Lower floating-point precision in high-fidelity turbulence simulations

Philipp Schlatter[1,2], **Martin Karp[2], Ronith Stanly[2]**, Timofey Mukha[4], Luca Galimberti[4,5], Siavash Toosi[1],
Hang Song[3], Lisandro Dalcin[4], Saleh Rezaeiravesh[6],
Niclas Jansson[2], Stefano Markidis[2], Matteo Parsani[4], Sanjeeb Bose[3,7], Sanjiva Lele[3]

1 Institute of Fluid Mechanics (LSTM) Friedrich-Alexander-Universität (FAU) Erlangen-Nürnberg, Germany
2 KTH Royal Institute of Technology, Sweden
3 Stanford University, USA
4 King Abdullah University of Science & Technology (KAUST)
5 Politecnico di Milano, Italy
6 University of Manchester, UK
7 Cadence Design Systems, Inc., USA

Team (red: present in Stanford)

- **KTH**
  - **Martin Karp**, CST
  - Stefano Markidis, CST
  - Niclas Jansson, PDC
  - Ronith Stanly, Mechanics
- **FAU - LSTM**
  - **Philipp Schlatter** (also affiliated with KTH)
  - Siavash Toosi
- **University of Manchester**
  - Saleh Rezaeiravesh
- **King Abdullah University of Science and Technology (KAUST)**
  - Timofey Mukha
  - Luca Galimberti
  - Lisandro Dalcin
  - Matteo Parsani
- **Center for Turbulence Research**
  - **Hang Song**
  - **Sanjeeb Bose**
  - **Sanjiva Lele**

Direct numerical simulation of a starting rotor

Adam Peplinski[1], Elham Bagheri[1], Ronith Stanly[1], Siavash Toosi[2],

Niclas Jansson[1], Timofey Mukha[3], Philipp Schlatter[1,2]

