From: pqc-forum@list.nist.gov on behalf of Yang Bolin <yangbolin@zju.edu.cn>
Sent: Tuesday, August 11, 2020 9:26 AM
To: pqc-forum
Subject: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

Dear editor:

When I was doing some analysis on the reference implementation code from the CRYSTALS-KYBER, I found that the NTT function in this implementation is different from the traditional way. The variant “len”, which also means the distance in NTT, in the last round of their ntt is 2, rather than 1. So I wonder this is written by mistake or on purpose.

--

From:	'Bas Westerbaan' via pqc-forum <pqc-forum@list.nist.gov>
Sent:	Tuesday, August 25, 2020 5:00 AM
To:	Yang Bolin
Cc:	pqc-forum
Subject:	Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

This is intended, see lower part of page 6 of the round 2 spec of Kyber.

(The chosen field F does not contain a 512th root of unity. Thus X^256+1 does not split completely, but it does factor into degree two polynomials. So strictly speaking you can't do a regular NTT, but you can do one which is close enough. Instead of an efficient isomorphism from F[x] / (X^256+1) to F^256, you get one from F[x] / (X^256+1) to \prod_i F[x]/(X^2+zeta_i), for some particular zeta_i that are powers of the chosen 256th root of unity, which still allows you to speed up multiplication.)

On Tue, Aug 11, 2020 at 3:26 PM Yang Bolin <yangbolin@zju.edu.cn> wrote:

> Dear editor:
>
> When I was doing some analysis on the reference implementation code from the CRYSTALS-KYBER, I found that the NTT function in this implementation is different from the traditional way. The variant “len”, which also means the distance in NTT, in the last round of their ntt is 2, rather than 1. So I wonder this is written by mistake or on purpose.
>
> --

From:	pqc-forum@list.nist.gov on behalf of D. J. Bernstein <djb@cr.yp.to>
Sent:	Tuesday, December 1, 2020 4:10 AM
To:	pqc-comments@list.nist.gov
Cc:	pqc-forum
Subject:	[pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER
Attachments:	signature.asc

NIST's email dated 9 Jun 2020 15:39:09 +0000 states "we feel that the CoreSVP metric does indicate which lattice schemes are being more and less aggressive in setting their parameters".

Almost all lattice submissions have reported their Core-SVP levels (pre-quantum and post-quantum---let's focus on pre-quantum here), in line with this statement and with previous statements from NIST that seemed to be encouraging use of Core-SVP.

Question: What number does "the CoreSVP metric" assign to round-3 Kyber-512?

This might seem to be answered by Table 4 of the round-3 Kyber submission, which lists 2^118 for "Core-SVP" for round-3 Kyber-512. I have a clarification question here:

* Is the round-3 Kyber submission claiming that round-3 Kyber-512 is
  2^118 in "the CoreSVP metric", the metric that NIST says it's using
  to compare how "aggressive" lattice schemes are, the same metric
  used in other submissions?

My current understanding is that the answer is "no", meaning that this part of the round-3 Kyber submission needs to be disregarded for NIST's announced comparison mechanism, and instead there needs to be a new statement of the round-3 Kyber-512 Core-SVP level.

Here's how I arrived at this understanding. Please correct me if I've misunderstood something.

The round-2 Kyber submission listed an even smaller number, 2^111, as "Core-SVP" for round-2 Kyber-512. This doesn't directly contradict the idea that round-3 Kyber-512 reaches 2^118: the round-3 submission identifies changes from round-2 Kyber-512 to round-3 Kyber-512; perhaps these changes increase the Core-SVP level.

A more detailed reading appears to say, however, that these changes in the cryptosystem are _not_ enough to reach Core-SVP 2^118, and that the only way the round-3 Kyber submission is claiming 2^118 is by _changing the metric_, despite continuing to use the words "Core-SVP".

Specifically, there's a mechanism used in the previous literature for claiming "Core-SVP" levels, and then there's a different, more generous, mechanism used in the round-3 Kyber submission for claiming this 2^118.
For clarity, let's define "Kyber3-Modified-Core-SVP" as the mechanism used in the round-3 Kyber submission. Here are two examples of differences between Core-SVP and Kyber3-Modified-Core-SVP:

* In the literature, Core-SVP is the _minimum_ of primal and dual
  attack analyses. See, e.g., Table 3 of the round-2 Kyber
  submission, listing 2^111 as "core-SVP (classical)" for round-2
  Kyber-512, on the basis of Table 4 saying 2^111 for dual and 2^112
  for primal.

Kyber3-Modified-Core-SVP says "Primal attack only". This is not the same metric. Presumably the original Core-SVP metric would produce lower numbers than Kyber3-Modified-Core-SVP for round-3 Kyber-512.

Previous analyses showed some NISTPQC submissions choosing parameters in ranges where the dual component of Core-SVP was far above the primal component. Round-3 Kyber appears to be handling dual attacks in a more worrisome way, by changing the metric to disregard those attacks. A cryptosystem with several different attacks around the same security level has several different risks of those attacks being improved.

More to the point, whether or not one thinks dual attacks are as much of an issue as Core-SVP indicates, omitting dual attacks changes the metric away from Core-SVP.

* In the literature, Core-SVP for RLWE/MLWE-based systems is defined by 2n full samples (public multiples plus errors), whether or not the systems actually apply further rounding to those samples. See, e.g., the round-2 Kyber submission.

Kyber3-Modified-Core-SVP says that it "adds 6 bits of Core-SVP hardness" by "accounting for" rounding. This "accounting" is a change of the metric. The wording "adds 6 bits" appears to admit that round-3 Kyber-512 has Core-SVP at most 2^112. (Maybe even the same 2^111 as round-2 Kyber-512; see above regarding dual attacks.)

Note that this contradicts the claim in the round-3 Kyber submission that its "estimates of the security strength" for its "parameter sets" are "based on the cost estimates of attacks against the underlying module-learning-with-errors (MLWE) problem". This also means that Theorem 1, despite being labeled "tight", cannot justify the claimed round-3 Kyber-512 security level.

Yes, rounding poses a difficulty for attacks---it's not as if the full samples are provided to the attacker!---but certain people have previously criticized other submissions for focusing on the actual cryptosystem attack problems rather than the RLWE/MLWE problems. Also, certain people have been claiming that it's a problem if cryptosystem parameters provide less security in other cryptosystems; it's not hard to imagine users skipping the rounding since the idea of saving bandwidth in this way was first published and patented by Ding, two years before Peikert announced it as new.

More to the point, even if one doesn't think that this change of metric is reflecting a real danger, this isn't Core-SVP.

Labeling Kyber-Modified-Core-SVP as "Core-SVP" is confusing, and I don't see how it can be justified. Example of how the submitters could easily have predicted before the round-3 submission deadline that this labeling would cause problems:

* Presumably, at the beginning of round 3, NIST would compile a comparison table listing "the CoreSVP metric" for all parameter sets in all round-3 lattice submissions, to see "which lattice schemes are being more and less aggressive in setting their parameters".

* For round-3 Kyber-512, presumably NIST would take 2^118, since that's labeled as "Core-SVP" in the submission.

* In the absence of clarification, NIST would never realize that this 2^118 was calculated in a different way, and that "the CoreSVP metric" actually assigns a smaller number to round-3 Kyber-512.

This is unfair to various other submissions that---whether or not arguing that Core-SVP is flawed---have been reporting Core-SVP as per NIST's requests for comparability. The Kyber submission should have been reporting the original Core-SVP metric too, and giving any new metric a new name to avoid confusion.

I also have some questions for NIST at this point:

* Has NIST already made its round-3 Core-SVP comparison table? If not, why not, and what's the schedule for making this table?

  Assuming the table has been made already: Can you please post it publicly for review?

* NIST claimed in September 2020 that its public statements were "sufficient for any submission team working in good faith to determine what parameter sets will be uncontroversial, controversial and unacceptable for the claimed security levels given the current state of knowledge."

  I doubt anyone will assert that the Kyber-512 parameter set is "uncontroversial". But where is NIST drawing the line between "controversial" and "unacceptable"? Which side of the line was round-2 Kyber-512 on? Which side of the line is round-3 Kyber-512 on? How do we determine the answers to these questions from publicly available information? Also, just to confirm, NIST agrees that no version of Kyber-512 qualifies as "uncontroversial"?

  If NIST is unable to promptly answer these questions, shouldn't it be admitting for the record that the September 2020 claim quoted above wasn't true when it was made, and still isn't true now? Shouldn't it also be posting an analysis of how it ended up making such a claim, so as to help prevent similar errors in the future?

Looking forward to clarifications, answers, and retractions as appropriate.

---Dan

P.S. I should note---with all due respect---that all available evidence is consistent with the theory that NIST's strategy for handling concerns regarding the Kyber-512 security level is to adjust the NISTPQC security criteria so as to continue accepting the latest version of Kyber-512 (rather than suffering the public-relations problems of rejecting it).

If this theory is correct then evidently Kyber-512 isn't "unacceptable".
But NIST hasn't endorsed this theory, and it doesn't seem plausible that this unannounced strategy would have been the basis for NIST's claim that we were all supposed to have been able to figure out the dividing lines between "unacceptable", "controversial", and "uncontroversial".

--

From: 'Martin R. Albrecht' via pqc-forum <pqc-forum@list.nist.gov>
Sent: Tuesday, December 1, 2020 5:45 AM
To: pqc-forum
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

I'm confused: Core-SVP is a methodology for estimating the cost of blockwise lattice reduction algorithms like BKZ not a methodology for setting up lattices from LWE.

D. J. Bernstein <djb@cr.yp.to> writes:
> NIST's email dated 9 Jun 2020 15:39:09 +0000 states "we feel that the
> CoreSVP metric does indicate which lattice schemes are being more and
> less aggressive in setting their parameters".
>
> Almost all lattice submissions have reported their Core-SVP levels
> (pre-quantum and post-quantum---let's focus on pre-quantum here), in
> line with this statement and with previous statements from NIST that
> seemed to be encouraging use of Core-SVP.
>
> Question: What number does "the CoreSVP metric" assign to round-3
> Kyber-512?
>
> This might seem to be answered by Table 4 of the round-3 Kyber
> submission, which lists 2^118 for "Core-SVP" for round-3 Kyber-512. I
> have a clarification question here:
>
>   * Is the round-3 Kyber submission claiming that round-3 Kyber-512 is
>   2^118 in "the CoreSVP metric", the metric that NIST says it's using
>   to compare how "aggressive" lattice schemes are, the same metric
>   used in other submissions?
>
> My current understanding is that the answer is "no", meaning that this
> part of the round-3 Kyber submission needs to be disregarded for
> NIST's announced comparison mechanism, and instead there needs to be a
> new statement of the round-3 Kyber-512 Core-SVP level.
>
> Here's how I arrived at this understanding. Please correct me if I've
> misunderstood something.
>
> The round-2 Kyber submission listed an even smaller number, 2^111, as
> "Core-SVP" for round-2 Kyber-512. This doesn't directly contradict the
> idea that round-3 Kyber-512 reaches 2^118: the round-3 submission
> identifies changes from round-2 Kyber-512 to round-3 Kyber-512;
> perhaps these changes increase the Core-SVP level.
>
> A more detailed reading appears to say, however, that these changes in
> the cryptosystem are _not_ enough to reach Core-SVP 2^118, and that
> the only way the round-3 Kyber submission is claiming 2^118 is by
> _changing the metric_, despite continuing to use the words "Core-SVP".

From: pqc-forum@list.nist.gov on behalf of Christopher J Peikert <cpeikert@alum.mit.edu>
Sent: Tuesday, December 1, 2020 9:11 AM
To: pqc-forum; pqc-comments@list.nist.gov
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

I'll echo Martin; the central objection here doesn't make any sense to me. Core-SVP can be applied to a variety of lattice problems, including R/M/LWE with or without rounding. Increasing the amount of rounding will, all else being equal, tend to increase the Core-SVP hardness. This is not a change in the Core-SVP metric itself; it is a change in the lattice problem being analyzed under that metric.

> Yes, rounding poses a difficulty for attacks---it's not as if the
> full samples are provided to the attacker!---but certain people
> have previously criticized other submissions for focusing on the
> actual cryptosystem attack problems rather than the RLWE/MLWE
> problems. Also, certain people have been claiming that it's a
> problem if cryptosystem parameters provide less security in other
> cryptosystems;

Is this referring to my message on 17 September 2020 (https://groups.google.com/a/list.nist.gov/g/pqc-forum/c/LHQ308jHVF4/m/VvHaHPGxBgAJ) ?

If so, it (again) misrepresents my position, which is: "Given that variants of NIST-approved algorithms are likely to be adopted for such applications, I think it's very important to consider the robustness of the underlying LWE/LWR problems to variations like [revealing many samples]."

I don't recall anyone "criticizing other submissions for focusing on the actual cryptosystem attack problems rather than the RLWE/MLWE problems." Please provide unambiguous references, so that readers can tell whether you are accurately representing others, or putting words in their mouths.

> it's not hard to imagine users skipping the rounding
> since the idea of saving bandwidth in this way was first published
> and patented by Ding, two years before Peikert announced it as new.

Incorrect. The idea of reducing ciphertext size (thus saving bandwidth) by rounding away some low bits -- which is exactly what Kyber does -- had publicly appeared by September 2009, predating Ding's 2012 patent application by more than two years.

See, e.g., Section 4.2 of https://web.eecs.umich.edu/~cpeikert/pubs/svpcrypto.pdf , especially the Encaps algorithm and the preceding discussion: "When using a large value of q ... the efficiency of the prior schemes is suboptimal, because the plaintext-to-ciphertext expansion factor ... is at least lg q. Fortunately, it is possible to improve their efficiency (without sacrificing correctness) by discretizing the LWE distribution more 'coarsely' using a relatively small modulus q'."

Sincerely yours in cryptography,
Chris
--

From: pqc-forum@list.nist.gov on behalf of D. J. Bernstein <djb@cr.yp.to>
Sent: Tuesday, December 1, 2020 10:24 AM
To: pqc-forum
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER
Attachments: signature.asc

'Martin R. Albrecht' via pqc-forum writes:
> I’m confused: Core-SVP is a methodology for estimating the cost of
> blockwise lattice reduction algorithms like BKZ not a methodology for
> setting up lattices from LWE.

Then what exactly do you believe "the CoreSVP metric" is that NIST is using to evaluate "which lattice schemes are being more and less aggressive in setting their parameters"?

By its own words, this "CoreSVP metric" is a "metric" for "parameters".
If the thing you're calling "Core-SVP" refuses to turn an LWE instance into a lattice, then how could NIST have been using it to evaluate "parameters"?

If you're saying you're already confused by NIST's statement, then it's even more clear that there's something that needs resolution here.

If you _aren't_ confused by NIST's statement, then what exactly are you saying you _are_ confused about in my message? Do you claim that it's clear that the "metric" NIST is using for "parameters" gives 2^111 for
round-2 Kyber-512 and 2^118 for round-3 Kyber-512?

---Dan

--
You received this message because you are subscribed to the Google Groups "pqc-forum" group.
To unsubscribe from this group and stop receiving emails from it, send an email to pqc-forum+unsubscribe@list.nist.gov.
To view this discussion on the web visit https://groups.google.com/a/list.nist.gov/d/msgid/pqc-forum/20201201152423.691135.qmail%40cr.yp.to.

From: pqc-forum@list.nist.gov on behalf of Vadim Lyubashevsky <vadim.lyubash@gmail.com>
Sent: Tuesday, December 1, 2020 11:02 AM
To: pqc-forum; pqc-comments@list.nist.gov
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

Dear Dan, all,

> * In the literature, Core-SVP is the _minimum_ of primal and dual
> attack analyses. See, e.g., Table 3 of the round-2 Kyber
> submission, listing 2^111 as "core-SVP (classical)" for round-2
> Kyber-512, on the basis of Table 4 saying 2^111 for dual and 2^112
> for primal.
>
> Kyber3-Modified-Core-SVP says "Primal attack only". This is not the
> same metric. Presumably the original Core-SVP metric would produce
> lower numbers than Kyber3-Modified-Core-SVP for round-3 Kyber-512.

The version of "core-SVP" hardness that you enjoin we use is not the version used by other candidates, which actually contradicts your intent for a fair comparison on the same metric.

Indeed, the Kyber core-SVP-dual attack assumes (following the original NewHope paper) very conservatively that the SVP oracle provides exponentially many short vectors of the same length as the shortest one. This assumption has not reached consensus; for example the LWE-estimator of Albrecht et al. assumes a single vector, and concludes that the dual attack costs about 2^30 times the cost of the primal attack. This version of core-SVP-dual is the metric used by Saber, for example. The NTRU Round 3 specs do not even mention the dual attack.

Dismissing the dual attack therefore *aligns* the Kyber metrics with the other schemes rather than diverging from them. Given recent developments, we agree that the conservative assumptions for the dual attack are too unrealistic (as discussed in the "beyond core-SVP" section); hence our alignment to the literature.

For completeness, we note that with the conservative assumption, the dual attack has a core-SVP cost of 2^117, against the 2^118 we report for the primal attack. This is analogous to the Round 2 version where the Core-SVP costs were 2^111 and 2^112, respectively.

We must say that we are actually heartened that a disagreement over 1 bit of security provokes such passion in you. It certainly points to the maturity of the science behind lattice cryptanalysis when this is what's left to discuss :).

> Note that this contradicts the claim in the round-3 Kyber
> submission that its "estimates of the security strength" for its
> "parameter sets" are "based on the cost estimates of attacks
> against the underlying module-learning-with-errors (MLWE) problem".
> This also means that Theorem 1, despite being labeled "tight",
> cannot justify the claimed round-3 Kyber-512 security level.

On page 1 of the Round 3 changelog, we state "Relying on the rounding noise to add error is akin to the LWR assumption, but our reliance on it is quite small. First, it only adds 6 bits of Core-SVP hardness, and second, we are adding noise and rounding, which presumably has less algebraic structure than just rounding. In short, without the LWR assumption, our new parameter set for Kyber512 still has 112 bits of core-SVP hardness as before, while with a weak version of the LWR assumption, it has 118 bits." We believe that we are being very clear with what we are claiming for Kyber512 (Kyber768 and Kyber1024 are still based purely on MLWE as before).

> Yes, rounding poses a difficulty for attacks---it's not as if the
> full samples are provided to the attacker!---but certain people
> have previously criticized other submissions for focusing on the
> actual cryptosystem attack problems rather than the RLWE/MLWE
> problems. Also, certain people have been claiming that it's a
> problem if cryptosystem parameters provide less security in other
> cryptosystems; it's not hard to imagine users skipping the rounding

If users skip rounding, they're not using Kyber. Maybe a non-rounded version of Kyber-like schemes could be useful for something (e.g. it's less efficient for ZK proofs to prove knowledge of large errors), but those would probably have many other differences and will have to be independently analyzed anyway.

> P.S. I should note---with all due respect---that all available evidence
> is consistent with the theory that NIST's strategy for handling concerns
> regarding the Kyber-512 security level is to adjust the NISTPQC security
> criteria so as to continue accepting the latest version of Kyber-512
> (rather than suffering the public-relations problems of rejecting it).

We are not sure what you mean here. What adjustments did NIST make? At the end of Round 2, NIST publicly recommended that we should increase security a bit by widening the error distribution, and we did. As far as we are aware, this is all that has happened.

Best,

Vadim
(on behalf of the Kyber team)

> --
> You received this message because you are subscribed to the Google Groups "pqc-forum" group.
> To unsubscribe from this group and stop receiving emails from it, send an email to pqc-forum+unsubscribe@list.nist.gov.
> To view this discussion on the web visit https://groups.google.com/a/list.nist.gov/d/msgid/pqc-forum/20201201091007.670723.qmail%40cr.yp.to.

--
You received this message because you are subscribed to the Google Groups "pqc-forum" group.
To unsubscribe from this group and stop receiving emails from it, send an email to pqc-forum+unsubscribe@list.nist.gov.
To view this discussion on the web visit https://groups.google.com/a/list.nist.gov/d/msgid/pqc-forum/CAE158z_GRd9bZmw4bEC0%2BrSNwi%2BeUgQU_8y5XhZh-tuYxEaENg%40mail.gmail.com.

From: D. J. Bernstein <djb@cr.yp.to>
Sent: Wednesday, December 2, 2020 12:16 PM
To: pqc-comments
Cc: pqc-forum
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER
Attachments: signature.asc

In https://nvlpubs.nist.gov/nistpubs/ir/2020/NIST.IR.8309.pdf, NIST repeatedly refers to "Core-SVP" as a number attached to each lattice parameter set (e.g. "DILITHIUM has the lowest CoreSVP security strength parameter set of any of the lattice schemes still in the process"). I've also quoted NIST stating its belief that "the CoreSVP metric does indicate which lattice schemes are being more and less aggressive in setting their parameters". All of these quotes have been visible for months, along with many indications of their importance.

However, when I question the round-3 Kyber submission's claim that
round-3 Kyber-512 has Core-SVP 2^118, suddenly people jump in to attack the whole notion that Core-SVP attaches a number to each parameter set.
Why did none of the same people speak up to object to NIST repeatedly pointing to Core-SVP as a metric for lattice parameter sets, a metric used by NIST for comparisons? Why do these objections to the Core-SVP concept appear only when Kyber's Core-SVP claim is challenged?

Christopher J Peikert writes:
> Core-SVP can be applied to a variety of lattice problems, including
> R/M/LWE with or without rounding.

Please clarify. You're saying that "Core-SVP" takes "lattice problems"
(e.g., RLWE with specific parameters) as input?

Martin seems to be saying that Core-SVP _doesn't_ take a lattice problem as input (he says it's "not a methodology for setting up lattices from LWE"; he seems to indicate that it's simply the function mapping beta to (3/2)^(beta/2), or (13/9)^(beta/2) post-quantum), so you appear to be contradicting him rather than, as you claim, echoing him.

Meanwhile NIST's description of Core-SVP as being attached to each "parameter set" certainly doesn't match the type of input that Martin is claiming. It also doesn't match the type of input that you're claiming---it uses extra rules to specify the "problem" for a parameter set, because omitting such rules would break the idea of comparing parameter sets according to their Core-SVP values.

> "Given that variants of NIST-approved algorithms are likely to be
> adopted for such applications, I think it's very important to consider
> the robustness of the underlying LWE/LWR problems to variations like
> [revealing many samples]."

Why do you believe that the message you're quoting isn't an example of "claiming that it's a problem if cryptosystem parameters provide less security in other cryptosystems"? You've put quite a bit of effort into claiming that you're being misrepresented, without making clear to readers what exactly you're disagreeing with.

As a reminder, the context for the quote that you give included claims such as "Importantly, it appears that the requisite assumptions may be broken in some cases, so the resulting mKEM would be insecure" and "if the number of recipients exceeds ... then instantiating the mKEM with NTRU Prime will be insecure." (Whether the claims were correct isn't relevant here.) Surely you aren't going to try to argue that a claim of insecurity isn't claiming a problem.

> I don't recall anyone "criticizing other submissions for focusing on
> the actual cryptosystem attack problems rather than the RLWE/MLWE
> problems." Please provide unambiguous references

Here's an example:

https://groups.google.com/a/list.nist.gov/g/pqc-forum/c/V1RNjpio5Ng/m/uniJDvESBAAJ

This says, e.g., "one cannot directly rely on the Ring-LWE hardness assumption to argue security of the encryption procedure", and readers who check the context won't believe you if you try to argue that this isn't being presented as criticism.

Even with the fog of confusion that's suddenly blanketing Core-SVP, it seems reasonably clear that one can't rely on RLWE/MLWE hardness assumptions to argue for round-3 Kyber-512's 2^118 Core-SVP security claim. Perhaps one can rely on those assumptions to argue for a 2^112 Core-SVP security claim, but it seems very likely that 2^112 Core-SVP is below the minimum security level allowed in NISTPQC (even if 2^118 is above the minimum, which is far from clear at this point).

> it's not hard to imagine users skipping the rounding
> since the idea of saving bandwidth in this way was first published
> and patented by Ding, two years before Peikert announced it as new.
> Incorrect.

What I wrote was correct. "In this way" isn't spelling out the details, but anyone who looks for the cited publications will find

* your 2014 paper presenting a compact noisy-DH encryption scheme with compressed reconciliation (and claiming as its "main technical innovation" a "low-bandwidth" reconciliation technique that "reduces the ciphertext length" of previous "already compact" schemes "nearly twofold, at essentially no cost") and

* Ding's patent two years earlier on a compact noisy-DH encryption scheme with compressed reconciliation (the same 2x improvement in ciphertext size compared to LPR).

