Cost Models

NIST Internal Presentation by Ray Perlner

Tuesday August 11, 2020

The Question that Started This

- Category 1 security is defined as:

"at least as hard to break as brute force key search against AES 128"

  - NIST estimates this costs 2^{143} classical gates worth of computation
  - This attack parallelizes almost perfectly and requires almost no communication between the threads
  - The threads only require about 2^8 bits of memory apiece
    - Although if we assume the attack is parallelized so that it has depth $\leq 2^{64}$, the total memory must be at least 2^{79}

- The Kyber team estimates an attack on Kyber512 may cost as little as 2^{136} classical gates worth of computation, but
  - While the above attack can be parallelized (at least somewhat), the threads must access a ***shared*** memory of size 2^{89}
  - Not sure this is correct, but I will mostly assume the attack parallelizes perfectly and memory accesses are from random locations
- Is Kyber512 category 1?

Classical or Quantum?

- For categories 1, 3, 5, barring nongeneric quantum speedup (e.g. Shor, Kuperberg)
  - The classical attack is almost always the main concern
  - This is because AES key search is about as Groverizable as an algorithm can be
- This started with a dispute about category 1, so we'll mostly focus on classical attacks

Models to Consider

- RAM model
- Nonlocal gate model
- Time × Area
- Time × Volume
- Custom models

RAM Model

- Based on assembly instructions for typical CPUs
  - Assume unit cost to read or write from any memory location, regardless of the memory size
  - Assume that the contents of a memory location can be used to address another memory location
  - Typically the cost is arrived at by counting instructions (which in turn can be costed in bit operations/gates)
  - Can consider multithreading in which case we can consider also
    - “Width” – Total memory size
    - “Depth” – Number of operations per thread
- This is the most common way to cost classical attacks and classical algorithms in general
- Not clear how well it translates to extremely large distributed computations (The closest thing to a realistic model of the attacks we’re talking about)

Nonlocal Gate Model

- Write the computation as a binary circuit with, say, AND and XOR gates.
  - Assume that gates can be performed between any two memory locations at unit cost
  - But, you can't use the contents of any memory location to decide which gate to perform later on
- When generalized to 2-qubit quantum gates, this is the most common way to cost quantum attacks
- To simulate a single threaded memory access, gates will have to touch every memory location
  - Cost is proportional to the size of the memory!
- For algorithms that can be parallelized, so memory per thread is small
  - Nonlocal Gate Model is equivalent to RAM model up to log factors in memory size: See https://arxiv.org/abs/1207.2307

Time × Area

- Same as nonlocal gate model, but
  - Assume memory locations are arranged 2-dimensionally
  - Gates can only be performed between nearest neighbor memory locations
  - Rather than counting gates, we multiply the number of memory locations by the circuit depth to compute cost
- Dan Bernstein's favorite model
- Makes sense if
  - heat dissipation per gate,
  - density of gates per memory location per time

  are constant as computation size scales
- Makes less sense otherwise
- Even if the algorithm parallelizes, random access to memory costs *square root* of memory size
- Local memory access is essentially unit cost

Time × Volume

- Same as nonlocal gate model, but
  - Assume memory locations are arranged ***3***-dimensionally
  - Gates can only be performed between nearest neighbor memory locations
  - Rather than counting gates, we multiply the number of memory locations by the circuit depth to compute cost
- Even if the algorithm parallelizes, random access to memory costs ***cube root*** of memory size
- Local memory access is essentially unit cost

Custom Models

- RAM model says memory access is essentially free

- Gate models say memory access is super expensive unless it can be parallelized
- Additionally, Time × Area treats data at rest the same as data being actively processed
- Even Time × Volume treats a long distance fiber optic link the same as a line of densely packed gates

- Can we find a middle ground between gate and RAM models?
  - Assume RAM access has logarithmic cost, but cube root latency?
  - Give each location where a gate can happen ~2^{50} memory locations for free?
  - Give separate, per-category hard limits for gate count, RAM size, total RAM queries, depth?

Part 2
How Hard are the Categories?

