Following mail discussions and running many test simulations it is found that certain default settings from gromacs 4.6 onward lead to incorrect/different liquid densities. Examples are:

<table>
<thead>
<tr>
<th>Substance</th>
<th>Gromacs-5.1 Density</th>
<th>Gromacs-5.1 Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caleman2012 gromacs-5.1</td>
<td>1375</td>
<td>1448</td>
</tr>
<tr>
<td>chloroform (GAFF)</td>
<td>1424</td>
<td>1438</td>
</tr>
<tr>
<td>1,1,2-trichloroethane (OPLS)</td>
<td>1389</td>
<td>1385</td>
</tr>
<tr>
<td>1-bromopropane (CGenFF)</td>
<td>2474</td>
<td>2570</td>
</tr>
<tr>
<td>dibromomethane (OPLS)</td>
<td>777</td>
<td>774</td>
</tr>
<tr>
<td>methanol (OPLS)</td>
<td>777</td>
<td>774</td>
</tr>
</tbody>
</table>

The worst cases are really off. Preliminary tests with other codes show that the gromacs-5.1 result above for chloroform is roughly correct. However the following updates should be made in gromacs defaults which are too weak:

- `verlet-buffer-tolerance = 1e-4 (default 5e-3)`
- `lincs-iter              = 2 (default 1)`
- `constraints             = h-bonds (give warning for all-bonds?)`
- `delta_t                 = 0.001 ps (give warning for combinations?)`
- `shake-tol               = 1e-5 (single precision, default 1e-4)`

Associated revisions

Revision 483a7a11 - 02/07/2017 11:37 AM - Berk Hess
Add mdp option check with decoupled modes

When atoms involved in an angle with constrained bonds have very different masses, there can be very weakly coupled dynamics modes. Default mdp settings are often not sufficiently accurate to obtain equipartitioning. This change adds a grompp check for this issue.

Part of #2071.

Change-Id: I64323154c38abe23b8d38d9d539d2a713a5c35e0

History

Gerrit received a related patchset '1' for Issue #2071.
Uploader: David van der Spoel (davidvanderspoel@gmail.com)
Change-ID: i503c3aaee328261542928f81b9849053d21b0e0c
Gerrit URL: https://gerrit.gromacs.org/6320

2 - 11/07/2016 04:35 PM - David van der Spoel

3 - 11/07/2016 05:38 PM - Szilárd Páll
To elaborate on the point I made on gerrit, while it's good to be on the safe side, AFAIK the current defaults are fine for most cases, and therefore the suggested changes will

- potentially send a the wrong message to users, especially without details of what the issues addressed here + carefully phrased release notes;
- make many runs unnecessarily slow: with a standard protein benchmark system (ADH, PME, nstlist=20) I got ~20% performance drop (slowdown of ~35% in nonbondeds, 20% in search, 11% in constraints).
Performance is completely irrelevant unless the simulation is accurate enough.

Similarly, fine in "most cases" is also irrelevant unless we have a very clear and simple way for the user to decide whether it is accurate enough.

If that is not the case, it's a fragile optimization that shouldn't be enabled by default, no matter how good it is for performance. The burden of checking needs to be on the user/developer who wants to push performance, rather than a non-expert user (and in this case even an expert) risking the default settings completely screwing up their results.

To the best of my knowledge, there is no universal measure for "accurate enough". If we'd apply the conservative approach suggested to protect most non-expert users from picking incorrect settings, our default build should be double-precision (in particular for constraints not much else helps) and most tolerances should be far tighter.

It's exactly performance the reason why the group scheme employed the optimizations that were clearly at the detriment of accuracy and I don't think it was clear how to improve it. With that baseline in mind, the current settings were devised to get orders of magnitude lower drift that is also controllable in a clearly documented manner. So calling now the Verlet-scheme related parameter choices "fragile optimizations" does not seem substantiated.

When it comes to LINCS and SHAKE parameters, as far as I understood from Berk's recent talk on the simulation meeting, it is pretty hard to improve accuracy in a system-independent manner (at least without fully or mostly double precision treatment).

I'll repeat some comments I made in an email discussion a few days ago.

