

Your Optimizing Compiler is Not Optimizing Enough. To Hell With Multiple Recursions!

Federico Bruzzone,¹ PhD Student

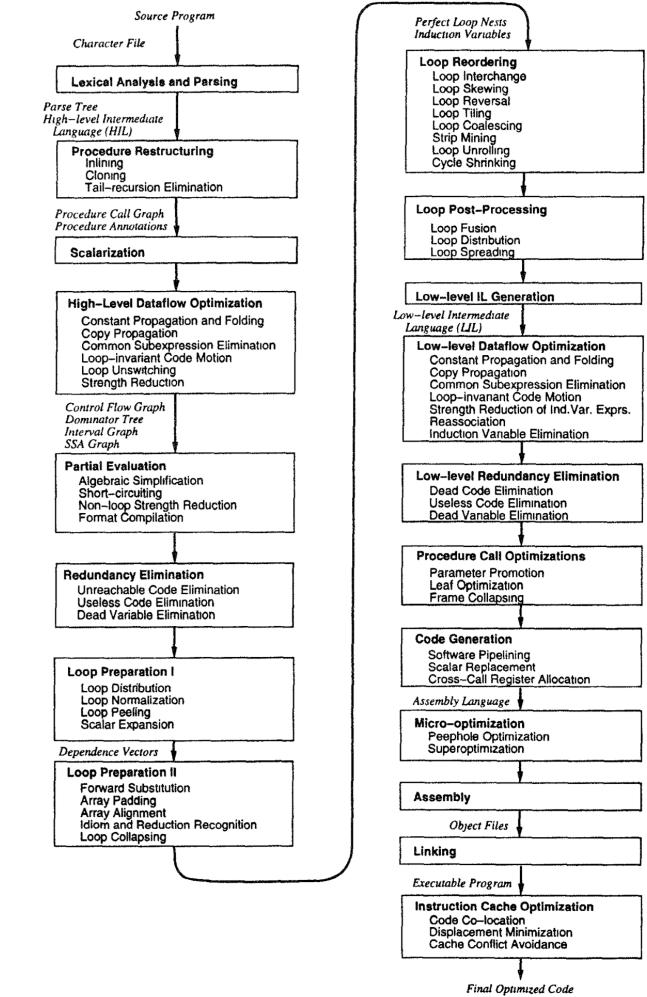
Milan, Italy – 27 November 2025

¹ADAPT Lab – Università degli Studi di Milano,
Website: federicobruzzone.github.io,
Github: github.com/FedericoBruzzone,
Email: federico.bruzzone@unimi.it

Compilers as Musical Compositions

Compilers are frequently perceived as intricate musical compositions—like the unfinished *J. S. Bach's Art of Fugue*—where mathematical precision and logical interplay guide each part.

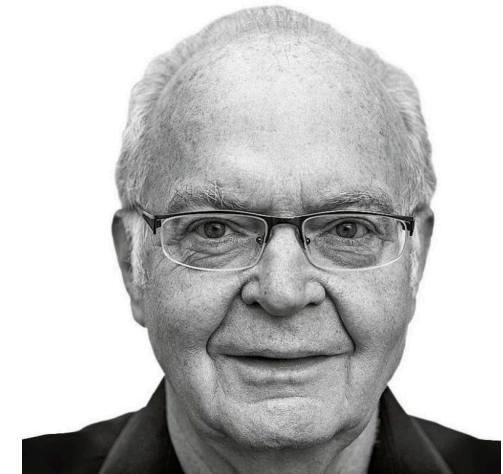
Every module enters in perfect timing, weaving together a structure that only the keenest ears can fully grasp.



Premature Optimizations

Donald E. Knuth warned in 1974 about the dangers of **premature optimization** in programming [2]:

We should forget about small efficiencies, say about 97% of the time: premature optimization is the root of all evil. Yet we should not pass up our opportunities in that critical 3%.



