nb3dfft

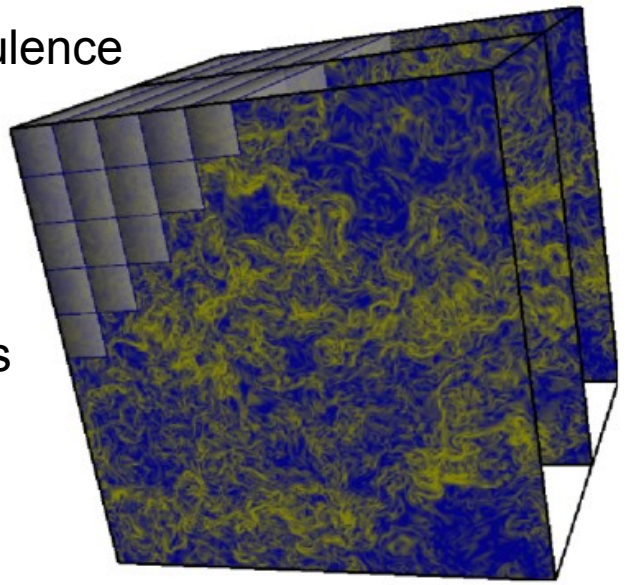
Jens Henrik Göbbert

Jülich Supercomputing Centre, Forschungszentrum Jülich GmbH, Germany

Motivation

Pseudo-Spectral Application

- isotropic homogeneous decaying/forced turbulence
- turbulent channel flow
- turbulent premixed flames
- scalar mixing



- 80% of the compute time required by 3d-FFTs
  - all-to-all communication
  - 2x data redistribution among processes

development of the improved 3d-FFT library ‚nb3dfft‘

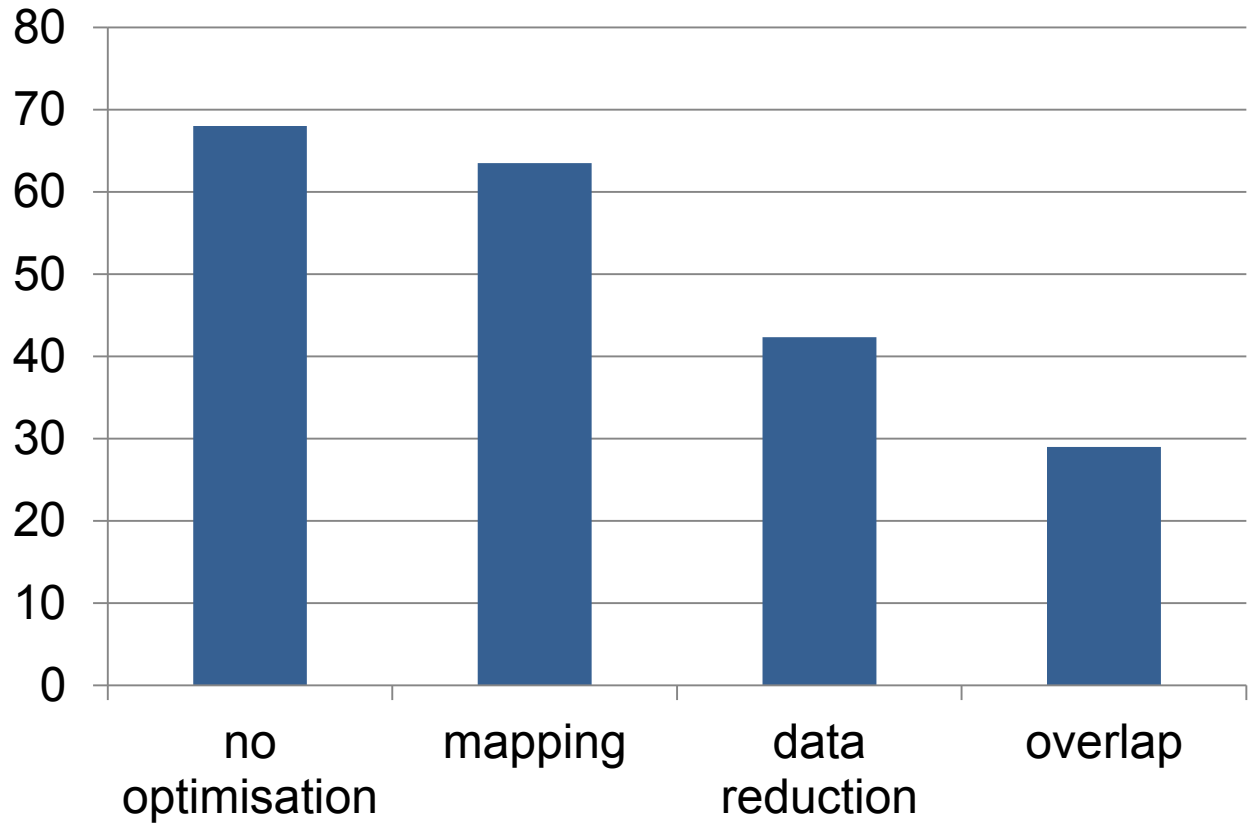
- based on P3DFFT (thanks to Dmitry Pekurovsky)
- data reduction by including the filter process (dealiasing)
- overlapping communication and computation of multiple 3D-FFTs