The 2012 version of the LPR paper (and talks going back to April 2010) presented a compact noisy-DH encryption scheme with reconciliation, but didn't compress the reconciliation. Quotient NTRU is much older and is as compact as compressed LPR, but it isn't a noisy-DH encryption scheme with reconciliation; some people have published one paper after another claiming that this is an important distinction, which led to a bunch of NISTPQC submissions using compact noisy-DH encryption with compressed reconciliation, and NTRU isn't prior art for a patent on that.

If it was so obvious that the LPR reconciliation could be compressed, why wasn't it already compressed in the LPR paper (which highlighted size as an issue), and why was your 2014 paper claiming that all previous schemes were 2x larger? It's hard enough to convince judges that things are obvious to people of ordinary skill in the art even when there _aren't_ subsequent papers claiming that those things are new!

Ding's patent isn't the only problem for Kyber (and SABER): there's the much earlier, February 2010, Gaborit--Aguilar Melchor patent that as far as I can tell covers the entire idea of compact noisy-DH encryption with reconciliation. This is what's usually called "LPR"---but the Eurocrypt
2010 (May 2010) version of the LPR paper, sent to Springer in February 2010, presented a much _bigger_ RLWE cryptosystem. The 2012 version of the paper switched to compact noisy-DH encryption with reconciliation.

If compact noisy-DH encryption with reconciliation was already an obvious cryptosystem from publications in 2009, then why do people credit it to 2010 LPR, and why did the original version of the LPR paper present a much bigger cryptosystem? Could it _possibly_ be that the authors didn't figure out the smaller cryptosystem until after sending the paper to Springer, and that what you claim years later to be obvious wasn't in fact obvious at the time? Why is a court supposed to believe that things are obvious when there are subsequent papers from acclaimed experts taking credit for these "innovations" or, even more extreme, presenting worse results?

> The idea of reducing ciphertext size (thus saving bandwidth) by
> rounding away some low bits -- which is exactly what Kyber does -- had
> publicly appeared by September 2009
  [...]
> See, e.g., Section 4.2 of https://web.eecs.umich.edu/~cpeikert/pubs/
> svpcrypto.pdf

That wasn't a "compact" cryptosystem, so it deviates from "what Kyber does" in a way that's essential for exactly the topic at hand.

Most patents, like most publications, are on combinations of previous ideas, and pointing out how various pieces appeared in previous work doesn't convince courts that the new combinations are obvious.

---Dan

From: Christopher J Peikert <cpeikert@alum.mit.edu>
Sent: Wednesday, December 2, 2020 2:34 PM
To: pqc-forum; pqc-comments
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

On Wed, Dec 2, 2020 at 12:15 PM D. J. Bernstein <djb@cr.yp.to> wrote:

> In https://nvlpubs.nist.gov/nistpubs/ir/2020/NIST.IR.8309.pdf, NIST repeatedly refers to "Core-SVP" as a number attached to each lattice parameter set (e.g. "DILITHIUM has the lowest CoreSVP security strength parameter set of any of the lattice schemes still in the process").

Dan, can you clarify whether you consider "amount of rounding" (e.g., number of low bits dropped) to be part of a lattice scheme's "parameter set"? I certainly do, and I think most others would too, since rounding is a form of additive "error."

As we know, different amounts of rounding will tend to yield different Core-SVP hardness numbers. Round-3 Kyber does a small amount of rounding that Round-2 Kyber didn't do; as one would expect, this slightly increased the associated Core-SVP hardness. What's the objection?

(If this thread exists only because of some semantic dispute about whether "amount of rounding" is part of the "parameter set" or not, I will be disappointed but not surprised.)

> However, when I question the round-3 Kyber submission's claim that round-3 Kyber-512 has Core-SVP 2^118, suddenly people jump in to attack the whole notion that Core-SVP attaches a number to each parameter set.

I don't see anybody disputing that notion. I see people (rightly) considering "amount of rounding" as part of Kyber's parameter set that is analyzed with Core-SVP.

> > I don't recall anyone "criticizing other submissions for focusing on the actual
> > cryptosystem attack problems rather than the RLWE/MLWE problems." Please
> > provide unambiguous references
>
> Here's an example:
>
> https://groups.google.com/a/list.nist.gov/g/pqc-forum/c/V1RNjpio5Ng/m/uniJDvESBAAJ
>
> This says, e.g., "one cannot directly rely on the Ring-LWE hardness assumption to argue security of the encryption procedure", and readers who check the context won't believe you if you try to argue that this isn't being presented as criticism.

Thanks for the reference. But I read it as a neutral observation -- as "Not sure what implications this remark has, though" makes clear -- so will file this as another example of you mischaracterizing others.

> \>   it's not hard to imagine users skipping the rounding
> \>   since the idea of saving bandwidth in this way was first published
> \>   and patented by Ding, two years before Peikert announced it as new.

> \> Incorrect.
>
> What I wrote was correct. "In this way" isn't spelling out the details,

What you wrote was not even close to correct, but it's neat to see you try to backtrack like this. "In this way" referred specifically to Round-3 Kyber rounding away low bits (of Z_q elements), which -- again -- is described in detail in the 2009 paper that predates that patent application by more than two years.

> \> The idea of reducing ciphertext size (thus saving bandwidth) by
> \> rounding away some low bits -- which is exactly what Kyber does -- had
> \> publicly appeared by September 2009
> [...]
> \> See, e.g., Section 4.2 of https://web.eecs.umich.edu/~cpeikert/pubs/
> \> svpcrypto.pdf
>
> That wasn't a "compact" cryptosystem, so it deviates from "what Kyber
> does" in a way that's essential for exactly the topic at hand.

No, Kyber's "compactness" -- obtained from its use of a (degree-256) polynomial ring -- is entirely orthogonal to its use of rounding, and hence irrelevant to your incorrect claim.

(There are plenty of constructions in the literature with every combination of "uses polynomial ring" or not, and "uses rounding" or not.)

> Ding's patent isn't the only problem for Kyber (and SABER): there's the
> much earlier, February 2010, Gaborit--Aguilar Melchor patent that as far
> as I can tell covers the entire idea of compact noisy-DH encryption with
> reconciliation.

If you really believe this, then you should lay out your reasoning, instead of making unjustified assertions.

Despite multiple attempts by different experts, I have never seen a demonstration of how Kyber/SABER could fall under that patent's claims -- every attempt failed to meet at least three central requirements.

So, I don't think anyone should credit the belief that the patent "covers the entire idea of compact noise-DH encryption with reconciliation," without a proper demonstration.

(Since this issue is quite far afield from the ostensible topic of this thread, I'll omit the technical details here, but am happy to share with whoever is interested.)

Sincerely yours in cryptography,
Chris

From: Vadim Lyubashevsky <vadim.lyubash@gmail.com>
Sent: Wednesday, December 2, 2020 2:47 PM
To: pqc-forum; pqc-comments
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

Hi Dan, all,

> The 2012 version of the LPR paper (and talks going back to April 2010)
> presented a compact noisy-DH encryption scheme with reconciliation, but
> didn't compress the reconciliation.

No version of the LPR paper presents "reconciliation". The LPR papers presented a *public key encryption* scheme. There is a very important distinction here. In a *key exchange scheme* using "reconciliation" (see e.g. Jintai's talk https://csrc.nist.gov/CSRC/media/Presentations/Ding-Key-Exchange/images-media/DING-KEY-EXCHANGE-April2018.pdf where the word reconciliation is explicitly used so there can be no confusion as to what it means), the users end up with a random shared key that neither party explicitly chose. In a public key encryption scheme, the sender chooses the message that he wants both parties to have. Of course one can trivially convert a key exchange scheme into an encryption scheme (by adding an xor with the message), but this is not what's happening in Kyber/Saber. You can see that there is a fundamental difference between the two approaches in the fact that there is slight bias in the shared key in Ding's scheme (see e.g. page 9 of https://patentimages.storage.googleapis.com/53/08/b7/b93d5b6b131e46/US9246675.pdf and consider when j=1 and t=2. Then an element y in the range [-(q-1)/2,(q-1)/2] is biased towards 0 when looking at y mod 2). So this bias would propagate itself into a public key encryption scheme if using an xor construction. But such a bias would simply never exist in a direct construction of a public key encryption scheme because the sender can pick the message from whatever distribution he wanted (e.g. an unbiased one). These are just different ways of doing things that achieve the same eventual goal (Ding's scheme is actually slightly more efficient as a CPA-KEM, by 1 bit per shared message bit; but is then equal in size to PKE with compression after being converted to a CCA-KEM).

Also notice that Jintai has a Ring-LWE encryption scheme (page 14 of https://patentimages.storage.googleapis.com/53/08/b7/b93d5b6b131e46/US9246675.pdf) which is like LPR and *does not* (unless I am reading something wrong) do any rounding / compression - so it just outputs two elements D1,D2 (which could be matrices over some ring).

-Vadim

From: D. J. Bernstein <djb@cr.yp.to>
Sent: Friday, December 4, 2020 12:06 PM
To: pqc-comments
Cc: pqc-forum
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER
Attachments: signature.asc

I would like the NISTPQC security requirements and security claims for
Kyber-512 to be stated clearly so that cryptanalysts who publish papers breaking Kyber-512---showing that it doesn't meet its security claims, or doesn't meet the NISTPQC security requirements, or both---don't have to worry about retroactive revisions of the claims and/or requirements.
This is an example of a scientific requirement called "falsifiability".

The gargantuan ambiguities in the NISTPQC security requirements are entirely under NIST's control, but the change in how round-3 Kyber calculates "Core-SVP" is a different story.

Each column on the "Estimate all the {LWE, NTRU} schemes!" page always led readers to believe that it was a clearly defined function mapping parameter sets to security claims. One of these functions from parameter sets to security claims was repeatedly labeled "Core-SVP" by NIST and others. None of the current commentators spoke up to claim that this was nonsense or that it had conflicting definitions.

Many submissions were already using Core-SVP, because they were simply copying what NewHope was doing (or at least they were trying to; we've seen failures, such as LightSABER miscalculating 2^125 when the correct calculation says 2^118). Other submissions pointed out inaccuracies in Core-SVP and made various efforts to reduce the inaccuracies, leading to a variety of different functions, usually larger than Core-SVP. The submissions prioritizing accuracy were then criticized for supposedly being less conservative and supposedly interfering with comparisons.

By the beginning of round 2, everyone was trying to report Core-SVP. For example, the round-2 NTRU Prime submission presented a detailed review of how the "Estimate" page computed this metric, and reported the output of this metric for each round-2 NTRU Prime parameter set. The submission _also_ presented the most comprehensive available survey of inaccuracies and potential inaccuracies in Core-SVP, and reported the results of alternative metrics aimed at addressing the three most glaring Core-SVP inaccuracies. All of the different metrics were clearly labeled.

How did NIST respond to this? By criticizing NTRU Prime for supposedly measuring "generic lattice security differently", and by asking whether the parameter sets "actually meet their claimed security categories"---a question that NIST did not ask regarding lower-security, more aggressive parameter sets in other proposals.

This brings me to Kyber-512. My current understanding is that the following three mechanisms, when applied to round-3 Kyber-512, produce the following "Core-SVP" numbers:

   * The mechanism used on the "Estimate" page: <=2^112 (see below).
   * The mechanism used in the round-2 Kyber submission: <=2^112.
   * The mechanism used in the round-3 Kyber submission: 2^118.

The reason for this change is that the round-3 Kyber submission switched to a new mechanism of mapping parameter sets to security levels, knowing that this mechanism is new, while continuing to prominently label the output as "Core-SVP". Procedurally, this labeling is an attack against NIST's announced use of "the CoreSVP metric" to compare "which lattice schemes are being more and less aggressive in setting their parameters".

Allowing this would not be fair to other submissions. Is NIST going to criticize Kyber for measuring "generic lattice security differently"? Is it going to ask Kyber to report Core-SVP by "the CoreSVP metric" that every other lattice submission was pretty much forced to use, meaning that round-3 Kyber-512 drops from 2^118 to <=2^112? Sure, Kyber argues that Core-SVP is an underestimate of _actual_ security, but submissions that already argued this in more detail were criticized for this and weren't given exemptions from NIST's announced comparison procedures.

Vadim Lyubashevsky writes:
> We must say that we are actually heartened that a disagreement over 1
> bit of security provokes such passion in you.

Given the attack literature, the claim that round-3 Kyber-512 meets the minimum NISTPQC security requirements is skating on _extremely_ thin ice. For example, regarding NIST's undefined set of "classical gates", round-3 Kyber-512 claims to be 8 bits harder to break than AES-128, plus _or minus_ 16 bits!

Clearly 1 bit of change, or the full 7 bits covered in my message, could push Kyber-512 below what NIST calls the "floor" of category 1. Clearly this is also why the Kyber submission spent so much effort on claiming these extra 7 bits: changing Kyber-512 to a new parameter set, changing to a different mechanism of computing its "Core-SVP" claims as noted above, and---in some ways the most amazing change---abandoning the tight link between Kyber and MLWE.

Many people seem to believe that the security levels of RLWE and MLWE are thoroughly understood (while the same people sometimes express doubts regarding the security levels of RLWR and MLWR). This supposed understanding, in turn, has been repeatedly portrayed as the core reason that users are supposed to trust the security claims for specific KEMs such as NewHope-512 and the three different versions of Kyber-512.
People who believe all this should be

* enthusiastic about the fact that, for RLWE/MLWE cryptosystems, the Core-SVP metric on parameter sets is actually an evaluation of the underlying RLWE/MLWE instances;

* disturbed by Kyber-512 choosing an MLWE size that doesn't seem to meet NISTPQC's minimum security requirements; and

* disturbed by Kyber-512 switching to a different metric---one that isn't evaluating the MLWE instances---to be able to claim to meet NISTPQC's minimum security requirements.

Maybe a complete attack analysis would show that a direct attack on Kyber-512 is below the "floor" anyway, in which case further security losses don't matter---but showing this would also need clarity from NIST regarding the definitions of the NISTPQC security requirements. We also don't have NIST's answer to the question of where Kyber-512 falls among "unacceptable" and "controversial" and "uncontroversial"; NIST claimed in September that we could all determine the dividing lines here from NIST's public statements.

> We believe that we are being very clear with what we are claiming for
> Kyber512

Does the round-3 Kyber submission claim that the MLWE instance inside round-3 Kyber-512 is as hard to break as AES-128?

This was claimed by the round-1 and round-2 Kyber submissions regarding round-1 Kyber-512 and round-2 Kyber-512, right? The category assignments and "Core-SVP" claims were hardness claims for those MLWE instances? Is the round-3 Kyber submission making this claim for round-3 Kyber-512?

If not, then is the Kyber team going to withdraw the statement in the
round-3 submission that the "estimates of the security strength" for
round-3 Kyber-512 are "based on the cost estimates of attacks against the underlying module-learning-with-errors (MLWE) problem"?

If not, how can this statement be reconciled with the 2^118 Core-SVP claim regarding round-3 Kyber-512?

Also, is the Kyber team withdrawing the claim that Theorem 1 is "tight"
for Kyber-512? If not, what exactly does this claim mean, and how can this be reconciled with the handling of Kyber-512?

If the MLWE instance inside Kyber-512 _is_ being claimed to meet the specified security level, then this claim needs to clearly stated so that cryptanalysts breaking it aren't faced with a subsequent "No, we never meant that". If the MLWE instance _isn't_ being claimed to meet the specified security level then various other claims regarding the relationship between Kyber and MLWE need to be clearly withdrawn.

Finally, just to confirm the numbers, am I correctly understanding the Kyber submission to be claiming that ignoring dual attacks increases "Core-SVP" for round-3 Kyber-512 from 2^111 to 2^112 (this is also a statement about the MLWE instance), and that accounting for rounding (which is what breaks the MLWE link) then increases 2^112 to 2^118? What is the "primal" block size that Kyber claims to be optimal for the MLWE problem, leading to the 2^112 claim? (Multiplying by 0.292 and rounding compresses multiple possible block sizes into one number, even if it's clear which rounding mechanism is being used.)

> If users skip rounding, they're not using Kyber.

I agree. However, if you take rounding into account for an RLWE/MLWE system, then you're breaking the supposedly tight, frequently advertised link between that system and the underlying RLWE/MLWE problem, and you're not using "the Core-SVP metric" that NIST said it's using for comparing parameters across lattice submissions.

> for example the LWE-estimator of Albrecht et al. assumes a single
> vector, and concludes that the dual attack costs about 2^30 times the
> cost of the primal attack

I agree that there's a conflict in the literature between multiple definitions of the dual component of Core-SVP, and that for some ranges of parameters this affects the final Core-SVP number, possibly by many bits (although not so many for Kyber). Thanks for pointing this out.

Perhaps this conflict has already corrupted NIST's use of Core-SVP to compare "which lattice schemes are being more and less aggressive in setting their parameters". Everyone should support (1) figuring out the effects of any previous confusion caused by conflicting definitions, and
(2) resolving the definitional conflict so that the conflict doesn't create unfair comparisons going forward.

Here are three possibilities for a unified definition:

(1) most favorable to the attacker, the assumption in round-2 Kyber:
assume sieving generates essentially as many short vectors as its
run time;

(2) assume some intermediate number of vectors;

(3) least favorable to the attacker, the "Estimate" assumption:

assume sieving generates only 1 short vector, the other vectors
being useless.

The round-3 Kyber submission argues briefly for #3. I'm puzzled by this argument, for several reasons.

First, is the Kyber team seriously claiming that dual attacks are as wimpy as indicated in #3? Isn't the actual situation for known attacks something in the #2 range---possibly much closer to #1 than to #3?

The statement in the submission that "most" vectors will be "sqrt(4/3) larger" doesn't contradict #2. "Most" doesn't mean "all but 1"; more importantly, these are probabilistic processes, and a larger number of longer vectors could do more damage than a smaller number of shorter vectors. Perhaps this can be quantified along the lines of analyses of Bleichenbacher's attack.

Second, regarding the claim that obtaining many short vectors is inconsistent with "dimensions for free": Let's assume, arguendo, that this claim is true and remains true, rather than being an artifact of the overload on cryptanalysts. Isn't the "dimensions for free" speedup asymptotically swamped by the many-vectors speedup? More precisely, isn't this subexponential vs. exponential? Is there a concrete claim that the asymptotics are misleading: that up through dimension d, covering all cryptographic dimensions, one should take "dimensions for free" and use only the shortest vector? Where is d quantified? Where is this claim justified?

In general, the way that Core-SVP was constructed was by taking the best asymptotics and replacing o(1) with 0 (a fundamentally flawed procedure for people who care about accuracy, but that's not the point here). The subexponential speedup from "dimensions for free" isn't visible in this process. The Kyber team wasn't calling earlier for the Core-SVP metric to be revised on the basis of "dimensions for free". Why are "dimensions for free" suddenly supposed to be taken into account in the definition?

Third, the whole push for every submission to use Core-SVP came from the general idea that Core-SVP is (relatively) simple but "conservative", meaning that it supposedly underestimates attack costs. Doesn't this mean that, if there's a conflict between two "Core-SVP" definitions, one definition more "conservative" and the other less "conservative", the less "conservative" definition should be eliminated?

> It certainly points to the maturity of the science behind lattice
> cryptanalysis when this is what's left to discuss :).

In case the above statement causes confusion despite the smiley, let me point to the graph in https://video.cr.yp.to/2020/0813/video.html as a quantified reason to be terrified of lattice-based cryptography. This graph reflects only one type of risk, and beyond this there are many further risks in lattice-based cryptography and specifically NISTPQC lattice KEMs, as illustrated by a variety of attack advances published _this year_, including very fast attacks against some supposedly safe systems. We haven't seen the end of this story.

For people trying to downplay lattice risks: Have you ever taken a quantitative approach to analysis of the risks of improved attacks, and used this quantification to compare lattices to other options? If not, aren't you concerned about the possibility of spreading misinformation, damaging allocation of precious cryptanalytic resources, damaging the processes for selecting cryptographic systems, and generally increasing security risks for cryptographic users?

Someone asking for clarity regarding Kyber-512's security claims shouldn't have this portrayed as implicitly denying other risks.

> P.S. I should note---with all due respect---that all available evidence
> is consistent with the theory that NIST's strategy for handling concerns
> regarding the Kyber-512 security level is to adjust the NISTPQC security
> criteria so as to continue accepting the latest version of Kyber-512

> (rather than suffering the public-relations problems of rejecting it).
> We are not sure what you mean here. What adjustments did NIST make?

For example, in August 2020, NIST made a preliminary (still not fully
defined) proposal to add memory costs, in violation of NIST's previously announced "minimum" criteria for metrics. For references and further analysis of this ongoing change in the NISTPQC evaluation criteria, see Section 5.5 of https://cr.yp.to/papers.html#categories.

Without this change from NIST, the Kyber statement

   We do not think that even a drop as large as 2^16 would be
   catastrophic, in particular given the massive memory requirements
   that are ignored in the gate-count metric

regarding round-3 Kyber-512 simply wouldn't be relevant to the NISTPQC evaluation process. Showing that known attacks use only 2^(151-16) =
2^135 "gates" (assuming NIST defines the "gates"!) would eliminate Kyber-512.

With NIST pushing enough extra "memory" factors into the definitions of the NISTPQC "categories", such attacks wouldn't eliminate Kyber-512.
Furthermore, if NIST never clearly defines the extra "memory" factors, or doesn't commit to leaving the definitions unchanged, then NIST is free to respond to even better attacks by inserting further "memory"
factors into the "category" definitions. A sufficiently large attack advance will put an end to this, presumably, but the minimum NISTPQC security requirements should have been nailed down years ago.

Kyber is far from unique in arguing that memory-access costs are important---of course they exist in reality!---but to exactly what extent are they included in the metrics that NIST uses to define the minimum security levels allowed in NISTPQC? The lack of an answer has led to the NISTPQC security requirements being continually interpreted in incompatible ways, with memory-is-expensive interpretations used to say that Kyber-512 is safe against known attacks, and memory-is-cheap interpretations used to criticize other submissions.

Most submissions have tremendous flexibility in parameter choices, and will be able to comfortably target whatever security levels NIST asks for---as soon as NIST _defines_ the security targets, rather than hiding behind pseudo-definitions with conflicting interpretations. Kyber is different, forced by its "framework" (which NIST has praised!) to make a big jump from 512 to 768. This limited flexibility should be treated negatively according to the NISTPQC criteria. If NIST were actually enforcing objective boundaries for its five "categories" then those boundaries would be likely to illustrate Kyber's limited flexibility.
Manipulating the boundaries can easily hide this, making Kyber look better than it actually is. See https://cr.yp.to/papers.html#categories.

---Dan

From: Christopher J Peikert <cpeikert@alum.mit.edu>
Sent: Friday, December 4, 2020 2:44 PM
To: pqc-forum; pqc-comments
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

Let's stipulate, consistent with NIST's statements, that Core-SVP can attach a number to each lattice scheme parameter set.

At the risk of repeating myself, it seems to me that Dan's entire objection about the Round-3 Kyber Core-SVP analysis is premised on *not* considering "amount of rounding" as part of a lattice scheme's "parameter set." If one *does* consider it as a parameter -- and one should! -- then I see no grounds for the objection. (Despite my request, Dan's long message did not offer any clarity on this central point; I think he should address it directly.)

For cryptanalytic purposes, ignoring rounding leaves out very important information, and can even produce perverse Core-SVP numbers.

For example, ignoring rounding would lead us to conclude that all of the NTRU Prime parameters have *trivial* Core-SVP hardness (~2^0), because NTRU Prime uses rounding alone for ciphertext "error"; without such rounding, the scheme would become trivially insecure to lattice attacks.

Of course, the NTRU Prime submission did *not* report trivial Core-SVP hardness, because the authors (Dan included) rightly included the rounding in their Core-SVP analysis. Obviously, other submissions should not be criticized for doing the same.

On Fri, Dec 4, 2020 at 12:06 PM D. J. Bernstein <djb@cr.yp.to> wrote:

> This brings me to Kyber-512. My current understanding is that the
> following three mechanisms, when applied to round-3 Kyber-512, produce
> the following "Core-SVP" numbers:
>
>   * The mechanism used on the "Estimate" page: <=2^112 (see below).
>   * The mechanism used in the round-2 Kyber submission: <=2^112.
>   * The mechanism used in the round-3 Kyber submission: 2^118.
>
> The reason for this change is that the round-3 Kyber submission switched
> to a new mechanism of mapping parameter sets to security levels,

I don't think this is accurate. Round-3 Kyber introduced rounding that was not present in its previous versions. The updated Core-SVP analysis reflected the existence of that rounding, presumably in a manner consistent with how other submissions had treated rounding. This is not a "new mechanism," it is the ordinary mechanism applied to new parameters.

> in some ways the most amazing change---abandoning the tight
> link between Kyber and MLWE.

Here there is an unstated premise that "MLWE" (and later, "the MLWE instance inside Kyber-512") does *not* include the rounding -- even though rounding is widely understood to be a form of error (the 'E' in MLWE). I don't agree with this premise, for the same reasons given above.

