Using the Cloud to Determine Key Strengths Triennial Update

Marguerite Delcourt[1,2], Thorsten Kleinjung[1], Arjen K. Lenstra[1], Shubhojyoti Nath[1,3], Dan Page[4], and Nigel P. Smart[5]

[1] EPFL IC IINFCOM LACAL, Station 14, CH-1015 Lausanne, Switzerland.
[2] EPFL IC IINFCOM LCA2, Station 14, CH-1015 Lausanne, Switzerland.
[3] Indian Institute of Technology Kanpur, India.
[4] Dept. Computer Science, University of Bristol, Merchant Venturers Building, Woodland Road, Bristol, BS8 1UB, United Kingdom.
[5] COSIC, Departement Elektrotechniek (ESAT), Kasteelpark Arenberg 10 - bus 2440, 3001 Leuven, Belgium.

Abstract. We develop a new methodology to assess cryptographic key strength using cloud computing, by calculating the true economic cost of (symmetric- or private-) key retrieval for the most common cryptographic primitives. Although the present paper gives the current year (2018), 2015, 2012 and 2011 costs, more importantly it provides the tools and infrastructure to derive new data points at any time in the future, while allowing for improvements such as of new algorithmic approaches. Over time the resulting data points will provide valuable insight in the selection of cryptographic key sizes. For instance, we observe that the past clear cost-advantage of total cost of ownership compared to cloud-computing seems to be evaporating.[6]

1 Introduction

An important task for cryptographers is the analysis and recommendation of parameters, crucially including key size and thus implying key strength, for cryptographic primitives; clearly this is of theoretic and practical interest, relating to the study and the deployment of said primitives respectively. As a result, considerable effort has been, and is being, expended with the goal of providing meaningful data on which such recommendations can be based. Roughly speaking, two main approaches dominate: use of special-purpose hardware designs, including proposals such as [14, 15, 39, 40] (some of which have even been realised), and use of more software-oriented (or at least less bespoke) record setting computations such as [4, 5, 18, 20]. The resulting data can then be extrapolated using complexity estimates for the underlying algorithms, and appropriate versions of Moore's Law, in an attempt to assess the longevity of associated keys. In [21] this results in key size recommendations for public-key cryptosystems that offer security comparable to popular symmetric cryptosystems; in [25] it leads to security estimates in terms of hardware cost or execution time. Existing work for estimating symmetric key strengths (as well as other matters) is discussed in [41].

It is not hard to highlight disadvantages in these approaches. Some special-purpose hardware designs are highly optimistic; they all suffer from substantial upfront costs and, as far as we have been able to observe, are always harder to get to work and use than expected. On the other hand, even highly speculative designs may be useful in exploring new ideas and providing insightful lower bounds. Despite being more pragmatic, software-oriented estimates do not necessarily cater adequately for the form or performance of future generations of general-purpose processors (although so far straightforward application of Moore's Law is remarkably reliable, with various dire prophecies, such as the "memory wall", not materialising yet). For some algorithms, scaling up effort (e.g., focusing on larger keys) requires no more than organisational skill

[6] We see this as a living document, which will be updated as algorithms, the Amazon pricing model, and other factors change. The original version appeared in May 2011 as `https://eprint.iacr.org/2011/254.pdf`, was updated in May 2012, and was due to the second, third, fifth, and sixth author. The first and fourth authors of the present December 2018 update contributed to the 2015 and 2018 additions, respectively; further updates will be posted on `http://eprint.iacr.org`.

combined with patience. Thus, record setting computation that does not involve new ideas may have little or no scientific value: for the purposes of assessing key strength, a partial calculation is equally valuable. For some other algorithms, only the full computation adequately prepares for problems one may encounter when scaling up, and overall (in)feasibility may thus yield useful information. Finally, for neither the hardware- nor software-oriented approach, is there a uniform, or even well understood, metric by which "cost" should be estimated. For example, one often overlooks the cost of providing power and cooling, a foremost concern for modern, large-scale installations.

Despite these potential disadvantages and the fact that papers such as [21] and [25] were already a decade old in 2011, their results have proved to be quite resilient and are widely used. This can be explained by the fact that standardisation of key size requires some sort of long term extrapolation: there is no choice but to take the inherent uncertainty and potential unreliability for granted. In this paper we propose to complement traditional approaches, using an alternative that avoids reliance on special-purpose hardware, one time experiments, or record calculations, and that adopts a business-driven and thus economically relevant cost model. Although extrapolations can never be avoided, our approach minimises their uncertainty because whenever anyone sees fit, he/she can use commodity hardware to repeat the experiments and verify and update the cost estimates. In addition we can modify our cost estimates as our chosen pricing mechanism alters over time.

The current focus on cloud computing is widely described as a significant long-term shift in how computing services are delivered. In short, cloud computing enables any party to rent a combination of computational and storage resources that exist within managed data centers; the provider we use is the Amazon Elastic Compute Cloud [3], however others are available. The crux of our approach is the use of cloud computing to assess key strength (for a specific cryptographic primitive) in a way that provides a useful relationship to a true economic cost. Crucially, we rely on the fact that cloud computing providers operate as businesses: assuming they behave rationally, the pricing model for each of their services takes into account the associated purchase, maintenance, power and replacement costs. In order to balance reliability of revenue against utilisation, it is common for such a pricing model to incorporate both long-term and supply and demand driven components. We return to the issue of supply and demand below, but for the moment assume this has a negligible effect on the longer-term pricing structure of Amazon in particular. As a result, the Amazon pricing model provides a valid way to attach a monetary cost to key strength. A provider may clearly be expected to update their infrastructures and pricing model as technology and economic conditions dictate; indeed we show the effect of this over the last 8 years or so. However, by ensuring our approach is repeatable, for example using commodity cloud computing services and platform agnostic software, we are able to track results as they evolve over time: in the present version of this document covering the period from 2011 to 2018. We suggest this, and the approach as a whole, should therefore provide a robust understanding of how key size recommendations should be made.

In Section 2 we briefly explain our approach and those aspects of the Amazon Elastic Compute Cloud that we depend on. In Section 3 we describe our analysis applied to a number of cryptographic primitives, namely DES, AES, SHA-2, RSA and ECC. Section 4 and Section 5 contain concluding remarks. Throughout, all monetary quantities are given in US dollars.

2 The Amazon Elastic Compute Cloud

Amazon Elastic Compute Cloud (EC2) is a web-service that provides computational and storage resource "in the cloud" (i.e., on the Internet) to suit the specific needs of each user. In this section we describe the current (per June 2018) EC2 hardware platform and pricing model, and compare it with the previous (per October 2015, May 2012, February 2011) pricing model, from a point of view that is relevant for our purposes.

EC2 Compute Units. At a high level, EC2 consists of numerous installations (or data centers) each housing numerous processing nodes (a processor core with some memory and storage) which can be rented by users. In an attempt to qualify how powerful a node is, EC2 uses the notion of an *EC2 Compute Unit*, or *ECU* for short, with $\mathfrak{u}$ indicating one ECU. One ECU provides the equivalent computational capacity of a 1.0-1.2 GHz Opteron or Xeon processor circa 2007, which were roughly state of the art when EC2 launched. When

new processors are deployed within EC2, they are given an ECU-rating; currently there are different types of core, rated from approximately 3u to 4u. The lack of rigour in the definition of an ECU (e.g., identically clocked Xeon and Opteron processors do not have equivalent computational capacity) is not a concern for our purposes.

Instances. An *instance* refers to a specified amount of dedicated compute capacity that can be purchased: it depends on the processor type (e.g., 32- or 64-bit), the number of (virtualised) cores (2, 4, 8, 16, 32 or 36), the ECU-rating per core, memory and storage capacity, and network performance. Back in 2012 there were thirteen different instances, in 2015 there were 38 different instances, and now there are 82 different instances which can be partitioned into the *instance types* that are here termed "general purpose" (std), "micro", "memory optimized" (high-mem.), "high-CPU", "cluster compute" (cl-C), "GPU" (cl-G), "storage optimized" and "dense storage". In almost all instance types, there are subinstances intended for High Performance Computing (HPC) applications, and come with 10 Gb ethernet connections; each type has two to five subinstances of different sizes, indicated by, "micro", "small", "medium", "Large (L)" and "Extra Large (EL)". Use of instances can be supported by a variety of operating systems, ranging from different versions of Unix and Linux through to Windows.

Pricing model. Instances are charged per instance-hour depending on their capacity and the operating system used. In 2007 this was done at a flat rate of $0.10 per hour on a 1.7 GHz processor with 1.75 GB of memory [1]; in 2008 the pricing model used ECUs charged at the same flat rate of $0.10 per hour per ECU [2]. Since then the pricing model has evolved [3]. Currently, instances can be purchased at four different price bands, "on-demand", "reserved", "spot" and "dedicated" charged differently according to which of ten different geographic locations one is using (US east coast, US west coast (Oregon or North California), Ireland, Frankfurt, Singapore, Tokyo, Sydney, Sao Paulo or AWS Govcloud).

On-demand pricing allows purchase of instance-hours as and when they are needed. After a fixed annual (or higher triennial) payment per instance, reserved pricing is significantly cheaper per hour than on-demand pricing: it is intended for parties that know their requirements, and hence can reserve them, ahead of when they are used. Spot pricing is a short-term, market determined price used by EC2 to sell off unused capacity. In 2015, the "reserved" instances (which is of more interest for our purposes) were further divided into no, partial and all upfront payments, and discount of at least 5% is available for those spending at least half a million dollars. A Dedicated Host is a physical EC2 server dedicated to a user. Dedicated Hosts can help one reduce costs by allowing the use of one's existing server-bound software licenses.