- We certainly don't want to trade performance for (measurable) accuracy.
- But I think we can always come up with cases that will fail, at least in single precision, with any accuracy setting.
- Ideally we would want to auto-tune the accuracy settings for the system and/or auto-detect when the accuracy is not enough, but that might require an enormous effort and might be impossible or we might still miss some cases.
- Given that some cases will fail, what extra computational cost are we willing to accept to have a few, but probably not all, problematic cases run well/better with default settings?

Concerning the Verlet buffer tolerance, that we can decrease for the 2017 release without much extra cost.

My question on the results reported in this issue is: are these with all-bonds constrained or h-bonds only?

I would think we should be able to get nearly all cases right with h-bonds only, but some (like chloroform) might be hard with all bonds constrained. We could consider warning about constraining all bonds, but there are very many setup out there that run fine with all bonds constrained.

For lincs with h-bonds the density varies between 1446 (correct) and 1467, 1.5%, roughly the accuracy I would like to be able to predict with. Shake with h-bonds varies between 1435 and 1468. Lincs with all-bonds varies between 1387 and 1635, Shake will all-bonds between 1312 and 1637. No constraints varies from 1444 to 1466.

I would consider 1.5% too much. It would be nice if we could get errors down to 0.1%. Could you mail me the input for the problematic systems so I can play around with them?

I have been trying a lot of things with chloroform but still don't fully understand exactly how the constraint errors behave/propagate. I have a way of stopping the LINCS rounding errors from accumulating, this reduces the energy drift, but actually makes the partitioning error larger. I can understand some of this, but don't understand why OpenMM doesn't seem to be affected by this (unless the 1e-4 relative error tolerance is not what was actually used).

Another issue is that constraining heavy atoms adds more rounding noise, but that is negligible with these small systems and a 2 fs time step. I already planned to add a check for this in the next release.

I would consider 1.5% too much. It would be nice if we could get errors down to 0.1%.

I would agree. In general, errors have to be as low as the experimental errors are. Quoted experimental errors for density measurements for organic liquids are around = 0.001 g/cm^3, indicating that we really should get down to 0.1%. Otherwise, we cannot say for sure if the simulation matches experiment or not.

It does appear that at high levels of constraint, in double precision, LINCS and SHAKE do agree with chloroform, at around 1448. So we probably want 1448 +/- 1 (or maybe 1449 +/- 1, I'm still looking through).

I have been trying a lot of things with chloroform but still don't fully understand exactly how the constraint errors behave/propagate. I have a way of stopping the LINCS rounding errors from accumulating, this reduces the energy drift, but actually makes the partitioning error larger. I can understand some of this, but don't understand why OpenMM doesn't seem to be affected by this (unless the 1e-4 relative error tolerance is not what was actually used).
Yeah, I haven't figured this out yet, either . . . I haven't had the time to look at it. We should check with them to try to understand the sensitivity.

#10 - 11/08/2016 06:34 AM - Michael Shirts
Looking for more patterns

1. for shake in double precision, the large errors seem to come when nsttcouple/nstpcoule/nstcalcenergy are greater than 1. When they are all 1, shake with all-bonds behaved quite well (if the verlet-drift was large) I suspect there is something happening off in the bookkeeping there.
2. lincs with 8 iterations just does not seem to be behaving in a very numerically stable way.
3. it would be interesting to see if md-vv (but not clear if it is yet interesting enough to run all of these tests until it's better automated).

#11 - 11/08/2016 06:52 AM - David van der Spoel
- File chloroform.tgz added
@Michael, I suspected the nsttcouple and nstpcouple bugs (2031 and 2032) to be the main culprit first, but found issues even without. Will try a couple of md-vv runs, however MTTK pressure control does not seem to work with constraints.

#12 - 11/08/2016 10:47 AM - Mark Abraham
I echo the sentiments that we would like default accuracy comparable with experiment, preferably by default, and that with defaults we should err on the side of being correct rather than maximally fast. However, some of these variables are coupled, and we should recognize that a single set of defaults need not hit any particular sweet spot.