Of course, it's also well known that MLWE with extra rounding is at least as hard as MLWE *without* such rounding (all other parameters being equal), so one could also claim a tight reduction from decision-MLWE w/o rounding to, say, breaking the CPA security of the pre-FO version of Kyber.

I agree that it would be good to get a precise statement from the Kyber team concerning what they mean by "MLWE," and the consequences.

> Many people seem to believe that the security levels of RLWE and MLWE
> are thoroughly understood (while the same people sometimes express
> doubts regarding the security levels of RLWR and MLWR).

Again, please provide unambiguous references, so the reader can check whether you are accurately representing what "many people" "seem to believe" and express.

(The reader may wish to check previous messages in this thread regarding this issue.)

Sincerely yours in cryptography,
Chris

From: Vadim Lyubashevsky <vadim.lyubash@gmail.com>
Sent: Saturday, December 5, 2020 2:34 AM
To: Christopher J Peikert
Cc: pqc-forum; pqc-comments
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

Hi Chris,

Thank you for distilling that email into one question.

> Here there is an unstated premise that "MLWE" (and later, "the MLWE instance inside Kyber-512") does *not* include the rounding -- even though rounding is widely understood to be a form of error (the 'E' in MLWE). I don't agree with this premise, for the same reasons given above.
>
> Of course, it's also well known that MLWE with extra rounding is at least as hard as MLWE *without* such rounding (all other parameters being equal), so one could also claim a tight reduction from decision-MLWE w/o rounding to, say, breaking the CPA security of the pre-FO version of Kyber.
>
> I agree that it would be good to get a precise statement from the Kyber team concerning what they mean by "MLWE," and the consequences.

Relying just on MLWE, Kyber512 has CoreSVP of 112 with tight reductions and everything as before. If you're willing to accept that rounding also adds noise which contributes to hardness -- which does appear to be true according to today's cryptanalysis (and something that several proposals base all of their security on) -- then Kyber512 has CoreSVP of 118. While we felt somewhat uncomfortable basing the security of our scheme entirely on the MLWR assumption, we felt OK risking (a maximum of) 6 bits on it.

Best,

Vadim
(On behalf of the Kyber team)

From:	'dustin...@nist.gov' via pqc-forum <pqc-forum@list.nist.gov>
Sent:	Tuesday, December 8, 2020 4:12 PM
To:	pqc-forum
Cc:	vadim....@gmail.com; pqc-forum; cpei...@alum.mit.edu
Subject:	Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

All,

CoreSVP was the most widely used technique for estimating the security of lattice schemes at the end of Round 2. The differences between teams' methodologies in computing CoreSVP were not large enough to affect our decisions or the 2nd Round report. If, by the end of the 3rd Round, there are other, widely agreed-upon estimation techniques that better approximate real attack costs, then we will consider those instead. This is consistent with our prior statement that "*We have not and will not specify a set of ‘rules’ that must be used to evaluate the security of every candidate without regard to whether using these ‘rules’ would accurately reflect the level of security of every candidate.*"

The Kyber Round 3 specification provides estimates for gate cost in the RAM model of the best-known classical attacks against their updated parameters. These estimates exceed our estimates for the cost of attacking AES at each security category. Additionally, the Kyber team claims known quantum speedups are too small to be relevant for assessing categories 1, 3, and 5. If this analysis is correct, then Kyber clearly meets the security categories defined in the CFP. If the analysis is found to be incorrect, or if new attacks arise, then we will re-examine the situation.

This email serves to respond to process questions that arose in this thread. The merit of technical claims is a research matter for the community to address.

NIST PQC Team

On Saturday, December 5, 2020 at 2:34:28 AM UTC-5 vadim....@gmail.com wrote:

> Hi Chris,
>
> Thank you for distilling that email into one question.
>
> > Here there is an unstated premise that "MLWE" (and later, "the MLWE instance inside Kyber-512") does *not* include the rounding -- even though rounding is widely understood to be a form of error (the 'E' in MLWE). I don't agree with this premise, for the same reasons given above.
> >
> > Of course, it's also well known that MLWE with extra rounding is at least as hard as MLWE *without* such rounding (all other parameters being equal), so one could also claim a tight reduction from decision-MLWE w/o rounding to, say, breaking the CPA security of the pre-FO version of Kyber.
> >
> > I agree that it would be good to get a precise statement from the Kyber team concerning what they mean by "MLWE," and the consequences.
>
> Relying just on MLWE, Kyber512 has CoreSVP of 112 with tight reductions and everything as before. If you're willing to accept that rounding also adds noise which contributes to hardness -- which does appear to be true according to today's cryptanalysis (and something that several proposals base all of their security on) -- then Kyber512 has CoreSVP of 118. While we felt somewhat uncomfortable basing the security of our scheme entirely on the MLWR assumption, we felt OK risking (a maximum of) 6 bits on it.

> Best,
>
> Vadim
> (On behalf of the Kyber team)

--
You received this message because you are subscribed to the Google Groups "pqc-forum" group.
To unsubscribe from this group and stop receiving emails from it, send an email to pqc-forum+unsubscribe@list.nist.gov.
To view this discussion on the web visit https://groups.google.com/a/list.nist.gov/d/msgid/pqc-forum/bd511ae8-4f2c-4d0d-b346-a674a3732ed5n%40list.nist.gov.

From: pqc-forum@list.nist.gov on behalf of D. J. Bernstein <djb@cr.yp.to>
Sent: Friday, December 11, 2020 10:08 AM
To: pqc-comments
Cc: pqc-forum
Subject: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER
Attachments: signature.asc

Quoting the first entry in https://ntruprime.cr.yp.to/faq.html (from
October):

> There are known patent threats against the
> "Product NTRU"/"Ring-LWE"/"LPR" lattice proposals: Kyber, SABER, and
> NTRU LPRime (ntrulpr). These proposals use a "noisy DH +
> reconciliation" structure that appears to be covered by U.S. patent
> 9094189 expiring 2032, and a 2x ciphertext-compression mechanism that
> appears to be covered by U.S. patent 9246675 expiring 2033. There are
> also international patents, sometimes with different wording.

In the rest of this message I'll elaborate on these patent issues. I'm filing this as an OFFICIAL COMMENT for Kyber since all signals from NIST appear to be leaning towards Kyber, but for these issues I don't see any relevant differences among the LPR-based NISTPQC proposals.

Let me start by emphasizing procedures, starting with the role of patents in NISTPQC. The word "critical" appears exactly once in the NISTPQC call for proposals:

> NIST believes it is critical that this process leads to cryptographic
> standards that can be freely implemented in security technologies and
> products.

This is in Section 2.D, "Intellectual Property Statements / Agreements / Disclosures". NIST appears to have tried to collect statements from submitters regarding their own patents on their own submissions; this is helpful, and seems authoritative, but it doesn't make clear that a submission "can be freely implemented". Sometimes submissions are covered by patents from other people.

Patents are also included in the call for proposals under evaluation criterion 4.C.3, "Adoption", which broadly considers all factors "that might hinder or promote widespread adoption of an algorithm or implementation", and names "intellectual property" as an example. Again the statements from submitters regarding their own patents on their own submissions are not sufficient for evaluating this.

NISTPQC has already established a track record of mistakes even within the technical areas of expertise of the submitters and evaluators. It's not reasonable to imagine that evaluations of patent threats will have a zero error rate. It's important to have procedures in place to recognize and correct errors in evaluations of patent threats, starting with a rule of detailed public analyses.

As an analogy, NISTPQC efficiency claims are subjected to detailed public reviews, even when it's clear that the specific claims matter for only a narrow (and shrinking) corner of the user base. When two patents have been identified that can each singlehandedly destroy >99% of the potential usage of Kyber et al. between now and the early 2030s, we should be putting a correspondingly careful, publicly reviewed effort into establishing the magnitude and boundaries of the threat.

NIST IR 8309 says that if "intellectual property issues threaten the future of KYBER and SABER" then "NTRU would be seen as a more appealing finalist"---but hides its _reasons_ for saying this. Readers are misled into thinking this is a purely hypothetical issue. Readers who already know better aren't being given the opportunity to see and comment on the NIST handling of patent issues. Given the (obviously intentional) lack of transparency regarding such an important issue, I've filed a FOIA request for metadata regarding NIST's secret patent discussions, after careful consideration of the potential consequences of such a request.

Patent problems, like efficiency problems, are sometimes solved. There are occasional rumors of efforts to solve NISTPQC patent problems. This is _not_ an argument against public evaluations of the problems that currently exist. We should publicly evaluate the dangers to users _and_ publicly evaluate the chance of the dangers going away. If they go away, great; if they don't, we know how bad they are; either way, we're putting due diligence into understanding the issues.

I was disappointed to see a recent non-NIST message M on this list where the ending of M

(1) is the patent equivalent of a map stating "there are no landmines
here" and

(2) sounds like it's relying on discussions that, like NIST's patent
discussions, were never even _trying_ to meet the minimum
requirement of being published.

#1 isn't a problem per se but #2 is a problem. The rest of this message is organized as a reply to the patent comments in M, in the same order as M. (There are also non-patent topics in M, which I'll address in the original thread.)

This message includes what might be the first detailed public chart matching up the LPR cryptosystem to patent 9094189, which was filed 18 February 2010 and wasn't known to the community until eight years later.
In theory, patents are published; in reality, millions of hard-to-read patents operate as a denial-of-service attack against the general public.

The patent-statement requirement deserves credit for bringing this submarine patent to the attention of the community---but it did so only because the patent holders happened to be on other submissions, and it's completely missing the public analyses needed to establish which submissions can be "freely implemented".

> What you wrote was not even close to correct,

What I wrote was correct exactly as stated: "it's not hard to imagine users skipping the rounding since the idea of saving bandwidth in this way was first published and patented by Ding, two years before Peikert announced it as new".

> but it's neat to see you try to backtrack like this.

There's no backtracking.

> "In this way" referred specifically to Round-3 Kyber rounding away low
> bits (of Z_q elements)

No. "The rounding" is referring specifically to Kyber's rounding, but "in this way" is generalizing to what Ding published and patented. See, e.g., claims 4 and 5 of U.S. patent 9246675.

It's particularly important in patent discussions to be clear about levels of generality. When a patent has a claim with limitations X+Y+Z, meaning that it's claiming anything simultaneously doing X and Y and Z, you can pretty much always

find X and Y and Z separately in the prior art, so someone doing X+Y+Z is doing a special case of the X prior art and a special case of the Y prior art and a special case of the Z prior art---but this doesn't invalidate the patent. Even having X+Y and X+Z and Y+Z as prior art doesn't invalidate the patent. Meanwhile doing
X+Y+Z+A+B+C doesn't escape the patent, since it includes doing X+Y+Z.

> which -- again -- is described in detail in the 2009 paper that
> predates that patent application by more than two years.

No. Readers who follow references to a compression idea "first published and patented by Ding, two years before Peikert announced it as new" will see that these are 2012 Ding and 2014 Peikert respectively, and will find the following statement in 2014 Peikert:

As compared with the previous most efficient ring-LWE cryptosystems
and KEMs, the new reconciliation mechanism reduces the ciphertext
length by nearly a factor of two, because it replaces one of the
ciphertext's two R_q elements with an R_2 element.

This reduction in ciphertext size is such an important part of the paper that it's also summarized as the culmination of the abstract, as a consequence of a claimed "innovation" in the paper:

One of our main technical innovations (which may be of independent
interest) is a simple, low-bandwidth _reconciliation_ technique that
allows two parties who ``approximately agree'' on a secret value to
reach _exact_ agreement, a setting common to essentially all
lattice-based encryption schemes. Our technique reduces the
ciphertext length of prior (already compact) encryption schemes
nearly twofold, at essentially no cost.

The basic difficulty here is that Ding had already published and
(unfortunately) patented essentially the same cryptosystem in 2012, using essentially the same technique. Courts don't care about minor
differences: the "doctrine of equivalents" asks whether the accused device performs "substantially" the same function in "substantially" the same way to obtain the same result. (This is the wording in U.S. courts, but similar principles apply internationally.)

Readers seeing this claim from 2014 Peikert can't logically rule out the possibility of essentially the same cryptosystem already appearing in
2009 Peikert, but anyone checking 2009 Peikert will see that there's nothing in that paper anywhere near this level of efficiency. The fact that one can point to various features of the cryptosystem that already appeared in the 2009 paper---rounding, noisy multiples, etc.---doesn't kill the patent.

Here's a summary of how a court will evaluate and reject the argument that this 2009 Peikert cryptosystem invalidates claim 4 of U.S. patent
9246675:

* As a preliminary step, the lawyers argue about what exactly the
  words in the patent mean. Definitions are settled in enough detail
  to evaluate the prior art (and the claimed infringement).

* The defendant's lawyers argue that the 2009 cryptosystem meets all
  the limitations of claim 4 of the patent, so the claimed invention
  isn't novel.

In response, the plaintiff's lawyers have experts testify that the 2009 cryptosystem doesn't have the "ring R_q=F_q[x]/f(x) with f(x)=x^n+1" from claim 4.

To eliminate equivalence arguments, the plaintiff's experts also testify that the level of efficiency reached by claim 4 of the patent (and mentioned in the patent) is just one R_q element for keys and slightly larger for ciphertexts, far better than the level of efficiency reached by the 2009 cryptosystem (and mentioned in the 2009 paper). This is a winning argument; patent courts understand the concept of efficiency.

* The defendant's laywers argue that the claimed invention is obvious: specifically, that something covered by claim 4 of the patent was, at the time of filing of the patent, obvious to someone of ordinary skill in the art, given the 2009 cryptosystem and other publications available before the patent was filed.

  In response, the plaintiff's lawyers have a slam dunk: if the cryptosystems covered by claim 4 were obvious to someone of ordinary skill in the art in 2012, how could they not have been obvious to the world-renowned author of a 2014 paper claiming that nothing before 2014 had achieved this level of efficiency? (Not to mention the reviewers of the paper.)

  The defendant's lawyers will try to escape this, maybe even paying the author to testify that the 2014 paper actually meant something else, and that really everything in the patent was obvious in 2009 or 2010 or 2011. The plaintiff's lawyers will spend money on other experts saying the opposite. Patent courts are facing this sort of battle all the time, and---to make a long story short---basically always rule against obviousness _unless_ there's a slam-dunk argument _for_ obviousness, which is the opposite of the situation here.

There are various sources of randomness in this process, but, given what 2014 Peikert says, the defendant's lawyers will expect to lose the obviousness argument, and will be desperate to find a pre-2012 lattice cryptosystem that's as small as Ding's cryptosystem. The 2009 cryptosystem clearly doesn't do the job here.

> No, Kyber's "compactness" -- obtained from its use of a (degree-256)
> polynomial ring -- is entirely orthogonal to its use of rounding

1. The word "No" here falsely claims a contradiction between the orthogonality statement and the correct statement it was responding to.

2. Regarding the orthogonality statement: Courts understand the idea of interchangeable parts, but they're also faced with a constant stream of defendants claiming that _clearly_ the components inside the patent are interchangeable, while being unable to point to prior art _saying_ that the components are interchangeable. The winning words for the plaintiff have been repeated thousands of times: the most important advances often come from people simply having

the cleverness to put together two things that hadn't been combined before, it's easy to claim in hindsight that this was obvious but much harder to see it the first time, etc.

> (There are plenty of constructions in the literature with every
> combination of "uses polynomial ring" or not, and "uses rounding" or
> not.)

Claim 4 of Ding's patent has one party sending one element of R_q = F_q[x]/(x^n+1)---let's call this the "key"---and the other party sending one element of R_q plus slightly more information---let's call this the "ciphertext".

I'm not aware of any prior lattice systems that are so small. Are you?

Original LPR was one R_q element for the key but two for the ciphertext.
As far as I can tell, compressed LPR was Ding's idea. Pointing to compressed versions of _larger_ cryptosystems isn't a scientifically valid argument to refuse to credit him, and, more to the point, won't work in court.

> > Ding's patent isn't the only problem for Kyber (and SABER): there's
> > the much earlier, February 2010, Gaborit--Aguilar Melchor patent
> > that as far as I can tell covers the entire idea of compact noisy-DH
> > encryption with reconciliation.
> If you really believe this, then you should lay out your reasoning,
> instead of making unjustified assertions.

Many readers will interpret the word "unjustified" here as saying that there was a request for justification. There had, however, been no such request before the above statement.

Anyway, now that there's a request for justification (however poorly worded the request might be), let's go through the details.

There are a few choices of details at this point, since there are some differences in the members of the patent family. For definiteness let's take European Patent 2537284; my reason to pick Europe instead of the U.S. here is that the European patent has already survived one round of litigation, whereas I haven't heard about any litigation yet regarding the U.S. patent. (I don't expect the U.S. patent to be invalidated, but, all else being equal, it's reasonable to estimate the ultimate invalidation chance as being even lower for a patent that some people have tried and so far failed to invalidate.)

Within this patent, let's look at Claim 19, and go through the exercise of matching up the limitations of the claim to the LPR cryptosystem:

* "Cryptographic method": yes;

* "according to any one of Claims 1 to 18": let's pick Claim 1;

* "in which the ring R is the ring F_q[x]/(X^n-1)"---LPR emphasizes different rings, but the patent description says that one can take other rings and gives various examples, so LPR will be covered by (European versions of) the doctrine of equivalents;

* "in other words the set of polynomials with coefficients in the body F_q with q elements for which the remainder by division with the polynomial (X^n - 1) is considered"---sure, this is what we assumed they meant anyway by F_q[x]/(X^n-1).

The remaining task is to match up the limitations of Claim 1 to the LPR cryptosystem:

* "Cryptographic method": yes;

* "for communicating a confidential piece of information m"---yes; at this point m could be either the key or a subsequent user message;

* "between a first electronic entity (A)"---let's assume that in the LPR example this will be the key generator;

* "and a second electronic entity (B)"---the sender;

* "comprising a distribution step and a reconciliation step"---we'll check details below;

* "the distribution step comprising several steps consisting in that"---we'll check details below;

* "on the one hand, the first entity (A): calculates a first syndrome S_A = X_A + f(Y_A)"---to match this to LPR, let's take Y_A as the LPR secret, f() as multiplying by some public random ring element, and X_A as the LPR error;

* "based on a first secret piece of information composed of two primary elements X_A and Y_A"---yes, the LPR secret Y_A and the LPR error X_A are both secrets;

* "belonging to a ring R"---yes;

* "and having a norm that is substantially small relative to an element f(X_A) or f(Y_A)"---yes, the LPR secret Y_A and the LPR error X_A are both practically guaranteed to be small compared to the big multiples f(X_A) and f(Y_A);

* "the ring R having addition and multiplication"---yes;

* "f being an internal composition law associating with any element X_I of the ring R, another element f(X_I) of the ring R"---yes, multiplying by a public ring element does this;

* "and having the property that, for any pair of elements X_I and Y_I of R, such that X_I and Y_I have a norm that is small relative to the elements f(X_I) and f(Y_I), then X_I.f(Y_I) - Y_I.f(X_I) has a small norm"---sounds like zero, which is definitely small;

* "and generates a first message composed from this first syndrome S_A"---yes, the LPR public key is sent as part of a message;

* "such that the said first syndrome S_A is accessible by the second entity (B)"---yes, the LPR sender sees the public key;

* "on the other hand, the second entity (B): calculates a second syndrome S_B = X_B + f(Y_B)"---yes, the LPR sender has another secret Y_B multiplied by the same public ring element, and another error X_B;

* "based on a second secret piece of information composed of two secondary elements X_B and Y_B"---yes, the LPR sender secret Y_B and the LPR sender error X_B are both secrets;

* "belonging to the ring R"---yes;

* "and having a norm that is substantially small relative to an element f(X_B) or f(Y_B)"---yes, same situation as the other side;

* "transmits to the first entity (A) a second message"---yes, this is the LPR ciphertext;

* "composed from the second syndrome S_B"---yes, the ciphertext includes the sender's noisy multiple S_B;

* "such that the said second syndrome S_B is accessible by the first entity (A)"---yes, the LPR receiver sees the ciphertext;

* "characterized in that, during this first distribution step, the first entity (A) and the second entity (B) respectively calculate a first intermediate value P_A and a second intermediate value P_B, such that: P_A = Y_A.S_B = Y_A.X_B + Y_A.f(Y_B)"---yes, the LPR receiver multiplies the secret Y_A by the ciphertext component S_B, and the second equation is just the definition of S_B;

* "and P_B = Y_B.S_A = Y_B.X_A + Y_B.f(Y_A)"---yes, the LPR sender multiplies the sender secret Y_B by the LPR public key S_A;

* "such that, during the reconciliation step, the first entity (A) is capable of recovering the confidential information"---yes, the LPR receiver recovers a confidential message in the end;

* "by an operation for decrypting a noisy message"---yes, the LPR receiver obtains and decrypts a noisy message;

* "composed by the second entity (B)"---yes, the noisy message comes from the LPR sender;

* "from, among others, the second intermediate value P_B"---yes, the LPR sender uses P_B in building the noisy message.

I think "LPR" remains the proper name for the cryptosystem, since scientifically the first publication wins, and as far as I know the first publication of the cryptosystem was in an April 2010 LPR talk.
However, under patent law, an April 2010 publication doesn't invalidate a February 2010 patent filing.

> Despite multiple attempts by different experts, I have never seen a

> demonstration of how Kyber/SABER could fall under that patent's claims

Within the above list, what's Kyber doing differently? The noisy message is shorter (see above re compressed LPR and the 2012 patent), but this isn't even a literal deviation from the claims of the 2010 patent. More to the point, the doctrine of equivalents says that one has to be doing something "substantially" different.

Maybe I should note that it's common for a patent on X+Y (e.g., the general LPR idea) to be followed by a patent on X+Y+Z (e.g., compressed LPR). Someone doing X+Y+Z is violating both. But a patent on X+Y is invalid if there's a _previous_ patent on X+Y+Z, since X+Y+Z is prior art for X+Y. Again, one has to be clear about levels of generality.

> -- every attempt failed to meet at least three central requirements.

Namely?

I'll say this with all due respect: My best guess is that whoever did that analysis was unaware of the doctrine of equivalents and excitedly reported the minus sign in X^n-1; and that "three" is exaggeration for rhetorical effect. I would love to see a convincing analysis concluding that Kyber and other LPR-type systems avoid the patent, but so far this sounds to me like a combination of ignorance and wishful thinking.

> So, I don't think anyone should credit the belief that the patent
> "covers the entire idea of compact noise-DH encryption with
> reconciliation," without a proper demonstration.

Patents are public, and the rules for interpreting patents are public.
Going through claim 19 and particularly claim 1 is tedious but hardly rocket science. One _almost_ doesn't even need to know the rules, except that people who don't know the rules will incorrectly think that the patent applies only to X^n-1 and not X^n+1.

Again, there are various sources of randomness in court cases, so it's hard to _guarantee_ that a user deploying an LPR-type cryptosystem will lose in court, but my assessment is that the risks are close to 100%.

> (Since this issue is quite far afield from the ostensible topic of
> this thread,

Patents arose briefly in the original message as a natural part of a careful analysis of the main topic of that message:

* The original message gave "two examples of differences between Core-SVP and Kyber3-Modified-Core-SVP". The second example started as follows: "Core-SVP for RLWE/MLWE-based systems is defined by 2n full samples (public multiples plus errors), whether or not the systems actually apply further rounding to those samples. See, e.g., the round-2 Kyber submission."

* Within the second example, there was a paragraph discussing two types of previous NISTPQC commentary consistent with preferring this Core-SVP definition over Kyber3-Modified-Core-SVP. The second half of the paragraph was as follows: "Also, certain people have been claiming that it's a problem if cryptosystem parameters provide less security in other cryptosystems; it's not hard to imagine users skipping the rounding since the idea of saving bandwidth in this way was first published and patented by Ding, two years before Peikert announced it as new."

Followups led to more detailed patent discussion, and I'm happy to split patents off into this separate thread, although it seems that NIST needs the "ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER" subject line for all official comments about Kyber no matter what the specific topic is.

> I'll omit the technical details here, but am happy to share with
> whoever is interested.)

Now that this is a separate thread, please elaborate! It would be great to have some way to kill or avoid these patents without abandoning the whole LPR idea. Please tell me that potential LPR users haven't pinned their hopes of patent avoidance to the choice of sign in X^n+-1. Please tell me that they haven't pinned their hopes to an unpublished analysis by an unnamed "expert" unaware of the doctrine of equivalents.

---Dan

From: Vadim Lyubashevsky <vadim.lyubash@gmail.com>
Sent: Friday, December 11, 2020 3:32 PM
To: pqc-forum; pqc-comments
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

Hi Dan, all,

I don't want to get into an argument because it never leads anywhere and because I obviously do not have the requisite competence to discuss patent law. I am just including some extra facts.