For all instances and price bands, Windows usage is more expensive than Linux/Unix and is therefore not considered. Similarly, 32-bit instances are not considered, because for all large computing efforts and price bands 32-bit cores are at least as expensive as 64-bit ones. To enable a longitudinal study we restrict to the five remaining relevant instances which were similarly available in 2011, 2012, 2015 and 2018, and which are most suited to our needs. Tables 1 and 2 list the relevant technical specifications and 2011 to 2018 pricing of these instances. For all five instances under consideration k-fold multiples can be purchased at k times the price listed, for any positive integer k. Although of course there is an upper bound on k due to the size of the installation of the cloud service, to simplify our cost estimates we ignore this upper bound, and assume that the provider will provide more capacity (at the same cost) as demand increases. Separate charges are made for data transfer and so on, and the tables do not take into account the 5% discount for large purchases. With Y the number of hours per year, the tables are illustrated on a per-ECU cost basis in Figure 1, clearly indicating that cryptanalytic costs have dropped considerably since 2011.
With, for a given instance, δ, α, τ, ϵ_α, ϵ_τ the four pricing parameters as indicated in Table 2, it turns out that we have:

$$\left.\begin{array}{ll} 2011: & 1.962(\tau+3\epsilon_\tau Y) < 3\delta Y < 2.032(\tau+3\epsilon_\tau Y) \\ 2012: & 2.351(\tau+3\epsilon_\tau Y) < 3\delta Y < 3.489(\tau+3\epsilon_\tau Y) \\ 2015: & 2.166(\tau+3\epsilon_\tau Y) < 3\delta Y < 2.680(\tau+3\epsilon_\tau Y) \\ 2018: & 1.880(\tau+3\epsilon_\tau Y) < 3\delta Y < 2.772(\tau+3\epsilon_\tau Y) \end{array}\right\} \quad (1)$$

That is, for any instance using on-demand pricing continuously for a triennial period was approximately two times in 2011 (then three times in 2012 and 2.4 times in 2015 and in 2018 best approximated as 2.7 times) as

Table 1: Instance technical specifications (with 2 GPUs in 2011-2015 and 4 GPUs in 2018).

instance	2011 and 2012				2015				2018			
	cores	ECUs	GB RAM		cores	ECUs	GB RAM		cores	ECUs	GB RAM	
			total	per core			total	per core			total	per core
gen. purp. L	2	4u	7.5	3.75	2	6.5u	7.5	3.75	2	6.5u	8	4
high-mem. EL	2	6.5u	17.1	8.55	4	13u	30.5	7.625	4	13.5u	30.5	7.625
high-CPU EL	8	20u	7	0.875	8	31u	15	1.875	8	34u	16	2
cluster EL	unspecified	33.5u	23	unknown	16	55u	30	1.875	16	68u	32	2
GPU EL		33.5u	22		32	104u	60	1.875	32	94u	244	7.625

Table 2: US east coast instance pricing, in US dollars, using 64-bit Linux/Unix.

Instance	February 2011						February 2012					
	ECUs	on-dem. per hour δ	reserved				ECUs	on-dem. per hour δ	reserved			
			fixed payment		per hour				fixed payment		per hour	
			1 yr α	3 yr τ	1 yr ϵ_α	3 yr ϵ_τ			1 yr α	3 yr τ	1 yr ϵ_α	3 yr ϵ_τ
gen. purp. L	4u	0.34	910	1 400	0.120	0.120	4u	0.32	780	1 200	0.060	0.052
high-mem. EL	6.5u	0.50	1 325	2 000	0.170	0.170	6.5u	0.45	1 030	1 550	0.088	0.070
high-CPU EL	20u	0.68	1 820	2 800	0.240	0.240	20u	0.66	2 000	3 100	0.160	0.140
cluster EL	33.5u	1.60	4 290	6 590	0.560	0.560	33.5u	1.30	4 060	6 300	0.297	0.297
GPU EL	33.5u	2.10	5 630	8 650	0.740	0.740	33.5u	2.10	6 830	10 490	0.494	0.494

Instance	October 2015						June 2018					
	ECUs	on-dem. per hour δ	reserved				ECUs	on-dem. per hour δ	reserved			
			fixed payment		per hour				fixed payment		per hour	
			1 yr α	3 yr τ	1 yr ϵ_α	3 yr ϵ_τ			1 yr α	3 yr τ	1 yr ϵ_α	3 yr ϵ_τ
gen. purp. L	6.5u	0.133	421	673	0.035	0.030	6.5u	0.0928	470	917	0	0
high-mem. EL	13u	0.350	1 082	2 066	0.066	0.052	13.5u	0.2660	1 370	2 628	0	0
high-CPU EL	31u	0.441	1 243	2 388	0.141	0.092	34u	0.3400	1 764	3 224	0	0
cluster EL	55u	0.840	2 608	4 063	0.209	0.180	68u	0.7680	3 982	7 292	0	0
GPU EL	104u	2.600	9 224	25 228	0.568	0.240	94u	2.2800	12 723	31 867	0	0

expensive as using reserved pricing with a triennial term for the entire triennial period. Furthermore, for all instances reserved pricing for three consecutive (or parallel) annual periods was approximately 1.3 times in 2011 (then 1.5 times in 2012 and 1.45 times in 2015 and 1.42 times in 2018) as expensive as a single triennial period:

$$\left.\begin{array}{ll} 2011: & 1.292(\tau+3\epsilon_\tau Y) < 3(\alpha+\epsilon_\alpha Y) < 1.305(\tau+3\epsilon_\tau Y) \\ 2012: & 1.416(\tau+3\epsilon_\tau Y) < 3(\alpha+\epsilon_\alpha Y) < 1.594(\tau+3\epsilon_\tau Y) \\ 2015: & 1.350(\tau+3\epsilon_\tau Y) < 3(\alpha+\epsilon_\alpha Y) < 1.547(\tau+3\epsilon_\tau Y) \\ 2018: & 1.197(\tau+3\epsilon_\tau Y) < 3(\alpha+\epsilon_\alpha Y) < 1.642(\tau+3\epsilon_\tau Y) \end{array}\right\} \quad (2)$$

There are more pricing similarities between the various instances. Suppose that one copy of a given instance is used for a fixed number of hours h. We assume that h is known in advance and that the prices do not change during this period (cf. remark below on future developments). Which pricing band(s) should be used to obtain the lowest overall price depends on h in the following way. For small h use on-demand pricing, if h is larger than a first cross-over point γ_α but at most one year, reserved pricing with one year term must be used instead. Between one year and a second cross-over point γ_τ one should use reserved pricing with one year term for a year followed by on-demand pricing for the remaining $h-Y$ hours, but for longer periods up to three years one should use just reserved pricing with a triennial term. After that the pattern repeats. This holds for all instances, with the cross-over values varying little among them, as shown below.

The first cross-over value γ_α satisfies $\delta\gamma_\alpha = \alpha+\epsilon_\alpha\gamma_\alpha$ and thus $\gamma_\alpha = \frac{\alpha}{\delta-\epsilon_\alpha}$. The second satisfies $\alpha+\epsilon_\alpha Y + \delta(\gamma_\tau - Y) = \tau+\epsilon_\tau\gamma_\tau$ and thus $\gamma_\tau = \frac{(\delta-\epsilon_\alpha)Y+\tau-\alpha}{\delta-\epsilon_\tau}$. For the 2011 prices we find $\gamma_\alpha \approx 4100$ and $\gamma_\tau \approx Y+2200$ in all price instances. But for 2012, 2015 and 2018 there are more variations between the cut-off points.