In particular, comparison with Caleman 2012 (using unbuffered group scheme and as far as I can see from the paper, unstated nst*) doesn’t seem able to defend the suggestion that verlet-buffer-tolerance is too weak. It might be, but we can’t say that yet. :-)

David, can you please upload the spreadsheet you have mentioned? I would also encourage people to avoid substantive discussion in private email. :-) We're not afraid of being wrong, and we gain from discussing possible dirty laundry in the open, whether we're right or wrong!

#13 - 11/08/2016 11:51 AM - David van der Spoel
- File simulation-results.xls added

#14 - 11/08/2016 01:19 PM - David van der Spoel
To make this absolutely clear: the verlet scheme is a big improvement, but it appears we don't completely understand the ins and outs of the different variables and how they work together. My plan is to redo all the simulations from Caleman2012 (and some more) once I have a reliable setup, and then compare the results to other codes more extensively.

#15 - 11/08/2016 01:45 PM - Berk Hess
What's the issue with methanol?
The results look consistent to me, there are only differences between constraining more or less bonds which could be correct, or?

#16 - 11/08/2016 01:58 PM - Berk Hess
I posted without looking carefully enough. The errorbars on the methanol results are too large to tell which settings work ok and which not.

For the other cases the common aspect seems to be a very heavy atom and a very light atom on both ends of an angle connected by constrained bonds. My hypothesis is that this leads to two (or more) angle vibration modes with a large difference in period and thus very weak and slow kinetic energy exchange which makes equipartitioning difficult.
I/we need to confirm this. But if this is the case, it's simple to add a check for this and the buffer and constraint accuracy settings in grompp.

#17 - 11/08/2016 02:10 PM - David van der Spoel
I actually added methanol (and water) to have some other examples without heavy atoms as well. I can easily test more if anyone has suggestions. Methanol seems to behave normally I would say. Deviations between constrained and unconstrained molecules can result from the unconstrained case in practice being a different model (e.g. in water with polarization through bond elongation). We should therefore not dwell too much on such differences.

#18 - 11/08/2016 02:36 PM - John Chodera
Michael Shirts wrote:
I would consider 1.5% too much. It would be nice if we could get errors down to 0.1%.

I would agree. In general, errors have to be as low as the experimental errors are. Quoted experimental errors for density measurements for organic liquids are around = 0.001 g/cm^3, indicating that we really should get down to 0.1%. Otherwise, we cannot say for sure if the simulation matches experiment or not.

I just want to correct Michael's figure here: We have an inexpensive Mettler-Toledo DM40, with manufacturer specified error an order of magnitude
smaller than what Michael quotes: 0.0001 g/cm^3. For water, this is a 0.01% error.


There are even models (DM45 and DM50) that push this to 0.00002 g/cm^3, which is 0.002%.

#19 - 11/08/2016 02:41 PM - John Chodera

David van der Spoel wrote:

To make this absolutely clear: the verlet scheme is a big improvement, but it appears we don't completely understand the ins and outs of the different variables and how they work together. My plan is to redo all the simulations from Caleman2012 (and some more) once I have a reliable setup, and then compare the results to other codes more extensively.

I wonder if this would be a good time to switch to using density measurements from the ThermoML Archive:
http://trc.nist.gov/ThermoML.html

Michael is very familiar with the data that lives there, and we found an enormous wealth of data there easily obtained in machine-readable format. As of the writing of our paper, there were 3592 neat liquid density measurements in the temperature range [270,330] K at 1 atm, and we have some tools available for retrieving and extracting the data and setting up simulation systems:
http://dx.doi.org/10.1021/acs.jpcb.5b06703
https://github.com/choderalab/thermopyl
https://github.com/choderalab/LiquidBenchmark

If the goal is to compare between codes, though, a much smaller set would of course be reasonable!

#20 - 11/08/2016 02:45 PM - David van der Spoel

Maybe we should not drift off into science yet (and not here), but there are other error sources in experimental numbers than the measurement, e.g. the purity. 0.1% statistical error a simulation of 1000 molecules seems ambitious.