> The basic difficulty here is that Ding had already published and (unfortunately) patented essentially the same cryptosystem in 2012, using essentially the same technique. Courts don't care about minor differences: the "doctrine of equivalents" asks whether the accused device performs "substantially" the same function in "substantially" the same way to obtain the same result. (This is the wording in U.S. courts, but similar principles apply internationally.)

I wrote about this in a previous email, but I will repeat it again. Jintai did not present a new *cryptosystem*. His KEM construction is quite different from LPR / Kyber / Saber with compression. In fact, as you can see from page 13 onward, where he actually needs a *cryptosystem* for something else, he uses LPR (or something similar based on Module-LWE, which also already appeared prior) *without any compression / rounding*. So given this, it seems hard to argue that his reconciliation-using KEM obviously implies Module-LWE encryption with compression. (I am not arguing the other direction that LWE with compression implies reconciliation because I don't care about invalidating anything).

> There are a few choices of details at this point, since there are some differences in the members of the patent family. For definiteness let's take European Patent 2537284;
> Within this patent, let's look at Claim 19, and go through the exercise of matching up the limitations of the claim to the LPR cryptosystem:

You did not list the prior work that the El Gamal-like LPR encryption scheme (i.e. with small secrets / errors) is actually derived from, and which precedes this patent. (And I know that you know it because we discussed it ... but it's possible you forgot because it was a while ago). I am referring to this paper which appeared at TCC 2010 (https://eprint.iacr.org/2009/576.pdf). Now observe the encryption scheme in section 3 (for the definition of the \odot symbol, it's easiest to look at the example on page 3. In particular, A \odot s = As+e, where e is the "carry" vector) and go through the checklist that you made comparing the claim and LPR. Everything in this cryptosystem -- in particular the fact that all the secrets and errors have small norms -- matches except for the "belonging to the ring R" part. If we wanted to use something like NewHope (where all elements indeed belong to a ring), then one could start arguing how relevant this distinction is. I suppose it could be relevant if one needed commutativity (which one doesn't for NewHope or LPR). But I don't even need to bother with that -- the elements in the equation As+e in Kyber / Saber are not ring elements! They are vectors and there is no commutativity -- i.e. As =/= sA. So the only possibly difference between the claim and the TCC paper is the algebraic properties that are induced by having A (and s and e) be a ring element, and Kyber / Saber don't use that.

There is one thing that you might nitpick at: A\odot s = As+e, but the e is not random in the cryptosystem. But in the talk I gave at TCC 2010 (https://www.dropbox.com/s/s4utim5y2jd3rov/subset_sum_tcc.ppt?dl=0) on slide 15, the cryptosystem is described generically with the only restriction being that all the errors and secrets are small. Now you might again nitpick that one can't really be sure that I gave this talk during TCC (Feb. 9 - 11, 2010). Luckily, Oded Goldreich was there and he "blogged" about the talk (by "blogged", I mean put a postscript file on his website http://www.wisdom.weizmann.ac.il/~oded/X/tcc10.ps) and dated it. He summarizes this talk in section 7 and mentions that the noise could be random. So what exactly do Kyber/Saber use from the patent instead of from the TCC paper/talk? If anything, it's clear that this patent did not mention very relevant prior art.

Best,
Vadim

From: Christopher J Peikert <cpeikert@alum.mit.edu>
Sent: Friday, December 11, 2020 5:58 PM
To: pqc-forum; pqc-comments
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

On Fri, Dec 11, 2020 at 10:08 AM D. J. Bernstein <djb@cr.yp.to> wrote:

> Quoting the first entry in https://ntruprime.cr.yp.to/faq.html (from October):
>
> There are known patent threats against the "Product NTRU"/"Ring-LWE"/"LPR" lattice proposals: Kyber, SABER, and NTRU LPRime (ntrulpr). These proposals use a "noisy DH + reconciliation" structure that appears to be covered by U.S. patent 9094189 expiring 2032, and a 2x ciphertext-compression mechanism that appears to be covered by U.S. patent 9246675 expiring 2033.

There are multiple problems with this paragraph.

The second clause is factually incorrect: Kyber and SABER (at least) do not even use a "2x ciphertext-compression mechanism," much less one described in the cited patent. The (less than 2x) compression mechanism they do use (1) was published more than two years prior to the cited patent, and (2) does not even appear in that patent. See below for details.

(To be clear, "ciphertext-compression mechanism" refers to the schemes' method of dropping [or "rounding away"] some low bits of Zq elements, which is what the above points are addressing.)

The first clause, which concludes that the other patent covers any scheme with a "noisy DH + reconciliation" structure, is far too imprecise and broad. The patent, like any other, is limited to specific methods that must meet particular requirements. I still have not seen a demonstration of how Kyber or SABER could be covered by the patent's claims (see below for more).

> \> What you wrote was not even close to correct,
>
> What I wrote was correct exactly as stated: "it's not hard to imagine users skipping the rounding since the idea of saving bandwidth in this way was first published and patented by Ding, two years before Peikert announced it as new".
>
> \> "In this way" referred specifically to Round-3 Kyber rounding away low
> \> bits (of Z_q elements)
>
> No. "The rounding" is referring specifically to Kyber's rounding, but "in this way" is generalizing to what Ding published and patented.

This is absurdly tortured interpretation, but another interesting attempt to backtrack. Read your own statement: "the idea of saving bandwidth in this way" can only be referring to its antecedent, which is Kyber's "rounding" (which does indeed save bandwidth).

"[T]he idea of saving bandwidth in this way" -- namely, rounding away some low bits of Zq elements, exactly as Kyber does -- was already published by September 2009, more than two years prior to Ding's patent submission (details appear in my prior messages).

> > which -- again -- is described in detail in the 2009 paper that
> > predates that patent application by more than two years.
>
> No. Readers who follow references to a compression idea "first published
> and patented by Ding, two years before Peikert announced it as new"

Readers don't need to follow the references to Ding's 2012 patent in order to see that your chronology and assignment of credit are wrong, because they can see that (a) the rounding method Kyber uses was first described by 2009, and (b) 2009 happened before 2012.

Moreover, the "rounding" method described in Ding's patent isn't what Kyber does. (Vadim has given many details on this.) So the attempt to shoehorn Ding's patent into a discussion of Kyber's rounding is an irrelevant distraction.

> anyone checking 2009 Peikert will see that there's
> nothing in that paper anywhere near this level of efficiency. The fact
> that one can point to various features of the cryptosystem that already
> appeared in the 2009 paper---rounding, noisy multiples, etc.---doesn't
> kill the patent.

Nobody claimed to "kill the patent," so this a strawman. I only killed your attribution of "the idea of saving bandwidth in this way."

However: the idea of applying rounding Zq-coefficients **of polynomials** (which are the source of Kyber's and SABER's efficiency) was also published by August 2011, several months before Ding's patent submission.
See https://eprint.iacr.org/2011/401.pdf and search for "ring-LWE."

So, even this attempt to make an artificial distinction on efficiency -- which is based only on how the Zq-elements were obtained prior to rounding them -- also fails.

> 2. Regarding the orthogonality statement: Courts understand the idea of
> interchangeable parts, but they're also faced with a constant stream of
> defendants claiming that _clearly_ the components inside the patent are
> interchangeable, while being unable to point to prior art _saying_ that
> the components are interchangeable.

If you're looking for even more prior art saying that large "unstructured" matrices are interchangeable with polynomials that compactly represent "structured" matrices (and are faster to compute with) -- the very principle giving Kyber/SABER their efficiency and compactness -- then look no further than this 2008 chapter by Micciancio and Regev, and references therein: https://cims.nyu.edu/~regev/papers/pqc.pdf .

Specifically see Section 4.2, which details how "The efficiency of lattice-based cryptographic functions can be substantially improved replacing general matrices by matrices with special structure."

Of course, as they write, "A fundamental question that needs to be addressed whenever a theoretical construction is modified for the sake of efficiency, is if the modification introduces security weaknesses."

For example, almost all of the LPR'10 paper is devoted to addressing this question for LWE; only one introductory paragraph informally sketches an example cryptosystem. This is because, as the paper says, "Most [LWE] applications can be made more efficient, and sometimes even practical for real-world usage, by adapting them to ring-LWE. This process is often straightforward..." So yes, this form of interchange was well established prior to the patent application.

> > > Ding's patent isn't the only problem for Kyber (and SABER): there's the
> > > much earlier, February 2010, Gaborit--Aguilar Melchor patent that as far
> > > as I can tell covers the entire idea of compact noisy-DH encryption with
> > > reconciliation.
> > If you really believe this, then you should lay out your reasoning,
> > instead of making unjustified assertions.
>
> Within this patent, let's look at Claim 19, and go through the exercise
> of matching up the limitations of the claim to the LPR cryptosystem:

This is not what I requested.

You concluded that "the entire idea of noisy-DH encryption with reconciliation" was covered by the patent.

I requested a justification of that broad claim, or even something more limited: "a demonstration of how Kyber/SABER could fall under that patent's claims."

You've possibly shown that the specific LPR cryptosystem is covered, which might be a problem for NTRU LPRime -- but that doesn't imply anything about Kyber/SABER. Saying that they are somehow "LPR-like," and hence covered, won't do. You need to match up the limitations of the patent's claims to Kyber/SABER themselves.

> > -- every attempt failed to meet at least three central requirements.
>
> Namely?
>
> I'll say this with all due respect: My best guess is that whoever did
> that analysis was unaware of the doctrine of equivalents and excitedly
> reported the minus sign in X^n-1; and that "three" is exaggeration for
> rhetorical effect.

No and no; the failures were not on minor things like signs, and there were indeed at least three major issues.

Here are some details. In an attempt to match up the patent's claims to Kyber/SABER, one would need to show (with respect to Claim 1):

1. what the elements X_A, Y_A, X_B, Y_B, P_A, P_B, etc. correspond to, and what *common ring* R they all belong to;

2. what "internal composition law" f on R satisfies the requisite "X_I * f(Y_I) - Y_I * f(X_I) has a small norm" property;

3. how both entities A and B use the *same* f in computing P_A and P_B using the provided equations.

Despite multiple attempts, I have never seen even one of these items successfully demonstrated for Kyber/SABER, much less all three at once.

(To be clear, this is not meant to be an exhaustive list of the problems with applying the patent to Kyber/SABER, and nothing I've written should be taken as endorsing either of the cited patents' validity.)

> Again, there are various sources of randomness in court cases, so it's
> hard to _guarantee_ that a user deploying an LPR-type cryptosystem will
> lose in court, but my assessment is that the risks are close to 100%.

You haven't even checked the patent claim's against Kyber/SABER, yet you make such a confident assessment? This is FUD.

Sincerely yours in cryptography,
Chris

From: Kirk Fleming <kpfleming@mail.com>
Sent: Friday, December 11, 2020 8:29 PM
To: pqc-forum
Cc: pqc-comments
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: Classic McEliece

Dan Bernstein (speaking for himself) wrote:
> As far as I can tell, the questions specific to Classic McEliece
> have already been answered.

In the formal sense that you responded, I guess. It's like asking the time and being told that it depends on the observer's frame of reference. Thanks Einstein. It is an answer but I still don't know if I'm late for class.

Kirk Fleming wrote:
>>>> I am filing this comment to request that the Classic McEliece
>>>> submitters justify the claimed security categories for their
>>>> parameters.

Dan Bernstein (on behalf of the Classic McEliece team) wrote:
>>> Can you please clarify how you think that the numbers already
>>> in the literature don't already justify the assignments in the
>>> submission?

Kirk Fleming wrote:
>> Let's try an easier question. Can you point to the paper cited
>> by your submission that gives numbers for the mceliece-4608-096
>> parameter set which clearly justify its assignment as category 3?

Dan Bernstein (speaking for himself) wrote:
> NIST has failed to define the metrics to be used, so there's no
> way for anyone to be sure what's included in "category 3".

To be absolutely clear, you are stating that the numbers in the literature already justify the assignments in the submission but that you can't point to any specific numbers because there is no way for anyone to be sure what's included in each category.

Okaaaay.

Let's try an even easier question. Why mceliece-4608-96?

The mceliece-6960-119 parameter set was taken from "Attacking and defending the McEliece cryptosystem". You could have chosen the mceliece-4624-95 parameter set from the same paper. Instead you chose a parameter set which had "optimal security within 2^19 bytes if n and t are required to be multiples of 32". (mceliece-4624-95 was chosen to be optimal without the restrictions on n and t)

mceliece-5856-64 and mceliece-4160-128 are also within 2^19 bytes. Both of these have smaller public keys than mceliece-4608-96 and mceliece-5856-64 has smaller ciphertexts. Please explain the metric you used to decide which of the three parameter sets had "optimal security".

Kirk

From: Robert Ransom <rransom.8774@gmail.com>
Sent: Sunday, December 13, 2020 4:25 AM
To: pqc-forum; pqc-comments
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

On 12/11/20, D. J. Bernstein <djb@cr.yp.to> wrote:

> This message includes what might be the first detailed public chart
> matching up the LPR cryptosystem to patent 9094189, which was filed 18
> February 2010 and wasn't known to the community until eight years later.
> In theory, patents are published; in reality, millions of hard-to-read
> patents operate as a denial-of-service attack against the general public.

> There are a few choices of details at this point, since there are some
> differences in the members of the patent family. For definiteness
> let's take European Patent 2537284; my reason to pick Europe instead
> of the U.S. here is that the European patent has already survived one
> round of litigation, whereas I haven't heard about any litigation yet
> regarding the U.S. patent. (I don't expect the U.S. patent to be
> invalidated, but, all else being equal, it's reasonable to estimate
> the ultimate invalidation chance as being even lower for a patent that
> some people have tried and so far failed to invalidate.)

The other reason that you picked the European patent rather than US
9,094,189 is that both of the independent claims in the U.S. patent, claims 1 and 21, are explicitly limited to rings of one of two forms:

* R = F_q[x]/(X^n - 1) "with n equal to 6000", or

* R = Z/pZ "with p equal to 2500".

Other constraints are specified in the patent for each of those two classes of rings. Note that I did not introduce any error or omit any formatting from the integer ring; the patent really does specify that p is equal to the number two thousand, five hundred.

I have not yet obtained the patent wrapper to find out whether the U.S. patent was limited to those two classes of rings for any particular reason.

From: D. J. Bernstein <djb@cr.yp.to>
Sent: Sunday, December 13, 2020 8:43 AM
To: pqc-comments
Cc: pqc-forum
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER
Attachments: signature.asc

Robert Ransom writes:
> both of the independent claims in the U.S. patent, claims 1 and 21,
> are explicitly limited to rings of one of two forms:
> * R = F_q[x]/(X^n - 1) "with n equal to 6000", or
> * R = Z/pZ "with p equal to 2500".

The doctrine of equivalents will cover other rings. The specific patent claim that I analyzed in detail in the message you're replying to _also_ names a ring that's different from the NISTPQC choices, and I explained in that message why this doesn't matter. No court will find a "substantial" difference between F_q[x]/(x^6000-1) and F_q[x]/(x^512+1) unless there was prior art forcing the choice of number and/or sign--- and how could there have been, since this patent filing predated LPR?

It's understandable that most people reading patent claims won't realize that courts go beyond the literal meaning and extend the coverage to include anything that's "substantially" the same. This is a perfect example of what I said about errors in NISTPQC evaluations, and about the importance of detailed public analyses.

> The other reason that you picked the European patent rather than US
> 9,094,189

Nope. The European patent already surviving a round of litigation is useful information, and something I already tweeted about in early July:

https://twitter.com/hashbreaker/status/1279677625410088963

One of the followups there asked about x^n-1, and I replied (briefly) that this doesn't matter given the doctrine of equivalents.

> I have not yet obtained the patent wrapper to find out whether the
> U.S. patent was limited to those two classes of rings for any
> particular reason.

The file might have interesting information, and a public analysis would be useful. Some care in allocating human resources is warranted, since there's a high risk of error in the analysis if it's coming from someone who doesn't know the basics of how courts interpret patents.

---Dan

From: Christopher J Peikert <cpeikert@alum.mit.edu>
Sent: Sunday, December 13, 2020 9:46 AM
To: pqc-forum; pqc-comments
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

> something I already tweeted about in early July:
>
> https://twitter.com/hashbreaker/status/1279677625410088963

Wait -- since July you've been claiming (with certainty) that the patent covers, e.g., Kyber and SABER, without ever showing how these systems are covered by the patent's specific limited claims?

It's far past time to do so, or retract the claim. As of now, it is baseless -- and in my opinion, should be seen as FUD.

Sincerely yours in cryptography,
Chris

From: D. J. Bernstein <djb@cr.yp.to>
Sent: Sunday, December 13, 2020 9:56 AM
To: pqc-comments
Cc: pqc-forum
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER
Attachments: signature.asc

Vadim Lyubashevsky writes:
> Everything in this cryptosystem -- in particular the fact that all the
> secrets and errors have small norms -- matches except for the
> "belonging to the ring R" part.

Mathematically, I don't understand what distinction you're trying to draw here. Simply zero-padding any vector or matrix produces a square matrix, and matrix-vector multiplication is exactly multiplication in the ring of square matrices.
If the goal were to kill Claim 1 of the
2010 patent then it would be worth going through this in detail. (A claim is invalidated by any example appearing in the prior art.)

My analysis focused, however, on Claim 19 of the patent, matching it up point by point to the subsequent LPR cryptosystem. Claim 19 uses a _polynomial_ ring, making it much more efficient than the matrix case.
So the matrix cryptosystem you mention doesn't match Claim 19, and it's worse in a way that a court will easily understand.

The defendant's lawyers can still try to argue that the LPR cryptosystem was _obvious_ to someone of ordinary skill in the art given the 2009 work, but, if it was obvious, then why does the last slide of your February 2010 presentation say "'Ideal Lattice' analogue" with a question mark under "Open Problems"? And why does the May 2010 Springer version of the LPR paper feature a larger, slower cryptosystem, replaced by the LPR cryptosystem only in a subsequent revision of the paper?

> If anything, it's clear that this patent did not mention very relevant
> prior art.

Do you think you can show that the applicants _knew_ about relevant publications that they didn't mention to the patent office? "Inequitable conduct" by patent holders can effectively invalidate an entire patent, so this could be a useful line to explore.

---Dan

From: Andrea Basso <a.basso@cs.bham.ac.uk>
Sent: Sunday, December 13, 2020 11:19 AM
To: pqc-forum; pqc-comments
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

Hi Dan, hi all,

I have looked into the first patent (US9094189B2) and I do not believe it covers SABER. I also suspect it does not cover Kyber either, but I have not investigated it much.

As a disclaimer, I have worked as a patent engineer in the past, but I do not have much experience with cryptographic patents or US patent law.
Also, I am a member of the SABER team (although I am speaking for myself), so there is an inherent conflict of interest, as for most other people who have participated in the debate.

Firstly, the US and European are slightly different, so it is better to treat them separately. Claim 1 of the US patent specifies that the ring R can either be F_q[x]/<x^n+1> or Z/pZ, and in the first case the value of the exponent n must be 6000 and the norm of X_A, X_B, Y_A and Y_B must be 35. Clearly, this does not apply to either Kyber nor SABER. Now, the doctrine of equivalents cannot be applied here because of prosecution history estoppel. That means that if claims are amended during prosecution, the subject matter that has been excluded cannot then be covered by the doctrine of equivalents. This is exactly the case here. The full history of the US prosecution of the patent can be found on the USPTO website
(https://globaldossier.uspto.gov/#/). One can see that the claim 1 initially filed in 2013 does not contain any mention of specific values of the exponent n nor of the norms. After the examiner complained that claim 1 was indefinite (see the non-final rejection from February 2014), the specific values requirements were introduced with the amended claims submitted in 2015. Thus, the US patent does not cover neither Kyber nor Saber.

For the EU patent, the situation is slightly less straightforward since claim 1 does not have the same limitations. However, the EU patent also does not cover SABER, mainly due to its reliance on LWR. Whether the function f(X) is interpreted as either f(X) = A*X or f(X) = p/q*A*X, the computations for the intermediate values P_A and P_B are clearly different. Once again, the differences and their effects are substantial enough that the doctrine of equivalents does not apply to this case.

One thing to note here is that while the application is done at the European Patent Office, once granted the patent is converted into individual patents for each European country. Thus prosecution only takes place at the national level. The patent, according to information on Espacenet
(https://worldwide.espacenet.com/patent/search/family/042753478/publication/EP2537284B1?q=EP2537284B1) , has lapsed in all countries but Germany, France, Switzerland and the UK.

At a first look, the requirements for the application of the doctrine of equivalents in these four countries are more stringent than in the US.
In particular, Germany and Switzerland seem to require the replace features be obvious in light of the original patent, which is clearly not the case here. The UK and France require the replace features have the same consequences as the original claims, and if MWLE-based systems fall under the patent (and that's a big if), MWLR have clearly different advantages/disadvantages. Thus, the EU patent also does not cover SABER in the four countries where it is active.

This is what I gathered on a first look. I suspect there are many other reasons why the patent does not cover SABER. If, instead, there is any mistake with this reasoning, please let me know. I also have not had a chance to look into the second patent, but at a glance it seems like it does not cover SABER either.

All the best,
Andrea

From: Christopher J Peikert <cpeikert@alum.mit.edu>
Sent: Sunday, December 13, 2020 11:35 AM
To: pqc-forum; pqc-comments
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

> My analysis focused, however, on Claim 19 of the patent, matching it up
> point by point to the subsequent LPR cryptosystem. Claim 19 uses a
> _polynomial_ ring, making it much more efficient than the matrix case.
> So the matrix cryptosystem you mention doesn't match Claim 19, and it's
> worse in a way that a court will easily understand.
>
> The defendant's lawyers can still try to argue that the LPR cryptosystem
> was _obvious_ to someone of ordinary skill in the art given the 2009
> work

Yes, the technique of improving efficiency by replacing "unstructured" matrices with "structured" ones corresponding to polynomials was well documented prior to the patent. Again, see this 2008 survey by Micciancio and Regev (in a book you edited!), and the many references therein: https://cims.nyu.edu/~regev/papers/pqc.pdf .

Section 4.2 details how "The efficiency of lattice-based cryptographic functions can be substantially improved replacing general matrices by matrices with special structure," and specifically considers replacing each square n-by-n matrix with a circulant or anti-circulant matrix. These correspond to polynomials in the rings Z_q[x]/(x^n-1) or Z_q[x]/(x^n+1).

Applying this mechanical transformation to the TCC'10 cryptosystem (which also precedes the patent) yields the "LPR" cryptosystem that you've claimed is covered by the patent.

> but, if it was obvious, then why does the last slide of your
> February 2010 presentation say "'Ideal Lattice' analogue" with a
> question mark under "Open Problems"?

Because, as Micciancio and Regev also write, "A fundamental question that needs to be addressed whenever a theoretical construction is modified for the sake of efficiency, is if the modification introduces security weaknesses."

Applying the mechanical transformation from unstructured matrices to structured ones/polynomials is obvious. Supporting the resulting system's *security* by, e.g., a connection to worst-case "ideal lattices" is highly non-obvious.

> And why does the May 2010 Springer
> version of the LPR paper feature a larger, slower cryptosystem, replaced
> by the LPR cryptosystem only in a subsequent revision of the paper?

Both versions of the LPR paper informally describe just one example cryptosystem, because applications are not the paper's focus. Almost all of the paper is devoted to formally defining Ring-LWE and giving a connection to worst-case ideal lattices (i.e., addressing the "security" question stated in MR'08).

Regarding applications, the paper says (Springer version): "This scheme and its security proof are a direct translation of the ‘dual’ scheme from [13] based on the standard LWE problem, and similarly direct adaptations are possible for most other LWE-based schemes..." This is essentially restating for LWE the general approach described in MR'08. Again, performing this direct adaptation on the TCC'10 cryptosystem yields the "LPR" cryptosystem.

The "larger, slower" example sketched in the Springer version is much more versatile, because it can be made identity-based and more. (For those who know the jargon, it is the Ring-LWE adaptation of the widely used "dual Regev" LWE system from GPV'08.) However, its full CPA-security proof based on Ring-LWE relies on a non-trivial "regularity" lemma that we ultimately decided to break out into a separate paper devoted to applications (the LPR'13 "toolkit" paper).

The smaller example system is much more feature-limited, but it has an elementary and self-contained CPA-security proof based on Ring-LWE, following the strategy from the TCC'10 work.

Again, both example systems are "direct adaptations ... of other LWE-based schemes," using a well documented approach, all of which preceded the patent application.

Sincerely yours in cryptography,
Chris

From:	D. J. Bernstein <djb@cr.yp.to>
Sent:	Sunday, December 13, 2020 12:38 PM
To:	pqc-comments
Cc:	pqc-forum
Subject:	Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER
Attachments:	signature.asc

Focusing here on the application to Kyber of Claim 19 of the 2010 patent that I mentioned. I already spelled out the application to LPR.