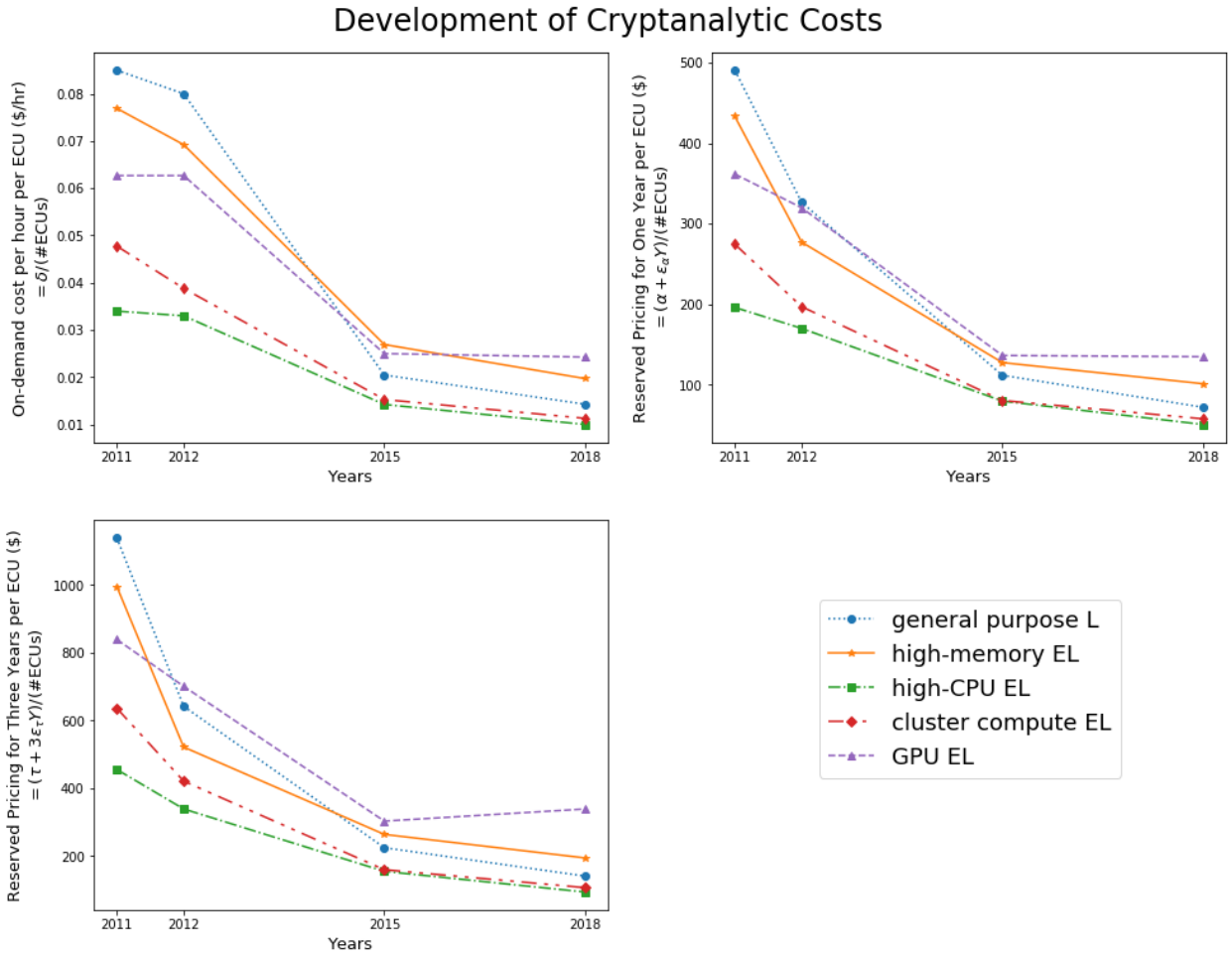

Fig. 1: Variation of EC2 pricing parameters over the years: on demand cost δ per hour, EC2 reserved pricing $\alpha + \epsilon_\alpha Y$ for one year, and EC2 reserved pricing $\tau + 3\epsilon_\tau Y$ for three years, all per ECU.

Our requirements and approach. For each cryptographic primitive studied, our approach hinges on the use of EC2 to carry out a negligible yet representative fraction of a certain computation. Said computation, when carried to completion, should result in a (symmetric- or private-) key or a collision depending on the type of primitive. For DES, AES, and SHA-2 this consists of a fraction of the symmetric-key or collision search, for RSA of small parts of the sieving step and the matrix step of the Number Field Sieve (NFS) integer factorisation method [23], and for ECC a small number of iterations of Pollard's rho method for the calculation of a certain discrete logarithm [32, 30].

Note that in each case, the full computation would require at least many thousands of years when executed on a single core[7]. With the exception of the NFS matrix step and disregarding details, each case allows embarrassing parallelisation with only occasional communication with a central server (for distribution of inputs and collection of outputs). With the exception of the NFS sieving and matrix steps, memory requirements are negligible. Substantial storage is required only at a single location. As such, storage needs

[7] The processor type can be left unspecified but quantum processors are excluded; to the best of our knowledge and consistent with almost all estimates (an exception can be found in [17]) of the last two decades, it will always take at least another decade before such processors are operational.

are thus not further discussed. Thus, modulo details and with one exception, anything we could compute on a single core in y years, can be calculated on n such cores in y/n years, for any $n > 0$.

Implementing a software component to perform each partial computation on EC2 requires relatively little upfront cost with respect to development time. In addition, execution of said software also requires relatively little time (compared to the full computation), and thus the partial computation can be performed using EC2's most appropriate pricing band for short-term use; this is the only actual cost incurred. Crucially, it results in a reliable estimate of the number of ECU years, the best instance(s) for the full computation, and least total cost (as charged by EC2) to do so (depending on the desired completion time). Obviously the latter cost(s) will be derived using the most appropriate applicable long-term pricing band.

Minimal average price. Let $\mu(h)$ denote the minimal average price per hour for a calculation that requires h hours using a certain fixed instance. Then we have

$$\mu(h) = \begin{cases} \delta & \text{for } 0 < h \leq \gamma_\alpha \quad \text{(constant global maximum)} \\ \frac{\alpha}{h} + \epsilon_\alpha & \text{for } \gamma_\alpha < h \leq Y \quad \text{(with a local minimum at } h = Y\text{)} \\ \frac{\alpha + \epsilon_\alpha Y + (h-Y)\delta}{h} & \text{for } Y < h \leq \gamma_\tau \quad \text{(with a local maximum at } h = \gamma_\tau\text{)} \\ \frac{\tau}{h} + \epsilon_\tau & \text{for } \gamma_\tau < h \leq 3Y \text{ (with the global minimum at } h = 3Y\text{).} \end{cases}$$

For $h > 3Y$ the pattern repeats, with each triennial period consisting of four segments, reaching decreasing local maxima at $h = 3kY + \gamma_\alpha$, decreasing local minima at $h = 3kY + Y$, decreasing local maxima at $h = 3kY + \gamma_\tau$, and the global minimum at $h = 3kY + 3Y$, for $k = 1, 2, 3, \ldots$. For all relevant instances, Figure 2 depicts the graphs of the minimal average prices for periods of up to six years, per ECU, in 2011, 2012, 2015 and 2018.

Consequences. Given the embarrassingly parallel nature and huge projected execution time of each full computation, the above implies that we can always reach lowest (projected) cost by settling for a three year (projected) completion time. Faster completion would become gradually more expensive until completion time $\frac{\tau}{\alpha}Y > Y$ is reached[8], at which point one should switch right away to a shorter completion time of one year; faster than one year again becomes gradually more expensive until γ_α is reached, at which point the cost has reached its (constant) maximum and the completion time is only limited (from below) by the number of available on-demand copies of the required instance. We stress yet again that this all refers to projected computation, none of which is actually completed: all we need is an estimate of the cost of completion, possibly depending on the desired completion time.

ECU years. In the remainder of the paper computational efforts are measured in "ECU years", i.e., the number of years required by the computation when using single ECU, and $\mathcal{Y}_\text{u}$ is used to indicate one ECU year. For relevant computations it is silently assumed that an effort of $k\mathcal{Y}_\text{u}$ may be done using ℓ ECUs for $\frac{k}{\ell}$ years, for any reasonable value of ℓ.

Table 3: Cost in US dollars per $\mathcal{Y}_\text{u}$ for large computing efforts, depending on completion time.

Instance	completion											
	as soon as possible				within one year				three years or more			
	2011	2012	2015	2018	2011	2012	2015	2018	2011	2012	2015	2018
high-mem. EL	673.85	606.46	235.85	172.60	432.95	277.06	127.70	101.48	331.67	173.83	88.01	64.84
high-CPU EL	297.84	289.08	124.62	87.60	196.12	170.08	79.94	51.88	151.79	112.99	51.67	31.61
cluster EL	418.39	339.94	133.79	98.94	274.50	198.86	80.71	58.56	212.01	140.35	53.29	35.75

Table 3 lists the data underlying the left most parts, the local minima at one year, and the global minima at three years of the graphs from Figure 2 for the three, for our purposes, most relevant EC2-instances: a large

[8] Note that $\frac{\tau}{\alpha} \approx 1.5$, so $\frac{\tau}{\alpha}Y$ is about a year and a half for all instances.

