An update on lattice cryptanalysis vol. 2

The cost of sieving, and the security margin of Kyber768

John Schanck
Mozilla
March 24, 2024

Concrete cryptanalysis — bit security paradigm

Arthur Ganson's "Behold the Big Bang"

Secure systems take more than 2^{128} turns of the crank to break.

Resource realism

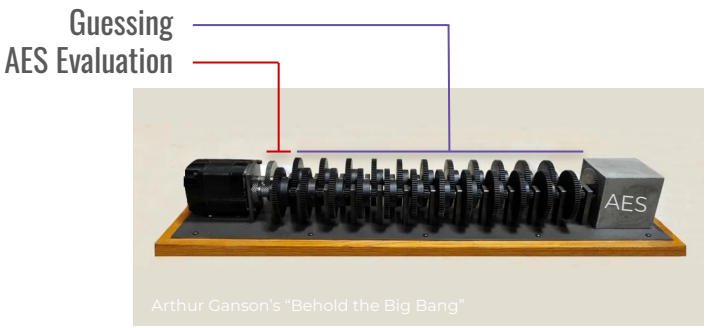
- Real attackers are constrained by:
  - Chip area (mass, system fits on earth),
  - Time (human scale),
  - Power (solar flux, other natural resources),
  - Physical law (locality, finiteness, reliability),

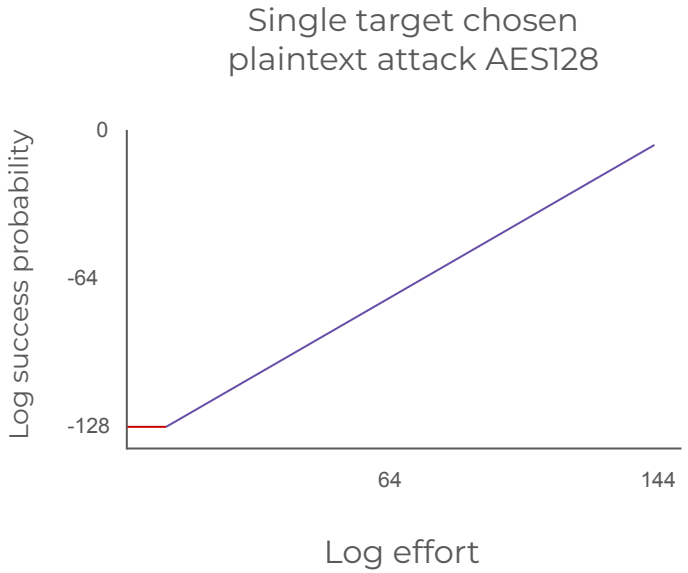
- Real attackers maximize their success probability subject to their constraints.

$$\max_{\text{attack}} \Pr[\text{success} \mid \text{attack, constraints}]$$

- $-\log_2$ of this is an operational definition of “security margin”

Concrete cryptanalysis — resource realist paradigm





Shape of the effort → success probability curve matters.
It defines "security margin" for various constraints.

The dual attack

Faster Dual Lattice Attacks for Solving LWE
with applications to CRYSTALS
Qian Guo[1], Thomas Johansson[1]

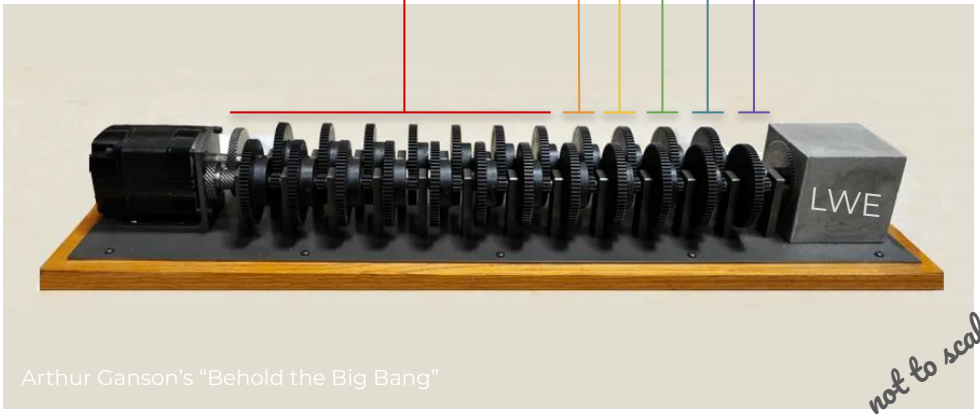
Report on the Security of LWE:
Improved Dual Lattice Attack
The Center of Encryption and Information Security – MATZOV*[1]
IDF