Christopher J Peikert writes:
> 1. what the elements X_A, Y_A, X_B, Y_B, P_A, P_B, etc. correspond to,
> and what *common ring* R they all belong to;

The simplest answer in court is the doctrine of equivalents.
Generalizing from r*r to r*m, while preserving the sizes of the communicated objects, is not a "substantial" difference.

But, just for fun, let's imagine an alternate universe without the doctrine of equivalents, and see how easy it is to dispense with these three alleged showstoppers.

Kyber-768 uses the ring ((\Z/q)[x]/(x^256+1))^(3x3). Kyber chooses various matrix entries to be zero, but there's nothing in the patent requiring the entries to be nonzero. Kyber _describes_ those matrices differently, as vectors, but the plaintiff's lawyers will provide a translation table from the operations on vectors to the identical operations on matrices. What was the difficulty supposed to be here?

> 2. what "internal composition law" f on R satisfies the requisite "X_I
> * f(Y_I)
> - Y_I * f(X_I) has a small norm" property;

Again the doctrine of equivalents is the easiest approach, but let's stick to the alternate universe and focus on the 3x3 matrix ring.

In a nutshell, the answer is the same as what I already gave for LPR: f multiplies by some public random ring element. This is also what the examples in the patent description do, with the small-norm difference being uniformly 0. (Obviously whoever wrote the patent was trying to state more generally what makes the system work.)

Presumably Dr. Peikert's objection here is to the patent notation in the non-commutative case: the system doesn't work without transpositions, or equivalently multiplying in a different order. The notation is sloppy even in the commutative case: e.g., the patent sometimes has f taking two inputs, as in "f(X_A,Y_A) = X_A.h + Y_A". But "Aha! The patent is sloppy, and will be punished for this!" is wishful thinking, much like "Aha! This NISTPQC submission is sloppy, and will be punished for this!"

What will actually happen is an initial hearing to establish the exact meaning of, e.g., this "small norm" requirement. The defendant's lawyers will propose definitions that try to make this sound as restrictive as possible---but they'll run into the "X_A.h + Y_A" example stated for any ring, and the court won't accept definitions that exclude this example.

The plaintiff's lawyers will propose a definition with the necessary transpositions filled in, will explain that this is exactly what makes the example work in the generality stated, and will pull out one example after another from the math/physics literature where it was left to the reader to fill in the "obvious" transpositions. I don't like this notation, but I've seen it many times (even Sage will automatically transpose sometimes!) and it's an easy winner in court.

> 3. how both entities A and B use the *same* f in computing P_A and P_B
> using the provided equations.

Let's again take the alternate universe, no doctrine of equivalents, and see what happens with the 3x3 matrix ring.

The transpose-as-necessary approach will end up with f multiplying the Alice inputs by a public random ring element h, and multiplying the Bob inputs by the same h on the opposite side---equivalently, transposing the Bob inputs, then multiplying by h transpose, then transposing back.

The objection here is that Alice's inputs aren't being handled by f the same way as Bob's inputs, so f is no longer _one_ function, but rather _two_ functions, and in our alternate universe this means we can't find _one_ function f that brings Kyber within this patent claim! Game over!

But wait a minute. Is it actually two functions? Let's look at the input matrices. Alice's inputs are 3x3 matrices of the form

```
  a1  0  0
  a2  0  0
  a3  0  0
```

while Bob's inputs are 3x3 matrices of the form

```
  b1 b2 b3
   0  0  0
   0  0  0
```

so we can simply define f to handle Alice's matrices in the Alice way, and handle Bob's matrices in the Bob way, and return anything on the overlapping inputs

```
   c  0  0
   0  0  0
   0  0  0
```

which the plaintiff's lawyers will have an expert testify will never actually occur in Kyber, so what's computed here is exactly what Kyber computes. What was the difficulty supposed to be here?

---Dan

From:	D. J. Bernstein <djb@cr.yp.to>
Sent:	Sunday, December 13, 2020 1:32 PM
To:	pqc-comments
Cc:	pqc-forum
Subject:	Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER
Attachments:	signature.asc

Andrea Basso writes:
> Now, the doctrine of equivalents cannot be applied here because of
> prosecution history estoppel. That means that if claims are amended
> during prosecution, the subject matter that has been excluded cannot
> then be covered by the doctrine of equivalents.

No. See _Festo v. Shoketsu_, 535 U.S. 722 (2002), where the Supreme Court specifically rejected this view.

The usual situation where prosecution-history estoppel kicks in is that a claim was amended _to avoid the prior art_, but that's not the situation here, according to your description of the history. In other situations, the patentee simply has to argue that the reason for the amendment doesn't surrender the particular equivalent in question.

> the EU patent also does
> not cover SABER, mainly due to its reliance on LWR

Huh? The notation is different, and some numbers are described in a different format (omitting some known-to-be-zero bits), but I don't see how this is relevant to anything in the patent claim. Rounding f(Y_A) to obtain S_A in a restricted set is simply a matter of choosing X_A appropriately; nothing in the patent claim rules out this choice.

> Germany and Switzerland seem to require the replace features be
> obvious in light of the original patent

The bits I've read about the German analysis sound very different from this, and are fairly close to the U.S. analysis, but I'll let someone more familiar with German patent law comment.

---Dan

From: Vadim Lyubashevsky <vadim.lyubash@gmail.com>
Sent: Sunday, December 13, 2020 7:26 PM
To: pqc-forum; pqc-comments
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

Hi Dan, all,

On Sun, Dec 13, 2020 at 3:56 PM D. J. Bernstein <djb@cr.yp.to> wrote:

> Vadim Lyubashevsky writes:
> > Everything in this cryptosystem -- in particular
> > the fact that all the secrets and errors have small norms -- matches
> > except for the "belonging to the ring R" part.
>
> Mathematically, I don't understand what distinction you're trying to draw here. Simply zero-padding any vector or matrix produces a square matrix, and matrix-vector multiplication is exactly multiplication in the ring of square matrices. If the goal were to kill Claim 1 of the 2010 patent then it would be worth going through this in detail. (A claim is invalidated by any example appearing in the prior art.)
>
> My analysis focused, however, on Claim 19 of the patent, matching it up point by point to the subsequent LPR cryptosystem. Claim 19 uses a _polynomial_ ring, making it much more efficient than the matrix case. So the matrix cryptosystem you mention doesn't match Claim 19, and it's worse in a way that a court will easily understand.

Sorry, I misunderstood that earlier you were referring to just LPR and not also to Kyber / Saber. I don't really care about the actual LPR cryptosystem (e.g. as used in NewHope or NTRULPrime), and was always just talking about Kyber/Saber. So let's just focus on those from now on.

Let's go back to the European Patent 2537284. Just to understand, you're saying that Claim 19 (which specifies the polynomial ring to use) and Claim 1 (in which the ring could actually be interpreted as a matrix of ring elements), covers Kyber / Saber. Let's ignore the inconsistency of Claim 19 being very specific with what the ring R is and then you changing the precise definition. So let's just look at claim 1.

I agree that one could interpret Kyber/Saber as a cryptosystem with just matrix multiplication (where, for efficiency, one really wouldn't do matrix multiplication). But that is not what claim 1 is doing! I know you already mentioned this, but I am going to restate one of the offending lines for clarity. Reading lines 53 - 55 on page 17 "for any pair of elements X and Y of R, such that X and Y have a norm that is small relative to the elements f(X) and f(Y), then X.f(Y) - Y.f(X) has a small norm". I guess you see that if f(X) is defined as MX, then for this to hold true, you would need XMY - YMX to be close to 0, which would only happen if we have commutativity (and this is certainly not true for Kyber / Saber). In your last email, you made a reference to something of this sort and said that it's wishful thinking to think that a patent will be punished for this. Really? Here is a pretty fun example. Look at slide 9 of https://www.wipo.int/edocs/mdocs/aspac/en/wipo_ip_bkk_19/wipo_ip_bkk_19_p_3.pdf and a more detailed explanation of this here: http://www.waltmire.com/2016/11/08/burnt-dough-difficulties-patent-drafting/#:~:text=Chef%20lost%20when%20it%20sued,%2C%20Inc.%2C%20358%20F.) So now, do you really think that "accidentally" implying commutativity will be forgiven? And this is on top of the fact that every single example of an instantiation (claims 19 -24) is a commutative ring. I think that it's quite clear that there is no accident or typo here --

claim 1 is simply about commutative rings, and this is how the authors always meant it to be. Your chain of events for how claim 1 can be made to fit Kyber / Saber sounds like the plaintiffs will allowed to rewrite their patent on the spot. With the amount of changes your hypothetical court is allowing, I can avoid every possible patent by converting El Gamal to Kyber / Saber.

I initially wrote a bit more explaining how interpreting (and rewriting) claim 1 to include Kyber / Saber would also imply that it includes the TCC 2010 cryptosystem, and therefore the claim would be thrown out because it patents prior art, as you said ... and so the only hope of the claimants is to actually insist that the patent only covers commutative rings and then maybe (big maybe) they can have a case just against the NewHope-type schemes. But I don't see how this matters anymore. Kyber / Saber simply do not fit into claim 1 as written.

Best,

Vadim

> ---Dan
>
> --
> You received this message because you are subscribed to the Google Groups "pqc-forum" group.
> To unsubscribe from this group and stop receiving emails from it, send an email to pqc-forum+unsubscribe@list.nist.gov.
> To view this discussion on the web visit https://groups.google.com/a/list.nist.gov/d/msgid/pqc-forum/20201213145628.246775.qmail%40cr.yp.to.

From: Christopher J Peikert <cpeikert@alum.mit.edu>
Sent: Monday, December 14, 2020 11:52 AM
To: pqc-forum; pqc-comments
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

To elaborate and expand on Vadim's most recent comments and others from this thread, let's see how a case attempting to show that Kyber (or alternatively, SABER) infringes on the patent would likely play out.

The defendant would argue that the patent is invalid on account of prior art (among other reasons). This argument is straightforward: a "person skilled in the art" -- let's call this skilled person Skip -- could have easily used existing publications to come up with a system that is covered by the patent's claims, as follows:

1. Take the TCC'10 cryptosystem (say, the one presented at the conference).

2. Improve its efficiency by mechanically applying the "unstructured matrices -> polynomials" transformation detailed in Section 4.2 of Micciancio-Regev'08. (See previous messages for details.)

3. Show how the resulting system is covered by Claim 1 of the patent (this is obvious).

I see no reason to dispute this argument, and every reason to accept it. (Notably, there has not been any quarrel with any part of this argument in this thread. Also, Step 2 does not require Skip to demonstrate anything about the *security* of the resulting system; only the construction matters.)

In the likely event that the court accepts this argument, the patent is invalidated and the defendant wins, case closed.

But let's suppose, as a hypothetical, that the court does not accept the above argument. This would require it to conclude that Skip -- despite his skill in the art and the detailed instructions provided in MR'08 -- could *not* have easily performed the above steps. The court clearly does not think much of Skip's abilities! So, for the rest of the case we would have to treat them as very limited.

Next, the plaintiff would argue that Kyber infringes on the patent, perhaps using the argument Dan laid out.

Since Skip's skills are so limited, all the "doctrine of equivalents" parts of that argument are highly questionable -- if Skip can't even apply the explicit directions in MR'08, how can he see how to interchange all these other purportedly equivalent objects? But let's focus on another serious issue:

> > 2. what "internal composition law" f on R satisfies the requisite "X_I * f(Y_I)
> > - Y_I * f(X_I) has a small norm" property;

> In a nutshell, the answer is the same as what I already gave for LPR: f
> multiplies by some public random ring element. This is also what the
> examples in the patent description do, with the small-norm difference
> being uniformly 0. (Obviously whoever wrote the patent was trying to
> state more generally what makes the system work.)
>
> Presumably Dr. Peikert's objection here is to the patent notation in the
> non-commutative case

The non-commutative case is indeed a major problem, but it's not just a matter of notation.

The court simply cannot accept Dan's argument that the patent is broad enough to cover non-commutative matrix rings, because the very same argument would also make the patent cover prior art like the TCC'10 cryptosystem and the ACPS'09 system, exactly as written (i.e., without any "structured matrix"/polynomial optimizations). Obviously, this is not tenable.

To see this, replace "Kyber-768 uses the ring ((\Z/q)[x]/(x^256+1))^(3x3)" with, e.g., "the TCC'10 system uses the ring (\Z/q)^(n x n)" and proceed identically from there.

(Similarly, if the court accepts the very broad "doctrine of equivalents" argument -- despite Skip's severe limitations -- it winds up with the same untenable conclusion that prior art is covered.)

I can anticipate an objection along the lines of "well, the TCC'10 system is less efficient than Kyber, so the non-commutativity argument wouldn't broaden the patent that far." But (a) So what? The patent doesn't even mention "efficiency" or "efficient," so it doesn't limit itself along that dimension; and (b) FrodoKEM demonstrates that a system very similar to the TCC'10 one can be efficient enough for real-world usage.

For the record, I don't find the argument for non-commutative rings persuasive in isolation either: in such a ring, order of multiplication matters by definition -- it is the very essence of non-commutativity. The patent explicitly requires X * f(Y) - Y * f(X) to have small norm for all appropriate X,Y. Allowing anyone -- much less our hapless Skip -- to reorder the multiplications takes us well outside both the text of the patent itself, and even what the authors knew how to do when they wrote it!

Sincerely yours in cryptography,
Chris

From: Andrea Basso <a.basso@cs.bham.ac.uk>
Sent: Monday, December 14, 2020 2:29 PM
To: pqc-forum; pqc-comments
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

Thank you for your reply Dan. As I mentioned, I am not too familiar with US patent law.

> The usual situation where prosecution-history estoppel kicks in is
> that a claim was amended _to avoid the prior art_, but that's not the
> situation here, according to your description of the history. In other
> situations, the patentee simply has to argue that the reason for the
> amendment doesn't surrender the particular equivalent in question.

I am not sure that this is the correct interpretation though. In Festo v. Shoketsu, the Supreme Court reiterated that estoppel applies when an amendment is made for substantial reason related to patentability.
Moreover, the SC decision states

> the patentee still might rebut the presumption that estoppel bars a
> claim of equivalence. The patentee must show that at the time of the
> amendment one skilled in the art could not reasonably be expected to
> have drafted a claim that would have literally encompassed the alleged
> equivalent.

In this case, it is clear that at the time of filing, the patentee might have drafted a claim covering different moduli, exponents and norms, and they would have been expected to do so if they intended to cover those cases.

> Huh? The notation is different, and some numbers are described in a
> different format (omitting some known-to-be-zero bits), but I don't
> see how this is relevant to anything in the patent claim. Rounding
> f(Y_A) to obtain S_A in a restricted set is simply a matter of
> choosing X_A appropriately; nothing in the patent claim rules out this choice.

There are different ways to define the function f, so depending on which one you choose there are different reasons why the patent does not apply. If the function f is taken to be f(X) = A*X, then the error X_A is X_A = round(p/q*A*X) - A*X, which is roughly (p-q)/q*A*x (if we ignore the rounding operation). This means that its norm would not be small when compared to either f(X) or f(A*X).

This is without considering that, as others have pointed out, there are also issues with commutativity and the interpretation of multiplication.

> The bits I've read about the German analysis sound very different from
> this, and are fairly close to the U.S. analysis, but I'll let someone
> more familiar with German patent law comment.

I am not particularly familiar with German patent law as well, but for instance see "Doctrine of Equivalents After Hilton Davis: A Comparative Law Analysi"s by Toshiko Takenaka https://core.ac.uk/download/pdf/267973723.pdf, which says

> The German Federal Supreme Court reaffirmed the test in Formstein.
> The Court declared that an accused embodiment makes equivalent use of
> a patented invention if one skilled in the art, having taken into
> consideration the disclosures made in the patent and his or her
> general knowledge, would have conceived of replacing the disputed
> element of the accused embodiment with the corresponding element in
> the claimed invention to obtain the same result.

All the best,
Andrea

On 2020-12-13 18:32, djb@cr.yp.to wrote:
> Andrea Basso writes:
>> Now, the doctrine of equivalents cannot be applied here because of
>> prosecution history estoppel. That means that if claims are amended
>> during prosecution, the subject matter that has been excluded cannot
>> then be covered by the doctrine of equivalents.
>

From: D. J. Bernstein <djb@cr.yp.to>
Sent: Thursday, December 17, 2020 10:20 AM
To: pqc-comments
Cc: pqc-forum
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER
Attachments: signature.asc

Executive summary: NIST promoted and relied on a particular metric that assigns security claims to lattice parameter sets. Round-3 Kyber-512 has switched to another metric so as to be able to claim several bits more security. Claims that there hasn't been a switch are simply not true.

Details: Let's start with the "Estimate all the {LWE, NTRU} schemes!"
page, along with its corresponding paper. Each column on the page is the output of a metric assigning a number, a claimed security level, to each parameter set for each lattice submission.

Within these metrics, one specific metric, the column labeled 0.292 beta, is generally known as the (pre-quantum) "Core-SVP" metric.
(There's also a column labeled 0.265 beta, generally known as the post-quantum Core-SVP metric. This is simply a constant times the pre-quantum Core-SVP metric, so it doesn't affect comparisons between lattice parameter sets. I'll focus on pre-quantum Core-SVP.)

NIST endorsed the Core-SVP metric as a comparison mechanism for lattice parameter sets: "we feel that the CoreSVP metric does indicate which lattice schemes are being more and less aggressive in setting their parameters". NIST IR 8309 used Core-SVP again and again as a comparison mechanism.

Each of the "Estimate" metrics for parameter sets, and in particular the Core-SVP metric, has the following two structural features:

  * The metric for RLWE/MLWE cryptosystem parameter sets is purely a
    function of the underlying RLWE/MLWE problems.

  * The metric for RLWR/MLWR cryptosystem parameter sets is purely a
    function of the underlying RLWR/MLWR problems.

The metric uses the underlying problem _even if_ the cryptosystem releases less information to the attacker than the underlying problem.
For example, as noted in Section 2.1 of the "Estimate" paper, the metric asks how secure full MLWE samples are _even if_ the MLWE cryptosystem actually removes some bits of the samples ("bit dropping").

Were any of the current commentators objecting to the "Estimate" page and the "Estimate" paper focusing on the underlying RLWE-etc. problems?
No. Were they objecting to NIST using Core-SVP for comparisons? No. When certain people criticized other submissions for focusing on the actual cryptosystem attack problems rather than the RLWE/MLWE problems, did these commentators speak up to defend those submissions as moving towards more accurate security analyses? No.

Meanwhile we've been hearing years of continual advertising of the RLWE/MLWE problems. We've been told, for example,

* that MLWE is a "standard" lattice problem;

* that the security of Kyber is "based on the hardness" of MLWE;

* that Kyber has a "security bound" that's "tight" assuming hardness of MLWE;

* that the "estimates of the security strength" for Kyber parameter sets are "based on the cost estimates of attacks" against MLWE;

etc. Would Kyber and NIST want to be in the position of advertising the supposed hardness of the MLWE problem as the foundation of security, and then choosing MLWE instances that _don't meet the minimum NISTPQC security level_? From this perspective, it makes perfect sense to use security metrics that, for MLWE systems such as Kyber, focus purely on how secure the MLWE instances are. Core-SVP does this.

Presumably this situation would have remained stable, except that
round-2 Kyber-512 didn't manage to get past its first security review:

* I filed a comment dated 30 May 2020 02:15:31 +0200 saying "I'm unable to verify, and I'm able to disprove parts of, the submission's argument that Kyber-512 meets category 1, the minimum security requirement in NIST's call for submissions, against the attacks that were already described in the submission document".

* Daniel Apon wrote "Thanks, Dan. I'll consider it."

* The Kyber team wrote a new security analysis and concluded "We agree that 136 and 141 are smaller than 143, but at the moment we do not consider this to be a sufficient reason to modify the Kyber-512 parameter set. The additional memory requirement of this attack strongly suggests that Kyber-512 is more secure than AES-128 in any realistic cost model."

* NIST's comments then left totally unclear whether NIST is going to take memory costs into account. NIST later claimed that everyone could see the boundaries between "unacceptable" and "controversial" and "uncontroversial" lattice parameter sets; this claim was false when it was made, and remains false today.

* The round-3 Kyber submission then modified the Kyber-512 parameter set, replacing the round-2 Kyber-512 with a new round-3 Kyber-512, despite the previous "we do not consider this to be a sufficient reason to modify the Kyber-512 parameter set" statement.

Compared to round-2 Kyber-512, round-3 Kyber-512 is less efficient, and _could_ be several bits more secure---but Core-SVP doesn't show this! So the round-3 Kyber submission also

* switched from Core-SVP to a modified metric for parameter sets, despite NIST's advertising of the unmodified Core-SVP metric;

* selected the details of the modified metric to _not_ simply be an evaluation of the MLWE problem, despite all the advertising of

  the MLWE problem; and

  * confusingly reused the name "Core-SVP" for the modified metric.

As a direct result of all these changes, the submission claimed "Core-SVP" 2^118 for round-3 Kyber-512.

When I challenged this change of metric, suddenly the aforementioned commentators started popping out of the woodwork to

  * claim to be confused at the idea of a metric comparing the
    underlying RLWE/MLWE/... problems, and to

  * try to make the reader believe that _of course_ Core-SVP accounts
    for further bits dropped in the cryptosystem.

Again, why weren't these commentators already objecting to the "Estimate" work, and to NIST using "Core-SVP" for comparisons? Why do the objections appear only when the objections seem to favor Kyber?

I _think_ the round-3 Kyber submission has dropped the claim that the MLWE instance underlying Kyber-512 is as hard to break as AES-128. I'm not totally sure about this---my clarification questions on this point remain unanswered---but if the claim has been dropped then are the same commentators going to object to Kyber-512's continued advertising of an MLWE problem below the minimum allowed NISTPQC security level? Is NIST going to object to this?

Christopher J Peikert writes:
> Dan, can you clarify whether you consider "amount of rounding" (e.g.,
> number of low bits dropped) to be part of a lattice scheme's
> "parameter set"?

Each scheme submitted to NISTPQC is required to specify its parameters, and in some cases these parameters include amounts of rounding, so in those cases the answer to this question is yes by definition.

> As we know, different amounts of rounding will tend to yield different
> Core-SVP hardness numbers.

No, not for RLWE/MLWE schemes. The Core-SVP metric for parameter sets, the metric that NIST says it feels "does indicate which lattice schemes are being more and less aggressive in setting their parameters", has always made various simplifications for the sake of supposedly being "conservative", and this is one of those simplifications.

> Round-3 Kyber does a small amount of rounding that Round-2 Kyber
> didn't do; as one would expect, this slightly increased the associated
> Core-SVP hardness.

No, it didn't. See above.

As a side note, the word "slight" is misleading in this context. The submission's analysis is consistent with the possibility of Kyber-512 being hundreds of times cheaper to break by known attacks than AES-128; the submission's best guess is the other way around, but for such an aggressive NISTPQC proposal one can't simply wave away a claimed several-bit change as a "slight" change.

> What's the objection?

Anyone reading my original message can see various questions that remained unanswered at the time of your message (and that obviously weren't answered by your message). For example: "Is the round-3 Kyber submission claiming that round-3 Kyber-512 is 2^118 in 'the CoreSVP metric', the metric that NIST says it's using to compare how 'aggressive' lattice schemes are, the same metric used in other submissions?"

For moving the analysis forward, one tries to understand the arguments leading to different conclusions, identify the points of agreement and the points of dispute, etc. It's not helpful to pretend that disputed conclusions aren't disputed: "as we know"; "as one would expect"; "what's the objection?"; etc.

> (If this thread exists only because of some semantic dispute about
> whether "amount of rounding" is part of the "parameter set" or not, I
> will be disappointed but not surprised.)

I'm unable to figure out what your logic is supposed to be here.

If a metric assigns numbers to parameter sets for an MLWE cryptosystem by looking only at the underlying MLWE problems, then by construction the metric will _not_ be influenced by parameters that don't affect those problems, such as what the "Estimate" paper calls "bit dropping".

> > However, when I question the round-3 Kyber submission's claim that
> > round-3 Kyber-512 has Core-SVP 2^118, suddenly people jump in to
> > attack the whole notion that Core-SVP attaches a number to each parameter set.
> I don't see anybody disputing that notion.

Martin Albrecht, in the message you claimed to be agreeing with: "I'm
confused: Core-SVP is a methodology for estimating the cost of blockwise lattice reduction algorithms like BKZ not a methodology for setting up lattices from LWE."

Each of the metrics I've been talking about, including what NIST calls "the CoreSVP metric" for parameter sets, is structured as follows:

```
  parameter set
    -> underlying LWE/... instance
      -> block size (plus dimension, for some of the metrics)
        -> cost
```

Martin's statement claims that Core-SVP is only the third step, going from block size to cost, and not even the second step, going from LWE to a block size (via a choice of lattice), never mind the first step.
This cannot be reconciled with NIST, in NIST IR 8309 and elsewhere, talking about Core-SVP as a metric for parameter sets.