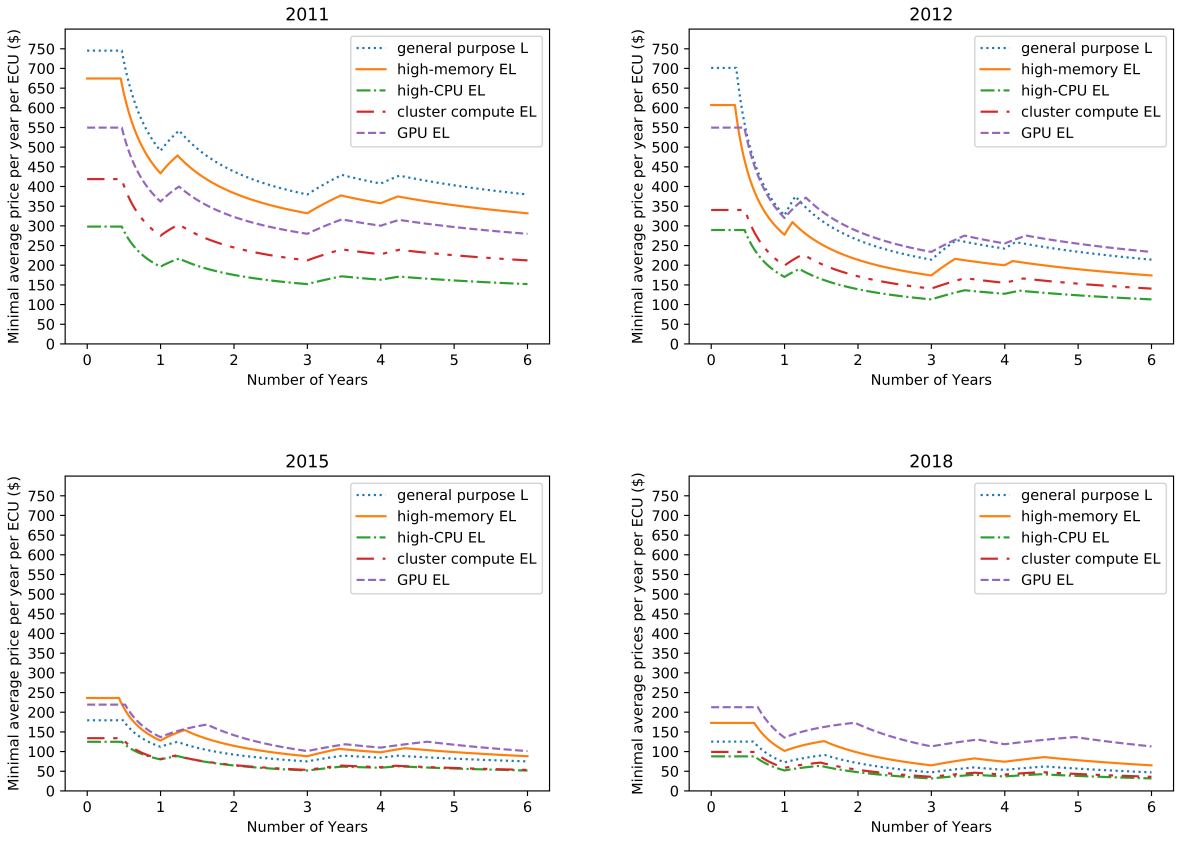

Fig. 2: Minimal average prices per year per ECU as a function of the number of years required by the calculation.

computational effort that would require $y_h \mathcal{Y}_u$ on a high-CPU EC2-instance would cost about y_h times \$31.61 if we can afford to wait three years until completion (and would require $\lceil \frac{y_h}{3 \cdot 34} \rceil$ three year reservations of that instance); for one year completion it increases to y_h times \$51.88 requiring $\lceil \frac{y_h}{34} \rceil$ one year reservations; or y_h times \$87.60 while trying to reserve as many on-demand instances as possible. If the same – or another – calculation requires $y_c \mathcal{Y}_u$ on a cluster compute EC2-instance, it would cost y_c times \$35.75 on $\lceil \frac{y_c}{3 \cdot 68} \rceil$ three year reserved cluster instances, y_c times \$58.56 on $\lceil \frac{y_c}{68} \rceil$ one year reserved cluster instances, or y_c times \$98.94 on as many on-demand instances as possible.

Note that cluster compute and high-memory EL instances are currently approximately 13% and 99% more expensive than high-CPU EL instances.

Accommodating future developments. As one can see prices and pricing models will change over time, and so may security assessment strategies and their interaction with advances in processor design and manufacture. In particular, one could imagine that if a party decided to use the Amazon cloud for key recovery or collision search then the increase in demand would induce Amazon to increase the instance costs. However, we assume that the effect of such supply and demand on the pricing is relatively constant over a long period of time. Thus, we assume the non-spot prices are a relatively accurate reflection of the actual economic cost to Amazon (bar a marginal profit) of providing the service.

The cost estimates produced by our approach are valid at the moment they are calculated (in the way set forth above), but cannot take any future developments into account. However, this problem can be mitigated

by adopting an open source model for the software components and using (as far as is sensible) platform agnostic programming techniques; for example, this permits the software to maintain alignment with the latest algorithmic and processor developments, and to add or remove primitives as and when appropriate (cf. [38]). Almost all our test software used has been made available on a web-site (which is currently inaccessible but) which will be updated, as years pass by, with the latest costs (the link will be given as soon as the data are accessible again).

EC2 versus Total Cost of Ownership (TCO). The approach set forth above associates a monetary cost to key strength, but does so at negligible actual cost; this is useful for many purposes. However, no key recovery nor collision is completed. The question remains, if one desires to complete a computation, whether doing so on EC2 is less expensive than acquiring a similar platform and operating it oneself.

TCO includes many costs that are hard to estimate. Nevertheless, the following may be useful. At moderate volume, a dual node server with two processors, each with twelve 1.9 GHz cores, and 32 GB of memory per node could be purchased for approximately \$8 000 in 2012. This implies that at that fixed cost approximately $2{\cdot}2{\cdot}12{\cdot}1.9 = 91.2$ℭ with 1.33 GB of memory per core can be purchased. At $\frac{20}{91.2}{\cdot}\$8000 \approx \$1\,750$ per 20ℭ this compares favourably to the fixed triennial payment $\tau = \$3\,100$ for the 20ℭ of EC2-instance high-CPU EL. Power consumption of the above server is estimated to be bounded by 600 Watts. Doubling this to account for cooling and so on, we arrive at approximately 265 Watts for 20ℭ, thus about a quarter kWh. At a residential rate of \$0.25 per kWh we find that, back in 2012, running our own 20ℭ for three years costs us $\$1\,750 + 3Y\frac{265}{1000} \cdot \$0.25 \approx \$3\,500$, as opposed to $\$3\,100 + 3Y \cdot \$0.14 \approx \$6\,779$ for EC2's high-CPU EL.

Since 2012, EC2's high-CPU EL three year cost has dropped by a factor of more than $3.5 \approx \frac{112.99}{31.61}$ per ECU (cf. Table 3), with cluster and high-memory instances per ECU dropping by factors of $3.9 \approx \frac{140.35}{35.75}$ and $2.7 \approx \frac{173.83}{68.84}$. Given the relative stability of hardware pricing, EC2 may turn out to become a serious contender for TCO. This is further illustrated by a current acquisition price of about \$10 000 for a 24 core, 2.5GHz, 256 GB RAM server (thus about 60ℭ and 10.66 GB RAM per core), with power consumption estimated to be bounded by 150 Watt, and thus an estimated three year ownership cost of $\$10\,000 + 3Y\frac{2{\cdot}150}{1000}{\cdot}\$0.25 \approx \$11\,971$. This is comparable to the cost $60{\cdot}3{\cdot}\$64.84 = \$11\,671.20$ of 60ℭ worth of high-memory EL three year instances (at 7.625 GB RAM per core; cf. tables 1 and 3).

3 Results

In this section we detail the application of our approach to five different cryptographic primitives: the block ciphers DES and AES, the cryptographic hash function SHA-2, and the public-key cryptosystems RSA and ECC. The first is chosen for historic reasons, whilst the others are the primitives of choice in many current applications.

For each of the five primitives, the fastest methods to recover the symmetric-key (DES and AES), to find a collision (SHA-2), or to derive the private-key (RSA and ECC) proceed very similarly, though realised using entirely different algorithms. In each algorithm, a huge number of identical computations are performed on different data; they can be carried out simultaneously on almost any number (and type) of client cores. With one exception (the NFS matrix step, as alluded to above), each client operates independently of all others, as long as there are central servers tasked with distributing inputs to clients and collecting their outputs. Furthermore, for each client the speed at which inputs are processed (and outputs produced, if relevant) is constant over time: a relatively short calculation per type of client along with a sufficiently accurate estimate of the number of inputs to be processed (or outputs to be produced, if relevant) suffices to be able to give a good indication of the total computational effort required.

The client-server approach has been common in a cryptographic context since the late 1980s, originally implemented using a variety of relatively crude application-specific, pre-cloud approaches [9, 24] (that continue to the present day [5, 19]), and later based on more general web-based services such as [8] that support collaborative compute projects (such as [29]). Thus, for each primitive under consideration, the problem of managing the servers is well understood. Additionally, the computational effort expended by said servers is dwarfed by the total computation required of the clients. As a result, in this section we concentrate on a series of experiments using the client software only: for each primitive, we execute the associated client

software for a short yet representative period of time on the most appropriate on-demand EC2 instance, use the results to extrapolate the total key retrieval cost using the corresponding reserved EC2 instance, and relate the result to a discussion of prior work.

We implemented a software component to perform each partial computation on EC2, focusing on the use of platform agnostic (i.e., processor non-specific) programming techniques via ANSI-C. In particular, we used processor-specific operations (e.g., to cope with carries efficiently) only in situations where an alternative in C was at least possible; in such cases we abstracted the operations into easily replaceable macros or used more portable approaches such as compiler intrinsics where possible. As motivated above, the goal of this was portability and hence repeatability; clearly one could produce incremental improvements by harnessing processor-specific knowledge (e.g., of instruction scheduling or register allocation), but equally clearly these are likely to produce improvement by only a (small) constant factor and may defeat said goal.

3.1 DES

We first examine DES (which is considered broken with current technology) to provide a base line EC2 cost against which the cost of other key retrieval efforts can be measured.