In the absence of either empirically measured or theoretically justified performance issues, programmers should **avoid** making optimizations based **solely** on assumptions about potential performance gains.

Optimizing Compilers

Compilers use information collected during analysis passes to guide transformations [3], [4].

Compiler optimizations² are such transformations (say *meaning-preserving mappings* [6]) applied to the input code to improve certain aspects—such as performance, resource utilization, and power consumption—without altering its observable behavior.

In accordance with the literature [7], [8], such compilers are referred to as **optimizing compilers**.

²A chronologically sorted list of papers on compiler optimization, from the works of 1952 through the techniques of 1994, is available at [5].

Optimizing Compilers

Compilers use information collected during analysis passes to guide transformations [3], [4].

Compiler optimizations³ are such transformations (say *meaning-preserving mappings* [6]) applied to the input code to improve certain aspects—such as performance, resource utilization, and power consumption—without altering its observable behavior.

In accordance with the literature [7], [8], such compilers are referred to as **optimizing compilers**.

³A chronologically sorted list of papers on compiler optimization, from the works of 1952 through the techniques of 1994, is available at [5].

Optimizing Compilers

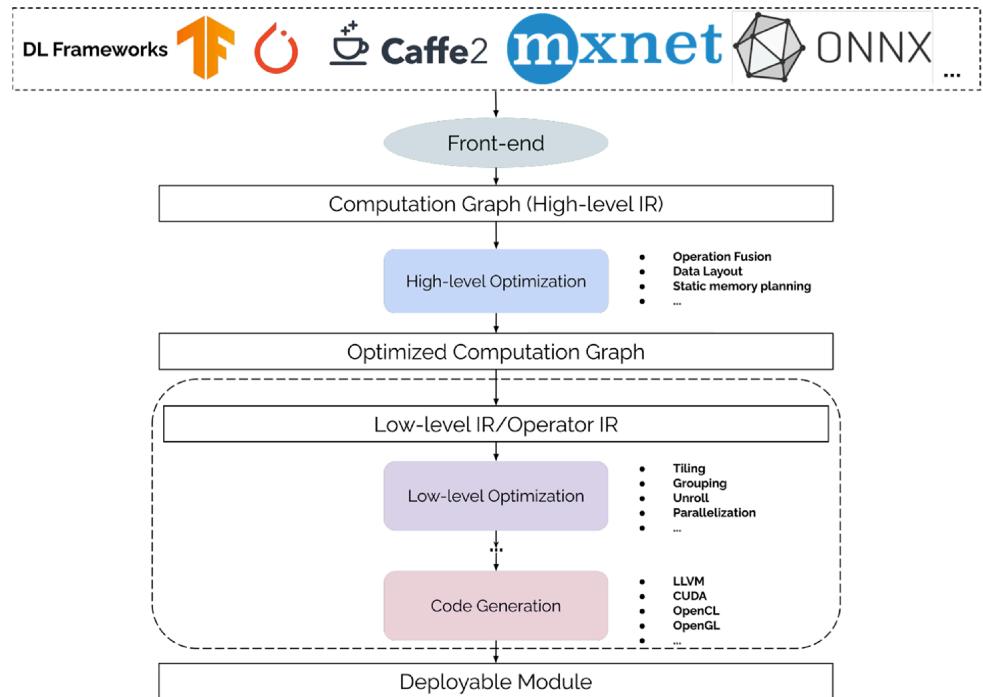
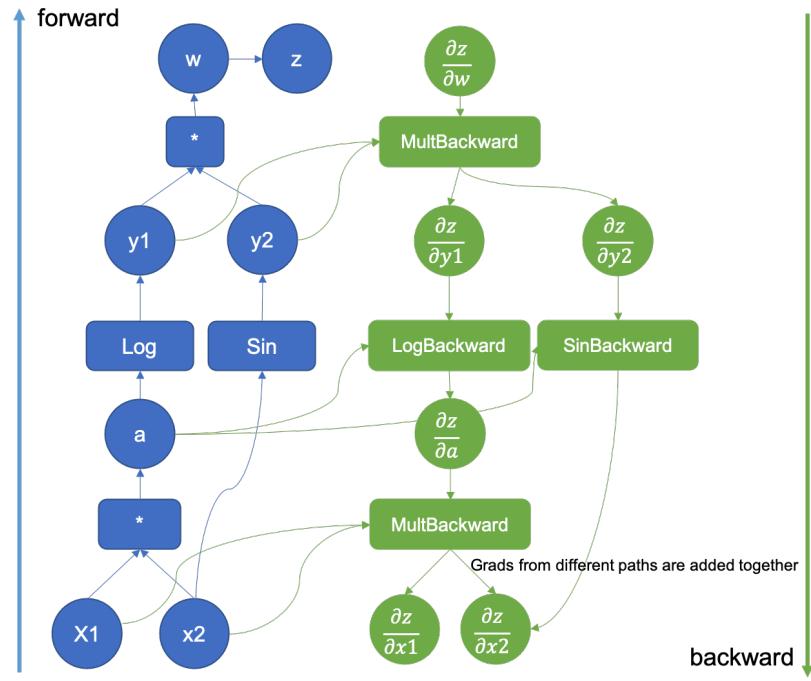
Compilers use information collected during analysis passes to guide transformations [3], [4].