[1] KTH Royal Institute of Technology, Stockholm, Sweden

[2] Friedrich–Alexander–Universität (FAU) Erlangen–Nürnberg, Germany

[3] King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

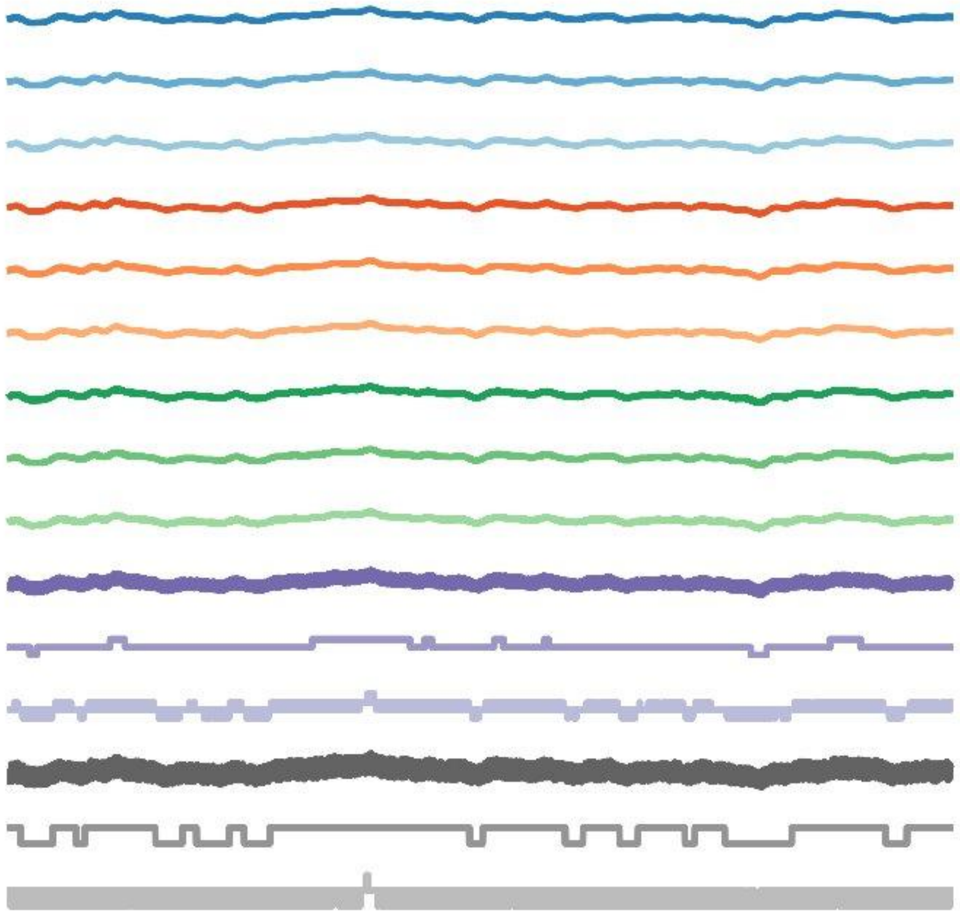
Outline

- **Motivation**
- **Floating-point numbers**
  - Formats
  - Rounding and relation to noise
- **Methodology**
- **Initial Case studies**
  - Lorenz system
- **DNS cases**
  - Turbulent channel flow at Ret=550
  - K-type Transition
  - Flow around a circular cylinder at ReD=3900
- **Summary and Outlook**

Link to the paper: https://arxiv.org/abs/2506.05150
accpeted in J. Comput. Phys. (2025)

Motivation

- Low-precision floating-point support in modern computer hardware - large performance improvements in particular for AI applications.
- **Opportunities for CFD:**
  - Reduce cost (less data movement and higher flop/s)
  - Cheaper chips (consumer vs server-grade)
  - Memory footprint (larger problem size per node)
- **De-facto standard is double precision** (64bit precision)
- **Single precision – shown promise for several cases**
  - Direct numerical simulation of homogenous isotropic turbulence (Homann et al., 2007, PK Yeung et al., 2015)
  - Certain parts of codes (e.g. pressure, preconditioners) run in single precision
  - Implicit LES - PyFR (Witherden and Jameson, 2020)
  - Ansys Fluent, OpenFOAM typically in single precision
- **When and how can single or lower floating point formats be utilized?**
  - And are there any commonalities across solvers integrating the Navier-Stokes equations?



Time series of channel flow simulation sampled with lower precision

Floating-point numbers

- **Defined by sets of bits**
  - Sign bit
  - Exponent bits
    - Range "How large number can be represented"
    - Base 2 in binary IEEE
  - Mantissa/significand bits
    - Precision - "Smallest relative difference that can be represented"
- **Machine epsilon, ε**
  - 2^(Mantissa bits +1)
- **Special numbers (Inf, NaN, and denormal)**
- **Quantization of real numbers**
  - Rounding

$$(-1)^{\text{sign}}\left(1+\sum_{i=1}^{52} b_{52-i}2^{-i}\right)\times 2^{e-1023}$$

Conventional IEEE double precision number (64 bit)

Possible to compile and run

emulated in software

Name	Bits	b	b_e	Machine epsilon ε
FP64 (double)	64	52	11	$2^{-53}\approx 10^{-16}$
FP32 (single)	32	23	8	$2^{-24}\approx 6\cdot 10^{-8}$
FP16 (half)	16	10	5	$2^{-11}\approx 5\cdot 10^{-4}$
bfloat16	16	7	8	$2^{-8}\approx 4\cdot 10^{-3}$
q43 (quarter)	8	3	4	$2^{-4}=0.0625$
q52 (quarter)	8	2	5	$2^{-3}=0.125$

Table: Number of mantissa bits b, exponent bits b_e. Machine epsilon, the relative difference is maximum ε, $|u_{\text{FP}}-u|<\varepsilon u$.