Prior work. As one of the oldest public cryptographic primitives, DES has had a considerable amount of literature devoted to its cryptanalysis over the years. Despite early misgivings and suspicions about its design and the development of several new cryptanalytic techniques, the most efficient way to recover a single DES key is still exhaustive search. This approach was most famously realised by *Deep Crack* developed by the EFF [14]. Designed and built in 1998 at a cost of $200 000, this device could find a single DES key in 22 hours. Various designs, often based on FPGAs, have been presented since specifically for DES key search. The most famous of these is COPACOBANA [15], which at a cost of $10 000 can perform single DES key search in 6.4 days on average. One can also extrapolate from suitable high-throughput DES implementations. For example [35] presents an FPGA design, which on a Spartan FPGA could perform an exhaustive single DES key search in 9.5 years. Using this design, in [41][p. 19] it is concluded that a DES key search device which can find one key every month can be produced for a cost of $750. Alternatively, using a time-memory trade-off [34], one can do a one-time precomputation at cost comparable to exhaustive key search, after which individual keys can be found at much lower cost: according to [33] a DES key can be found in half an hour on a $12 FPGA, after a precomputation that takes a week on a $12 000 device.

DES key search. In software the most efficient way to implement DES key search is to use the bit-sliced implementation method of Biham [6]. In our experiments we used software developed by Matthew Kwan[9] for the RSA symmetric-key challenge eventually solved by the DESCHALL project in 1997. Our choice was motivated by the goal of using platform agnostic software implementations of the best known algorithms.

Given a message/ciphertext pair the program searches through all the keys trying to find the matching key. On average one expects to try 2^{55} keys (i.e., one half of the key space) until a match is found. However in the bit-slice implementation on a 64-bit machine one evaluates DES on the message for 64 keys in parallel. In addition there are techniques, developed by Rocke Verser for the DESCHALL project, which allow one to perform an early abort if one knows a given set of 64 keys will not result in the target ciphertext. For comparison using processor extensions, we implemented the same algorithm on 128 keys in parallel using the SSE instructions on the x86 architecture.

DES key search on EC2. Using EC2 we found that a complete DES key search requires $97\mathcal{Y}_u$ using a vanilla 64-bit C implementation, and $51\mathcal{Y}_u$ using the 128-bit SSE implementation. With the high-CPU EL figures from Table 3 this leads to Table 4. The values are so low that the Amazon bulk discount has not been applied, however we see that the cost of obtaining a DES key has fallen by 38% since 2015.
In comparing these to earlier figures for special-purpose hardware, one needs to bear in mind that once a special-purpose hardware device has been designed and built, the additional cost of finding subsequent keys after the first one is essentially negligible (bar the maintenance and power costs). Thus, unless time-memory

[9] Available from `http://www.darkside.com.au/bitslice/`.

Table 4: Estimated EC2 DES key retrieval costs in thousand US dollars.

implementation technique	effort	2011			2012			2015			2018		
		ASAP	1 year	3 years	ASAP	1 year	3 years	ASAP	1 year	3 years	ASAP	1 year	3 years
vanilla C	$97\mathcal{Y}_u$	28.9	19.0	14.7	28.0	16.5	11.0	12.1	7.8	5.0	8.5	5.0	3.1
SSE version	$51\mathcal{Y}_u$	15.2	10.0	7.7	14.7	8.7	5.8	6.4	4.1	2.6	4.5	2.6	1.6

Table 5: Estimated EC2 AES-128 key retrieval costs in million US dollars.

effort	2012			2015			2018		
	ASAP	1 year	3 years	ASAP	1 year	3 years	ASAP	1 year	3 years
$10^{24}\mathcal{Y}_u$	$\approx 10^{20}$			$\approx 10^{20}$	$\approx 10^{19}$	$\approx 10^{19}$	$\approx 10^{19}$		

trade-off key search is used, the cost-per-key of specialised hardware is lower than using EC2. We repeat that our thesis is that dedicated hardware gives a point estimate, whereas our experiments are repeatable. Thus as long as our costs are scaled by an appropriate factor to take into account the possibility of improving specialised hardware, our estimates (when repeated annually) can form a more robust method of determining the cost of finding a key.

We end with noting that whilst few will admit to using DES in any application, the use of three-key triple DES (or 3DES) is widespread, especially in the financial sector. Because the above cost underestimates the cost of 2^{55} full DES encryptions (due to the early abort technique), multiplying it by 2^{112} lower bounds the cost for a full three-key 3DES key search (where the factor of 3 incurred by the three distinct DES calls can be omitted by properly ordering the search).

3.2 AES

Prior work. Since its adoption around fifteen years ago, AES has become the symmetric cipher of choice in many new applications. It comes in three variants, AES-128, AES-192, and AES-256, with 128-, 192-, and 256-bit keys, respectively. The new cipher turned out to have some unexpected properties in relation to software side-channels [31], which in turn triggered the development of AES-specific instruction set extensions [16]. Its strongest variant, AES-256, was shown to have vulnerabilities not shared with the others [7]. These developments notwithstanding, the only known approach to recover an AES key is by exhaustive search. With an AES-128 key space of 2^{128} this is well out of reach and therefore it is not surprising that there seems little work on AES specific hardware to realise a key search. One can of course extrapolate from efficient designs for AES implementation. For example using the FPGA design in [42] which on a single Spartan FPGA can perform the above exhaustive key search in $4.6 \cdot 10^{23}$ years, the authors of [41] estimate that a device can be built for $\$2.8 \cdot 10^{24}$ which will find an AES-128 key in one month.

AES key search on EC2. In software one can produce bit-slice versions of AES using the extended instruction sets available on many new processors [26]. However, in keeping with our principle of simple code, which can be run on multiple versions of today's computers as well as future computers, we decided to use a traditional AES implementation in our experiments: we found that a complete AES key search requires approximately $10^{24}\mathcal{Y}_u$. The resulting costs, using the high-CPU EL instances from Table 3 and rounded to the nearest order of magnitude, are listed in Table 5.

Again our comments for DES concerning the cost of specialised hardware versus our own estimates apply for the case of AES, although in the present case the estimates are more closely aligned. However, in the case of AES the new Westmere 32nm Intel core has special AES instructions [16]. It may be instructive to perform our analysis on the EC2 service, once such cores are available on this service[10], but the key retrieval cost will

[10] As of April 2012 none of the 64-bit instances we ran on the EC2 service had the Westmere 32nm Intel core on them.

remain astronomical: whilst a 3-to-10 fold performance improvement for AES encryption using Westmere has been reported, our own experiments on our local machines only show a two fold increase in performance for key search.

3.3 SHA-2

Prior work. The term SHA-2 denotes a family of four hash functions; SHA-224, SHA-256, SHA-384 and SHA-512. We shall be concentrating on SHA-256 and SHA-512; the SHA-224 algorithm only being introduced to make an algorithm compatible with 112-bit block ciphers and SHA-384 being just a truncated version of SHA-512. The three variants SHA-256, SHA-384 and SHA-512 were standardised by NIST in 2001, with SHA-224 being added in 2004, as part of FIPS PUB 180-2 [27]. The SHA-2 family of algorithms is of the same algorithmic lineage as MD4, MD5 and SHA-1.

Cryptographic hash functions need to satisfy a number of security properties; for example preimage-resistance, collision resistance, etc. The property which appears easiest to violate for earlier designs, and which generically is the least costly to circumvent, is that of collision resistance. Despite the work on cryptanalysis of the related hash functions MD4, MD5 and SHA-1 [44–47, 43], the best known methods to find collisions for the SHA-2 family still are the generic ones. Being of the Merkle-Damgård family each of the SHA-2 algorithms consists of a compression function, which maps b-bit inputs to h-bit outputs, and a chaining method. The chaining method is needed to allow the hashing of messages of more than b bits in length. The input block size $b = 512$ for SHA-256 but $b = 1024$ for SHA-384 and SHA-512. The output block size h is given by the name of the algorithm, i.e., SHA-h.

SHA-2 collision search. The best generic algorithm for collision search is the parallel "distinguished points" method of van Oorschot and Wiener [30]. In finding collisions this method can be tailored in various ways; for example one could try to obtain two meaningful messages which produce the same hash collision. In our implementation we settle for the simplest, and least costly, of all possible collision searches; namely to find a collision between two random messages.

The collision search proceeds as follows. Each client generates a random element $x_0 \in \{0,1\}^h$ and then computes the iterates $x_i = H(x_{i-1})$ where H is the hash function. When an iterate x_d meets a given condition, say the last 32-bits are zero, we call the iterate a distinguished point. The tuple (x_d, x_0, d) is returned to a central server and the client now generates a new value $x_0 \in \{0,1\}^h$ and repeats the process. Once the server finds two tuples (x_d, x_0, d) and $(y_{d'}, y_0, d')$ with $x_d = y_{d'}$ a collision in the hash function can be obtained by repeating the two walks from x_0 and y_0.

By the birthday paradox we will find a collision in roughly $\sqrt{\pi \cdot 2^{h-1}}$ applications of H, with n clients providing an n-fold speed up. If $1/p$ of the elements of $\{0,1\}^h$ are defined to be distinguished then the memory requirement of the server becomes $O(\sqrt{\pi \cdot 2^{h-1}}/p)$.

SHA-2 collision search on EC2. We implemented the client side of the above distinguished points algorithm for SHA-256 and SHA-512. This resulted in the $\mathcal{Y}_{\mathrm{u}}$ values listed in Table 6 and the associated collision finding costs, where again we use high-CPU EL instances and round to the nearest order of magnitude. Note, that SHA-256 collision search matches the cost of AES-128 key retrieval, as one hopes would happen for a well designed hash function of output twice the key size of a given well designed block cipher.