Compiler optimizations⁴ are such transformations (say *meaning-preserving mappings* [6]) applied to the input code to improve certain aspects—such as performance, resource utilization, and power consumption—without altering its observable behavior.

In accordance with the literature [7], [8], such compilers are referred to as **optimizing compilers**.

⁴A chronologically sorted list of papers on compiler optimization, from the works of 1952 through the techniques of 1994, is available at [5].

Machine Learning Framework are Just Optimizing Compilers⁵



⁵TensorFlow XLA, NVIDIA CUDA Compiler (NVCC), MLIR, and TVM all use LLVM [9]. Li *et al.*, [10] compiled a survey on ML compilers.

Peephole Optimizations in x86-64 (cf. [11], [12])

```
1 ; x = x * 2, s.t. x: i32          asm
2 mov     eax, dword ptr [rbp - 4]
3 imul    eax, 2
4 mov     dword ptr [rbp - 4], eax
```

Peephole Optimizations in x86-64 (cf. [11], [12])

```
1 ; x = x * 2, s.t. x: i32          asm
2 mov     eax, dword ptr [rbp - 4]
3 imul    eax, 2
4 mov     dword ptr [rbp - 4], eax
```

The optimized version replaces the multiplication by 2 with a **more efficient** binary shift operation.

```
1 ; x = x << 1, s.t. x: i32          asm
2 mov     eax, dword ptr [rbp - 4]
3 shl     eax
4 mov     dword ptr [rbp - 4], eax
```

Peephole Optimizations in x86-64 (cf. [11], [12])

```
1 ; x = x * 2, s.t. x: i32
2 mov     eax, dword ptr [rbp - 4]
3 imul    eax, 2
4 mov     dword ptr [rbp - 4], eax
```

asm

```
1 ; x = x + 0, s.t. x: i32
```

```
2 mov     eax, dword ptr [rbp - 4]
3 add     eax, 0
4 mov     dword ptr [rbp - 4], eax
```

asm

The optimized version replaces the multiplication by 2 with a **more efficient** binary shift operation.

```
1 ; x = x << 1, s.t. x: i32
2 mov     eax, dword ptr [rbp - 4]
3 shl     eax
4 mov     dword ptr [rbp - 4], eax
```

asm

Peephole Optimizations in x86-64 (cf. [11], [12])

```
1 ; x = x * 2, s.t. x: i32
2 mov     eax, dword ptr [rbp - 4]
3 imul    eax, 2
4 mov     dword ptr [rbp - 4], eax
```

asm

```
1 ; x = x + 0, s.t. x: i32
```

```
2 mov     eax, dword ptr [rbp - 4]
3 add     eax, 0
4 mov     dword ptr [rbp - 4], eax
```

asm

The optimized version replaces the multiplication by 2 with a **more efficient** binary shift operation.

```
1 ; x = x << 1, s.t. x: i32
2 mov     eax, dword ptr [rbp - 4]
3 shl     eax
4 mov     dword ptr [rbp - 4], eax
```

asm

```
1 mov     eax, dword ptr [rbp - 4]
2 mov     dword ptr [rbp - 4], eax
```

asm

The optimized version removes the **unnecessary** addition operation.