Why we wrote our own 3D-FFT

It is all about saving compute time

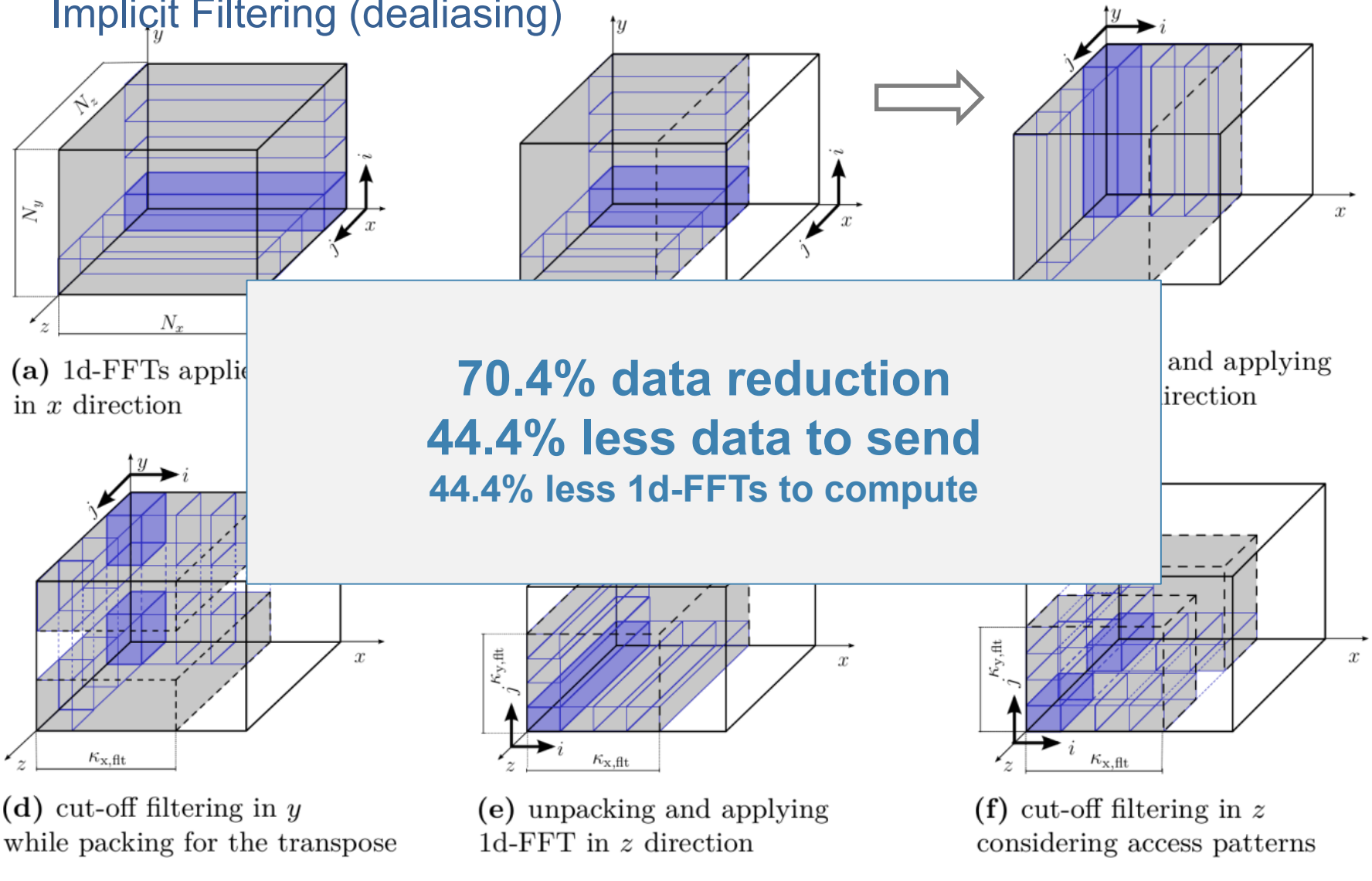
JÜLICH FORSCHUNGSZENTRUM

