

Exact Computation of the Catalan Number $C(2,050,572,903)$

Mahesh Ramani
`mahesh.ramani.iyer@gmail.com`

November 2025

Abstract

We describe a two-phase algorithm for computing exact Catalan numbers at record-breaking scale. Our method computes the Catalan number

$$C(n) = \frac{(2n)!}{n! (n+1)!},$$

for $n = 2,050,572,903$ (a result with 1,234,567,890 decimal digits) by operating on prime-exponent data rather than forming giant factorials. In Phase 1, a parallel segmented sieve enumerates primes up to $2n$, and Legendre's formula is used to compute each prime's exponent in $C(n)$. We group primes by exponent and write this factorization to disk. In Phase 2, a balanced-product-tree method reconstructs the integer from the factorization, using chunking to limit memory. This pipeline avoids infeasible intermediate factorials. To the best of our knowledge, this computation produces the largest exact Catalan number ever computed. We report performance statistics (wall times, memory) on a single-machine run, and provide implementation details (scripts `catalan.py`, `catalanreconstructor.py`) to ensure reproducibility. The final result's SHA-256 hash is recorded for verification. Verification strategies include independent reconstruction from the factor file and modular checks on low-order bits.

1 Introduction

Catalan numbers appear in many combinatorial problems (e.g. counting balanced parentheses, binary trees, polygon triangulations) and grow exponentially with n . By definition

$$C(n) = \frac{(2n)!}{n! (n+1)!},$$

which cannot be computed directly for large n due to the astronomical size of $(2n)!$. Traditional recursive or dynamic programming methods become infeasible for n on the order of billions. Instead, we use a prime-factorization approach: by Legendre's formula, one can compute the exponent of each prime $p \leq 2n$ in the numerator and denominator, thereby

determining the exact prime-power factorization of $C(n)$. Grouping primes by exponent and reconstructing via balanced product trees yields the final integer.

We implement this approach in two phases. Phase 1 (in `catalan.py`) enumerates primes up to $2n$ using a parallel segmented sieve and computes each prime's exponent in the factorization of $C(n)$. The primes are grouped by exponent e and written to a factorization file `catalan_[n].factorization.txt`, where $[n]$ represents the value of n . In Phase 2 (in `catalanreconstructor.py`), the file is read and processed group-by-group: for each exponent e , we compute

$$G_e = \prod_{p \in \text{group}} p$$

(using a memory-efficient balanced product tree) and then compute $P_e = G_e^e$. The final Catalan number is $\prod_e P_e$, computed again by a balanced product tree. All large multiprecision arithmetic is done with `gmpy2` (GMP). Intermediate results are written to disk (text and binary files) so that Phase 1 and Phase 2 can be run on separate machines if needed.

This paper presents the mathematical foundations (Section 2), algorithmic details (Section 3), and implementation (Section 4), along with pseudocode for each phase. In Section 5 we report experimental results and resource usage for the computation $n = 2,050,572,903$. We claim this to be, *to the best of our knowledge*, the largest Catalan number ever computed exactly. Section ?? discusses bottlenecks and future improvements. We conclude by describing verification strategies, including the provided SHA-256 hash of the final result, to enable independent validation.

2 Mathematical Foundation

Let $v_p(m)$ denote the exponent of prime p in the prime-factorization of integer m (the p -adic valuation). Legendre's formula states that for positive integers m and prime p ,

$$v_p(m!) = \sum_{k=1}^{\infty} \left\lfloor \frac{m}{p^k} \right\rfloor,$$

a finite sum since $\lfloor m/p^k \rfloor = 0$ for $p^k > m$. Using this, we can express the p -adic valuation of the Catalan number:

$$v_p(C(n)) = v_p((2n)!) - 2v_p(n!) - v_p(n+1).$$

Thus each prime $p \leq 2n$ contributes an exponent

$$e_p = v_p((2n)!) - 2v_p(n!) - v_p(n+1),$$

and $C(n) = \prod_{p \leq 2n} p^{e_p}$. Note $v_p(n+1)$ is nonzero only for primes dividing $n+1$. Because $C(n)$ is known to be an integer (e.g. by combinatorial interpretation or Kummer's theorem), all e_p computed as above are nonnegative. Working with these exponents avoids ever forming the enormous factorials $(2n)!$, $n!$, etc.

3 Algorithm Overview

Our computation proceeds in two phases. First we generate the factorization of $C(n)$ (in prime-exponent form), and then we reconstruct the integer from that factorization.