#21 - 11/08/2016 02:46 PM - John Chodera

Berk Hess wrote:

- Ideally we would want to auto-tune the accuracy settings for the system and/or auto-detect when the accuracy is not enough, but that might require an enormous effort and might be impossible or we might still miss some cases.

I think this is a great idea. We've done some work on measuring error in a property-independent way [http://dx.doi.org/10.1103/PhysRevX.3.011007], but identifying a range of settings where the physical property of interest is insensitive to the parameters is probably optimal.

If this can't easily be done, there would be great value in a publication that explores parameter sensitivity for a variety of liquids to both (1) recommend a set of "best practices" parameters, and (2) demonstrate why this parameter set is recommended. Even if the gromacs defaults are not changed, this manuscript could be used as a starting point for anyone wanting to do density calculations.

#22 - 11/08/2016 02:53 PM - John Chodera

David van der Spoel wrote:

Maybe we should not drift off into science yet (and not here).

Fair enough!

but there are other error sources in experimental numbers than the measurement, e.g. the purity. 0.1% statistical error a simulation of 1000 molecules seems ambitious.

I could be wrong here, but I don't believe a 0.1% purity error leads directly to a 0.1% density error.

#23 - 11/08/2016 03:13 PM - Szilárd Páll

David van der Spoel wrote:

To make this absolutely clear: the verlet scheme is a big improvement, but it appears we don't completely understand the ins and outs of the different variables and how they work together.

Indeed, we have been discussing the different effects in our meetings/seminars, in particular how can more automation make things robust to reduce the hand-waving in the choice or simulation parameters.

It would be a good to continue and extend this discussion e.g. on the GROMACS dev meeting. We have the opportunity to do that already tomorrow if people can participate!
A related topic that we've discussed a number of times is introducing an additional mixed-precision mode that could provide most of the benefits of double precision for a cost a lot closer to single. This is particularly relevant because on many GPUs, especially the ones with good perf/buck ratio that, having to use double is simply not an option. However, without understanding how the different algorithm affect accuracy, it did not seem to be worth using more hand-waving and increasing the complexity of the choice the user has to make. Any interest in this direction?

John Chodera wrote:

I just want to correct Michael's figure here: We have an inexpensive Mettler-Toledo DM40, with manufacturer specified error an order of magnitude smaller than what Michael quotes: 0.0001 g/cm³. For water, this is a 0.01% error.

Does Ken think that the uncertainty's really that low after their systematic studies at NIST? I thought that it was generally a bit higher than the machine precision (things like sample purity, etc. affecting results).

Regardless, I think we can agree that 1.5% is not good, and the difference should be at LEAST no higher than 0.1%

John Chodera wrote:

I just want to correct Michael's figure here: We have an inexpensive Mettler-Toledo DM40, with manufacturer specified error an order of magnitude smaller than what Michael quotes: 0.0001 g/cm³. For water, this is a 0.01% error.

Does Ken think that the uncertainty's really that low after their systematic studies at NIST? I thought that it was generally a bit higher than the machine precision (things like sample purity, etc. affecting results).

That's the manufacturer-specified accuracy. My understanding was that Ken thought that the repeatability may indeed be as specified, but it was critical to calibrate this to well-characterized samples (e.g. toluene), and there were issues like sample purity that degraded the accuracy. Inflating the error by a factor of ten seems a bit drastic, though.

David van der Spoel wrote:

I have another piece of information, possibly the solution to the problem. In the Caleman paper we use PME grid spacing of 0.12 nm, another "default" value. Changing this to 0.1 nm gives me a stable simulation using double precision (rho = 1449 g/l), lincs on all-bonds and dt = 1fs. 0.12 nm gives a way too large density. Since PME contributes to the virial, could the systematic drift be caused by this?

"default" value. Changing this to 0.1 nm gives me a stable simulation using double precision (rho = 1449 g/l), lincs on all-bonds and dt = 1fs. 0.12 nm gives a way too large density. Since PME contributes to the virial, could the systematic drift be caused by this?

