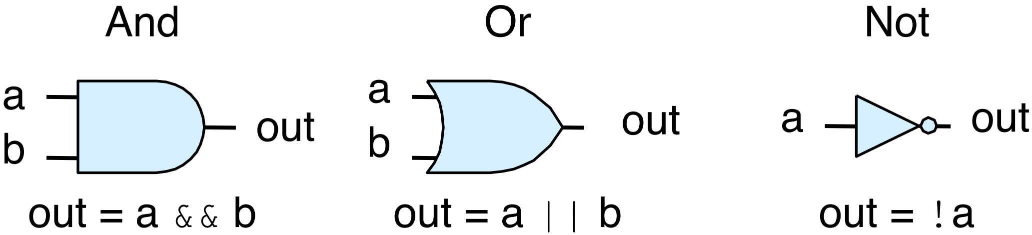


Figure: Source: CS:APP

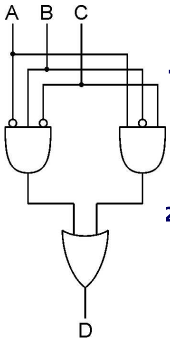
The set of logic gates {NOT, AND, OR} is universal

- Any truth table can be expressed as sum of products form.
- Write each row with output 1 as a product (minterm).
- Sum the products (minterm).
- Forms a disjunctive normal form (DNF).
- $D = \overline{A}B\overline{C} + A\overline{B}C$
- Always only needs NOT, AND, OR gates.
- Supplementary slides example...

Logical Completeness

Can implement ANY truth table with AND, OR, NOT.

A	B	C	D
0	0	0	0
0	0	1	0
0	1	0	1
0	1	1	0
1	0	0	0
1	0	1	1
1	1	0	0
1	1	1	0



. AND combinations that yield a "1" in the truth table.

2. OR the results of the AND gates.

Sum of products
OR of AND clauses

16

The set of logic gates {NOT, AND, OR} is universal

- Any truth table can be expressed as sum of products form.
- Write each row with output 1 as a product (minterm).
- Sum the products (minterm).
- Forms a disjunctive normal form (DNF).
- $D = \overline{A}B\overline{C} + A\overline{B}C$
- Always only needs NOT, AND, OR gates.
- Supplementary slides example...

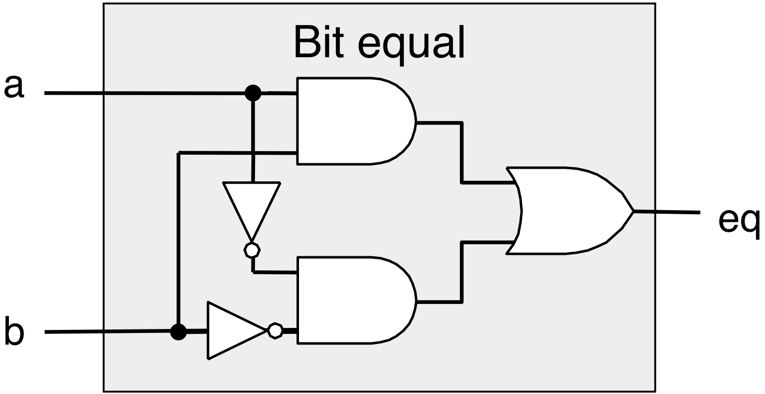
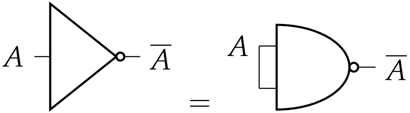