Million core/h required to compute 6144^3 project



nb3dfft Optimizations

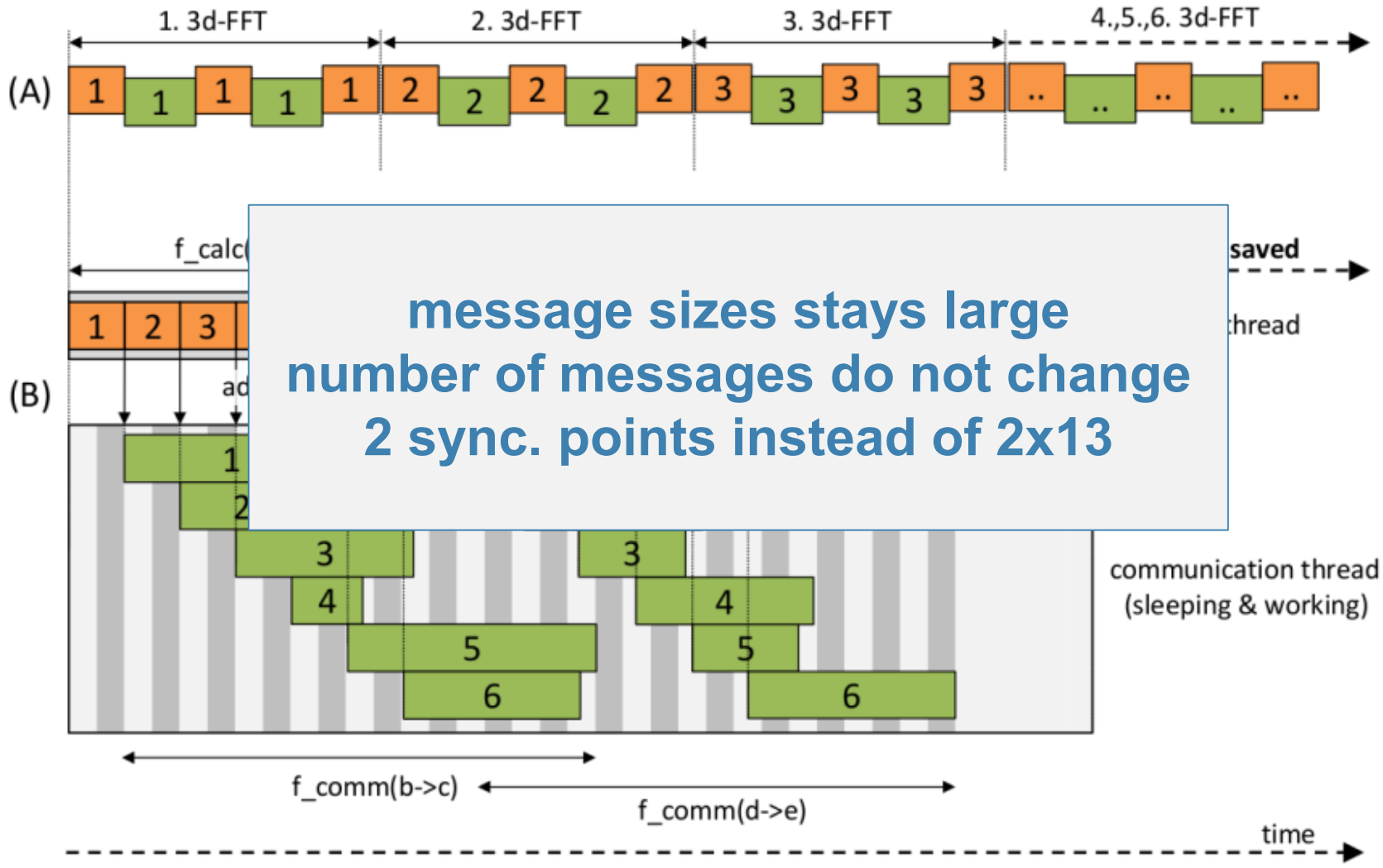
Implicit Filtering (dealiasing)