Peephole Optimizations in x86-64 (cf. [11], [12])

```
1 ; x = x * 2, s.t. x: i32
2 mov     eax, dword ptr [rbp - 4]
3 imul    eax, 2
4 mov     dword ptr [rbp - 4], eax
```

asm

```
1 ; x = x + 0, s.t. x: i32
```

```
2 mov     eax, dword ptr [rbp - 4]
3 add     eax, 0
4 mov     dword ptr [rbp - 4], eax
```

asm

The optimized version replaces the multiplication by 2 with a **more efficient** binary shift operation.

```
1 ; x = x << 1, s.t. x: i32
2 mov     eax, dword ptr [rbp - 4]
3 shl     eax
4 mov     dword ptr [rbp - 4], eax
```

asm

```
1 mov     eax, dword ptr [rbp - 4]
2 mov     dword ptr [rbp - 4], eax
```

asm

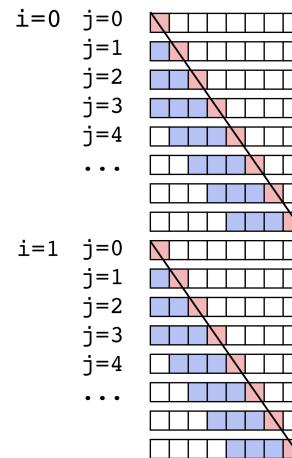
The `mov` instructions are redundant and can be **pruned** as well!

Loop Nest Optimizations – Loop Tiling (cf. [13], [14])

```
for (int i=0; i<n; ++i) {  
    for (int j=0; j<m; ++j) {  
        c[i][j] = a[i] * b[j];  
    }  
}
```

The vector **b** **may not** fit into a line of CPU cache, causing multiple cache misses during the inner loop.

It implies multiple **fetches** from the main memory, which is **slow**.



[A. Vladimirov, Session 10](#)

Loop Nest Optimizations – Loop Tiling (cf. [13], [14])

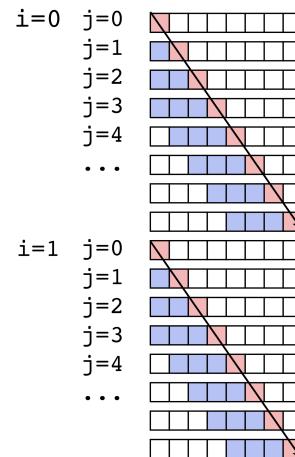
```
for (int i=0; i<n; ++i) {  
    for (int j=0; j<m; ++j) {  
        c[i][j] = a[i] * b[j];  
    }  
}
```



The vector **b** **may not** fit into a line of CPU cache, causing multiple cache misses during the inner loop.

It implies multiple **fetches** from the main memory, which is **slow**.

```
int TILE_SIZE = 16;
```



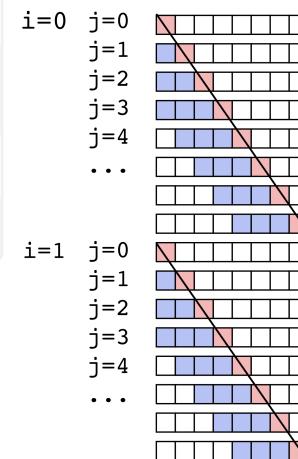
A. Vladimirov, Session 10

Loop Nest Optimizations – Loop Tiling (cf. [13], [14])

```
for (int i=0; i<n; ++i) {
    for (int j=0; j<m; ++j) {
        c[i][j] = a[i] * b[j];
    }
}
```

The vector **b** **may not** fit into a line of CPU cache, causing multiple cache misses during the inner loop.