> I see people (rightly) considering "amount of rounding" as part of
> Kyber's parameter set that is analyzed with Core-SVP.

Do you claim that the "Estimate" metrics for parameter sets look at the number of dropped bits? If so, how do you reconcile this with the "Estimate" paper specifically saying the opposite?

Do you claim that what NIST calls "the CoreSVP metric" for parameter sets, and said it's using for comparisons, looks at the dropped bits?

Do you claim that what NIST calls "the CoreSVP metric" pre-quantum _isn't_ the "Estimate" metric labeled 0.292 beta? (Similarly 0.265 beta post-quantum.) If so, what do you claim this metric is?

> Thanks for the reference. But I read it as a neutral observation

I was hoping to eliminate this pointless tangent by writing "readers who check the context won't believe you if you try to argue that this isn't being presented as criticism", but, since you insist:

* The message starts by citing D'Anvers. Anyone who follows up on this can see that D'Anvers identified an apparently unfixable error in the round-1 Kyber security "proof", and that this error was acknowledged and led to modifications in round-2 Kyber.

* The message continues by claiming that the D'Anvers observation "also applies to NTRU LPRime". Anyone who knows the context reads this as criticism.

* After the "cannot directly rely" sentence that I quoted, the message continues by claiming, e.g., that the "modelling" is less "adequate" than it could be. Reading this as criticism doesn't require knowing the Kyber context.

* Followup messages from the same author claimed, e.g., that the email was supposed to help "clarify" that "NTRU LPRime does not enjoy a security proof that is analogous to that of the LPR scheme". Still claiming this isn't criticism, Dr. Peikert?

I'll add a few side notes for readers who are in fact neutral and would like to understand the technical content, or rather lack of such, in the above criticism:

* No parameter set under consideration for NISTPQC standardization has a worst-case-to-average-case proof. Such proofs are the core of the "provable security" advertising for lattices, but are too loose to say anything about any of the NISTPQC parameter sets.

* If those theorems are disregarded then the provable-security picture for LPR etc. boils down to a trivial key-ciphertext split. Section 7 of https://cr.yp.to/papers.html#latticeproofs explains how easy these splits are to prove, uses NTRU LPRime as an example, and then explains how these splits damage security review.

* If, however, proposed cryptosystems are allowed to claim security on the basis of theorems that in fact apply only to larger cryptosystems, then _all_ of the lattice submissions have the same worst-case-to-average-case reductions. See Section 9 of https://cr.yp.to/papers.html#latticeproofs.

* When you see someone trying to make you think that LWE has a better proof picture than LWR by criticizing LWR >= LWE, and criticizing the worst-case-to-average-case reduction for LWR, as needing larger cryptosystems, it's useful to ask whether they aren't giving equal air time to criticizing the worst-case-to-average-case reduction for LWE as needing larger cryptosystems.

* I asked the author of the criticism quoted above for clarification
  (e.g., "A proof for anything like Round5 or Saber would not qualify
  as 'enjoy a security proof that is analogous', because the starting
  assumption is hardness of Ring-LWR/Module-LWR instead of
  Ring-LWE?") and he refused to answer.

As for the claim that the "modelling" is less "adequate" than it could be, the model was always clearly labeled as a simplification. Equating this to an outright proof error in Kyber is indefensible. One can always generically criticize simplified models for their lack of accuracy, or generically criticize more accurate models for their complications, but there's a fascinating pattern in how such criticisms are allocated across different NISTPQC submissions:

* The submission mentioned above presented a simplified attack model
  (again, clearly labeled as such) and was criticized for this.

* When round-1/round-2 _Kyber_ used a simplified security metric, it
  was praised for supposedly being "conservative".

* NIST's criticism of measuring "generic lattice security
  differently" was regarding the same _non-Kyber_ submission, which
  beyond reporting Core-SVP had also gone to a ton of work to report
  more accurate estimates (clearly labeled as separate estimates).

* Now that round-3 _Kyber_ switches to a more complicated security
  metric (and confusingly labels this as if it weren't a change),
  it's praised for the increased accuracy.

Whether or not the _intention_ of these inconsistencies is to promote Kyber, it's troubling to see NIST's four-year failure to take even the most basic procedural steps to ensure consistent evaluation of submissions. When NISTPQC evaluation criteria are being inconsistently applied, NIST's response should be to promptly clarify the criteria--- not to stall or to make up excuses for the situation. NIST should also, for each criterion, insist on doing a cross-submission comparison, both to test the criterion's clarity and to support doing the comparisons that NIST claimed from the outset it would be doing.

Look at how many cross-submission classifications and tables there were in the 1268-word-per-submission document

https://nvlpubs.nist.gov/nistpubs/jres/104/5/j45nec.pdf

regarding 15 submissions to the AES competition. Compare this to how few such comparisons there are in the 572-word-per-submission document

https://nvlpubs.nist.gov/nistpubs/ir/2020/NIST.IR.8309.pdf

regarding 26 more complicated submissions to NISTPQC. The AES document explicitly goes through various criteria and tries to classify each submission under each criterion, while the NISTPQC document leaves it entirely to the reader to figure out how comments about each submission map to criteria---and to figure out how many criterion-submission pairs are missing. NIST promised to be more transparent after Dual EC, but instead seems to have become even _less_ transparent, and one wonders whether NIST is actively trying to hide NISTPQC evaluation errors.

> -- as "Not sure what implications this remark has, though" makes clear

For readers who aren't starting from denial of the fact that the message is criticism, the statement "Not sure what implications this remark has"
sounds like "I haven't identified an attack exploiting X". This doesn't contradict the message being criticism of X.

> -- so will file this as another example of you mischaracterizing others.

Dr. Peikert has indeed repeatedly accused me of mischaracterizations, and has also repeatedly pointed to his own pattern of accusations. The rhetorical device here is similar to the "As we know" etc.

---Dan

From: pqc-forum@list.nist.gov on behalf of D. J. Bernstein <djb@cr.yp.to>
Sent: Thursday, December 17, 2020 1:31 PM
To: pqc-forum
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER
Attachments: signature.asc

'dustin...@nist.gov' via pqc-forum writes:
> If this analysis is correct, then Kyber clearly meets the security
> categories defined in the CFP.

Let's assume, arguendo, that the analysis in the round-3 Kyber submission is correct. I still don't see how NIST reaches its conclusion that Kyber "clearly meets" the NISTPQC security categories.

The round-3 Kyber submission says 2^151 "gates" for Kyber-512, but it also says that this 151 is +-16 given the "known unknowns". This has nothing to do with the possibility of new attacks: it's saying that, because of _known_ question marks about the analyses of _known_ attacks, these attacks could be using anywhere between 2^135 and 2^167 "gates".
For comparison, NIST says AES-128 key search is 2^143 "gates".

So where is NIST's "clearly meets" claim coming from? Is NIST claiming that Kyber-512 "clearly" needs at least 2^143 "gates" to break? This is a stronger claim than what I see in the submission's analysis. Where is this claim coming from?

Also, which "gates" is NIST talking about? The literature has many different definitions of gate sets, often producing very different numbers, so leaving any ambiguity here is begging for errors.

Even if Kyber's "gates" are defined somewhere _and_ match NIST's "gates", a cryptanalyst beating 2^151 "gates" is going to have people responding that Kyber only claimed 2^135 through 2^167, so why is NIST treating the analysis as claiming >=2^143? Did NIST miss the +-16?

Or is NIST saying that 2^135 "gates" is "clearly" okay because of the (unquantified) memory-access costs? This is a different story, and the details would be particularly interesting for other submissions trying to figure out the meaning of NIST's ill-defined security requirements.

> This email serves to respond to process questions that arose in this thread.

There are various procedural questions and requests for NIST in this thread that remain unanswered. For example: "Has NIST already made its round-3 Core-SVP comparison table? If not, why not, and what's the schedule for making this table? Assuming the table has been made already: Can you please post it publicly for review?"

Is it correct to interpret "serves to respond" as a refusal by NIST to answer these questions? If not, what _does_ it mean? Surely NIST isn't trying to claim that NIST _has_ answered these questions.

> The merit of technical claims is a research matter for the community
> to address.

NIST claimed that "the public statements we've made on the forum and in our report are sufficient for any submission team working in good faith to determine what parameter sets will be uncontroversial, controversial and unacceptable for the claimed security levels given the current state of knowledge". This is a claim about what's clear _now_ (more precisely, what was clear at the time), not a claim about future research.

My questions challenging this claim have ranged from content questions (e.g., "Which side of the line was round-2 Kyber-512 on? Which side of the line is round-3 Kyber-512 on?") to procedural questions (e.g., "How do we determine the answers to these questions from publicly available information?") Is NIST now making the bold claim that round-3 Kyber-512 is "uncontroversial"? Or merely that it isn't "unacceptable"? Meanwhile there's no answer regarding round-2 Kyber-512---which is surprising, given that the official evaluation criteria for NISTPQC include "maturity of analysis".

More broadly, the transparency principles from

https://www.nist.gov/system/files/documents/2017/05/09/VCAT-Report-on-NIST-Cryptographic-Standards-and-Guidelines-Process.pdf

ask for much more detailed information than what NIST has been releasing so far in NISTPQC, and they aren't limited to procedural information.

---Dan

From: Christopher J Peikert <cpeikert@alum.mit.edu>
Sent: Thursday, December 17, 2020 1:54 PM
To: pqc-forum; pqc-comments
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

On Thu, Dec 17, 2020 at 10:20 AM D. J. Bernstein <djb@cr.yp.to> wrote:

> Executive summary: NIST promoted and relied on a particular metric that
> assigns security claims to lattice parameter sets. Round-3 Kyber-512 has
> switched to another metric so as to be able to claim several bits more
> security. Claims that there hasn't been a switch are simply not true.

This repeated claim that the CoreSVP metric was "switched" in Round-3 Kyber is based entirely on the premise that the CoreSVP metric can account for random error, or deterministic error ("rounding"), but *not* a combination of the two.

This is a silly premise on its face. Moreover, the pointers to the literature that have been offered in support of the premise actually undermine it. (See below.)

As far as I can tell, every technical objection raised by Dan in this thread is founded on his own faulty definition of the CoreSVP metric, which equates the metric itself with specific modeling choices the "Estimate" project may have made in estimating the metric for LWE-like problems and parameters. Without this faulty definition, the objections fall apart completely.

> Details: Let's start with the "Estimate all the {LWE, NTRU} schemes!"
> page, along with its corresponding paper. Each column on the page is the
> output of a metric assigning a number, a claimed security level, to each
> parameter set for each lattice submission.

Yes, though let's be clear: the "Estimate" project did not define the CoreSVP metric, nor did it set the rules for how it can be applied to LWE-like problems. (It has never claimed to do either of these things.)

Instead, the project estimated the values of the metric for various parameter sets, by mapping those parameters to lattice problems, optimizing the block size beta, etc.

> Each of the "Estimate" metrics for parameter sets, and in particular the
> Core-SVP metric, has the following two structural features:
>
> * The metric for RLWE/MLWE cryptosystem parameter sets is purely a
>   function of the underlying RLWE/MLWE problems.
>
> * The metric for RLWR/MLWR cryptosystem parameter sets is purely a
>   function of the underlying RLWR/MLWR problems.
>
> The metric uses the underlying problem _even if_ the cryptosystem
> releases less information to the attacker than the underlying problem.

No, the CoreSVP metric does not "use the underlying problem _even if_ ..." The "Estimate" project may have chosen to compute CoreSVP estimates for the underlying problems, but that's irrelevant to the definition of CoreSVP.

> For example, as noted in Section 2.1 of the "Estimate" paper, the metric
> asks how secure full MLWE samples are _even if_ the MLWE cryptosystem
> actually removes some bits of the samples ("bit dropping").

That's not at all what Section 2.1 says, and what it does say goes against your central premise and conception of CoreSVP.

First, Section 2.1 does not even mention CoreSVP at all, so it can't possibly note that "the [CoreSVP] metric asks how secure..." In fact, CoreSVP doesn't appear until 8 pages later, and the term appears only once in the entire paper!

(The reader may wish to recall this misrepresentation of the literature when similar claims are made in the future.)

Section 2.1 merely recalls variants of the LWE problem that are relevant to many proposed cryptosystems. Here is what it says about rounding / "bit dropping":

"There is also a class of LWE-like problems that replace the addition of a noise term by a deterministic rounding process... We can interpret this as a LWE instance... where [the error] e is chosen from a uniform distribution on the set { -q/(2p) + 1, ... , q/(2p) } [Ngu18]. The same ideas apply to the other variants of LWE that use deterministic rounding error..."

In summary: deterministic rounding is a form of error, and we can consider it as such in variants of LWE. This is fully consistent with the accounting for combined error and rounding when estimating CoreSVP. Then:

"Due to the way decryption works this bit dropping can be quite aggressive, and thus the noise in the second sample can be quite large. In the case of Module-LWE, a ciphertext in transit produces a smaller number of LWE samples, but n samples can still be recovered from the public key. In this work, we consider the n and 2n scenarios for all schemes. We note that, for many schemes, n samples are sufficient to run the most efficient variant of either attack."

In summary: the ciphertext noise can be "quite large" due to bit-dropping, but in many cases accounting for this won't give a better final CoreSVP number for the full scheme because there's a better attack on the public key itself. This is saying quite clearly that one may account for combined error and rounding when estimating CoreSVP!

A few final points in case there is any remaining confusion:

> \> As we know, different amounts of rounding will tend to yield different
> \> Core-SVP hardness numbers.
>
> No, not for RLWE/MLWE schemes. The Core-SVP metric for parameter sets,
> the metric that NIST says it feels "does indicate which lattice schemes
> are being more and less aggressive in setting their parameters", has
> always made various simplifications for the sake of supposedly being
> "conservative", and this is one of those simplifications.

To be clear, this reply is based entirely on a bizarre definition of CoreSVP that is not supported by any of the literature (and certainly not the cited paper, as one can see from the text I've quoted above).

> \> (If this thread exists only because of some semantic dispute about whether
> \> "amount of rounding" is part of the "parameter set" or not, I will be
> \> disappointed but not surprised.)
>
> I'm unable to figure out what your logic is supposed to be here.

> If a metric assigns numbers to parameter sets for an MLWE cryptosystem
> by looking only at the underlying MLWE problems,

To be clear, the CoreSVP metric is not limited to doing this ("looking only at..."), so this is a false premise.

> > -- so will file this as another example of you mischaracterizing others.
>
> Dr. Peikert has indeed repeatedly accused me of mischaracterizations

Thanks to your description of Section 2.1 in the "Estimate" paper, now we have another good example!

Sincerely yours in cryptography,
Chris

From: Mike Hamburg <mike@shiftleft.org>
Sent: Thursday, December 17, 2020 4:40 PM
To: Christopher J Peikert
Cc: pqc-forum; pqc-comments
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

Hi all,

I'd like to add my 2c to this discussion.

I have no interest in mudslinging and accusations, nor do I want to join the battle about patents or whether NIST is unfairly favoring Kyber. Nor indeed do I have an opinion on how 2^118 CoreSVP compares to NIST's stated goal of 2^143 gates, or even what a "gate" is. But I do have an opinion about security "metrics".

In my view, CoreSVP is simply a shortcut to modeling the difficulty of a lattice reduction problem. The shortcut is:

1) Model the problem as a distribution of lattices in some dimension, and requirements the size of the reduced basis vector(s);
2) Estimate the blocksize that BKZ would need to solve the lattice problem using your favorite model of BKZ's output, generally some sort of GSA;
3) Estimate the number of nodes visited by the SVP solver, using your favorite estimate. The most popular seems to be 2^(0.292 beta).

This shortcut enables us to estimate the security of a given cryptosystem, complete with parameters, as follows:

1) Prove or at least convincingly argue, with some tightness and under some assumptions (ROM etc) that an adversary must solve a certain underlying LWE / MLWE / RLWE / NTRU instance to break the system.

2) For each of several standard attacks against the underlying problem (e.g. primal, dual, hybrid):
2a) Compute the lattice problem that the attack must solve:
2b) Estimate the CoreSVP difficulty of that problem.
2c) Add any large additional work that must be performed (e.g. collision finding for hybrid attacks, brute force to prepare CCA queries, etc) and divide by the probability of success.
2d) Tune the parameters of the attack, if applicable, to minimize this result

3) Estimate that the security of the system is about the minimum difficulty across all the standard attacks.

There are many ways for authors to diverge at each of these steps: they could take a different metric from 2^(0.292 beta), or a different model of BKZ's output, or a different model of the underlying problem for the cryptosystem. They could use a loose proof or a tight one. They could assume that BKZ returns only one useful vector, or more than one, which may affect the size and thus the blocksize required, etc. So CoreSVP estimations are not always consistent or perfectly comparable. Furthermore, it could be that the techniques themselves don't reflect the security of systems in a consistent way, e.g. that somehow CoreSVP techniques misjudge the true security of schemes by an amount that depends whether they have { large / small, uniform / binomial, fixed-weight / i.i.d, etc } noise. And as the participants in this thread know well, the tightness of different schemes' proofs varies somewhat, especially in the QROM.

However, it's better that everyone at least use a consistent estimation procedure ("metric") in the places where their underlying models are the same. For example, it's best if everyone is using 0.292 instead of some using this and others using 0.2975. Otherwise we will see differences in estimates that differ even more from the true security difference of cryptosystems. That is, having everyone evaluate their scheme using mostly shared steps in their methodology might lead to flawed comparisons, but it's better than if everyone uses completely different methodology. Of course, non-Core-SVP evaluations are also useful, especially if the same evaluation is applied to several systems for comparison purposes.

When NIST talks about "CoreSVP security of parameter sets", I do not believe they mean that there is a function "CoreSVP(.)" which takes an entire parameterized cryptosystem and somehow returns a number. I also do not believe they mean whatever is recorded in the latest "Estimate all the LWE" webpage. Instead, I think they mean the authors' estimated security levels, which the authors of those schemes have convincingly argued using the above CoreSVP methods, with convincing choices at the steps which allow choices. It is therefore imprecise for them to call it a "metric". But in any case, I don't think it's makes sense to go from "NIST's use of CoreSVP is imprecise" to "Kyber team's use of CoreSVP is meaningless or dishonest" or even that "NIST's use of CoreSVP is meaningless or dishonest".

In the case of Kyber, Round 2 Kyber's security estimation section discards the rounding and sets up the underlying lattice problems as MLWE, but Round 3 additionally uses what might be called an MLWE+R problem, by considering the rounding as well. In either case, the authors then estimate their security using a CoreSVP method. "Estimate All the LWE" needs to simplify its approach so as not to get bogged down in the details of every cryptosystem, and thus also models Kyber as MLWE only and not MLWE+R.

Kyber's increase in the security estimate from Round 2 to Round 3 comes at the cost of a less convincing argument. There is the issue that MLWE+R is even less well-studied than MLWE, and that Kyber is assuming that the rounding noise simply combines with the additive noise. But every one of Kyber's remaining competitors uses a rounding problem as the foundation of security, so if rounding noise turns out to be much weaker than additive noise, those other systems are more likely to break, whereas Kyber itself would not be defeated by such a breakthrough. So maybe this isn't a huge problem.

However — perhaps I'm missing something — it appears to me that Kyber's security proof only assumes MLWE. Therefore, basing the estimate on MLWE+R requires the additional assumption that an attacker would have to attack the system this way. This seems like a reasonable assessment, but should be supported by precise definitions and a proof. Furthermore, the dual attack should be evaluated. So while I think an overall 2^118 CoreSVP could probably be convincingly argued for Kyber, the existing argument falls short.

If these two issues were fixed, and the general writing in that section made clearer, I think that the MLWE+R evaluation of Kyber would be a better comparison point to the other candidates. This is because it would better reflect the relative difficulty of currently-known lattice attacks on those systems. That is, it would bring the model of the difficulty of attacking Kyber more in line with the models of the difficulty of attacking Saber and NTRU.

Regards,
— Mike

From: daniel.apon <daniel.apon@nist.gov>
Sent: Monday, December 21, 2020 8:36 PM
To: pqc-forum
Cc: mi...@shiftleft.org; pqc-forum; pqc-comments; cpei...@alum.mit.edu
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

Hi Mike,

"Therefore, basing the estimate on MLWE+R requires the additional assumption that an attacker would have to attack the system this way. This seems like a reasonable assessment, but should be supported by precise definitions and a proof. Furthermore, the dual attack should be evaluated. So while I think an overall 2^118 CoreSVP could probably be convincingly argued for Kyber, the existing argument falls short."

I second this. This is quite a decent point.

As far as I'm personally concerned, this is a reasonable response to every question directed toward NIST in this thread to date.

Happy holidays all,
--Daniel Apon
On Thursday, December 17, 2020 at 4:40:48 PM UTC-5 mi...@shiftleft.org wrote:

> Hi all,
>
> I'd like to add my 2c to this discussion.
>
> I have no interest in mudslinging and accusations, nor do I want to join the battle about patents or whether NIST is unfairly favoring Kyber. Nor indeed do I have an opinion on how 2^118 CoreSVP compares to NIST's stated goal of 2^143 gates, or even what a "gate" is. But I do have an opinion about security "metrics".
>
> In my view, CoreSVP is simply a shortcut to modeling the difficulty of a lattice reduction problem. The shortcut is:
>
> 1) Model the problem as a distribution of lattices in some dimension, and requirements the size of the reduced basis vector(s);
> 2) Estimate the blocksize that BKZ would need to solve the lattice problem using your favorite model of BKZ's output, generally some sort of GSA;
> 3) Estimate the number of nodes visited by the SVP solver, using your favorite estimate. The most popular seems to be 2^(0.292 beta).
>
> This shortcut enables us to estimate the security of a given cryptosystem, complete with parameters, as follows:
>
> 1) Prove or at least convincingly argue, with some tightness and under some assumptions (ROM etc) that an adversary must solve a certain underlying LWE / MLWE / RLWE / NTRU instance to break the system.
>
> 2) For each of several standard attacks against the underlying problem (e.g. primal, dual, hybrid):
> 2a) Compute the lattice problem that the attack must solve:

From: daniel.apon <daniel.apon@nist.gov>
Sent: Monday, December 21, 2020 8:52 PM
To: pqc-forum
Cc: Apon, Daniel C. (Fed); mi...@shiftleft.org; pqc-forum; pqc-comments; cpei...@alum.mit.edu
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

P.S. Regarding Dan's question: *"Is NIST now making the bold claim that round-3 Kyber-512 is "uncontroversial"?"*

I'd point out that controversy can be instigated by anyone at anytime by one's simply posting a 30-paragraph email to the pqc-forum, and that whether non-NIST entities post such messages to the pqc-forum is entirely outside of NIST's control.

So, the question of whether a particular parameter set falls on the side of "controversial" vs. "uncontroversial" is -- unfortunately -- entirely meaningless from NIST's point of view.

On Monday, December 21, 2020 at 8:36:06 PM UTC-5 daniel.apon wrote:

> Hi Mike,
>
> *"Therefore, basing the estimate on MLWE+R requires the additional assumption that an attacker would have to attack the system this way. This seems like a reasonable assessment, but should be supported by precise definitions and a proof. Furthermore, the dual attack should be evaluated. So while I think an overall 2^118 CoreSVP could probably be convincingly argued for Kyber, the existing argument falls short."*
>
> I second this. This is quite a decent point.
>
> As far as I'm personally concerned, this is a reasonable response to every question directed toward NIST in this thread to date.
>
> Happy holidays all,
> --Daniel Apon
> On Thursday, December 17, 2020 at 4:40:48 PM UTC-5 mi...@shiftleft.org wrote:
>
>> Hi all,
>>
>> I'd like to add my 2c to this discussion.
>>
>> I have no interest in mudslinging and accusations, nor do I want to join the battle about patents or whether NIST is unfairly favoring Kyber. Nor indeed do I have an opinion on how 2^118 CoreSVP compares to NIST's stated goal of 2^143 gates, or even what a "gate" is. But I do have an opinion about security "metrics".
>>
>> In my view, CoreSVP is simply a shortcut to modeling the difficulty of a lattice reduction problem. The shortcut is:
>>
>> 1) Model the problem as a distribution of lattices in some dimension, and requirements the size of the reduced basis vector(s);
>> 2) Estimate the blocksize that BKZ would need to solve the lattice problem using your favorite model of BKZ's output,

From:	Peter Schwabe <peter@cryptojedi.org>
Sent:	Wednesday, December 23, 2020 6:31 PM
To:	Apon, Daniel C. (Fed)
Cc:	pqc-forum; mi...@shiftleft.org; pqc-comments; cpei...@alum.mit.edu
Subject:	Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

"'daniel.apon' via pqc-forum" <pqc-forum@list.nist.gov> wrote:
> Hi Mike,

Hi Mike, hi Daniel, hi all,

> *"Therefore, basing the estimate on MLWE+R requires the additional
> assumption that an attacker would have to attack the system this way.
> This seems like a reasonable assessment, but should be supported by
> precise definitions and a proof. Furthermore, the dual attack should be evaluated.
> So while I think an overall 2^118 CoreSVP could probably be
> convincingly argued for Kyber, the existing argument falls short." *
>
> I second this. This is quite a decent point.