That's really surprising that a spacing change from 0.12 nm to 0.1 nm causes that much of a problem. A change in grid spacing shouldn't make the simulation converge energy less, as it makes the grid a worse representation of the Fourier space electrostatic potential, but I THINK it should still have forces and energies that match each other (if not the results from finer spacings).

I was more contemplating that a larger grid introduces noise (in the best case that the deviation is not systematic as Michael suggests), but if it is not sufficiently accurate it does not match the analytical Ewald corrections any longer. This in turn might drive angle fluctuations. This is pure speculation obviously, not sure how to prove this.

Gerrit received a related patchset '89' for Issue #2071.
Uploader: David van der Spoel (davidvanderspoel@gmail.com)
Gerrit URL: https://gerrit.gromacs.org/4310
Here is another example of constraint issues + pressure coupling problems, now in a lipid membrane where the area per lipid may decrease by over 10% with "poor settings". When using the same settings used by Amber (Stochastic dynamics, Berendsen thermostat) the same area is reproduced. Using a long tau_p and Parrinello Rahman, coupled with h-bond constraints is also fine. In Phospholipids there is also a discrepancy between the mass of the P and other atoms. Some more simulations with even long tau_p are in the works.

Yes, some conditions are caught but not all. I think I communicated the missing ones, but can summarize again since I have since done more tests with e.g. the Andersen thermostat.

Updated spreadsheet with simulation settings and results. In red the simulations that are incorrect. The column S has the density and column T is one if grompp with the new patch generates a tpr, and 0 if it does not. Ideally the red cases would be caught by grompp.

Executive summary for chloroform:
Single precision, shake or lincs, all-bonds constrained typically "fails" with too low density
For the group with too high density it is difficult to discern anything at all. Maybe a PCA could help?

Berk's patches have added some code to generate new warnings, but the issue is still under consideration.

Here is an updated version of the lipid results. Briefly, using all bonds constrained means the lipid area is too low by ~3.5 Angstrom^2. Using constraints = h-bonds and sufficiently large tau_p yield area per lipid < 1 Angstrom less than the reference (SD T-coupling).

If the problem is indeed the deviation from equipartition then one would think SD could solve it, however with constraints = all-bonds even that is not enough.

What are tau-t and tau-p for all simulations?
Is there reason to assume a shorter tau-T would give less problems with equipartitioning?

#44 - 03/17/2017 06:00 PM - Mark Abraham
Note we are approaching end of life for the 5.1 branch, so I will bump this issue to a later branch in the next month or two.

#45 - 09/12/2017 11:58 AM - Mark Abraham
- Target version changed from 5.1.5 to 2016.5

We need to decide what work remains to do here

#46 - 09/12/2017 04:21 PM - David van der Spoel
I intend to get back to this within the near future by putting a person to work on this.

#47 - 12/30/2017 06:57 PM - Erik Lindahl
What is the status of this?

#48 - 12/31/2017 03:34 PM - David van der Spoel
Waiting for manpower. The idea was to test multiple packages for reproducability but it is hard to automate all the necessary conversions.

#49 - 01/05/2018 03:32 PM - Mark Abraham
- Tracker changed from Bug to Task
- Target version deleted (2016.5)
- Affected version deleted (4.6)

This is good to do to check that we're ok, but there's not a known bug.

**Files**

<table>
<thead>
<tr>
<th>File</th>
<th>Size</th>
<th>Date</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>chloroform.tgz</td>
<td>110 KB</td>
<td>11/08/2016</td>
<td>David van der Spoel</td>
</tr>
<tr>
<td>simulation-results.xls</td>
<td>141 KB</td>
<td>11/08/2016</td>
<td>David van der Spoel</td>
</tr>
<tr>
<td>LIPID14_validation_gromacs.xls</td>
<td>8.5 KB</td>
<td>12/10/2016</td>
<td>David van der Spoel</td>
</tr>
<tr>
<td>simulation-results.xls</td>
<td>197 KB</td>
<td>02/07/2017</td>
<td>David van der Spoel</td>
</tr>
<tr>
<td>LIPID14_validation_gromacs.xls</td>
<td>43 KB</td>
<td>02/07/2017</td>
<td>David van der Spoel</td>
</tr>
</tbody>
</table>