It implies multiple **fetches** from the main memory, which is **slow**.



A. Vladimirov, Session 10

```
int TILE_SIZE = 16;
```

The inner loop works on a **tile** of **b** that fits into the cache.

```
for (int jj=0; jj<m; jj+=TILE_SIZE)
    for (int i=0; i<n; ++i)
        for (int j=jj; j<min(jj + TILE_SIZE, m); ++j)
            c[i][j] = a[i] * b[j];
```

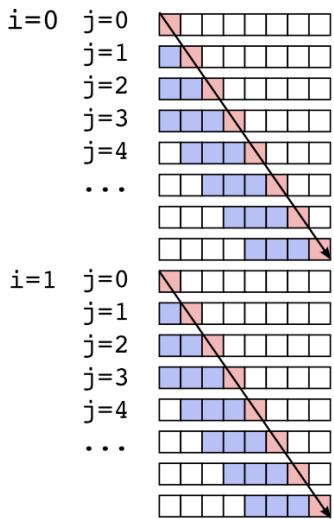
But, the values for the array **a** will be read m / TS times!

```
for (int ii = 0; ii < n; ii += TILE_SIZE_I)
    for (int jj = 0; jj < m; jj += TILE_SIZE_J)
        for (int i = ii; i < MIN(n, ii + TILE_SIZE_I); i++)
            for (int j = jj; j < MIN(m, jj + TILE_SIZE_J); j++)
                c[i][j] = a[i] * b[j];
```

Loop Tiling Visualization⁶

Original:

```
for (i=0; i<m; i++)
    for (j=0; j<n; j++)
        ...=....*b[j];
```



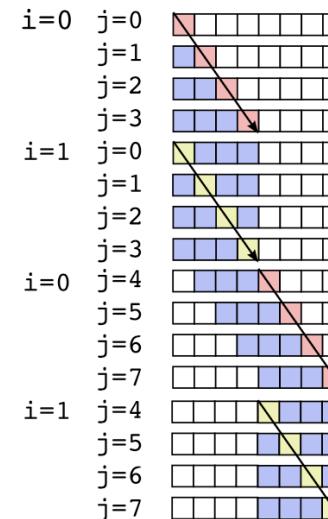
█ - cached, LRU eviction policy
█ - cache miss (read from memory, slow)
█ - cache hit (read from cache, fast)

Cache size: 4
 TILE=4
 (must be tuned to cache size)

Cache hit rate without tiling: 0%
 Cache hit rate with tiling: 50%

Tiled:

```
for (jj=0; jj<n; jj+=TILE)
    for (i=0; i<m; i++)
        for (j=jj; j<jj+TILE; j++)
            ...=....*b[j];
```

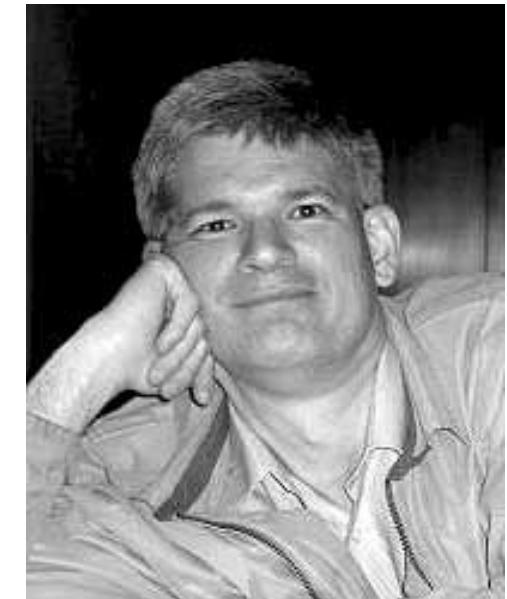


⁶[A. Vladimirov, Session 10 — Access to Caches and Memory](#)

Tail Call/Recursion Optimization (cf. [3], [15], [16])

Guy L. Steele, Jr. in 1977 observed that **tail-recursive procedure calls** can be optimized to avoid growing the call stack [17]:

In general, procedure calls may be usefully thought of as GOTO statements which also pass parameters, and can be uniformly coded as [machine code] JUMP instructions.