Rounding

- **Round to nearest (RTN)**
  - Conventional rounding
  - Worst case rounding of addition of n numbers ~εn
- **Stochastic rounding (SR)**
  - Round up/down with certain probability
  - Mean-independent
  - Increasing interest due to avoiding stagnation (Croci, 2022)
  - Round-off for n numbers scales as ~sqrt(n)ε (Connolly, Higham, and Mary, 2020)
- **Noise**
  - Independent perturbations
  - Model quantization through modifying PDF of compressed signal (Widrow, 1996)

$$P(X = x_1) = \frac{x_2 - x}{x_2 - x_1}$$
$$P(X = x_2) = \frac{x - x_1}{x_2 - x_1}$$
$$E(X) = x_1 P(x_1) + x_2 P(x_2) = x$$

Rounding procedure for a real number x between x_2 and x_1 with SR

$$\tilde{u} = \mathrm{round}_{\mathrm{RTN,FP}}(u)$$
$$\tilde{u} = \mathrm{round}_{\mathrm{SR,FP}}(u)$$
$$\tilde{u} = (1 + \epsilon\delta)u, \quad \delta \sim U^n_{[-1,1]}$$

Perturbing a value u by rounding to a FP format with machine epsilon ε with SR, RTN and adding uniform noise.

- State rounding
- Specific terms rounding (e.g. convective term)

CPFloat for emulation (Fasi and Mikaitis, 2023)