Landauer's Limit [Landauer 1961]

- Current computation architectures are irreversible (you can't reconstruct inputs from outputs)
  - E.g. if x AND y = 0, inputs could have been (0,0) (0,1) or (1,0)
- Landauer's limit:
  - At temperature T any computation that destroys a bit of information about the input must produce at least $kT\ \ln(2)$ waste heat to get rid of the entropy
  - k is Boltzmann's constant $k \approx 1.3 \times 10^{-23}\mathrm{J/K}$
- For reference
  - The Earth's thermal radiation carries away about 2^{151} bits of entropy per year
    - If we assume each irreversible gate loses a bit of entropy this means at most 2^{151} gates per year AND
    - The maximum depth of an irreversible computation at a sensible temperature like 300K is about 2^{70} per year
    - Compare with 2^{143} classical gates for category 1 and the middle value 2^{64} for MAXDEPTH from CFP

Brownian Computation [Bennett 1973]

- For constant-ish overhead we can convert an irreversible circuit into a reversible circuit
  - This is a standard technique in quantum computing
- Now we aren't bound by Landauer Limit. How much better can we do?
- Ballistic computation [Fredkin, Toffoli 1982]
  - Model computation with perfectly elastic billiard balls and reflectors
  - Says computation is basically free
  - But is generally considered unrealistic due to sensitivity to small perturbations
- Brownian computation
  - Let reversible computations be driven forward and backwards by thermal noise
  - Induce a biasing energy $\varepsilon < kT$ per gate
  - Gates will flip back and forth an average of kT / ε times, slowing down the computation
  - But the net cost per gate will only be ε

Implications of Brownian Computation

- If you're willing to compute slower and make up the difference with parallelism, reversible gates use proportionally less energy than irreversible gates
- Reversible computation with $\varepsilon = kT$ is probably worse than irreversible computation by some constant factor, but it's hard to compute, so we'll ignore it
- How much might we gain?
  - We guessed an irreversible computation with 2^{151} gates and depth 2^{70} might be the maximum feasible on Earth per year
  - This implies a circuit width around 2^{81}
  - Let's assume we can get 1 bit of memory for each atom in the Earth's crust
  - There are about 2^{159} of those
  - Now we can (maybe) do 2^{190} gates per year with a depth of 2^{31} per year
    - Compare to Category 3: 2^{207} classical gates, small value of MAXDEPTH: 2^{40}

Unpowered Brownian Computation "Classical Grover" [Perlner, Liu 2017]

- For AES key search and similar, there's another way to use classical reversible computation
- Make tiny computing units that randomly explore the AES key space driven by thermal noise, and only dissipate energy when they find the correct key
- Can be made to induce transition to final state in their neighbors
- Energy cost is basically the same as for Quantum Grover Search
- Memory can be reduced to a reasonable amount by speeding up computation
  - Speed is proportional to temperature
  - Computation at arbitrarily high temperatures is probably at least as hard as quantum computation, so we can maybe ignore this option for the purpose of dealing with classical security levels

What about Fundamental Limits on Memory?

- No particular reason moving memory without processing it should consume energy
- Routing to the correct memory location should have at least logarithmic energy cost in the memory size, but that's about it
- In weak gravity regime (no black holes), a constant maximum memory density per volume is probably realistic
  - Given maximum signaling speed c, this implies maximum throughput scales with area
  - And minimum latency scales with *cube root* memory size
- For black holes, it's information per unit area, but I don't think we're dealing with systems of that scale
- Note that initializing a unit of memory is itself an irreversible computation, so Landauer limit applies

OK,
But What about Real Technology?

Note: Commercial products are mentioned here for informational purposes. This is not an endorsement by NIST of any of the mentioned products.

How Much Does Computation Cost?

- We can assume any computation as large as we're talking about will use custom hardware
- As a proxy for what optimized hardware looks like we can look at bitcoin mining
- Example specs (Antminer s19 Pro)
  - 3250 W
  - 110 Th/ Sec
  - ~$3000 New (I saw a used one for $180)
- Note that at 10¢/ kWh, power becomes the dominant cost after ~1yr
  - (Shorter than the expected hardware life)
- ***Using the same 10¢/ kWh 2^{143} gates costs about $\$2^{64}$***
- The hash is double SHA2: ~2^{19} gates
- Earth gets 1.7×10^{17}W of power from the Sun
  - About 1/4 gets eaten up by the atmosphere
  - If we blanket the Earth's surface with solar panels (Typical efficiency 20%)
  - That leaves 2.5×10^{16}W
  - ***This is enough to perform 2^{143} gates in 500 years***

How Much Memory can we Get for $\$2^{64}$ over 500 Years

- Will estimate based on Hard Drives (based on SEAGATE ST16000NM001G EXOS X16 16TB)
- 16TB = 2^{47} bits
  - costs about $400
  - Lasts about 5 years
  - Needs about 10W of power
    - This adds up to about $45 over the lifetime of the drive
- Memory budget:

$$\frac{\$2^{64}}{500y} \times \frac{5y}{\$445} \times 2^{47} \text{ bits} \approx 2^{96.5} \text{ bits}$$

- Alternative budgets
  - 1TB Flash drive $200, 2.5 W, 5y: $2^{93.5}$ bits
  - Power only: $2^{99.5}$ bits
  - Flash, power only: $2^{97.5}$ bits
- Note these numbers are all greater than 2^{89}

What about Memory Bandwidth?