Goebbert, J. H.; Gauding, Michael; Ansorge, Cedrick; Hentschel, Bernd; Kuhlen, Torsten and Pitsch,Heinz
Direct Numerical Simulation of Fluid Turbulence at Extreme Scale with psOpen.
Advances in Parallel Computing Vol. 27 pp. 777-785 (2016).

nb3dfft Optimizations

Overlapping communication & computation



message sizes stays large
number of messages do not change
2 sync. points instead of 2x13

Overlapping Communication & Computation

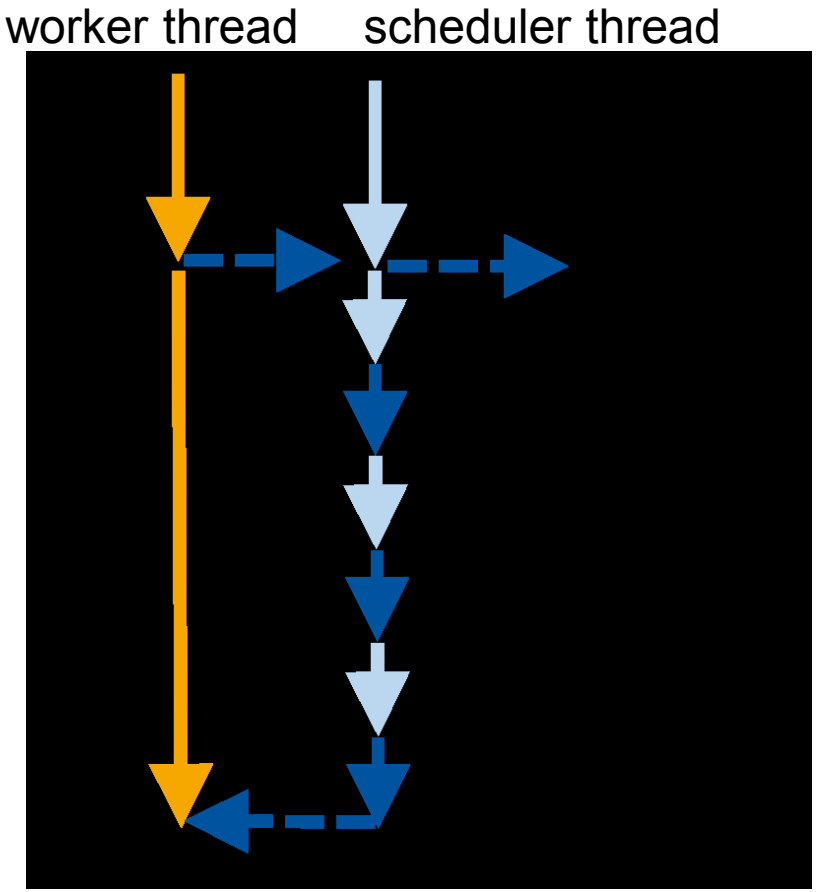
worker + scheduler thread

nb3dfft relies on the MPI 3.0 for
- non-blocking collective operations

and OpenMP for starting
- scheduler + worker threads.