Lorenz system as illustration

Rounding and perturbation of state

Lorenz system - RTN

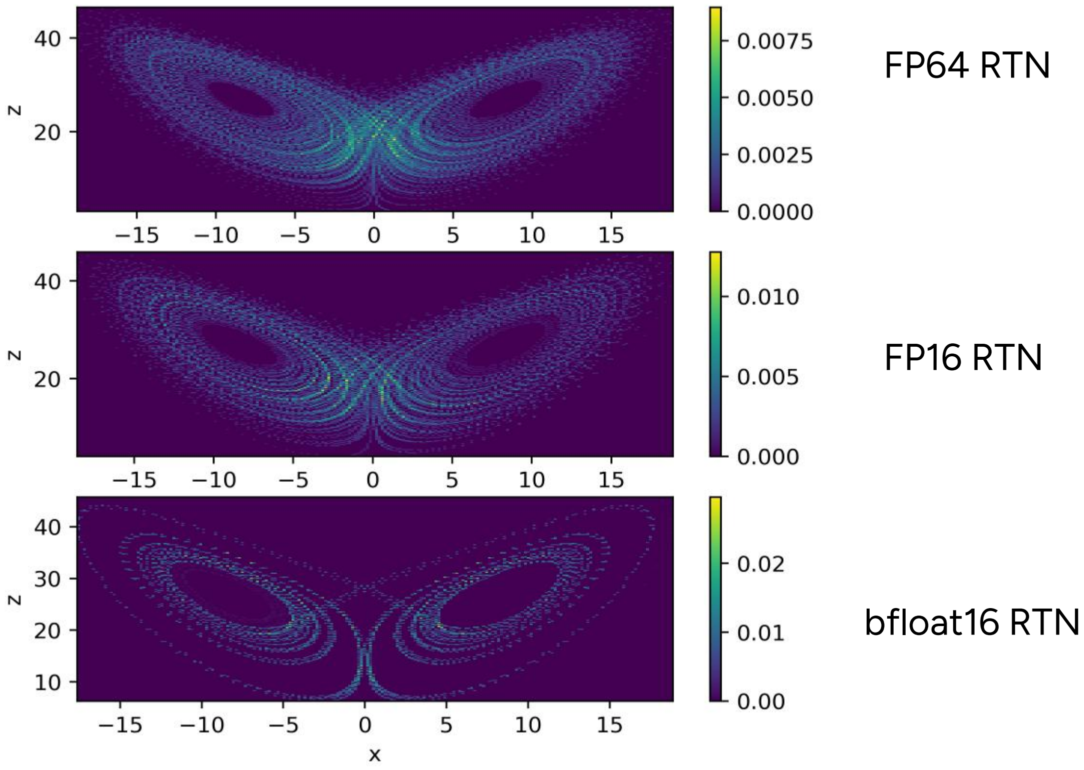
- **Low dimensional chaotic system**

$$\frac{d\mathbf{u}}{dt} = (\sigma(y-x), x(\rho - z) - y, xy - \beta z)$$

- **Discretize with RK4**
- **Integrate perturbed state,**

$$\mathbf{u}^{i+1} = \widetilde{\mathbf{F}(\mathbf{u}^i)}$$

- **Low precision RTN -> Periodic orbits**



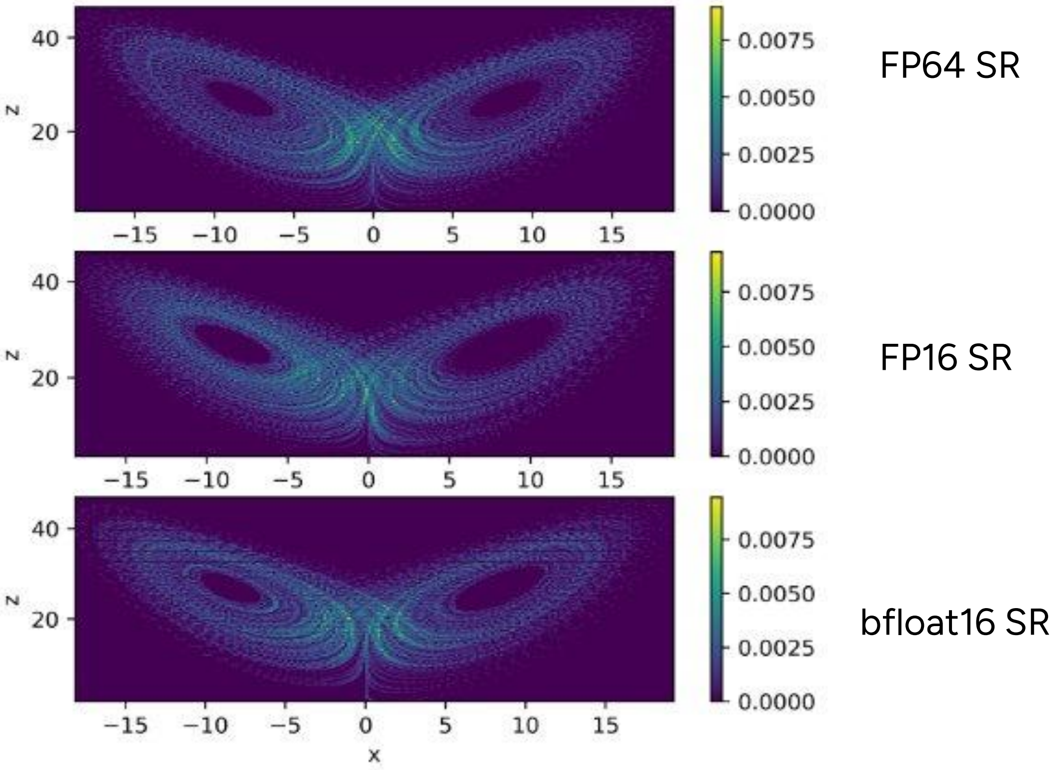
Histogram of x,z variables of perturbed Lorenz system

Lorenz system - SR

- **Low precision SR**
  - **Better coverage of phase-space**
- **First order moment of *z* (RTN, SR, Noise)**

	$\langle \bar{z} \rangle$
FP64	23.55, 23.55, 23.55
FP32	23.55, 23.55, 23.55
FP16	23.50, 23.55, 23.55
bfloat16	24.14, 23.59, 23.58
Q43	37.28, 22.77, 24.73
Q52	12.49, 25.51, 25.56