Shortest Vector from Lattice Sieving:
a Few Dimensions for Free
Léo Ducas

$\beta^* = \beta - O(\beta/\log \beta)$

Guessing
FFT
BKZ-β
SVP-β
Sieve-β*
Near neighbor search

	β*	Memory (bits)	RAM ops.
Kyber512	349	2^{87}	2^{123}
Kyber768	538	2^{127}	2^{179}
Kyber1024	761	2^{174}	2^{245}



https://github.com/malb/lattice-estimator commit 00ec72c. dual_hybrid. MATZOV reduction model.

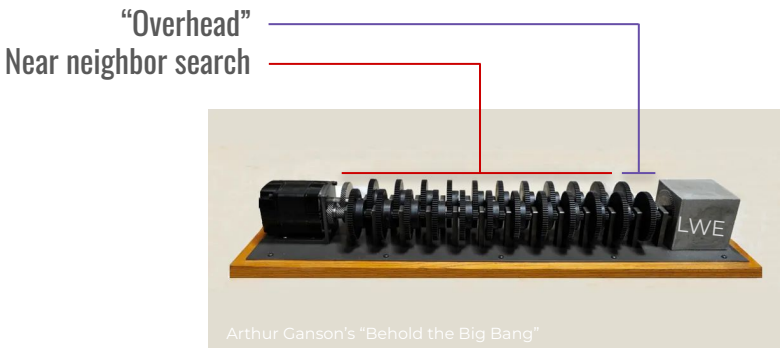
Outline

Lattice attacks on Kyber have:

1. poor effort → success probability scaling,
2. which gets worse when the attacker is memory constrained,
3. and even worse when we factor in data movement costs.

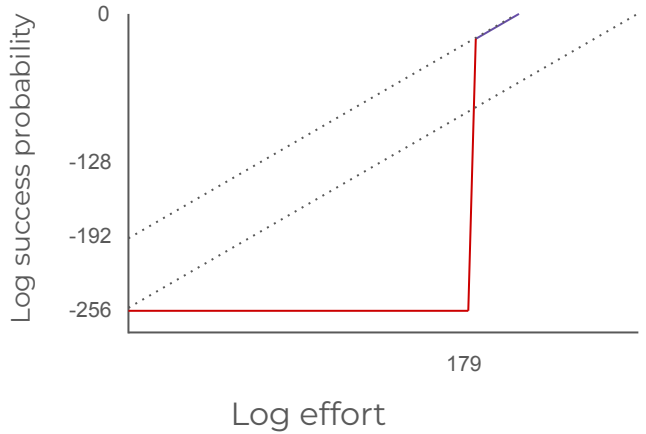
Close with open problems in the analysis.

Poor effort → success probability scaling



Small number of iterations
Huge cost per-iteration

Dual attack on Kyber768 using 2-sieve



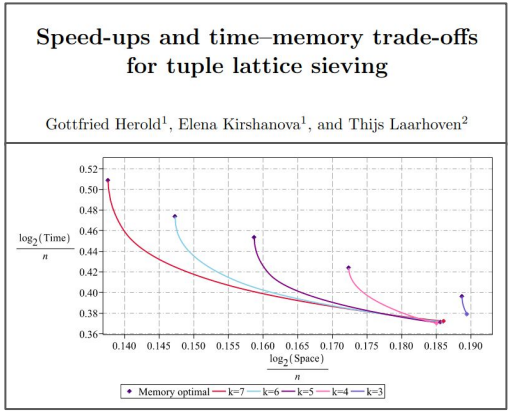
Outline

Lattice attacks on Kyber have:

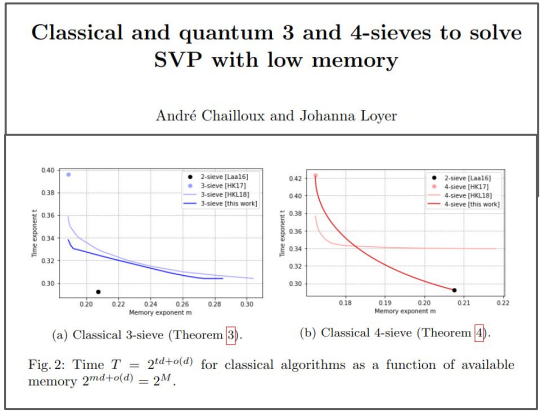
1. poor effort → success probability scaling,
2. which gets worse when the attacker is memory constrained,
3. and even worse when we factor in data movement costs.

Close with open problems in the analysis.

Memory constraints lead to worse effort → success probability scaling

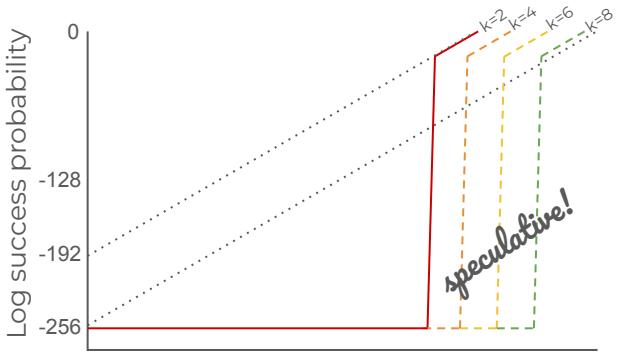


https://eprint.iacr.org/2017/1228



https://eprint.iacr.org/2023/200

Dual attack on Kyber768 using k-sieve



For Kyber768 (d=538), I suspect you need a memory exponent ~0.16 (k = 8?) if you are constrained to $<2^{100}$ bits.

Fermi approximation:
Is the attacker memory constrained?

Annual Silicon* Industry Trends

	2019	2020	2021	2022	2023
Area Shipments (MSI)	11,810	12,407	14,165	14,713	12,602
Revenues ($Billion)	11.2	11.2	12.6	13.8	12.3

Source: SEMI (www.semi.org), February 2024

**Data cited in this release include polished silicon wafers, including those used as virgin test wafers, as well as epitaxial silicon wafers, and non-polished silicon wafers shipped by the wafer manufacturers to end users. Shipments are for semiconductor applications only and do not include solar applications.*

Facts:

- Industry consumed ~2^{43} mm^2 of wafers in 2022.
- 3D NAND density is ~2^{34} bits/mm^2, or 2^{43} bits/g.

  2^{43} mm^2 · 2^{34} bits / mm^2 = 2^{77} bits.
- The moon has a mass of 2^{86} g.

  2^{86} g · 2^{43} bits / g = 2^{129} bits.

Conclusion: Yes.

- Need density–production product to scale by 2^{50} to store the 2^{127} bit database needed for a 2-sieve attack on Kyber768.

Outline

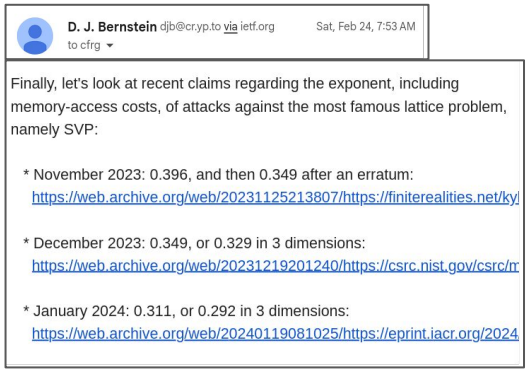
Lattice attacks on Kyber have:

1. poor effort → success probability scaling,
2. which gets worse when the attacker is memory constrained,
3. and even worse when we factor in data movement costs.

Close with open problems in the analysis.

Does memory-access add exponential cost?