3.1 Phase 1: Factorization Generation (`catalan.py`)

We enumerate all primes $p \leq 2n$ using a parallel segmented Sieve of Eratosthenes. The sieve is implemented in Python, first finding all primes $\leq \sqrt{2n}$ and then sieving disjoint intervals in parallel (using `multiprocessing.Pool`). For each prime p found, we compute

$$a_p = v_p((2n)!), \quad b_p = v_p(n!), \quad c_p = v_p(n+1),$$

using Legendre's formula. Then $e_p = a_p - 2b_p - c_p$. For primes $n < p \leq 2n$, the term $v_p((2n)!)-2v_p(n!)$ simplifies to either 1 or 0 (since $\lfloor 2n/p \rfloor - 2\lfloor n/p \rfloor$ is 1 iff p appears once in the binomial $\binom{2n}{n}$); the code handles this case automatically via the same floor-sum formula. Next, the script factors $n+1$ by trial division using the prime list and subtracts those valuations c_p from each e_p .

We then group the primes by their exponent e_p . Let G_e denote the set of all primes with exponent e . In our implementation, we write the factorization to the text file `catalan_2050572903_factorization.txt` in the format:

```
# Prime factorization of Catalan(n)
# exponent=E count=K
p1 p2 p3 ...  (primes with exponent E)
...
```

Each block lists all primes with the same exponent E . In parallel, the code accumulates a high-precision approximation of $\log_{10} C(n)$ to estimate the digit count (which is then output as metadata).

3.2 Phase 2: Reconstruction (`catalanreconstructor.py`)

In Phase 2 we read the factorization file and reconstruct $C(n)$. The file is parsed group by group. For each exponent e with prime group G_e , we compute the product

$$Q_e = \prod_{p \in G_e} p.$$

This is done using a memory-efficient balanced product tree: we multiply the list of primes by repeatedly pairing and multiplying subproducts (and periodically garbage-collecting) to limit intermediate sizes. Once Q_e is obtained, we compute $P_e = Q_e^e$ using fast exponentiation (via `gmpy2`). Finally, the Catalan number is $\prod_e P_e$. This final product is again computed by a balanced product tree over all P_e values. The result is written to `catalan_2050572903_reconstructed.bin` (binary format) and optionally to a newline-formatted decimal text for inspection.

Algorithm 1 Factorization Generation for Catalan(n) (in `catalan.py`)

Require: Integer n .

- 1: Compute all primes $\leq 2n$ via parallel segmented sieve.
 - 2: Factor $m = n + 1$ by trial division with primes $\leq n + 1$.
 - 3: **for** each prime $p \leq 2n$ **do**
 - 4: Compute $a = \sum_{k \geq 1} \lfloor 2n/p^k \rfloor$.
 - 5: Compute $b = \sum_{k \geq 1} \lfloor n/p^k \rfloor$.
 - 6: Set $e = a - 2b$.
 - 7: **if** p divides $n + 1$ with exponent c **then**
 - 8: Set $e = e - c$.
 - 9: **end if**
 - 10: **if** $e > 0$ **then**
 - 11: Add p to group G_e .
 - 12: **end if**
 - 13: **end for**
 - 14: Write each group G_e to `catalan_n_factorization.txt`.
 - 15: Estimate and output digit count of $C(n)$.
-

Algorithm 2 Integer Reconstruction for Catalan(n) (in `catalanreconstructor.py`)

Require: Factorization groups G_e from file.

- 1: Initialize list $\mathcal{P} = []$.
 - 2: **for** each exponent e with prime list G_e **do**
 - 3: Compute $Q_e = \text{BalancedProductTree}(G_e)$.
 - 4: Compute $P_e = Q_e^e$ using fast pow.
 - 5: Append P_e to \mathcal{P} .
 - 6: **end for**
 - 7: Compute $C(n) = \text{BalancedProductTree}(\mathcal{P})$.
 - 8: Write $C(n)$ to `catalan_n_reconstructed.bin`.
-

Here $\text{BalancedProductTree}(S)$ denotes a recursive divide-and-conquer multiplication of elements in set S . In practice, the primes in a large group G_e or the list \mathcal{P} may be huge, so the code multiplies in chunks (e.g. 100,000 primes at a time) and combines partial products, with periodic garbage collection to constrain RAM usage.

4 Implementation and Reproducibility

The code is implemented in Python 3 using standard libraries. Key files are:

- `catalan.py`: Phase 1 factorization generation.
- `catalanreconstructor.py`: Phase 2 integer reconstruction.

Both use the `gmpy2` library (which wraps GMP) for multiprecision arithmetic. The code also uses `multiprocessing` for parallel sieve segments.

To reproduce our result, one should run `catalan.py` with $n = 2,050,572,903$, which will produce `catalan_2050572903_factorization.txt` and an estimated digit count. Then run `catalanreconstructor.py` on that factorization file to produce `catalan_2050572903_reconstructed.bin`. We provide these output files for reference. All scripts include parameters (segment size, chunk size) with default values that worked for this run; these can be tuned for other hardware.

We list the software stack and hardware used in our experiments:

- CPU: AMD Ryzen 9 4900 (8-core, 16-thread)
- RAM: 16 GB
- GPU: NVIDIA GPU (6 GB, not used in current implementation)
- OS: Windows
- Python: CPython 3.x (64-bit)
- Libraries: `gmpy2`, GMP library for big-integer arithmetic.