From Recursive to Iterative Functions (cf. [18])

$$f(x) = \begin{cases} b(x_0) & \text{if } x = x_0 \\ a(x, f(d(x))) & \text{otherwise} \end{cases}$$

s.t. a , b , and so on may denote any pieces of code.

From Recursive to Iterative Functions (cf. [18])

$$f(x) = \begin{cases} b(x_0) & \text{if } x = x_0 \\ a(x, f(d(x))) & \text{otherwise} \end{cases}$$

s.t. a, b , and so on may denote any pieces of code.

To transform recursive function f into iterative form, we need to:

1. Identifies an increment \oplus to the argument of f , i.e., $x' = x \oplus y$ such that $x = \text{prev}(x')$, where prev is based on the arguments of the recursive call. In this case, $\text{prev}(x) = d(x)$ and, if d^{-1} exists, $x \oplus y = d^{-1}(x)$, can be plugged in for y .
2. Derives an incremental program $f'(x, r)$ that computes $f(x)$ using an accumulator r of $f(\text{prev}(x))$.
3. Forms an iterative version that initializes using the base case of f and iteratively applies f' until reaching the desired argument.

```

 $f(x) = \{$ 
   $x_1 = x_0; r = b(x_0);$ 
   $\text{while } (x_1 \neq x)\{$ 
     $x_1 = d^{-1}(x_1);$ 
     $r = a(x_1, r);$ 
   $\}$ 
   $\text{return } r;$ 
 $\}$ 
```

From Recursive to Iterative Functions (cf. [18])

$$f(x) = \begin{cases} b(x_0) & \text{if } x = x_0 \\ a(x, f(d(x))) & \text{otherwise} \end{cases}$$

s.t. a, b , and so on may denote any pieces of code.

To transform recursive function f into iterative form, we need to:

1. Identifies an increment \oplus to the argument of f , i.e., $x' = x \oplus y$ such that $x = \text{prev}(x')$, where prev is based on the arguments of the recursive call. In this case, $\text{prev}(x) = d(x)$ and, if d^{-1} exists, $x \oplus y = d^{-1}(x)$, can be plugged in for y .
2. Derives an incremental program $f'(x, r)$ that computes $f(x)$ using an accumulator r of $f(\text{prev}(x))$.
3. Forms an iterative version that initializes using the base case of f and iteratively applies f' until reaching the desired argument.

```
f(x) = {
    x1 = x0; r = b(x0);
    while (x1 ≠ x) {
        x1 = d⁻¹(x1);
        r = a(x1, r);
    }
    return r;
}
```

Note that, when a is in the form $a(a_1(x), y)$ and a is associative, we do not need d^{-1} and x_1 .

Tail-recursive Factorial Function

```
int fact(int n) {  
    if (n == 0) {  
        return 1;  
    }  
    return n * fact(n - 1);  
}
```

C++

The replacement of $n * ((n - 1) * (n - 2))$ by $(n * (n - 1)) * (n - 2)$ is valid due to the **associativity** of multiplication.

$$f(x) = \{$$
$$r = b(x_0);$$
$$\text{while } (x \neq x_0) \{$$
$$r = a(r, a_1(x));$$
$$x = d(x);$$
$$\}$$
$$\text{return } r;$$
$$\}$$

Tail-recursive Factorial Function

```
int fact(int n) {  
    if (n == 0) {  
        return 1;  
    }  
    return n * fact(n - 1);  
}
```

C++

The replacement of $n * ((n - 1) * (n - 2))$ by $(n * (n - 1)) * (n - 2)$ is valid due to the **associativity** of multiplication.

$$f(x) = \{$$

$$r = b(x_0);$$

$$\text{while } (x \neq x_0) \{$$

$$r = a(r, a_1(x));$$

$$x = d(x);$$

$$\}$$

$$\text{return } r;$$

$$\}$$

Note that, (i) when dealing with IEEE754 numbers, multiplication is **not** strictly associative, and (ii) the latter *might be* slower due to multiply bigger numbers.