Subject of intense discussion for 2-sieves.
More work needed for k-sieves



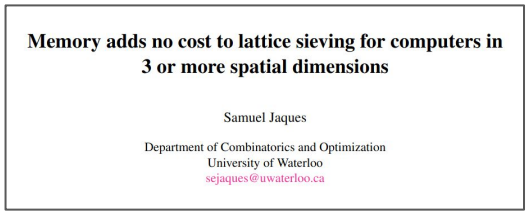
D. J. Bernstein djb@cr.yp.to via ietf.org Sat, Feb 24, 7:53 AM
to cfrg

Finally, let's look at recent claims regarding the exponent, including memory-access costs, of attacks against the most famous lattice problem, namely SVP:

* November 2023: 0.396, and then 0.349 after an erratum:
https://web.archive.org/web/20231125213807/https://finiterealities.net/ky

* December 2023: 0.349, or 0.329 in 3 dimensions:
https://web.archive.org/web/20231219201240/https://csrc.nist.gov/csrc/n

* January 2024: 0.311, or 0.292 in 3 dimensions:
https://web.archive.org/web/20240119081025/https://eprint.iacr.org/2024



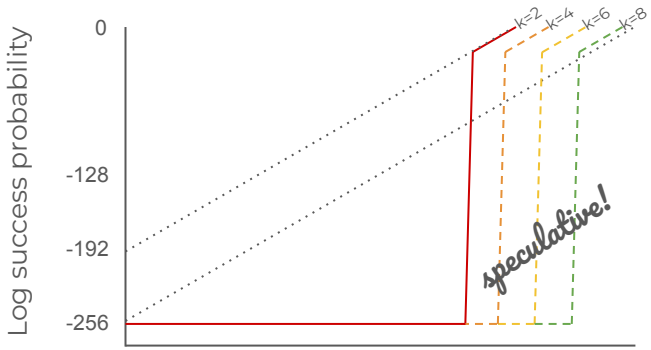
Memory adds no cost to lattice sieving for computers in 3 or more spatial dimensions

Samuel Jaques

Department of Combinatorics and Optimization
University of Waterloo
sejaques@uwaterloo.ca

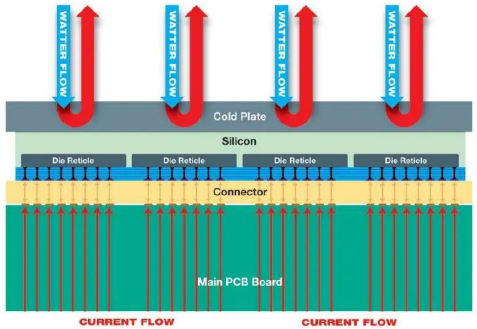
https://eprint.iacr.org/2024/080

Dual attack on Kyber768 using k-sieve



Current understanding: all curves move right by (small) exponential factor on 2D mesh architecture.

Why consider 2D mesh computers?



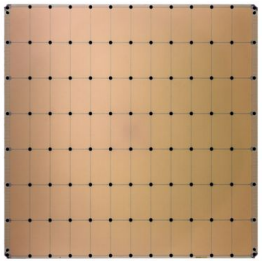
A Complexity Theory for VLSI

by

C. D. Thompson

August 1980

In particular, the area A and time T taken by any VLSI chip using any algorithm to perform an N-point Fourier transform must satisfy $AT^2 \geq cN^2 \log^2 N$, for some fixed $c > 0$. A more general result for both sorting and Fourier transformation is that $AT^{2x} = \Omega(N^{1+x} \log^{2x} N)$, for any x in the range $0 \leq x \leq 1$. Also, the energy dissipated by a VLSI chip during the solution of either of these problems is at least $\Omega(N^{3/2} \log N)$. The tightness of these bounds is demonstrated by the existence of nearly optimal circuits for both sorting and Fourier transformation. The circuits based on the shuffle-exchange interconnection pattern are fast but large: $T = O(\log^2 N)$ for Fourier transformation, $T = O(\log^3 N)$ for sorting; both have area A of at most $O(N^2/\log^{1/2} N)$. The circuits based on the mesh interconnection pattern are slow but small: $T = O(N^{1/2} \log\log N)$, $A = O(N \log^2 N)$.