Figure: Source: CS:APP

The NAND gate is universal

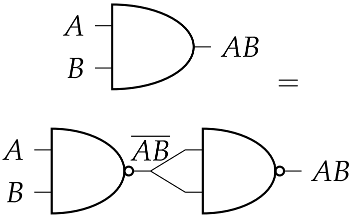
NOT gate as a single NAND gate



A	$\overline{A}$	AA	$\overline{AA}$
0	1	0	1
1	0	1	0

Table: $\overline{A} = \overline{AA}$

AND gate as two NAND gates



A	B	AB	$\overline{AB}$	$\overline{\overline{AB}}$
0	0	0	1	0
0	1	0	1	0
1	0	0	1	0
1	1	1	0	1

Table: $AB = \overline{\overline{AB}}$

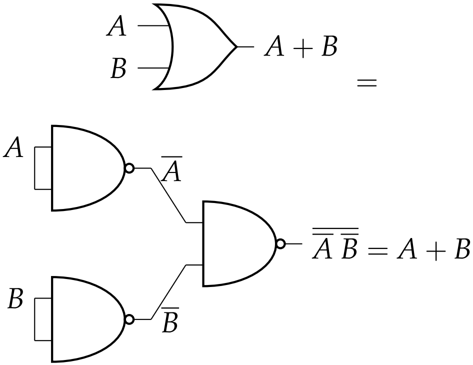
The NAND gate is universal

De Morgan's Law

A	B	$\overline{A}$	$\overline{B}$	$\overline{A}\,\overline{B}$	$A+B$	$\overline{A+B}$
0	0	1	1	1	0	1
0	1	1	0	0	1	0
1	0	0	1	0	1	0
1	1	0	0	0	1	0

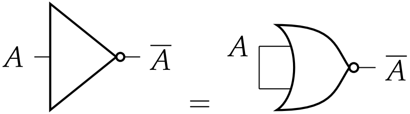
Table: $\overline{A}\,\overline{B} = \overline{A+B}$

OR gate as three NAND gates



The NOR gate is universal

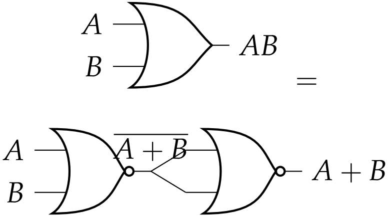
NOT gate as a single NOR gate



A	$\overline{A}$	$A+A$	$\overline{A+A}$
0	1	0	1
1	0	1	0

Table: $\overline{A} = \overline{A+A}$

OR gate as two NOR gates



A	B	$A+B$	$\overline{A+B}$	$\overline{\overline{A+B}}$
0	0	0	1	0
0	1	1	0	1
1	0	1	0	1
1	1	1	0	1

Table: $A+B = \overline{\overline{A+B}}$

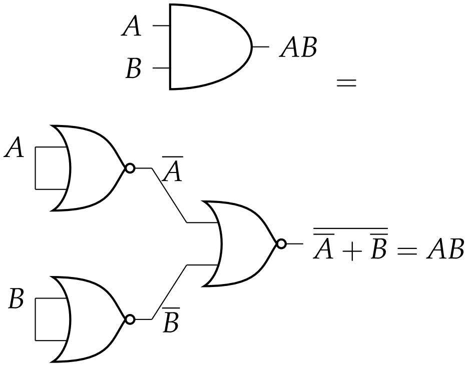
The NOR gate is universal

De Morgan's Law

A	B	$\overline{A}$	$\overline{B}$	$\overline{A}+\overline{B}$	AB	$\overline{AB}$
0	0	1	1	1	0	1
0	1	1	0	1	0	1
1	0	0	1	1	0	1
1	1	0	0	0	1	0

Table: $\overline{A}+\overline{B}=\overline{AB}$

AND gate as three NOR gates



Combinational vs. sequential logic

Combinational logic

- No internal state nor memory
- Output depends entirely on input
- Examples: NOT, AND, NAND, OR, NOR, XOR, XNOR gates, decoders, multiplexers.

Sequential logic

- Has internal state (memory)
- Output depends on the inputs and also internal state
- Examples: latches, flip-flops, Mealy and Moore machines, registers, pipelines, SRAMs.