- Both Flash and HDD typically advertise a maximum data transfer rate of about 250 MB per second
  - Might not be looking at the right number, since I always thought flash was faster than HDD
  - Maybe has to do with contiguous memory addresses, but I recall a comment in "Lattice Sieve Kernel" saying that they expected a bunch of the needed addresses to be clustered in memory
- Anyway,
  - If we assume 16TB HDD, this means we can access each memory location 2^{18} times in 500 years

$$2^{18} \times 2^{96.5} = 2^{114.5}$$

  - If we assume 1TB Flash, this means we can access each memory location 2^{22} times in 500 years

$$2^{22} \times 2^{93.5} = 2^{115.5}$$

  - Better numbers may be possible using smaller drives

What about the Network?

- Heat dissipation considerations demand that processors be distributed fairly evenly around the globe
- If memory is truly random access, this means processors will need to access data an average of 10000 kilometers away
- How much does it cost to send data 10000 km?
  - This question is surprisingly hard to answer

What ISPs (Allegedly) Pay for Bandwidth

- According to a Netflix report from 2011, the marginal cost of sending a GB of data was less than a penny
- According to this random web article from 2015: https://broadbandnow.com/report/much-data-really-cost-isps/

The price has recently fallen an average of something like 25% per year (15-50%)

- Extrapolating to 2020 this works out to about a dollar per TB
- So maybe, based on rumor and hearsay it is feasible to send

$$\mathbf{2^{64} \times 2^{43} = 2^{107}\ bits}$$

for $\$2^{64}$

Just for Kicks:
A System that is Almost as Good

- Forget long distance fiber optics
  - Send everything to a local data center
  - Let the data center pool your data on 16TB hard disks
    - less than 1 day to write, based on maximum data transfer rate 250MBps
  - And mail the hard disks 1 week, $128 international shipping to another data center
    - Which reads and disaggregates the data
- Total data sent:

$$\$\mathbf{2^{64}} \times \frac{\mathbf{2^{47}bits}}{\$\mathbf{128}} = \mathbf{2^{104}bits}$$

- Maximum number of accesses in series

$$500\ \mathrm{y} \times \frac{365\mathrm{d}}{\mathrm{y}} \times \frac{1\ \mathrm{access}}{9\mathrm{d}} \approx 2^{16}\ \mathrm{accesses}$$

- Which is fine as long as the memory size is *larger* than 2^{88}

What if we Buy our own Equipment?

- A $600, 5 watt device can deliver 100Gbps on 10 km of single mode fiber

(Cisco QSFP-100G-LR4-S Compatible 100GBASE-LR4 QSFP28 1310nm 10km DOM Optical Transceiver Module)

- A 12 strand single-mode fiber optic cable bundle costs $1.23 per foot

(Corning 12 Strand Singlemode Outdoor Figure 8 w/Messenger Fiber Optic Cable - Black (Per Foot))

- Assume the transceiver lasts 5 years and the fiber lasts 25
- Installation costs for fiber (labor, permitting) could easily dominate these costs (up to $80k per km), but we'll ignore it, because
  - It's hard to calculate
  - We can probably amortize it away by laying a lot of fiber in the same place
  - Replacing installed fiber is almost certainly cheaper, and we'll need to do that ~20 times more often

- Net cost of 100 Gbps over 10000 km for 500 years:

$$10000km \times 500y \times \left(\frac{\$600}{5y \times 10km} + \frac{.005kW}{10km} \times \frac{\$0.1}{kWh} \times \frac{8760h}{y} + \frac{\$1.23}{ft \times 12 \times 25y} \times \frac{3281\ ft}{km}\right)$$

$$= \$60\text{M (Transceivers)} + \$2\text{M(Power)} + \$67\text{M (Fiber)} = \$129\text{M}$$

- So how much memory access can we afford?

$$\$2^{64} \times 2^{9} \times \frac{2^{37}\textbf{bits}}{s \times \$2^{27}} \times \frac{2^{25}s}{y} = 2^{108}$$

A Note about Hardware Cost

- The cost of periodically replacing fiber and transponders represents 98.5% of our budget for long distance memory access!
- If we get rid of that cost, we can afford 2^{114} memory accesses instead of 2^{108}
- We could potentially get up to 2^{113} just by making the hardware last longer
- No particular reason to think we're close to any hard physical limits on this

What if we Mass Produce our own Equipment?