Cerebras WSE-3
4 Trillion Transistors
46,225 mm² Silicon

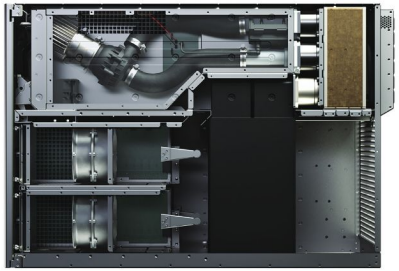


Figure 4: This side view shows the water movement assembly (top), and the air movement infrastructure — fans and a heat exchanger (bottom half).

https://fuse.wikichip.org/news/3010/a-look-at-cerebras-wafer-scale-engine-half-square-foot-silicon-chip/2/
https://8968533.fs1.hubspotusercontent-na1.net/hubfs/8968533/Datasheets/WSE-3%20Datasheet.pdf

Fermi approximation:
Cost of memory with mesh routing

Facts:

- 2022 silicon wafer supply → $2^{27.5}$ WSE-3s
  - $2^{47.5}$ cores,
  - 2^{66} bits of memory,
  - 2^{85} bits/s mesh bandwidth,
  - Sort 2^{66} bits of small data in a few minutes,
  - 4 TW of power.
- Annual global electricity supply ~30000 TWh

  30000 TWh / 3600 s = 8.3 TW

Conclusion:

- Already energy and chip-area constrained for a 2^{66} bit mesh sort. Factor 2^{61} away from Kyber768 2-sieve size.

Annual Silicon* Industry Trends

	2019	2020	2021	2022	2023
Area Shipments (MSI)	11,810	12,407	14,165	14,713	12,602
Revenues ($Billion)	11.2	11.2	12.6	13.8	12.3

Source: SEMI (www.semi.org), February 2024

**Data cited in this release include polished silicon wafers, including those used as virtual test wafers, as well as epitaxial silicon wafers, and non-polished silicon wafers shipped by the wafer manufacturers to end users. Shipments are for semiconductor applications only and do not include solar applications.*

Breakdown of global electricity supply and emissions, 2021-2026

TWh	2021	2022	2023	2026	Growth rate 2021-2022	Growth rate 2022-2023
Total Generation	28 426	29 124	29 734	32 694	2.5%	2.1%

Cerebras Wafer-Scale Engine

Fabrication process
5nm

Silicon area
46,225mm²

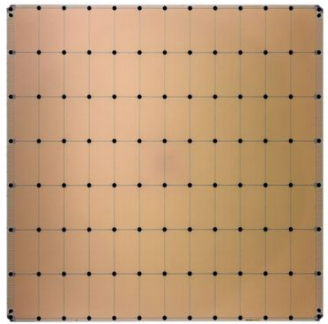
Transistors
4 Trillion

AI-optimized cores
900,000

Memory (on-chip)
44GB

Memory bandwidth
21PB/s

Fabric bandwidth
214Pb/s



Cerebras WSE-3
4 Trillion Transistors
46,225 mm² Silicon

24kW (youtube, techtechpotato)

Outline

Lattice attacks on Kyber have:

1. poor effort → success probability scaling,
2. which gets worse when the attacker is memory constrained,
3. and even worse when we factor in data movement costs.

Close with open problems in the analysis.

Open questions

- Re-evaluate FFT distinguisher step of the dual attack with memory / interconnect constraints.
- Compute non-asymptotic cost tables for k-sieves.
- Complete “resource realist” analysis of lattice attacks.
  - … with memory constraints.
  - … with energy or operation constraints.
- Determine best attack on Kyber768 for constrained adversaries.
    - Seed guessing?
    - Decryption failure attacks?
    - Combinatorial / hybrid attacks?

Takeaways

- Lattice attacks have poor effort → success scaling.
- 2-sieve memory is unobtainable.
- K-sieving reduces memory but adds exponential cost.
- Interconnect and chip area constraints add further cost, even if only subexponential.
- While there's a significant amount of analysis left to be done, it's not unreasonable to think that Kyber768 is as secure as AES-256.