Tail-recursive Factorial Function

```
int fact(int n) {  
    if (n == 0) {  
        return 1;  
    }  
    return n * fact(n - 1);  
}
```

C++

The replacement of $n * ((n - 1) * (n - 2))$ by $(n * (n - 1)) * (n - 2)$ is valid due to the **associativity** of multiplication.

```
f(x) = {  
    r = b(x0);  
    while (x ≠ x0) {  
        r = a(r, a1(x));  
        x = d(x);  
    }  
    return r;  
}
```

Note that, (i) when dealing with IEEE754 numbers, multiplication is **not** strictly associative, and (ii) the latter *might be* slower due to multiply bigger numbers.

```
define i32 @fact(int)(i32 %n) {  
entry:  
    %cmp3 = icmp eq i32 %n, 0  
    br il %cmp3, label %return, label %if.else ; label %if.else.preheader  
; if.else.preheader -> vector.ph -> vector.body -> middle.block -> if.else.preheader7  
if.else:  
    %n.tr5 = phi i32 [ %sub, %if.else ], [ %n.tr5.ph, %if.else.preheader7 ]  
    %acc.tr4 = phi i32 [ %mul, %if.else ], [ %acc.tr4.ph, %if.else.preheader7 ]  
    %sub = add nsw i32 %n.tr5, -1  
    %mul = mul nsw i32 %n.tr5, %acc.tr4  
    %cmp = icmp eq i32 %sub, 0  
    br il %cmp, label %return, label %if.else  
return:  
    %acc.tr.lcssa = phi i32 [ 1, %entry ], [ %4, %middle.block ], [ %mul, %if.else ]  
    ret i32 %acc.tr.lcssa  
}
```

What About Fibonacci? Is it the root of all evil?

```
int fib(int n) {  
    if (n <= 1) {  
        return n;  
    }  
    return fib(n - 1) + fib(n - 2);  
}
```

C++


```
int fib(int n) {  
    int fib_n_1 = 0;  
    int a = 0; // fib(0)  
    int b = 1; // fib(1)  
    for (int i = 2; i <= n - 1; ++i) {  
        int temp = a + b;  
        a = b;  
        b = temp;  
    }  
    fib_n_1 = b; // fib(n-1)  
  
    int fib_n_2 = 0;  
    a = 0; // fib(0)  
    b = 1; // fib(1)  
    for (i = 2; i <= n - 2; ++i) {  
        int temp = a + b;  
    }  
}
```

C++

*Your Optimizing Compiler is **Not** Optimizing Enough. To Hell With Multiple Recursions!*

```
a = b;  
b = temp;  
}  
fib_n_2 = b; // fib(n-2)  
  
return fib_n_1 + fib_n_2;  
}
```

What About Fibonacci? Is it the root of all evil?

```
int fib(int n) {  
    if (n <= 1) {  
        return n;  
    }  
    return fib(n - 1) + fib(n - 2);  
}
```

C++

```
int fib_iter_gen(int n) {  
    int fib_n_1 = 0;  
    int a = 0; // fib(0)  
    int b = 1; // fib(1)
```