- According to https://www.nature.com/articles/s41467-020-16265-x
  - Using a single chip, a transmission rate of 44TBps can be achieved
  - Over a single strand of standard fiber optic cable
  - At a distance of 75km
- If these can be mass produced at $10,000 apiece, our 2^{108} RAM query budget becomes 2^{116}
- Even if they cost something ridiculous like $200,000, we get 2^{112}

Why Communication May Cost More than it Should

- At current market rates it seems like sending 2^{107} bits, long distance, costs the same as 2^{143} gates
- But something more like 2^{115} seems possible if any of the following happen:
  - Hardware (Fiber and Transponders) gets much cheaper
    - If raw material cost for fiber is the problem, we can switch to multicore fibers
  - Hardware gets a bit cheaper and lasts much longer
  - Transponders can be upgraded to experimentally-demonstrated bandwidths without much cost increase
- What might be preventing/delaying this?
  - Non-competitive market
  - Parts optimized for compatibility with older equipment, not efficiency
  - Demand for capacity increases not large enough to cover R&D costs for high end equipment
  - No demand for parts to last more than a few years due to expectation that hardware will keep improving

Back to Models

- Based on current technology, bounds comparable to category 1's 2^{143} classical gates are something like:
  - $2^{95} - 2^{100}$ memory
  - $2^{105} - 2^{120}$ RAM queries
- If RAM model were right
  - 2^{143} memory, 2^{143} queries
- If Time × Area
  - 2^{95} memory, 2^{95} queries OR
  - 2^{76} memory, 2^{105} queries
- If Time × Volume
  - 2^{107} memory, 2^{107} queries OR
  - 2^{100} memory, 2^{110} queries

- RAM model clearly underestimates memory costs
- Time × Area almost certainly overestimates memory costs
- Time × Volume is pretty close, but
  - It may still be an overestimate
  - No clear theoretical basis
    - Looks like a case of overestimates and underestimates cancelling
  - WILL overstate costs when processor to memory ratio is <<1

Questions for Discussion

- Let's say we're ok with saying attacks requiring less than 2^{143} classical gates are fine at category 1 if they require e.g.:
  - 2^{80} depth
  - 2^{100} memory
  - 2^{120} queries to shared memory
  - How do we extend this to categories 3, 5?
- How confident are we that cryptanalysts can give accurate, optimized concrete costs for attacks in Time × Volume or custom cost models?
  - Can we use more than 2^{89} memory claimed for Kyber512 to improve memory access locality?
  - Can we prove or disprove conjectures like Ducas 2018?

Conjecture / Open Question

There exist a sieving circuit with:

$$A = 2^{.2075n+o(n)} \text{ and } T \leq 2^{.142n+o(n)}.$$

Hint

- **[Becker Gama Joux 2015]** with only on level of filtration
- 3 or 4 layers of 2-dimensions should suffice.
- Keep shift-registers not fully saturated, for easier on-the-fly insertion.

References

[Shor 1994] https://arxiv.org/pdf/quant-ph/9508027.pdf

[Kuperberg 2003] https://arxiv.org/abs/quant-ph/0302112

[Kuperberg 2011] https://arxiv.org/pdf/1112.3333.pdf

[Grover 1996] https://arxiv.org/pdf/quant-ph/9711043.pdf

[Beals et al. 2012] https://arxiv.org/abs/1207.2307

[Landauer 1961] https://ieeexplore.ieee.org/document/5392446

[Bennett 1973] https://www.math.ucsd.edu/~sbuss/CourseWeb/Math268_2013W/Bennett_Reversibiity.pdf

[Bennett 1982] http://www.pitt.edu/~jdnorton/lectures/Rotman_Summer_School_2013/thermo_computing_docs/Bennett_1982.pdf

[Fredkin Toffoli 1982] https://link.springer.com/article/10.1007/BF01857727

[Perlner Liu 2017] https://arxiv.org/abs/1709.10510

[Cooper 2015] https://broadbandnow.com/report/much-data-really-cost-isps/

[Corcoran et al. 2020] https://www.nature.com/articles/s41467-020-16265-x

[Ducas 2018] https://eurocrypt.iacr.org/2018/Slides/Monday/TrackB/01-01.pdf