Table 6: Estimated EC2 SHA-2 collision search costs in million US dollars.

algorithm	effort	2012			2015			2018		
		ASAP	1 year	3 years	ASAP	1 year	3 years	ASAP	1 year	3 years
SHA-256	$10^{24}\mathcal{Y}_{\mathrm{u}}$		$\approx 10^{20}$		$\approx 10^{20}$	$\approx 10^{19}$	$\approx 10^{19}$		$\approx 10^{19}$	
SHA-512	$10^{63}\mathcal{Y}_{\mathrm{u}}$		$\approx 10^{59}$		$\approx 10^{59}$	$\approx 10^{58}$	$\approx 10^{58}$		$\approx 10^{58}$	

3.4 RSA

NFS background. As the oldest public key algorithm, RSA (and hence integer factorisation) has had a considerable amount of research applied to it over the years. The current best published algorithm for factoring integers is Coppersmith's variant [12] of the Number Field Sieve method (NFS) [23]. Based on loose heuristic arguments and asymptotically for $n \to \infty$, its expected run time to factor n is

$$L(n) = \exp((1.902 + o(1))(\log n)^{1/3}(\log\log n)^{2/3}),$$

where the logarithms are natural. For the basic version, i.e., not including Coppersmith's modifications, the constant "1.902" is replaced by "1.923". To better understand and appreciate our approach to get EC2-cost estimates for NFS, we need to know the main steps of NFS. We restrict ourselves to the basic version.

Polynomial selection. Depending on the RSA modulus n to be factored, select polynomials that determine the number fields to be used.

Sieving step. Find elements of the number fields that can be used to derive equations modulo n. Each equation corresponds to a sparse k-dimensional zero-one vector, for some $k \approx \sqrt{L(n)}$, such that each subset of vectors that sums to an all-even vector gives a 50% chance to factor n. Continue until at least $k + t$ equations have been found for a small constant $t > 0$ (hundreds, at most).

Matrix step. Find at least t independent subsets as above.

Square root. Try to factor n by processing the subsets (with probability of success $\geq 1 - (\frac{1}{2})^t$).

Because $L(n)$ number field elements in the sieving step have to be considered so as to find $k + t \approx \sqrt{L(n)}$ equations, the run time is attained by the sieving and matrix steps with memory requirements of both steps, and central storage requirements, behaving as $\sqrt{L(n)}$. The first step requires as little or as much time as one desires – see below. The run time of the final step behaves as $\sqrt{L(n)}$ with small memory needs.

The set of number field elements to be sieved can be parcelled out among any number of independent processors. Though each would require the same amount $\sqrt{L(n)}$ of memory, this sieving memory can be optimally shared by any number of threads; smaller memories can be catered for as well at small efficiency loss. Although all clients combined report a substantial amount of data to the server(s), the volume per client is low. The resulting data transfer expenses are thus not taken into account in our analysis below. The matrix step can be split up in a small number (dozens, at most) of simultaneous and independent calculations. Each of those demands fast inter-processor communication and quite a bit more memory than the sieving step (though the amounts are the same when expressed in terms of the above L-function).

It turns out that more sieving than necessary to obtain the required number of equations (which is easy, as sieving is done on independent processors) leads to a smaller k (which is advantageous, as it makes the matrix step less cumbersome). It has been repeatedly observed that this effect diminishes, but the trade-off has not been analysed yet. In fact the optimal amount of oversieving depends on the cluster configuration.

Unlike DES, AES, or ECC key retrieval or SHA-2 collision finding methods, NFS is a multi-stage method which makes it difficult to estimate its run time. As mentioned, the trade-off between sieving and matrix efforts is as yet unclear and compounded by the different platforms (with different EC2 costs) required for the two calculations. The overall effort is also heavily influenced by the properties of the set of polynomials that one manages to find in the first step. For so-called *special* composites finding the best polynomials is easy: in this case the special number field sieve applies (and the "1.902" or "1.923" above is replaced by "1.526"). For generic composites such as RSA moduli, the situation is not so clear. The polynomials can trivially be selected so that the (heuristic) theoretical NFS run time estimate is met. As it is fairly well understood how to predict a good upper bound for the sieving and matrix efforts given a set of polynomials, the overall NFS-effort can easily be upper bounded. In practice, however, this upper bound is too pessimistic and easily off by an order of magnitude. It turns out that one can quickly recognise if one set of polynomials is "better" than some other set, which makes it possible (combined with smart searching strategies that have been developed) to efficiently conduct a search for a "good" set of polynomials. This invariably leads to substantial savings in the subsequent steps, but it is not yet well understood how much effort needs to be invested in the search to achieve lowest overall run time.

The upshot is that one cannot expect that for any relevant n a shortened polynomial selection step will result in polynomials with properties representative for those one would find after a more extensive search. For the purposes of the present paper we address this issue by simply skipping polynomial selection, and by restricting the experiments reported below to a fixed set of moduli for which good or reasonable polynomials are known – and to offer others the possibility to improve on those choices. The fixed set that we consider consists of the k-bit RSA moduli RSA-k for $k = 768, 896, 1024, 2048$ as originally published on the now obsolete RSA Challenge list [36]. That we consider only these moduli does not affect general applicability of our cost estimates, if we make two assumptions: we assume that for RSA moduli of similar size NFS requires a similar effort, and that cost estimates for modulus sizes between 768 and 2048 bits other than those above follow by judicious application of the L-function. Examples are given below. We find it hazardous to attach significance to the results of extrapolation beyond 2048.

Prior work. Various special purpose hardware designs have been proposed for factoring most notably TWINKLE [39], TWIRL [40] and SHARK [13]. SHARK is speculated to do the NFS sieving step for RSA-1024 in one year at a total cost of one billion dollars. With the same run time estimate but only ten million dollars to build and twenty million to develop, plus the same costs for the matrix step, TWIRL would be more than two orders of magnitude less expensive. Not everyone agrees, however that TWIRL can be built and will perform as proposed.

In 1999 NFS was used to factor RSA-512 [10] using software running on commodity hardware. Using much improved software and a better set of polynomials the total effort required for this factorisation would now be about 3 months on a single 2.2GHz Opteron core; this is now considered to be a strightforward calculation and it is not included in our estimates. In 2009, this same software version of NFS (again running on regular servers) was used to factor RSA-768 [18]. The total effort of this last factorisation was less than 1700 years on a single 2.2GHz Opteron core with 2 GB of memory: about 40 years for polynomial selection, 1500 years for sieving, 155 years for the matrix, and on the order of hours for the final square root step. The matrix step was done on eight disjoint clusters, with its computationally least intensive but most memory demanding central stage done on a single cluster and requiring up to a TB of memory for a relatively brief period of time. According to [18], however, half the amount of sieving would have sufficed. Combined with the rough estimate that this would have doubled the matrix effort and based on our first assumption above, 1100 years on a 2.2GHz core will thus be our estimate for the factoring effort of *any* 768-bit RSA modulus. Note that the ratio $\frac{1100}{0.25} = 4400$ of the efforts for RSA-768 and RSA-512 is of the same order of magnitude as $\frac{L(2^{768})}{L(2^{512})} \approx 6150$ (twice omitting the "$o(1)$"), thus not providing strong evidence against our second assumption above.

Factoring on EC2. Based on the sieving and matrix programs used in [18] we developed two simplified pieces of software that perform the most relevant sieving and matrix calculations without outputting any other results than the time required for the calculations. Sieving parameters (such as polynomials defining the number fields, as described above) are provided for the fixed set of moduli RSA-768, RSA-896, RSA-1024 and RSA-2048. For RSA-768 they are identical to the parameters used to derive the timings reported above, for both steps resulting in realistic experiments with threading and memory requirements that can be met by EC2. For RSA-896 our parameters are expected to be reasonable, but can be improved, and RSA-896 is small enough to allow meaningful EC2 sieving experiments. For the largest two numbers our parameter choices allow considerable improvement (at possibly substantial computational effort spent on polynomial selection), but we do not expect that it is possible to find parameters for RSA-1024 or RSA-2048 that allow realistic sieving experiments on the current EC2: for RSA-1024 the best we may hope for would be a sieving experiment requiring several hours using on the order of 100 GB of memory, increasing to several years using petabytes of memory for RSA-2048.

The simplified matrix program uses parameters (such as the k value and the average number of non-zero entries per vector) corresponding to those for the sieving. It produces timing and cost estimates only for the first and third stage of the three stages of the matrix step. The central stage is omitted. For RSA-768 this results in realistic experiments that can be executed using an EC2 cluster instance (possibly with the exception of storage of the full matrix). For the other three moduli the estimated sizes are beyond the capacity

of EC2. It is even the case that at this point it is unclear to us how the central stage of the RSA-2048 matrix step should be performed at all: with the approach used for RSA-768 the cost of the central stage would by far dominate the overall cost, whereas for the other three moduli the central stage is known or expected to be negligible compared to the other two.

Tables 7 and 8 specify the most suitable ECU instance for each program and the four moduli under consideration so far, and the resulting timings and EC2 2012, 2015 and 2018 cost estimates. The estimates include 20% and 10% discounts wherever appropriate for 2012 and 2015 respectively, i.e., for one and three year costs of over two million dollars, and a 10% discount for 2018 for one and three year costs of over four million dollars. The figures in italics (RSA-896 matrix step, both steps for RSA-1024 and RSA-2048) are crude L-based extrapolations[11].