The **worker** passes any non-blocking MPI operation to the **scheduler.**

The **scheduler** executes the MPI commands and iteratively calls MPI_TEST to proceed.



Overlapping Communication & Computation

application optimizations

JÜLICH FORSCHUNGSZENTRUM

1: **procedure** ITERATE VELOCITY $(n \to [n+1])$, PASSIVE SCALAR $([n-1] \to n)$
2: **input (velocity):** $\widehat{u}^{n}_{1|2|3},\ \widehat{G}^{n-1}_{4|5|6},\ \widehat{f}^{n-1}_{\boldsymbol{u}}$
3: **input (passive scalar):** $\widehat{\phi}^{n-1}_{17},\ J^{n-1}_{19},\ \widehat{J}^{n-2}_{18},\ \widehat{u}^{n-1}_{16}$
4: **for all** wave numbers $(\boldsymbol{\kappa})$ **do:** ▷ compute vorticity
5: $\widehat{\omega}^{n}_{x,10} \leftarrow \kappa_z \widehat{u}^{n}_{y,2} - \kappa_y \widehat{u}^{n}_{z,3}$
6: $\widehat{\omega}^{n}_{y,11} \leftarrow \kappa_x \widehat{u}^{n}_{z,3} - \kappa_z \widehat{u}^{n}_{x,1}$
7: $\widehat{\omega}^{n}_{z,12} \leftarrow \kappa_y \widehat{u}^{n}_{x,1} - \kappa_x \widehat{u}^{n}_{y,2}$
8:
9:
10:
11: **for all** wave numbers $(\boldsymbol{\kappa})$ **do:** ▷ advance passive scalar
12: $\widehat{\phi}^{n}_{17} \leftarrow \mathrm{F}_{\mathrm{adv}}(\widehat{\phi}^{n-1}_{17},\ \widehat{J}^{n-1}_{19},\ \widehat{J}^{n-2}_{18},\ \widehat{u}^{n-1}_{16})$
13: $\widehat{J}^{n-1}_{19} \overset{\text{zero copy}}{\rightleftharpoons} \widehat{J}^{n-2}_{18}$ ▷ swap memory
14: **for all** grid points (N) **do:** ▷ compute $\boldsymbol{G}^n = \boldsymbol{\omega}^n \times \boldsymbol{u}^n$
15: $G^{n}_{x,13} \leftarrow u^{n}_{y,8} \cdot \omega^{n}_{z,12} - u^{n}_{z,9} \cdot \omega^{n}_{y,11}$
16: $G^{n}_{y,14} \leftarrow u^{n}_{z,9} \cdot \omega^{n}_{x,10} - u^{n}_{x,7} \cdot \omega^{n}_{z,12}$
17: $G^{n}_{z,15} \leftarrow u^{n}_{x,7} \cdot \omega^{n}_{y,11} - u^{n}_{y,8} \cdot \omega^{n}_{x,10}$
18: $\widehat{\boldsymbol{\nabla}\phi}^{n}_{10|11|12} \leftarrow \widehat{\phi}^{n}_{17}$ ▷ 3x compute derivatives
19:
20:
21: $J^{n}_{19} \leftarrow \boldsymbol{u}^{n}_{7|8|9} \cdot \nabla \phi^{n}_{20|21|22}$ ▷ compute convective term
22: $\widehat{\boldsymbol{G}}^{n}_{13|14|15} \overset{\text{zero copy}}{\rightleftharpoons} \widehat{\boldsymbol{G}}^{n-1}_{4|5|6}$ ▷ swap memory
23: $\widehat{u}^{n}_{16} \overset{\text{zero copy}}{\rightleftharpoons} \widehat{u}^{n}_{2}$ ▷ swap memory
24: $\widehat{f}^{n}_{\boldsymbol{u}} \leftarrow \mathrm{F}(\widehat{f}^{n-1}_{\boldsymbol{u}})$ ▷ compute forcing energy
25: **for** wave numbers $(|\kappa_{\mathrm{ijk}}| < k_f = 2\sqrt{2})$ **do:** ▷ add forcing energy
26: $\widehat{\boldsymbol{u}}^{\mathrm{force}}_{1|2|3} \leftarrow \mathrm{F}_{\mathrm{force}}(\widehat{\boldsymbol{u}}^{n}_{1|16|3},\ \widehat{f}^{n}_{\boldsymbol{u}})$
27: **for all** wave numbers $(\boldsymbol{\kappa})$ **do:** ▷ advance velocity
28: $\widehat{\boldsymbol{u}}^{\mathrm{step1}}_{1|2|3} \leftarrow \mathrm{F}_{\mathrm{step1}}(\widehat{\boldsymbol{u}}^{\mathrm{force}}_{1|2|3},\ \widehat{\boldsymbol{G}}^{n}_{4|5|6},\ \widehat{\boldsymbol{G}}^{n-1}_{13|14|15})$
29: **for all** wave numbers $(\boldsymbol{\kappa})$ **do:** ▷ apply projection tensor
30: $\widehat{\boldsymbol{u}}^{n+1}_{1|2|3} \leftarrow \mathrm{F}_{\mathrm{step2}}(\widehat{\boldsymbol{u}}^{\mathrm{step1}}_{1|2|3})$
31: $\langle \widehat{\boldsymbol{u}}^{n+1}_{1|2|3} \rangle = 0$ ▷ set mean velocity to zero
32: **output (velocity):** $\widehat{\boldsymbol{u}}^{n+1}_{1|2|3},\ \widehat{\boldsymbol{G}}^{n}_{4|5|6},\ \boldsymbol{u}^{n}_{7|8|9},\ \widehat{u}^{n}_{16}$
33: **output (passive scalar):** $\widehat{\phi}^{n}_{17},\ J^{n}_{19},\ \widehat{J}^{n-1}_{18}$