5 Results

We ran the two phases sequentially on a single machine with the above specifications. The key statistics of the run are summarized in Table 1. Notably, the input $n = 2\,050\,572\,903$ yielded a Catalan number with exactly 1 234 567 890 decimal digits (as targeted). The final binary result file (`catalan_2050572903_reconstructed.bin`) has size 1.413 GB and SHA-256 checksum `dac68f4ee35db8e9400e68bd6140e6cbcc6fb8ce81059318400e2c44e45ae4`. The total wall-clock time was 1549.13 seconds, with 179.37 s in Phase 1 and 1369.77 s in Phase 2. Peak memory usage was about 15.4 GB during Phase 2; Phase 1 used much less (not tightly measured).

Table 1: Computation statistics for $C(2,050,572,903)$.

| Statistic | Value |
|---------------------------------|---------------------------------------|
| n | 2,050,572,903 |
| Decimal digits of $C(n)$ | 1,234,567,890 |
| Bit length of $C(n)$ | 4,101,145,759 |
| Size of result file | 1.413 GB |
| SHA-256 of output | found above |
| Phase 1 runtime (wall) | 179.37 s |
| Phase 2 runtime (wall) | 1369.77 s |
| Total wall time | 1549.13 s |
| Peak memory (Phase 1 / Phase 2) | 4 / 15.4 GB |
| Hardware | AMD Ryzen 9 4900, 16 GB RAM, 6 GB GPU |
| Software | Python 3, <code>gmpy2</code> , GMP |

These results demonstrate the feasibility of computing Catalan numbers in the billion-digit range with commodity hardware. In particular, we believe (to the best of our knowledge) that $C(2\,050\,572\,903)$ is the largest exact Catalan number computed so far. We provide the factorization file and final output hash so that others can independently verify the result.

6 Discussion

The two-phase pipeline—factorization followed by reconstruction—is central to the algorithm’s feasibility at this scale. Phase 1, the factorization, can be executed on machines with modest RAM, as its primary tasks are sieving for primes and computing Legendre’s formula exponents. The dominant bottleneck, both in terms of computational time and memory, is Phase 2: the reconstruction. The large-integer multiplications required to combine the prime powers consume the majority of the runtime and dictate the hardware requirements for extreme-scale computations.

Disk I/O can also become a limiting factor during reconstruction, particularly when intermediate products are offloaded to disk to manage memory. For this reason, high-throughput storage such as NVMe or other solid-state drives is strongly preferred over traditional spinning disks to mitigate this potential bottleneck. Ultimately, the performance of Phase 2 is bound by the available RAM and the efficiency of the underlying multiprecision arithmetic library.

6.1 Time Complexity

- **Sieve:** Enumerating primes up to $2n$ costs approximately $\mathcal{O}(n \log \log n)$ work. This phase benefits significantly from parallelization, as demonstrated in our implementation.
- **Exponent Computation:** For each prime p , the computation of its exponent via Legendre’s formula involves a sum of floor divisions. The cost of this step is subdominant compared to the reconstruction phase for the large values of n targeted by this work.
- **Reconstruction:** This phase is dominated by large-integer multiplications. Let B be the bit-length of the final integer, where $B = \Theta(n \log n)$. If $M(B)$ denotes the cost of multiplying two B -bit integers, the reconstruction cost is approximately $\mathcal{O}(M(B) \log P)$, where P is the number of distinct prime factors. In practice, using GMP’s advanced multiplication algorithms, this behaves asymptotically like $\Theta(n(\log n)^2)$ bit-operations.

6.2 Space Complexity

The peak memory usage is dictated by the largest intermediate integer held in memory during the balanced product-tree reductions in Phase 2. The size of this integer, and thus the space complexity, is determined by the bit-length of the final Catalan number, which is $\Theta(n \log n)$ bits. The storage for the prime numbers generated during the sieve requires

$\mathcal{O}(n/\log n)$ machine words. The two-phase design effectively trades increased disk I/O for a reduction in peak RAM requirements by storing the complete factorization on disk before reconstruction begins.

7 Conclusion

We have presented a practical and reproducible method for computing exact Catalan numbers at a scale previously unreported. The successful computation of $C(2,050,572,903)$ demonstrates the effectiveness of the prime factorization approach, which circumvents the impossibly large factorial computations inherent in the definition of Catalan numbers. The decoupling of factorization and reconstruction allows for a flexible use of hardware, where the more demanding reconstruction phase can be performed on a machine equipped with sufficient memory and processing power. The primary bottlenecks remain the large-integer multiplications and potential I/O limitations during reconstruction. We have provided the implementation scripts and the resulting factorization data to encourage independent verification and further research into optimizing and scaling these methods for even larger values of n .