Table 7: Estimated EC2 NFS sieving step costs in million US dollars.

modulus	effort	instance	2012			2015			2018		
			ASAP	1 year	3 years	ASAP	1 year	3 years	ASAP	1 year	3 years
RSA-768	$1650\mathcal{Y}_u$	high-CPU EL	0.48	0.28	0.19	0.21	0.13	0.09	0.14	0.09	0.05
RSA-896	$1.5 \cdot 10^5\mathcal{Y}_u$	high-mem EL	91	33	21	35	17	11	26	14	9
RSA-1024	$2 \cdot 10^6\mathcal{Y}_u$	high-mem EL	*≈ 1200*	*≈ 440*	*≈ 280*	*≈ 470*	*≈ 230*	*≈ 160*	*≈ 350*	*≈ 180*	*≈ 120*
RSA-2048	$2 \cdot 10^{15}\mathcal{Y}_u$	high-mem EL	(*cost estimates for RSA-1024 multiplied by 10^9*)								

Table 8: Estimated EC2 NFS matrix step costs in million US dollars, using the cluster EL instance.

modulus	effort	2012			2015			2018		
		ASAP	1 year	3 years	ASAP	1 year	3 years	ASAP	1 year	3 years
RSA-768	$680\mathcal{Y}_u$	0.23	0.14	0.10	0.09	0.05	0.04	0.07	0.04	0.02
RSA-896	*$30000\mathcal{Y}_u$*	*10.2*	*4.8*	*3.4*	*4.0*	*2.2*	*1.4*	*3.0*	*1.8*	*1.1*
RSA-1024	*$8 \cdot 10^5\mathcal{Y}_u$*	*≈ 270*	*≈ 130*	*≈ 90*	*≈ 110*	*≈ 60*	*≈ 40*	*≈ 80*	*≈ 40*	*≈ 30*
RSA-2048	(*effort and cost estimates for RSA-1024 multiplied by 10^9*)									

The rough cost estimate for factoring a 1024-bit RSA modulus in one year is of the same order of magnitude as the SHARK cost, without incurring the SHARK development cost and while including the cost of the matrix step.

3.5 ECC

Prior Work. The security of ECC (Elliptic Curve Cryptography) relies on the hardness of the Elliptic Curve Discrete Logarithm Problem (EC-DLP). In 1997 Certicom issued a series of challenges of different security levels [11]. Each security level is defined by the number of bits in the group order of the elliptic curve. The Certicom challenges were over binary fields and large prime fields, and ranged from 79-bit to 359-bit curves. The curves are named with the following convention. ECCp-n refers to a curve over a large prime field with a group order of n bits, ECC2-n refers to a similar curve over a binary field, and ECC2K-n refers to a curve with additional structure (a so-called Koblitz curve) with group order of n bits over a binary field.

The smaller "exercise" challenges were solved quickly: in December 1997 and February 1998 ECCp-79 and ECCp-89 were solved using 52 and 716 machine days on a set of 500 MHz DEC Alpha workstations, followed in September 1999 by ECCp-97 in an estimated 6412 machine days on various platforms from different contributors. The first of the main challenges were solved in November 2002 (ECCp-109) and in

[11] We note the huge discrepancy between EC2-factoring cost and the monetary awards that used to be offered for the factorizations of these moduli [36].

April 2004 (ECC2-109). Since then no challenges have been solved, despite existing efforts to do so; a case in point is [5], which has not finished yet and the status of which is unknown to us.

ECC key search. The method for solving EC-DLP is based on Pollard's rho method [32]. Similar to SHA-2 collision search, it searches for a collision between two distinguished points. We first define a deterministic "random" walk on the group elements; each client starts such a walk, and then when they reach a distinguished point, the group element and some additional information is reported to a central server. Once the servers received two identical distinguished points one can solve the EC-DLP using the additional information. We refer to [30] for details.

ECC key search on EC2. Because most deployed ECC systems, including the recommended NIST curves, are over prime fields, we focus on elliptic curves defined over a field of prime order p. We took two sample sets: the Certicom challenges [11] which are defined over fields where p is a random prime (ECC-p-X), and the curves over prime fields defined in the NIST/SECG standards [28, 37], where p is a so-called generalised Mersenne prime (secpX-r1), thereby allowing more efficient field arithmetic. Of the latter we took the random curves over fields of cryptographically interesting sizes listed in the table below.

All of the curves were analysed with a program which used Montgomery arithmetic for its base field arithmetic. The NIST/SECG curves were also analysed using a program which used specialised arithmetic, saving essentially a factor of two. Table 9 summarises the costs of key retrieval using high-CPU EL instances in May 2012, October 2015 and June 2018, rounded to the nearest order of magnitude for the larger p. We present the costs for the small curves for comparison with the effort spent in the initial analysis over a decade ago. Note that general orders of magnitude correlate with what we expect in terms of costs related to AES, SHA-2, etc.

Table 9: Estimated EC2 ECC key retrieval costs in million US dollars.

curve name	effort	2012			2015			2018		
		ASAP	1 year	3 years	ASAP	1 year	3 years	ASAP	1 year	3 years
ECCp-109	$300\mathcal{Y}_u$	0.087	0.051	0.034	0.037	0.024	0.016	0.026	0.016	0.009
ECCp-131	$10^6\mathcal{Y}_u$	≈ 290	≈ 170	≈ 110	≈ 120	≈ 80	≈ 50	≈ 90	≈ 50	≈ 30
ECCp-163	$10^{10}\mathcal{Y}_u$	$\approx 10^6$			$\approx 10^6$			$\approx 10^6$		$\approx 10^5$
ECCp-191	$10^{15}\mathcal{Y}_u$	$\approx 10^{11}$			$\approx 10^{11}$			$\approx 10^{11}$		$\approx 10^{10}$
ECCp-239	$10^{22}\mathcal{Y}_u$	$\approx 10^{18}$			$\approx 10^{18}$			$\approx 10^{18}$		$\approx 10^{17}$
ECCp-359	$10^{40}\mathcal{Y}_u$	$\approx 10^{36}$			$\approx 10^{36}$			$\approx 10^{36}$		$\approx 10^{35}$
secp192-r1	$10^{15}\mathcal{Y}_u$	$\approx 10^{11}$			$\approx 10^{11}$			$\approx 10^{11}$		$\approx 10^{10}$
secp224-r1	$10^{20}\mathcal{Y}_u$	$\approx 10^{16}$			$\approx 10^{16}$			$\approx 10^{16}$		$\approx 10^{15}$
secp256-r1	$10^{25}\mathcal{Y}_u$	$\approx 10^{21}$			$\approx 10^{21}$			$\approx 10^{21}$		$\approx 10^{20}$
secp384-r1	$10^{44}\mathcal{Y}_u$	$\approx 10^{40}$			$\approx 10^{40}$			$\approx 10^{40}$		$\approx 10^{39}$

4 Extrapolate

Comparing the current pricing model to EC2's 2008 flat rate of \$0.10 per hour per ECU, we find that prices have dropped by a factor of almost 28 (namely $\frac{31.61}{24 \cdot 365 \cdot 0.1}$). This shaves off almost five bits of the security of block ciphers over about a decade, following closely (but one bit less than) what one would expect based on Moore's law. The most interesting contribution of this paper is that our approach allows anyone to measure and observe at what rate key erosion continues in the future. A trend that may appear after doing so for a number of years could lead to a variety of useful insights – not just concerning cryptographic key size selection but also with respect to the sanity of cloud computing pricing models. Of particular interest is our observation that EC2 is beginning to become competitive with TCO.

In future versions of this paper these issues will be further elaborated upon in this section.

5 Can one do better?

If by better one means can one reduce the overall costs of breaking each cipher or key size, then the answer is yes. This is for a number of reasons: Firstly one could find a different utility computing service which is cheaper; however we selected Amazon EC2 so as to be able to repeat the experiment each year on roughly the same platform. Any price differences which Amazon introduce due to falling commodity prices, or increased power prices, are then automatically fed into our estimates on a year basis. Since it is unlikely that Amazon will cease to exist in the near future we can with confidence assume that the EC2 service will exist in a year's time.

Secondly, we could improve our code by fine tuning the algorithms and adopting more efficient implementation techniques. We have deliberately tried not to do this. We want the code to be executed every few years on the platforms which EC2 provides, therefore highly specialised performance improvements have not been considered. General optimisation of the algorithm can always be performed and to enable this we have made the source code available on a public web site, the link to which will be made available as soon as possible. However, we have ruled out aggressive optimisations as they would only provide a constant improvement in performance and if costs to break a key are of the order of 10^{20} dollars then reducing this to 10^{18} dollars is unlikely to be that significant in the real world.

Finally, improvements can come from algorithmic breakthroughs. Although for all the algorithms we have discussed algorithmic breakthroughs have been somewhat lacking in the last decade or so, we intend to incorporate them in our code if they occur.

Acknowledgements. The work of the second and third author was supported by the Swiss National Science Foundation under grant numbers 200021-119776, 206021-128727, and 200020-132160. The work in this paper was partially funded by the European Commission through the ICT Programme under Contract ICT-2007-216676 ECRYPT II. The sixth author's research was also partially supported by a Royal Society Wolfson Merit Award and by the ERC through Advanced Grant ERC-2010-AdG-267188-CRIPTO.