  - 1000 ensembles, 1000 time units
  - Aligns with Oliver et al., 2014



Histogram of x,z variables of perturbed Lorenz system

Observations with perturbed state

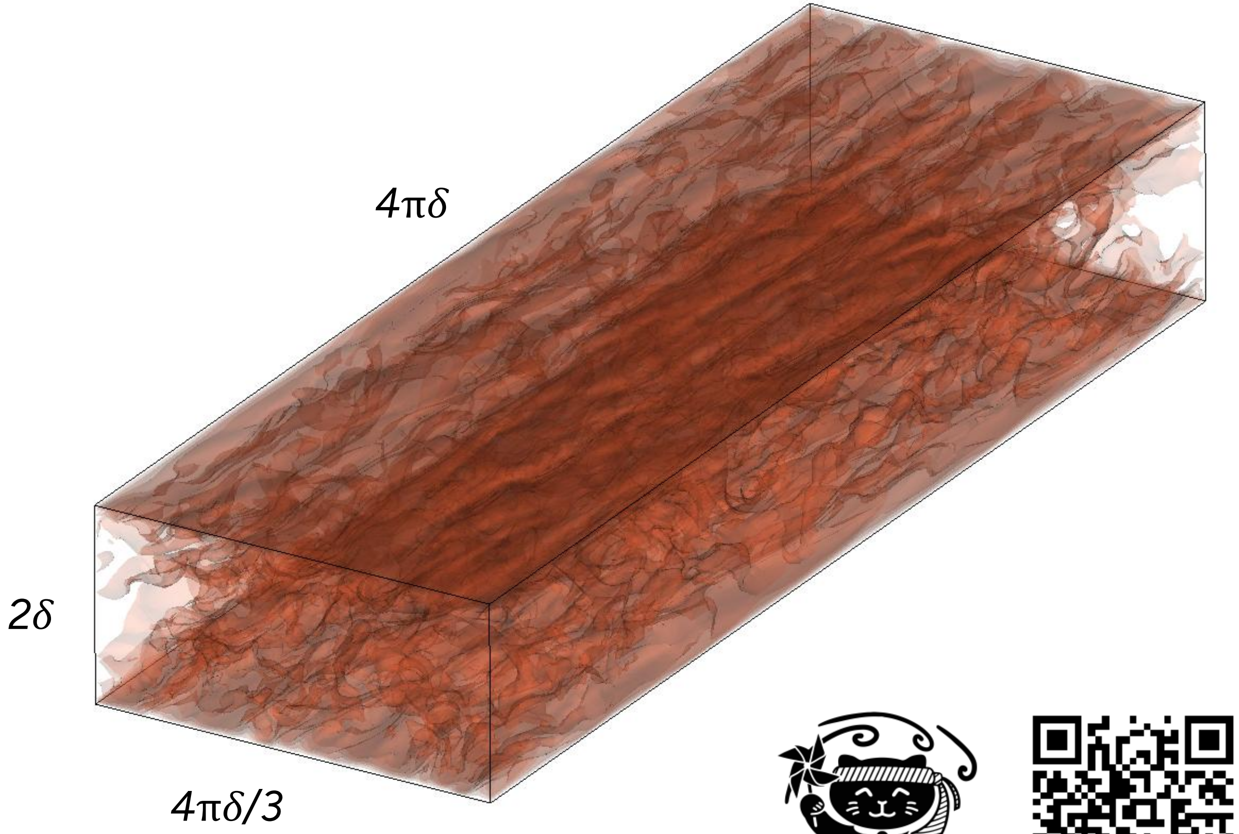
- **SR outperforms RTN**
- **Stagnation for small Δt**
- **Few DoFs -> Periodic "nonphysical" orbits**
  - "Nonphysical" shadowing orbits (Chandramoorthy, Wang 2021) – Can be found for high precision too
  - ***But…*** Shadowed trajectories seem to often work for our simulations
  - Reliability of simulations of chaotic systems (Corless 1994)

- **Difference - Significantly more DoFs in turbulence simulations**
  - Probability of being in periodic orbits scale with ~ ε^{DoFs}
  - Technically, scales with correlation dimension ~ DoFs
  - (Grebogi, Ott, Yorke 1988), (Beck, Roepstorff 1987)
  - -> Probability of being in basin of turbulent attractor is large