C++

```
// Cal up to n-1
for (int i = 2; i <= n - 1; +
+i) {
    int temp = a + b; //
    fib(i) = fib(i-2) +
    fib(i-1)

    a = b;           //
    Update fib(i-2) to
    fib(i-1)

    b = temp;        // Update
    fib(i-1) to fib(i)
}

fib_n_1 = b; // fib(n-1)

int fib_n_2 = 0;
a = 0; // fib(0)
```

```
b = 1; // fib(1)
// Call up to n-2
for (i = 2; i <= n - 2; ++i)
{
    int temp = a + b; //
    fib(i) = fib(i-2) +
    fib(i-1)
    a = b;           //
    Update fib(i-2) to
    fib(i-1)
    b = temp;        // Update
    fib(i-1) to fib(i)
}
fib_n_2 = b; // fib(n-2)
```

```
return fib_n_1 + fib_n_2; //  
fib(n)  
}
```

Bibliography

- [1] D. F. Bacon, S. L. Graham, and O. J. Sharp, “Compiler transformations for high-performance computing,” *ACM Computing Surveys (CSUR)*, vol. 26, no. 4, pp. 345–420, 1994.
- [2] D. E. Knuth, “Structured Programming with go to Statements,” *ACM Comput. Surv.*, vol. 6, no. 4, pp. 261–301, Dec. 1974.
- [3] K. D. Cooper and L. Torczon, *Engineering a Compiler*, no. . Morgan Kaufmann, 2022.
- [4] K. Kennedy and J. R. Allen, *Optimizing compilers for modern architectures: a dependence-based approach*. Morgan Kaufmann Publishers Inc., 2001.
- [5] F. Bruzzone, “Papers on Compiler Optimizations: Analysis and Transformations (1952-1994).” <https://github.com/FedericoBruzzone/papers-on-compiler-optimizations>, 2025.
- [6] R. Paige, “Future directions in program transformations,” *SIGPLAN Not.*, vol. 32, no. 1, pp. 94–98, Jan. 1997.
- [7] F. E. Allen, “Program Optimization,” *Annual Review of Automatic Programming*, vol. 5. Pergamon Press, pp. 239–307, 1966.

- [8] W. A. Wulf, “The design of an optimizing compiler,” 1973.
- [9] C. Lattner and V. Adve, “LLVM: A compilation framework for lifelong program analysis & transformation,” in *International symposium on code generation and optimization, 2004. CGO 2004.*, 2004, pp. 75–86.
- [10] M. Li *et al.*, “The deep learning compiler: A comprehensive survey,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 3, pp. 708–727, 2020.
- [11] W. M. McKeeman, “Peephole optimization,” *Commun. ACM*, vol. 8, no. 7, pp. 443–444, July 1965.
- [12] A. S. Tanenbaum, H. van Staveren, and J. W. Stevenson, “Using Peephole Optimization on Intermediate Code,” *ACM Trans. Program. Lang. Syst.*, vol. 4, no. 1, pp. 21–36, Jan. 1982.
- [13] M. Wolfe, “More iteration space tiling,” in *Proceedings of the 1989 ACM/IEEE Conference on Supercomputing*, in Supercomputing '89. Reno, Nevada, USA: Association for Computing Machinery, 1989, pp. 655–664.
- [14] M. E. Wolf and M. S. Lam, “A data locality optimizing algorithm,” *SIGPLAN Not.*, vol. 26, no. 6, pp. 30–44, May 1991.

- [15] A. V. Aho, R. Sethi, and J. D. Ullman, *Compilers: Principles, Techniques, and Tools*. Reading, Massachusetts: Addison Wesley, 1986.
- [16] S. Muchnick, *Advanced compiler design implementation*. Morgan kaufmann, 1997.
- [17] G. L. Steele, “Debunking the “expensive procedure call” myth or, procedure call implementations considered harmful or, LAMBDA: The Ultimate GOTO,” in *Proceedings of the 1977 Annual Conference*, in ACM '77. Seattle, Washington: Association for Computing Machinery, 1977, pp. 153–162.
- [18] Y. A. Liu and S. D. Stoller, “From recursion to iteration: what are the optimizations?,” in *Proceedings of the 2000 ACM SIGPLAN workshop on Partial evaluation and semantics-based program manipulation*, 1999, pp. 73–82.