nb3dfft Optimizations

Overlapping communication & computation

JÜLICH FORSCHUNGSZENTRUM

- message sizes stay large
- number of messages do not change
- 2 synchronization points instead of 2x13

Performance increase by overlapping comm & comp

	1024^3	**2048^3**	**4096^3**	**6144^3**
1024 nodes	15.50%	21.60%	-	-
2048 nodes	-	26.63%	29.63%	-
4096 nodes	-	17.70%	19.78%	-
8192 nodes	-	11.64%	22.69%	17.53%

Conclusion & Summary

JÜLICH FORSCHUNGSZENTRUM

Getting pseudo-spectral codes for DNS **scale on large systems** is challenging

- 3d-FFTs require lots of MPI_alltoall() calls, which cannot be avoided
- for best performance key functions have to rewritten

3d-FFT can be optimized, if it is not thought as a single function call,
but as part of the algorithm

- Moving the required **filter operation** of the main algorithm into the 3d-FFT reduces the data to be send by 44.4% and the 1d-FFTs by 44.4%

- Calling multiple 3d-FFT for different fields at the same time allows **overlapping of communication and computation**
  -> ~20% performance gain

nb3dfft is available under GPL License at
https://gitlab.version.fz-juelich.de/goebbert/nb3dfft

Appendix

nb3dfft

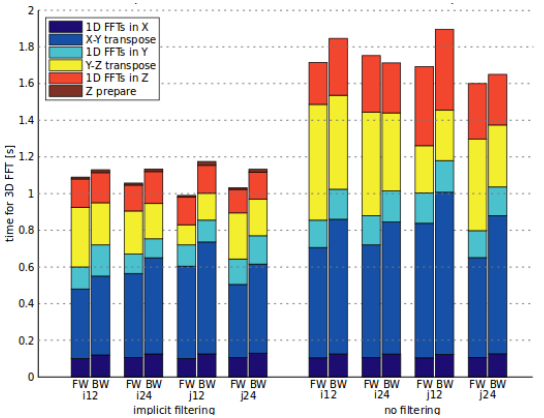
nb3dfft Optimizations

Implicit Filtering

Performance Measurements

RWTH Compute Cluster

- 672 MPI processes on 56 nodes
- domain size 2048^3



JUQUEEN

- 16.384 MPI processes on 8192 cores
- domain size 4096^3