Channel flow Re_τ=180

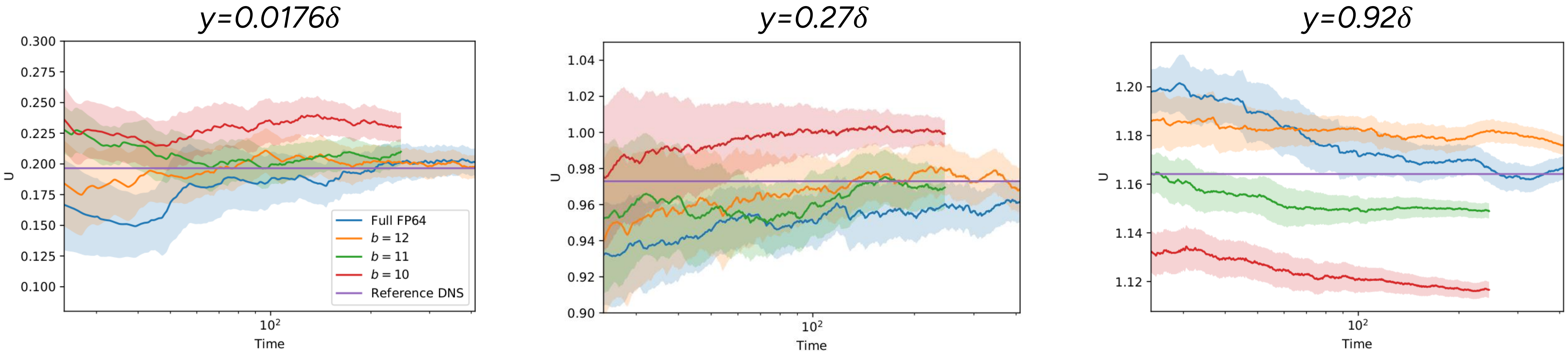
- **Non-dimensionalized**
  - Bulk velocity, U_b
  - Channel half height, δ
  - $Re_b = U_b\delta/\nu$
- **Spectral element discretization**
  - 18^3 elements, polynomial order 7
  - 127^3 unique grid points
- **How is uncertainty in time affected?**
  - Vary mantissa bits manually
  - Only RTN

Velocity magnitude isosurfaces in channel flow at Re_τ=180



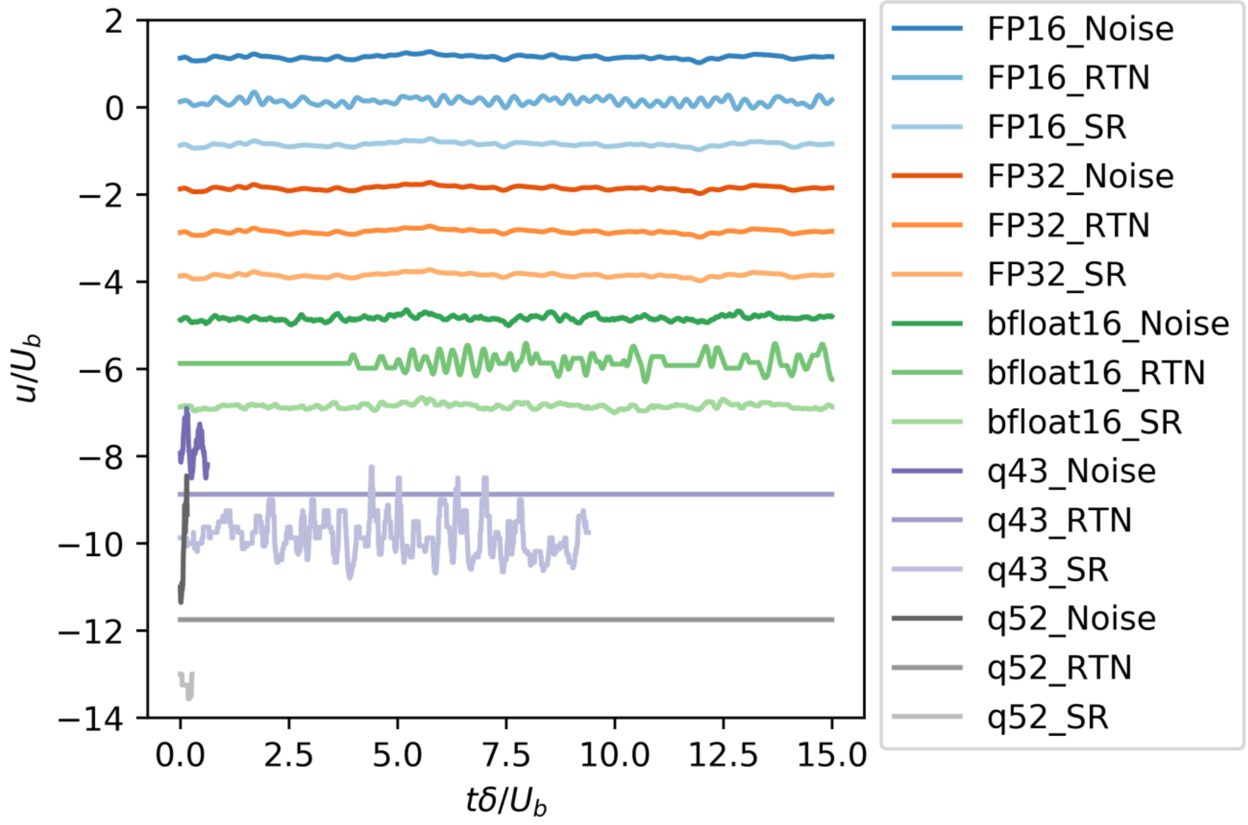
Convergence of Mean Flow with RTN and vary *b*