References

1. Amazon Elastic Compute Cloud *Limited Beta*, July 2007, `http://web.archive.org/web/20070705164650rn_2/www.amazon.com/b?ie=UTF8&node=201590011`
2. Amazon Elastic Compute Cloud *Beta*, May 2008, `http://web.archive.org/web/20080501182549rn_2/www.amazon.com/EC2-AWS-Service-Pricing/b?ie=UTF8&node=201590011`.
3. Amazon Elastic Compute Cloud (Amazon EC2), `http://aws.amazon.com/ec2/`.
4. F. Bahr, M. Boehm, J. Franke and T. Kleinjung, *Subject: RSA200*. Announcement, 9 May 2005.
5. D.V. Bailey, L. Batina, D.J. Bernstein, P. Birkner, J.W. Bos, H.-C. Chen, C.-M. Cheng, G. van Damme, G. de Meulenaer, L.J.D. Perez, J. Fan, T. Güneysu, F. Gurkaynak, T. Kleinjung, T. Lange, N. Mentens, R. Niederhagen, C. Paar, F. Regazzoni, P. Schwabe, L. Uhsadel, A. Van Herrewege and B.-Y. Yang, *Breaking ECC2K-130*. Cryptology ePrint Archive, Report 2009/541, `http://eprint.iacr.org/2009/541`, 2009.
6. E. Biham. A fast new DES implementation in software. *Fast Software Encryption – FSE 1997*, Springer-Verlag, LNCS 1297, 260–272, 1997.
7. A. Biryukov, D. Khovratovich and I. Nikolić. Distinguisher and related-key attack on the full AES-256. *Advances in Cryptology – Crypto 2009*, Springer-Verlag, LNCS 5677, 231–249, 2009.
8. The BOINC project, `http://boinc.berkeley.edu/`.
9. T.R. Caron and R.D. Silverman. Parallel implementation of the quadratic sieve. *J. Supercomputing*, **1**, 273–290, 1988.
10. S. Cavallar, B. Dodson, A. K. Lenstra, P. Leyland, P. L. Montgomery, B. Murphy, H. te Riele, P. Zimmermann, et al. Factoring a 512-bit RSA modulus. *Advances in Cryptology – Eurocrypt 2000*, Springer-Verlag, LNCS 1807, 1–18, 2000.
11. Certicom Inc. *The Certicom ECC Challenge*. `http://www.certicom.com/index.php/the-certicom-ecc-challenge`.
12. D. Coppersmith. Modifications to the number field sieve. *J. of Cryptology*, **6**, 169–180, 1993.
13. J. Franke, T. Kleinjung, C. Paar, J. Pelzl, C. Pripalta and C. Stahlke. SHARK: A realizable special hardware sieving device for factoring 1024-bit integers. *Cryptographic Hardware and Embedded Systems – CHES 2005*, Springer LNCS 3659, 119–130, 2005.

14. J. Gilmore (Ed.). *Cracking DES: Secrets of Encryption Research, Wiretap Politics & Chip Design.* Electronic Frontier Foundation, O'Reilly & Associates, 1998.
15. T. Güneysu, T. Kasper, M. Novotný, C. Paar, and A. Rupp. Cryptanalysis with COPACOBANA. *IEEE Transactions on Computers*, **57**, 1498–1513, 2008.
16. S. Gueron. Intel's new AES instructions for enhanced performance and security. *Fast Software Encryption – FSE 2009* Springer LNCS 5665, 51–66, 2009.
17. G. Kalai. The quantum computer puzzle. Available at `http://www.ams.org/journals/notices/201605/rnoti-p508.pdf`. 2016.
18. T. Kleinjung, K. Aoki, J. Franke, A.K. Lenstra, E. Thomé, J.W. Bos, P. Gaudry, A. Kruppa, P.L. Montgomery, D.A. Osvik, H. te Riele, A. Timofeev and P. Zimmermann. Factorization of a 768-bit RSA modulus. *Advances in Cryptology – Crypto 2010*, Springer LNCS 6223, 333–350, 2010.
19. T. Kleinjung, J.W. Bos, A.K. Lenstra, D.A. Osvik, K. Aoki, S. Contini, J. Franke, E. Thomé, P. Jermini, M. Thiémard, P. Leyland, P.L. Montgomery, A. Timofeev and H. Stockinger. A heterogeneous computing environment to solve the 768-bit RSA challenge. *Cluster Computing*, **15**, 53–68, 2012.
20. T. Kleinjung, C. Diem, A.K. Lenstra, C. Priplata, C. Stahlke. Computation of a 768-bit prime field discrete logarithm. *Advances in Cryptology – Eurorypt 2017*, Springer LNCS 10210, 185–201, 2017.
21. A.K. Lenstra. Unbelievable security; matching AES security using public key systems. *Advances in Cryptology – Asiacrypt 2001*, Springer-Verlag LNCS 2248, 67–86, 2001.
22. A.K. Lenstra. Key Lengths. Chapter 114 of *The Handbook of Information Security*, Wiley 2005.
23. A.K. Lenstra and H.W. Lenstra, Jr. (eds.). *The development of the number field sieve.* Lecture Notes in Math. 1554, Springer-Verlag, 1993.
24. A.K. Lenstra and M.S. Manasse. Factoring by electronic mail. *Advances in Cryptology – Eurocrypt'89*, Springer-Verlag LNCS 434, 355–371, 1989.
25. A.K. Lenstra and E.R. Verheul, Selecting Cryptographic Key Sizes. *J. of Cryptology*, **14**, 255–293, 2001.
26. M. Matsui and J. Nakakima. On the power of bitslice implementation on Intel Core2 processor. *Cryptography Hardware and Embedded Systems – CHES 2007*, Springer-Verlag LNCS 4727, 121–134, 2007.
27. NIST. *Secure Hash Signature Standard (SHS) – FIPS PUB 180-2.* Available at `http://csrc.nist.gov/publications/fips/fips180-2/fips180-2.pdf`
28. NIST. *Digital Signature Standard (DSS) – FIPS PUB 186-2.* Available at `http://csrc.nist.gov/publications/fips/fips186-2/fips186-2-change1.pdf`
29. NFS@home,`http://escatter11.fullerton.edu/nfs`.
30. P.C. van Oorschot and M.J. Wiener. Parallel collision search with cryptanalytic applications. *J. of Cryptology*, **12**, 1–28, 1999.
31. D.A. Osvik, A. Shamir and E. Tromer. Efficient Cache Attacks on AES, and Countermeasures. *J. of Cryptology*, **23**, 37–71, 2010.
32. J. Pollard. Monte Carlo methods for index computation mod p, *Math. Comp.*, **32**, 918–924, 1978.
33. J.-J Quisquater and F. Standaert. Exhaustive key search of the DES: Updates and refinements. SHARCS 2005 (2005).
34. J.-J Quisquater and F. Standaert. Time-memory tradeoffs. *Encyclopedia of Cryptography and Security*, Springer-Verlag, 614–616, 2005.
35. G. Rouvroy, F.-X. Standaert, J.-J. Quisquarter and J.-D. Legat. Design strategies and modified descriptions to optimize cipher FPGA implementations: Fact and compact results for DES and Triple-DES. *ACM/SIGDA - Symposium on FPGAs*, 247–247, 2003.
36. The RSA challenge numbers, formerly on `http://www.rsa.com/rsalabs/node.asp?id=2093`, now on for instance `http://en.wikipedia.org/wiki/RSA_numbers`.
37. SECG. *Standards for Efficient Cryptography Group.* SEC2: Recommended Elliptic Curve Domain Parameters version 1.0, `http://www.secg.org`.
38. `http://csrc.nist.gov/groups/ST/hash/sha-3/`.
39. A. Shamir. Factoring large numbers with the TWINKLE device. Manuscript, 2000.
40. A. Shamir and E. Tromer. Factoring large numbers with the TWIRL device. *Advances in Cryptology – Crypto 2003*, Springer-Verlag LNCS 2729, 1–26, 2003.
41. N.P. Smart (Ed.). *ECRYPT II: Yearly report on algorithms and keysizes (2009-2010).* Available from `http://www.ecrypt.eu.org/documents/D.SPA.13.pdf`.
42. F.-X. Standaert, G. Rouvroy, J.-J. Quisquarter and J.-D. Legat. Efficient implementation of Rijndael encryption in reconfigurable hardware: Improvements and design tradeoffs. *Cryptography Hardware and Embedded Systems – CHES 2003*, Springer-Verlag LNCS 2779, 334–350, 2003.

43. M. Stevens, A. Sotirov, J. Appelbaum, A. Lenstra, D. Molnar, D.A. Osvik and B. de Weger. Short chosen-prefix collisions for MD5 and the creation of a rogue CA certificate. *Advances in Cryptology – Crypto 2009*, Springer-Verlag LNCS 5677, 55–69, 2009.
44. X. Wang, D. Feng, X. Lai and H. Yu. *Collisions for Hash Functions MD4, MD5, HAVAL-128 and RIPEMD*. Cryptology ePrint Archive, Report 2004/199, `http://eprint.iacr.org/2004/199`, 2004.
45. X. Wang, A. Yao and F. Yao, *New Collision Search for SHA-1*. Crypto 2005 Rump session, `www.iacr.org/conferences/crypto2005/r/2.pdf`.
46. X. Wang, Y.L. Yin and H. Yu. Finding Collisions in the Full SHA-1. *Advances in Cryptology – Crypto 2005*, Springer-Verlag LNCS 3621, 17–36, 2005.
47. X. Wang and H. Yu. How to Break MD5 and Other Hash Functions. *Advances in Cryptology – Eurocrypt 2005*, Springer-Verlag LNCS 3494, 19–35, 2005.