We fully agree and will send a more detailed version of the LWE+R argument to the list, most likely toward the end of January (we're all somewhat busy until then). As a spoiler, there almost certainly won't be any exciting new insights over what is already informally stated in the current document. In particular, the claimed 118 bits of CoreSVP security comes from the analysis of the LWE problem in the public key generation; while the combination of error+rounding noise in the ciphertext generation actually gives an even larger CoreSVP security level.

All the best and happy holidays to everybody!

The Kyber team

From:	D. J. Bernstein <djb@cr.yp.to>
Sent:	Friday, January 1, 2021 7:19 AM
To:	pqc-comments
Cc:	pqc-forum
Subject:	Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER
Attachments:	signature.asc

Main goal of this message: Unify notation and terminology for the systems under discussion, to provide a framework for evaluating claims regarding what's the same, what's different, etc. This message is organized around a case study, namely a puzzling claim of a dividing line between Ding's patented work and newer compressed-LPR proposals.

The unification in https://cr.yp.to/papers.html#latticeproofs is narrower in that it's limited to round-2 NISTPQC submissions (although it also covers Quotient NTRU) and in that it focuses on the aspects of the cryptosystems that arise in security proofs. To the extent that there's an overlap in the coverage, the notation is synchronized, and follows the known idea of using ECDH-like notation for noisy DH.

Vadim Lyubashevksy writes (email dated 2 Dec 2020 20:47:11 +0100):
> No version of the LPR paper presents "reconciliation".

Let's say Alice sends A = aG+e, and Bob sends B = Gb+d, where lowercase letters are small quantities. Alice computes aB = aGb+ad. Bob computes Ab = aGb+eb. Can we agree that "noisy DH" is a good name for this?

At this point Alice and Bob have computed something that's _similar_ but not _identical_. People who say "reconciliation" in the noisy-DH context are referring to an extra step of Alice and Bob computing the _same_ secret with the help of further communication.

The normal reconciliation pattern is that Bob sends C = Ab+M+c, where M is in a sufficiently limited set to allow error correction: for example, each position in M is restricted to the set {0,floor(q/2)}. Alice now finds C-aB = M+c+eb-ad, and corrects errors to find M. Bob also finds M.

Example 1: Take the variables in R_q = F_q[x]/(x^n+1) where n is a power of 2, and take C as Ab+M+c.

This cryptosystem, with some further restrictions on distributions etc., appeared in the revised version of the LPR paper (not with the name "reconciliation", I agree) and in talk slides as early as April 2010.
This is pretty much always what people mean by "the LPR cryptosystem", modulo the question of which number fields are considered; I'm focusing here on the commonly used case of power-of-2 cyclotomic fields.

Example 2: Special case of Example 1, with the extra requirement of C being in a more restricted set, so that C can be sent in less space.

This is what I call "compressed LPR". The LPR cryptosystem was in the revised version of the LPR paper, but compressed LPR wasn't; see below for further discussion of the history. (Another compression mechanism, orthogonal to what I'm covering here, is to send, e.g., only 256 coefficients of C, or 256 sums of coefficients of C.)

Example 3: Special case of Example 2, where C is a deterministic function of Ab+M: e.g., C is obtained by rounding Ab+M to a more restricted set.

Example 4: Special case of Example 3, where M is the output of rounding -Ab to have each position in {0,floor(q/2)}. This puts each position of Ab+M between about -q/4 and q/4.

For definiteness let's say that C is then obtained by rounding each position of Ab+M to {-ceil(q/8),ceil(q/8)}. The difference c = C-Ab-M is then between about -q/4 and about q/4.

Notice that each position of the shared secret M is now communicating one bit of information about Ab, namely whether that position of Ab is closer to 0 or to q/2. One can also tweak the M range to communicate more/fewer/different bits of Ab, as the next example illustrates.

Example 5: Another special case of Example 3, with the following extra rules:

  * Choose each position of M is in {0,1} to limit each position of
    Ab+M to _even_ numbers.

  * Round Ab+M to C where each position is 0 or 2 floor(q/4), using
    an _even_ difference c between about -q/4 and about q/4.

  * Choose _even_ d and e.

Now decoding M from M+c+eb-ad is simply reducing mod 2. The secret M obtained in the end is exactly the secret Ab mod 2.

(Readers coming at all this from an FHE background will recognize the difference between Example 5 and Example 4 as being analogous to the difference between 2009 DGHV and 2000 Cohen/2003 Regev. But pointing to inefficient prior art won't win a patent case; see below.)

I've chosen this progression of examples to first make C deterministic and then make M deterministic. Some people prefer C being randomized, for example by first adding a random error and then rounding, but independently of this one can still make M deterministic, so that M communicates information about Ab as in Examples 3, 4, and 5. One can also make A and/or B deterministic; these choices are independent.

All of these examples have the following features:

  (P) There's a univariate polynomial quotient ring R_q.
  (A) Alice sends a noisy DH element A = aG+e in the ring.
  (B) Bob sends a noisy DH element B = Gb+d in the ring.
  (C) Bob sends C = Ab+M+c in the ring.
  (D) Alice recovers the shared secret M.

A+B are noisy DH, allowing Alice and Bob to _approximately_ agree on a shared secret in the ring. There was already a 2009 publication of a P+A+B system.

C+D, in the context of A+B, is reconciliation, _exactly_ agreeing on a shared secret in the ring. As far as I know, the first publication of a P+A+B+C+D system was in April 2010, the aforementioned talk re LPR.

All of the compressed-LPR examples, everything starting from Example 2, also have the following feature:

(S) Bob squeezes C into less space than an R_q element.

Example 1, original LPR, doesn't have this feature. As far as I know, the first publication of a P+A+B+C+D+S system, and the first publication of a compressed-LPR system, was the 2012 Ding cryptosystem, which is essentially Example 5.

How do we stop _all_ P+A+B+C+D+S systems from being covered by Ding's patent? Here are some ideas:

* The non-novelty/obviousness argument: i.e., the argument that the claimed invention was already in the prior art, or at least that the differences are such that the claimed invention "as a whole" was obvious to people of ordinary skill in the art.

  If you spend time reading through patent cases then you'll see that this is one of the main points of dispute, but that even glaringly obvious ideas have a considerable chance of being ruled unobvious in court. The basic problem here is that patents come to court with a presumption of validity (both novelty and unobviousness), and overcoming this presumption requires "clear and convincing" evidence. The "clear and convincing" rule applies even when the patent office didn't consider the literature in question; see _Microsoft v. i4i_, 564 U.S. 91 (2011).

  Yes, this system is tilted in favor of patent holders. Welcome to the real world.

  In the case of Ding's patent, we aren't talking about the most obvious ideas. The plaintiff's lawyer will pull out one expert witness after another saying that efficiency improvements now claimed in retrospect to be obvious weren't obvious at the time; and will then pull out the big gun, 2014 Peikert. How is the defendant's lawyer supposed to argue that saving space compared to LPR was obvious in 2012 to people of ordinary skill in the art, given that 2014 Peikert claimed that saving space compared to LPR was new, the result of an "innovation" in 2014 Peikert?

  If someone can find a publication before April 2012 that saves space compared to LPR, doing not just P+A+B+C+D but also S, fantastic! But so far each allegedly important piece of prior art that I've seen is missing something. In analyses of patent threats, it's a gigantic mistake to gloss over the difference between "there's prior art for X+Y" and "there's prior art for X and there's prior art for Y".

* The non-infringement argument. This time it's the defendant's lawyer trying to argue that there's a "substantial" difference between the patent claims and what the defendant is doing. The defendant is starting in trouble at this point: a key is being exchanged, and there's a rounding mechanism saving space compared to LPR, so what exactly is the "substantial" difference from what's claimed in the patent?

  "Substantial" and "unobvious" are different legal standards (even

though the analyses overlap), and they're starting from different
baselines---the prior art _plus_ the patent, vs. the prior art
before the patent. Furthermore, courts do _not_ presume
non-infringement; whichever side has the preponderance of evidence
regarding infringement wins.

For an academic paper, throwing around a few sentences, even a footnote, can suffice to make the reader believe that the paper has some important distinction from prior work. For a patent lawsuit, the plaintiff and defendant typically spend millions, and every word is scrutinized by all
sides:

* A claimed distinction that isn't crystal clear will fail.

* A claimed distinction that isn't _correct_ will fail.

* A claimed distinction that's clear and correct, but that isn't more
"substantial" than any of the other clear correct distinctions that
have been rejected in patent cases, will also fail.

Efficiency improvements are easy to understand and are constantly ruled "substantial" and "unobvious", but this is exactly what's in Ding's
favor: his system squeezes C into less space than the prior art. It's normal in patent cases for defendants to try to avoid a patented efficiency improvement by interpolating between the prior art and the efficiency improvement, and it's normal for the patentee to win.

> In a *key exchange scheme* using "reconciliation" (see e.g.
> Jintai's talk
> https://csrc.nist.gov/CSRC/media/Presentations/Ding-Key-Exchange/
> images-media/DING-KEY-EXCHANGE-April2018.pdf where the word
> reconciliation is explicitly used so there can be no confusion as to
> what it means),

I see nothing in the talk spelling out the level of generality of the word "reconciliation". The examples of "reconciliation" mentioned in the talk seem consistent with what I wrote above regarding how the word is used in this context.

A patent case will settle on definitions on the words in the allegedly infringed patent claim, and then the plaintiff's lawyers will go through the details of matching up the defendant's device to the limitations in the claim, and explaining why any differences are not "substantial".
Mere differences in _terminology_ are not "substantial"; at best they force the plaintiff's lawyers to spend more time stripping away the differences and showing what the defendant is in fact doing.

So it's not as if avoiding the word "reconciliation" is going to save anybody from the 2010 patent (or the 2012 patent, which doesn't even use the word). I find the word useful in clarifying what's going on. Clarity is important so that readers can efficiently see what the technology is actually doing, in particular as a starting point for evaluating the patent threats.

> the users
> end up with a random shared key that neither party explicitly chose.

I understand that you're proposing to use the word "reconciliation"
differently from what I said above, but I'm unable to figure out what you think it means. (Again, a claimed distinction that isn't crystal clear will fail in court.) For example:

* If Bob takes M as RNG output, is he "explicitly" choosing M?

* Does it matter whether the RNG is a true RNG or a PRNG?

* Is Bob still "explicitly" choosing M if he passes RNG output through his own hash function? Does it matter what kind of hash function this is?

* Can the hash function also take external inputs? How about Alice's public key?

* Is Bob choosing M as bits from Ab not an example of "explicitly" choosing M? Why not? (This is what Examples 4 and 5 do, and you want those to not qualify as "explicitly" chosen, right?)

The general theme of these questions is directly related to typical practices in NISTPQC submissions and in applications.

Beyond my questions about what this "explicitly" dividing line is supposed to mean, I'm unable to figure out why "reconciliation" is supposed to be a sensible name for this line. The normal English usage of "reconciliation" fits the idea of having Alice and Bob communicate so as to come to exact agreement, whereas it doesn't seem to have anything to do with the extent to which they _chose_ the outputs. (Financial auditors reconciling accounts are taking data from other people, but they're still engaging in a reconciliation process.)

Also, since it seems helpful to be able to refer to the process of Alice and Bob turning their aGb+ad and aGb+eb into an exactly shared secret, what would you suggest calling this process, if not "reconciliation"?

> In a public key encryption scheme, the sender chooses the message that
> he wants both parties to have.

I agree that the PKE definition takes a message as input and communicates this message. This differs from the data flow in the KEM definition, and in various other key-exchange definitions.

Each of the examples above can be turned into a PKE or into a KEM, with the addition of various labels and steps that I haven't described (e.g., labeling (B,C) as "ciphertext"), and choices that I haven't described (e.g., for each variable that isn't otherwise constrained, whether it's generated randomly or as a function of another variable).

I agree that the choices made in turning Example 4 or 5 into a PKE or a KEM have to be consistent with the deterministic choice of M. Also, the choices made in turning Example 3 into a PKE or a KEM have to be consistent with the deterministic choice of C.

> Of course one can trivially convert a key exchange scheme into an
> encryption scheme (by adding an xor with the message), but this is not
> what's happening in Kyber/Saber.

I don't see how the difference here is less "trivial" than the conversion that you're calling "trivial". Let's go step by step through the details.

Given any of the above procedures to share a secret---let me relabel the secret as X here to avoid confusion---Bob can also append X xor U to the ciphertext where U is the actual user message, at which point Alice finds U. You're calling this transformation "trivial".

Let's take q as a power of 2 to simplify notation, and let's relabel the 0,1 bits in X and U as positions {0,q/2} in elements of R_q, so X xor U is the same as U-X. The whole transformation is still "trivial", right?
It's not a question of how the bits in X and U are labeled?

In Example 4, X already had positions {0,q/2} without any relabeling.
Let's apply this transformation, and call the result Example 6:

* Alice sends A = aG+e.

* Bob sends (Gb+d,Ab+X+c,U-X), with the above restrictions on X and c.

* Alice subtracts a(Gb+d) from Ab+X+c to obtain X+c+eb-ad, rounds to obtain X, and adds U-X to obtain U.

The incorporation of U-X into the ciphertext, to send U instead of just X, is "trivial", right?

Since X and U have positions in {0,q/2}, rounding to obtain X and then adding U-X is the same as first adding U-X and then rounding. Let's call the result Example 7:

* Alice sends A = aG+e.

* Bob sends (Gb+d,Ab+X+c,U-X), with the above restrictions on X and c.

* Alice adds U-X to Ab+X+c, subtracts a(Gb+d) to obtain U+c+eb-ad, and rounds to obtain U.

Is this not also a "trivial" step? It's not as _generic_ as the previous transformation, but does this make it _nontrivial_?

At this point there's nothing using Ab+X+c and U-X separately, so we might as well simply have Bob do the addition. Let's call this Example 8:

* Alice sends A = aG+e.

* Bob sends (Gb+d,Ab+U+c).

* Alice subtracts a(Gb+d) from Ab+U+c to obtain U+c+eb-ad, and rounds to obtain U.

But wait a minute. How is this different from Example 2? Exactly which of these examples are covered by "reconciliation"? Was the supposedly important distinction between "reconciliation" and not "reconciliation" crossed by something as _trivial_ as having Bob add two R_q elements instead of sending them to Alice to add?

Patent courts are continually faced with conflicting narratives. Lawyers on one side hype differences between the patent and what the defendant is doing, while downplaying differences from the prior art. Lawyers on the other side downplay differences between the patent and what the defendant is doing, while hyping differences from the prior art. This drives the _choices_ of what each side labels as "trivial", "obvious", etc.; these choices are challenged by the other side, and the court processes then dig into the details.

> You can see that there is a fundamental difference between the two
> approaches

If there's a "fundamental" difference, then I would expect each of my questions to be very easily answered. We would all see a crystal-clear definition of the dividing line, and we would all be able to check that the dividing line implies these answers, and then we could start _hoping_ that this dividing line would hold up in court.

Right now I'm having trouble seeing how the dividing line meets even minimal scientific standards of clarity.

> in the fact that there is slight bias in the shared key in Ding's
> scheme

Now I'm even more puzzled.

I agree that there's a slight bias in the shared secret in the first scheme Ding published. But one can easily remove the bias by tweaking Ding's scheme, for example by being careful about the exact ranges of variables in in Example 4. Are you saying that this tweak crosses the "fundamental" line between a "key exchange scheme" using "reconciliation" and a "public key encryption scheme" not using "reconciliation"?

2014 Peikert specifically says it's unbiased. Does this mean it isn't "key exchange" and isn't using "reconciliation"?

Also, people normally hash shared secrets for various reasons, one of those reasons being to remove biases. Does replacing the shared secret M with H(M) cross this "fundamental" line?

The only way I can see the defendant trying to use this is through attacking the words "similar rounding" in the patent. Here's how I'd expect the procedures to play out in court:

* There will be preliminary arguments about what the words in the patent mean---including what qualifies as "similar" rounding. Each side proposes definitions. In the U.S. (since 1996), the battle between definitions is resolved by the judge rather than the jury, which makes it somewhat more predictable.

* Of course the defendant's lawyers will _want_ the minimum possible interpretation: namely, nothing beyond the specific rounding schemes that Ding gave as examples. But they have zero hope of getting this: this would effectively eliminate the "similar" claim, and courts have a general rule against interpreting text in a way that makes it content-free.

* Meanwhile the plaintiff's lawyers will _want_ the maximum possible interpretation---without bumping into the prior art---so they'll propose broad definitions that focus on user-visible features of the rounding mentioned in the patent: "similar" rounding is rounding that "allows us to get a common key" and has good "communication and computation efficiency". There's no rule stopping them from getting this.

* There will then be a dispute about what "good" means specifically for the rounding. The plaintiff's lawyers (knowing that they need to avoid covering LPR) will propose "reduces space", and maybe the defendant's lawyers will counter-propose "produces only one bit of

information", which will be a tough sell---what's supposed to be so
special about one bit?

* The defendant's lawyers will propose much narrower definitions of
"similar" rounding, trying to cut it down by adding one dividing
line after another. For example, the defendant's lawyers will say
that rounding isn't "similar" unless it's biased.

* The judge will ask what "biased" means, and why this is supposed to
be important for the patent. The plaintiff's lawyers will ask how
the details of the defendant's argument that the rounding is
"biased" are supposed to be reconciled (ahem) with the patent
saying "We can also choose q to be even positive number and things
need slight modification."

* In the end the judge will select the simpler, and much broader,
interpretation proposed by the plaintiff's lawyers.

There could have been questions from the patent examiner about words in the patent and how they relate to prior art, which could have resulted in narrowing the patent claims, and then the judge will follow this.
However, I would expect any questions about "similar" rounding to have forced more precise claim wording, and none of the prior art that I've seen would have triggered such questions. Of course, it would be good to download the patent history and read through it for anything helpful.

> Also notice that Jintai has a Ring-LWE encryption scheme (page 14 of
> https://
> patentimages.storage.googleapis.com/53/08/b7/b93d5b6b131e46/US9246675.
> pdf) which is like LPR and *does not* (unless I am reading something
> wrong) do any rounding / compression - so it just outputs two elements
> D1,D2 (which could be matrices over some ring).

It's certainly common sense to ask why the patent didn't give a compressed version of this example. Structurally, one can try to use this question in court in an inequivalence argument, the same way that the following questions support unobviousness arguments:

* If the LPR cryptosystem was already obvious in February 2010, then
why did the original May 2010 Springer version of the LPR paper
highlight a bigger, slower cryptosystem?

* If it was already obvious in April 2012 that one can compress LPR,
then why wasn't this in the April 2012 revision of the LPR paper?

As the Supreme Court put it long ago:

But it is plain from the evidence, and from the very fact that it was
not sooner adopted and used, that it did not, for years, occur in
this light to even the most skillful persons. It may have been under
their very eyes, they may almost be said to have stumbled over it;
but they certainly failed to see it, to estimate its value, and to
bring it into notice. ... Now that it has succeeded, it may seem very
plain to any one that he could have done it as well. This is often
the case with inventions of the greatest merit. It may be laid down

> as a general rule, though perhaps not an invariable one, that if a new combination and arrangement of known elements produce a new and beneficial result, never attained before, it is evidence of invention.

Regarding "beneficial", new levels of efficiency are easy for courts to understand, so it's hard to imagine getting anywhere with the whole idea of killing the 2010 and 2012 patents by claiming analogies to older cryptosystems that are indisputably less efficient. This would be true even if the plaintiff's lawyers _didn't_ have the gift of 2014 Peikert claiming that nothing before 2014 Peikert was smaller than LPR.

For the same reason, if there's something that isn't literally within a patent claim, and if it's possible to construct an efficiency argument saying that the differences are "substantial", then the court will listen. But how would this work for the patents at issue here?

The relevant claims of the 2012 patent say "key exchange". What we see in KEMs, TLS, etc. is exchanging a key. Any deviations from the claims are regarding smaller points, such as details of the rounding method, and then the equivalence analysis asks whether _those_ differences are "substantial". I'm not seeing how efforts to distinguish "key exchange" from "encryption" are relevant here.

For the 2010 patent, the literal scope is broader (a "cryptographic method" for "communicating a confidential piece of information" etc.). I've sent a separate message comparing LPR point by point to the claim limitations; the only difference requiring an equivalence analysis was X^n-1 vs. X^n+1. I've also sent a separate message regarding the idea that switching from LPR to Kyber avoids "three central requirements" in the patent; two parts of this idea are simply false, and I explained how easy it will be for the lawyers to work around the third part, even in a fantasy world without the doctrine of equivalents.

---Dan

From:	pqc-forum@list.nist.gov on behalf of D. J. Bernstein <djb@cr.yp.to>
Sent:	Saturday, January 2, 2021 12:24 PM
To:	pqc-comments
Cc:	pqc-forum
Subject:	Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER
Attachments:	signature.asc

Efforts to accurately analyze lattice security are becoming more and more complicated. After many correction factors, the state-of-the-art estimates still don't exactly match experiments even for the simplest attacks, and there's no reason to think that small errors will remain small when they're extrapolated to cryptographic sizes. Meanwhile the algorithms themselves are becoming more and more complicated as people find new speedups. Each new speedup poses a new analysis challenge.

Comparing parameter sets across proposals requires further work to automate estimates. This work is error-prone, as illustrated by the miscalculation of the "Estimate" numbers for SABER. The difficulty of automation is magnified when estimates are complicated and unstable.

There are many simplifications in the "Estimate" work, and in particular in "the CoreSVP metric" for parameter sets, at the expense of accuracy.
The simplifications could reverse comparisons between proposals; could mislead people regarding cost comparisons to AES; and clearly contribute to a dangerous lack of awareness of ongoing advances in lattice attacks.
Do we recognize these issues as problems, and insist on doing better as a prerequisite for making decisions regarding lattice proposals? Or do we praise the simplifications, saying that the simplifications are conservative, that it's better to have a simplified comparison than nothing at all, and that this comparison tells us which lattice schemes are being more and less aggressive in setting their parameters?

One of the worst imaginable ways to answer these questions is to make an ad-hoc decision between praising accuracy and promoting a simplified metric, depending on whether the answer seems to favor submission X:

 * If a submission that _isn't_ X contains the most detailed survey in
 the literature of inaccuracies and potential inaccuracies in
 Core-SVP, and contains the state-of-the-art fixes for the most
 glaring inaccuracies (while clearly labeling different metrics),
 don't reward it for this increased accuracy; instead criticize it
 for supposedly measuring "generic lattice security differently".

 * If submission X, as part of its effort to rescue a bleeding-edge
 parameter set, switches from Core-SVP to a different metric (while
 confusingly labeling the result as "Core-SVP" and not prominently
 announcing the results of the previous metric), praise it for the
 increased accuracy.

Why does NIST expect the public to believe that this discrepancy comes from something other than NSA pushing X? Has NIST been following the transparency rules from the NIST VCAT Dual EC report? Systematically answering clarification questions regarding its evaluation criteria and NISTPQC processes? Announcing procedures for resolving tensions between different criteria? Generally putting itself in a position of having decisions transparently forced by its previously announced criteria?
Recognizing that earning public trust requires limiting its own power?

Nope, none of the above.

Christopher J Peikert writes (email dated 4 Dec 2020 14:44:09 -0500):
> it seems to me that Dan's entire objection about the Round-3 Kyber
> Core-SVP analysis is premised on *not* considering "amount of
> rounding" as part of a lattice scheme's "parameter set."

No. There's no dispute here regarding the contents of the parameter sets. My questions and objections are regarding mechanisms for turning parameter sets into claimed security levels for comparisons: for example, about "the Core-SVP metric" being promoted when this seemed to favor Kyber, and then suddenly being swept under the rug as part of a campaign to rescue Kyber-512.

The fact that the metrics under discussion are simplified, and in particular that (before round-3 Kyber!) they ignore "bit-dropping", is explicit in, e.g., Section 2.1 of the "Estimate" paper, and in various submissions using Core-SVP, and in my first message in this thread:

   In the literature, Core-SVP for RLWE/MLWE-based systems is defined
   by 2n full samples (public multiples plus errors), whether or not
   the systems actually apply further rounding to those samples. See,
   e.g., the round-2 Kyber submission.

I have no idea how anyone could think that anything I wrote in this thread can be viewed as "*not* considering 'amount of rounding' as part of a lattice scheme's 'parameter set.'"

  [preceding the above:]
> At the risk of repeating myself,

Indeed, this bizarre part-of-parameter-set argument already appeared in email dated 2 Dec 2020 14:34:00 -0500, which I answered in email dated
17 Dec 2020 16:20:01 +0100, but the same argument was already repeated in email dated 4 Dec 2020 14:44:09 -0500, which I'm answering now.

> (Despite my request, Dan's long message did not offer any clarity on
> this central point; I think he should address it directly.)

One would expect this sort of comment to be attached to an exact quote of the supposedly unanswered request. My best guess is that Dr. Peikert is referring to a question in his email dated 2 Dec 2020 14:34:00 -0500, but the message he's replying to here is clearly labeled as replying to an earlier message (email dated 1 Dec 2020 17:01:37 +0100), so the suggestion of non-responsiveness is incorrect.

I'm reminded of Swift's "Truth comes limping after" quote. In general, one wonders how the NISTPQC discussions could be structured to reward more careful analysis and downgrade less careful analysis, compensating for the efficiency advantage of sloppiness.

> For cryptanalytic purposes, ignoring rounding leaves out very
> important information, and can even produce perverse Core-SVP numbers.
> For example, ignoring rounding would lead us to conclude that all of
> the NTRU Prime parameters have *trivial* Core-SVP hardness (~2^0)

"Each of the 'Estimate' metrics for parameter sets, and in particular the Core-SVP metric, has the following two structural features: The metric for RLWE/MLWE cryptosystem parameter sets is purely a function of the underlying

RLWE/MLWE problems. The metric for RLWR/MLWR cryptosystem parameter sets is purely a function of the underlying RLWR/MLWR problems."

In particular, Core-SVP for an RLWR/MLWR system _doesn't_ ignore rounding, and _doesn't_ treat the rounding the same way as zero noise.

If "ignoring rounding" is meant to include redefining Core-SVP for RLWR/MLWR systems to use a zero-noise problem instead of the underlying RLWR/MLWR problem, then I agree that this redefinition would produce very low security claims for such systems, completely missing the main problem that the systems say is hard. Nobody would be able to defend such a definition.

What happened with Core-SVP, and in round-2 Kyber, was very different.
Core-SVP---with its focus upon RLWE, RLWR, etc., and all its other simplifications---was not merely defensible, but was actively promoted.
Other submissions were criticized for supposedly measuring "generic lattice security differently", for focusing on the actual cryptosystem attack problems rather than the RLWE/MLWE problems, etc.

> Of course, the NTRU Prime submission did *not* report trivial Core-SVP
> hardness, because the authors (Dan included) rightly included the
> rounding in their Core-SVP analysis. Obviously, other submissions
> should not be criticized for doing the same.

It's not the same. The "Estimate" metrics evaluate the underlying RLWE/MLWE problems for RLWE/MLWE systems, and evaluate the underlying RLWR/MLWR problems for RLWR/MLWR systems. Describing this focus on the underlying problems as "ignoring rounding" is simply not correct; the case distinctions matter, and are built into the definitions.

The same distinctions appear again and again in security discussions, for example in NIST trying to make people believe that RLWR/MLWR isn't as good as RLWE/MLWE, and in Kyber claiming that its security is "based on the hardness" of MLWE, and in Kyber specifically advertising the supposed security advantages of not relying on MLWR.

> This brings me to Kyber-512. My current understanding is that the
> following three mechanisms, when applied to round-3 Kyber-512, produce
> the following "Core-SVP" numbers:
> * The mechanism used on the "Estimate" page: <=2^112 (see below).
> * The mechanism used in the round-2 Kyber submission: <=2^112.
> * The mechanism used in the round-3 Kyber submission: 2^118.
> The reason for this change is that the round-3 Kyber submission switched
> to a new mechanism of mapping parameter sets to security levels, I
> don't think this is accurate.

"This" being what, precisely? Quoting several items makes the topic of dispute unclear, and the text below doesn't help clarify.

> Round-3 Kyber introduced rounding that was not present in its previous
> versions.

I was careful to refer specifically to "round-3 Kyber-512", precisely because it's not the same as round-2 Kyber-512. The mechanism of mapping parameter sets to security levels _also_ changed, as I said.

> The updated Core-SVP analysis reflected the existence of that
> rounding, presumably in a manner consistent with how other submissions
> had treated rounding.

Is this supposed to be disputing the statement that the mechanism used in the round-3 Kyber submission assigns 2^118 to round-3 Kyber-512? Or that the mechanism used in the "Estimate" page assigns <=2^112 to
round-3 Kyber-512, as does the mechanism used in the round-2 Kyber submission? Or that there was a change of mechanism from the round-2 Kyber submission to the round-3 Kyber submission?

> This is not a "new mechanism," it is the ordinary mechanism applied to
> new parameters.

If deterministic mechanism K2 outputs <=2^112, and deterministic mechanism K3 outputs 2^118 on the same input (namely round-3 Kyber-512), then obviously it is not true that K2 and K3 are the same mechanism. The different numbers directly contradict the claims of consistency.

> Here there is an unstated premise that "MLWE" (and later, "the MLWE
> instance inside Kyber-512") does *not* include the rounding

When an MLWE submission is completely clear in naming its MLWE problem, defining the parameters for that problem, advertising that problem as the foundation for its security, etc., I don't think it's reasonable to describe the premise as being "unstated".

> I agree that it would be good to get a precise statement from the
> Kyber team concerning what they mean by "MLWE," and the consequences.

I'm not sure which statement of mine this is supposedly agreeing with.
If there's some lack of clarity in Kyber's definition of MLWE or of its parameter sets, then that's probably worth filing a separate formal comment about, but so far this looks like nothing more than a wild mischaracterization of the topic of discussion.

>   Many people seem to believe that the security levels of RLWE and MLWE
>   are thoroughly understood (while the same people sometimes express
>   doubts regarding the security levels of RLWR and MLWR).
> Again, please provide unambiguous references, so the reader can check
> whether you are accurately representing what "many people" "seem to
> believe" and express.

NIST IR 8309: "While reductions to MLWR from MLWE exist, they are not concretely applicable to SABER, which is a mild concern. ... NIST encourages additional research regarding ... the concrete differences between the security of MLWE and MLWR for the proposed parameter sets."

The same report also repeatedly leads the reader to believe that NIST is confident in the lattice finalists. NIST has also been going to amazing lengths to not kick out Kyber-512; proactively kicking it out now would be an embarrassment to them and to NSA, but this is nothing compared to the PR disaster that they'll be facing if Kyber-512 is later shown to flunk the minimum NISTPQC security requirements. Their comments and behavior make no sense if they aren't collectively confident in the
Kyber-512 security level. Note that I said "seem to believe", not "say"; words have meanings.

---Dan

--

From: daniel.apon <daniel.apon@nist.gov>
Sent: Sunday, January 3, 2021 5:34 PM
To: pqc-forum
Cc: D. J. Bernstein; pqc-forum; pqc-comments; pqc-comments
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

Hi Dan,

Are you proposing a new attack against Kyber?

If so, would you please write it in LaTeX and post the PDF either to this forum list or on ePrint? (Or better-- submit it to a peer-reviewed conference or journal?)

Speaking for myself,
--Daniel Apon

On Saturday, January 2, 2021 at 12:24:48 PM UTC-5 D. J. Bernstein wrote:

> Efforts to accurately analyze lattice security are becoming more and more complicated. After many correction factors, the state-of-the-art estimates still don't exactly match experiments even for the simplest attacks, and there's no reason to think that small errors will remain small when they're extrapolated to cryptographic sizes. Meanwhile the algorithms themselves are becoming more and more complicated as people find new speedups. Each new speedup poses a new analysis challenge.
>
> Comparing parameter sets across proposals requires further work to automate estimates. This work is error-prone, as illustrated by the miscalculation of the "Estimate" numbers for SABER. The difficulty of automation is magnified when estimates are complicated and unstable.
>
> There are many simplifications in the "Estimate" work, and in particular in "the CoreSVP metric" for parameter sets, at the expense of accuracy. The simplifications could reverse comparisons between proposals; could mislead people regarding cost comparisons to AES; and clearly contribute to a dangerous lack of awareness of ongoing advances in lattice attacks. Do we recognize these issues as problems, and insist on doing better as a prerequisite for making decisions regarding lattice proposals? Or do we praise the simplifications, saying that the simplifications are conservative, that it's better to have a simplified comparison than nothing at all, and that this comparison tells us which lattice schemes are being more and less aggressive in setting their parameters?
>
> One of the worst imaginable ways to answer these questions is to make an ad-hoc decision between praising accuracy and promoting a simplified metric, depending on whether the answer seems to favor submission X:
>
> * If a submission that _isn't_ X contains the most detailed survey in the literature of inaccuracies and potential inaccuracies in

From:	D. J. Bernstein <djb@cr.yp.to>
Sent:	Monday, January 4, 2021 8:40 AM
To:	pqc-comments
Cc:	pqc-forum
Subject:	Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER
Attachments:	signature.asc

Executive summary: Ten questions appear below regarding Kyber-512's (apparently contradictory) security claims. These questions are quoted from email to pqc-forum dated dated 4 Dec 2020 18:06:07 +0100. These questions remain unanswered today, as far as I can tell.

Vadim Lyubashevsky writes (email dated 5 Dec 2020 08:33:45 +0100):
> Thank you for distilling that email into one question.

The chain of replies leading up to this was

```
   #1: Vadim Lyubashevsky (email dated 1 Dec 2020 17:01:37 +0100)
-> #2: D. J. Bernstein (email dated 4 Dec 2020 18:06:07 +0100)
-> #3: Christopher J Peikert (email dated 4 Dec 2020 14:44:09 -0500)
-> #4: Vadim Lyubashevsky (email dated 5 Dec 2020 08:33:45 +0100)
```

where #1 stated "We believe that we are being very clear with what we are claiming for Kyber512" and #2, in direct reply to this, stated the ten questions quoted below regarding Kyber-512's security claims.

Seven of these ten questions are simply yes-no questions regarding the security claims. Some of the claims in the submission sound to me like they're contradicting each other, and the answers to these questions will help clarify the situation. In case the apparent contradictions persist, there are two conditional questions asking for explanations.
The last question is simply asking for a claimed block size to ease verification.

Question #9 is saying "just to confirm" for something that I think is _almost_ completely clear from previous statements, but not perfectly clear, which is why I asked the question. For the other six yes/no questions I had (and have) medium-confidence guesses based on what I've read, but, again, those guesses seem to contradict each other.

One of the reasons that I follow traditional email-handling practice of

* directly quoting questions (with all necessary context), and
* directly answering each question,
* in a reply to the message stating the questions,

is that this practice is helpful for readers (and authors!) trying to track what has been answered and what hasn't. #3 visibly didn't follow this practice: it makes a claim (which I've disputed separately) of a premise for my "entire objection", but neither quotes nor claims to be "distilling" most of my questions. Even if everything in #3 is blindly trusted as stated (which, procedurally, seems inappropriate for a process aiming to set cryptographic standards), I don't know how #4 arrives at the idea that replying to a question in #3 is a reply to all the questions in #2.

More to the point, I'd like to see answers to these questions. If a cryptanalyst puts in the work to show that the MLWE instance inside round-3 Kyber-512 takes fewer "gates" to break than the minimum allowed NISTPQC security level (whatever exactly "gates" means), then the cryptanalyst shouldn't have to worry that the Kyber team is going to say "We never claimed that this MLWE instance was that hard to break". On the contrary, the claim should be made clear to _encourage_ analysis.

If Kyber-512 _isn't_ claiming that its MLWE instance meets the minimum allowed NISTPQC security level, then this would appear to be a change from round-1 and round-2 Kyber---quite a dramatic change given how much advertising we've seen for the MLWE problem. This should be recognized explicitly, and scored negatively under NISTPQC's "maturity of analysis"
criterion. Also, in this case it would appear that some claims regarding the relationship between Kyber and MLWE will have to be withdrawn. Maybe I'm missing some way the claims can be maintained; this is what two of my questions are about.

---Dan

From: Christopher J Peikert <cpeikert@alum.mit.edu>
Sent: Monday, January 4, 2021 12:59 PM
To: pqc-forum; pqc-comments
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

Dear all, happy new year.

I'd like to point out that Dan's most recent posts on this topic are not responsive to:

1. my Dec 17 email, showing that his central claim that round-3 Kyber "switched from Core-SVP to a modified metric" is premised on a nonsensically limited conception of Core-SVP; and

2. Mike Hamburg's Dec 17 email, accurately summarizing the steps of a CoreSVP analysis, which round-3 Kyber's analysis is consistent with.

(Perhaps Dan will reply to these later.) So, there is not much new for me to say on this topic, but I will reiterate the key points.

The "switched from Core-SVP" claim is premised on the idea that Core-SVP can account for deterministic error ("rounding"), but *only* if no random error is added before the rounding. If there is, then Core-SVP can account *only* for the random error, not the rounding. This is nonsensical on its face, and nothing in the literature suggests that Core-SVP analysis must be limited in this way.

What is the source of this misconception? My conclusion is that Dan is conflating "the Core-SVP metric" itself (a somewhat misleading phrase, as Mike explains) with an optional simplification made in the "Estimate" work -- namely, that when modeling "LWE+R" as a lattice problem it ignores the rounding. (Importantly, the rounding cannot make the problem any easier.) Below I'll point out a few instances of this conflation in Dan's Jan 2 message.

However, "Estimate" did not introduce or define Core-SVP, nor the modeling of LW* as lattice problems, and its simplified modeling of LWE+R certainly isn't binding on others' Core-SVP analyses. So, this limited conception of Core-SVP is wholly unjustified, nullifying the premise of the "switched from Core-SVP" claim.

(For the record, I haven't expressed an opinion on round-3 Kyber's concrete Core-SVP estimates. My messages on this topic merely serve to debunk the "switched from Core-SVP" claim and the like.)

> Christopher J Peikert writes (email dated 4 Dec 2020 14:44:09 -0500):
> > it seems to me that Dan's entire objection
> > about the Round-3 Kyber Core-SVP analysis is premised on *not* considering
> > "amount of rounding" as part of a lattice scheme's "parameter set."
>
> No. There's no dispute here regarding the contents of the parameter
> sets. My questions and objections are regarding mechanisms for turning
> parameter sets into claimed security levels for comparisons...

OK, thanks for that clarification. My understanding of your objection is now this: the amount of rounding is part of the parameter set, but Core-SVP can't account for it if there is also some random error -- yet it *can* account for rounding if there is *no* random error. Again, this doesn't make any sense to me.

On Sat, Jan 2, 2021 at 12:24 PM D. J. Bernstein <djb@cr.yp.to> wrote:

> There are many simplifications in the "Estimate" work, and in particular
> in "the CoreSVP metric" for parameter sets, at the expense of accuracy.

This is one example of the above-described conflation. Again, "Estimate" did not introduce or define "the Core-SVP metric," nor does it claim to -- it cites prior work (the original NewHope work) in the one place it uses the term "Core-SVP."

> The fact that the metrics under discussion are simplified, and in
> particular that (before round-3 Kyber!) they ignore "bit-dropping", is
> explicit in, e.g., Section 2.1 of the "Estimate" paper, and in various
> submissions using Core-SVP, and in my first message in this thread:
>
>> In the literature, Core-SVP for RLWE/MLWE-based systems is defined
>> by 2n full samples (public multiples plus errors), whether or not
>> the systems actually apply further rounding to those samples. See,
>> e.g., the round-2 Kyber submission.

This is another example of the conflation. Core-SVP analysis has never been *defined to* or *required* ignoring rounding; doing so is an option because rounding cannot make the problem any easier.

> > Of course, the NTRU Prime submission did *not* report trivial Core-SVP
> > hardness, because the authors (Dan included) rightly included the rounding in
> > their Core-SVP analysis. Obviously, other submissions should not be criticized
> > for doing the same.
>
> It's not the same. The "Estimate" metrics evaluate the underlying
> RLWE/MLWE problems for RLWE/MLWE systems, and evaluate the underlying
> RLWR/MLWR problems for RLWR/MLWR systems.

This is another example of the conflation. "Estimate"'s choices don't limit Core-SVP.

Sincerely yours in cryptography,
Chris

From: wanlipeng <wanlipeng@iie.ac.cn>
Sent: Monday, January 18, 2021 7:30 AM
To: pqc-comments
Cc: pqc-forum
Subject: ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

I'm confused by your documentation and code implementation. In your document, h2i + h2i1X = (f2i + f2i1X)(g2i + g2i1X) mod X^2-zeta^(2 br7(i) +1). The modulo X^2-zeta^(2br7(i) + 1) means X^2 = zeta^(2br7(i) + 1). While in your code, the multiplier becomes *zetas[64+i]* and *-zetas[64+i]*, and I guess the *zetas* table means Mont*zetas^(br(i)). So why? I couldn't find any explanation in your code or document.

From: Vadim Lyubashevsky <vadim.lyubash@gmail.com>
Sent: Sunday, January 31, 2021 5:44 PM
To: Peter Schwabe
Cc: Apon, Daniel C. (Fed); pqc-forum; mi...@shiftleft.org; pqc-comments; cpei...@alum.mit.edu
Subject: Re: [pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

Dear all,

On Fri, Dec 25, 2020 at 7:44 AM Peter Schwabe <peter@cryptojedi.org> wrote:

> "'daniel.apon' via pqc-forum" <pqc-forum@list.nist.gov> wrote:
> > Hi Mike,
>
> Hi Mike, hi Daniel, hi all,
>
> > *"Therefore, basing the estimate on MLWE+R requires the additional
> > assumption that an attacker would have to attack the system this way. This
> > seems like a reasonable assessment, but should be supported by precise
> > definitions and a proof. Furthermore, the dual attack should be evaluated.
> > So while I think an overall 2^118 CoreSVP could probably be convincingly
> > argued for Kyber, the existing argument falls short." *
> >
> > I second this. This is quite a decent point.
>
> We fully agree and will send a more detailed version of the LWE+R argument to the list, most likely toward the end of January (we're all somewhat busy until then). As a spoiler, there almost certainly won't be any exciting new insights over what is already informally stated in the current document. In particular, the claimed 118 bits of CoreSVP security comes from the analysis of the LWE problem in the public key generation; while the combination of error+rounding noise in the ciphertext generation actually gives an even larger CoreSVP security level.

We have now added this more detailed analysis to the updated specification document at https://pq-crystals.org/kyber/data/kyber-specification-round3-20210131.pdf. It begins on page 21 in the subsection titled "The impact of the deterministic noise caused by Compress_q on Kyber512." The tldr summary is that if you were OK with the intuition in the earlier version of the document, then there is nothing particularly new here. We simply wrote out the details more formally.

As for the dual attack, it was already mentioned in Section 5.2.1 why we only consider the primal one. In particular, just judging by the Core-SVP numbers, the dual attack is 1 bit cheaper, but the overhead for it (versus the primal one) is significantly higher.

Best,
Vadim
(On behalf of the CRYSTALS-Kyber team)

All the best and happy holidays to everybody!
The Kyber team

From:	赵运磊 <ylzhao@fudan.edu.cn>
Sent:	Thursday, May 12, 2022 6:19 AM
To:	pqc-comments
Cc:	pqc-forum
Subject:	ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER

Dear Kyber team and dear all in PQC community:

Recently, we made a systematic optimization of the Kyber algorithm, and proposed an optimized version referred to as OSKR.

We note that with the AKCN mechanism proposed in the KCL proposal (in the first round submissions of NIST-PQC), on the same parameters, OSKR has more efficient decryption process and has lower error probability simultaneously.

We study how to encapsulate 512-bit key with OSKR-1024. By proposing a hybrid-NTT (HNTT) technique, OSKR-1024 not only encapsulates 512-bit key, but its implementation can be more efficient than Kyber-1024. Also thanks to the HNTT technique, all the three parameter sets can be implemented in modular and unified way.

The paper is available from: https://arxiv.org/abs/2109.02893

All my best
Yunlei

From:	pqc-forum@list.nist.gov on behalf of D. J. Bernstein <djb@cr.yp.to>
Sent:	Wednesday, July 27, 2022 2:32 AM
To:	pqc-comments
Cc:	pqc-forum
Subject:	[pqc-forum] ROUND 3 OFFICIAL COMMENT: CRYSTALS-KYBER
Attachments:	signature.asc

NIST's round-3 report claims a better FO proof picture for Kyber than it does for NTRU. This is wrong---exactly the opposite of what the literature says on this topic.

Here's what the report says about FO proofs for Kyber:

> The security proofs hold tightly in the ROM [169, 170] and
> non-tightly in the QROM. Yet under various other natural assumptions,
> KYBER may also achieve a tight security reduction in the QROM [184].

Here's what the report says about FO proofs for NTRU:

> The NTRU KEMs have tight CCA-security reductions to the underlying
> PKEs in the ROM, and non-tight security reductions in the QROM.
> Making some additional non-standard assumptions, one of the QROM
> security proofs can be made tight.

This portrays NTRU as having a worse security-proof picture than Kyber:
NTRU needs "non-standard assumptions" for a tight QROM proof, whereas Kyber "may" have a tight QROM proof under "natural" assumptions.

In fact, the situation for years has been that the literature has better FO proofs---better tradeoffs between tightness and the strength of the PKE assumption---for deterministic PKEs than for randomized PKEs:

* ROM, deterministic PKE: https://eprint.iacr.org/2018/526 Theorem 14.3 obtains IND-CCA2 very tightly from the standard minimal PKE security assumption, OW-CPA. (The techniques are older, but this paper is designed to support proof verification and identifies errors in some previously claimed theorems.)

* ROM, randomized PKE: https://eprint.iacr.org/2017/604 has a proof that's (almost as) tight, but assumes that the PKE is IND-CPA. IND-CPA is still standard, but it's stronger and more complicated than OW-CPA, and has received less attention from cryptanalysts. (See generally Section 6 of https://eprint.iacr.org/2019/691.)

* QROM, deterministic PKE: https://eprint.iacr.org/2019/590 obtains IND-CCA2 tightly from OW-CPA. I'm assuming here that sqrt(epsilon) is allowed as tight, unlike a number-of-queries loss factor.

* QROM, randomized PKE: All tight proofs in the literature make non-standard assumptions, such as the "disjoint simulatability"

> assumption from the paper [184] that NIST cites. This is stronger
> than standard assumptions; declaring that it's "natural" doesn't
> make cryptanalysis magically appear, and doesn't tell us whether
> the security levels are as high as desired.

The Kyber PKE (like other GAM/LPR variants) is randomized, so it definitely can't use the better proofs. The NTRU PKE is deterministic (since round 2), so presumably the better proofs apply. Someone should check the details of this application, but the risk of an error here doesn't justify NIST making claims that are out of whack with the applicable literature.

NIST's report thus needs an erratum to say that, oops, the report said that NTRU needs a "non-standard assumption" for a tight QROM proof and didn't say this about Kyber, whereas in fact the literature indicates that Kyber needs a non-standard assumption for a tight QROM proof while NTRU doesn't.

If NIST _isn't_ allowing sqrt(epsilon) as "tight", then the report needs to clarify the "tight" dividing line. An erratum is still required for the misinformation that Kyber has a better FO proof picture than NTRU:
in fact, Kyber has a worse FO proof picture than NTRU.

This is important because this Kyber proof gap could be hiding a big security loss. See https://eprint.iacr.org/2021/912 for examples where FO IND-CCA2 security is far below OW-CPA security of the underlying PKE.

This is exactly the "derandomization" risk described in Sections 3.8 and
5.8 of https://ntruprime.cr.yp.to/latticerisks-20211031.pdf, which was filed before NIST's deadline for round-3 input and which, unfortunately, NIST doesn't seem to have read. But simply reading through the previous FO proofs is sufficient to see that NIST's report gets this security comparison backwards.

---D. J. Bernstein

P.S. This is unrelated to the objections that have been raised to the handling of hashing in Kyber's FO security proofs. Qualitatively, those objections are identifying an error in the proofs, which of course is worrisome in a security analysis that NIST's report calls "thorough".
However, the idea that someone is going to find a collision in these hash functions is very far down any reasonable list of post-quantum security risks; and plugging in known indifferentiability results closes the proof gap at the expense of a quantitatively minor loss of tightness. Derandomization is a much bigger issue.

P.P.S. Kyber has had a new version in every round, and presumably one should wait to see the next version before filing comments on it, so I'm filing this is a round-3 comment. However, unless there's a radical change in Kyber, I would expect the same comment to continue to apply.

P.P.P.S. This comment is of course also regarding NTRU, which NIST's report says NIST could still select. The underlying issues are also applicable to the split between deterministic PKEs and randomized PKEs in other submissions, although unfortunately NIST's report is structured in a way that obfuscates such comparisons.