- **Bias in the middle of channel**
  - Low precision fails to capture different scales
  - Lower turbulence intensity
- **Near-wall region – Less sensitive**
  - Larger velocity gradients
- SC23 Workshops: Karp, Liu, Stanly, Jansson, Rezaeiravesh, Schlatter, and Markidis



Channel flow Re_τ=180

- **Compare different FP formats and rounding/noise for state**
- **Stagnation of quarter precision RTN**
- **Presented at DLES 2024:**
  - Karp, Jansson, Rezaeiravesh, Markidis, and Schlatter 2024 (Proceedings soon)



Timeseries of streamwise velocity in middle of channel for perturbed simulation, offset in y by u/Ub=1

DNS cases

Higher Re channel flow, K-type transition, flow around a circular cylinder

Different solvers, Perturbation of state and convective term, and full single-precision

Overview

- **Compare impact of floating-point precision**
  - Solvers with different discretizations and methods
  - Terms in the Navier-Stokes equations
    - State and convective term
  - Flow cases – Turbulence, transition, and separation
  - Highlight different aspects for the different cases and solvers

- **Center for Turbulence Research (CTR) Summer Program**
  - 4 weeks at CTR, Stanford, Biannual – Started 1987
  - Today ~50 projects and ~100 people
  - Dedicated hosts from CTR
  - Weekly status reports together with other participants

- **Collaborators:** R. Stanly, H. Song, T. Mukha, L. Galimberti, S. Toosi, S. Rezaeiravesh, M. Munch, N. Jansson, S. Markidis, L. Dalcin, M. Parsani, S. Bose, S. Lele, and P. Schlatter
  - KTH, KAUST, UoM, CTR, FAU

Solvers

- **PadeLibs** – Compact finite differences (compressible NS), (Song et al. 2024)
- **Neko** – Incompressible Navier-Stokes, spectral elements (Incompressible NS), (Jansson et al. 2024)
- **SSDC** – High-order discontinuous Galerkin (compressible NS), (Parsani, 2020)
- **SIMSON** – Pseudo-spectral solver (incompressible NS), (Chevalier et al. 2007)
- State, convective term, and full single/double

$$\frac{\partial u_j}{\partial x_j} = 0$$

$$\frac{\partial u_i}{\partial t} + \boxed{u_j \frac{\partial u_i}{\partial x_j}} = -\frac{\partial p}{\partial x_i} + \frac{1}{\mathrm{Re}} \frac{\partial^2 u_i}{\partial x_j \partial x_j}$$

Incompressible

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x_j}(\rho u_j) = 0$$

$$\frac{\partial \rho u_i}{\partial t} + \frac{\partial}{\partial x_j}(\boxed{\rho u_i u_j} + p\delta_{ij}) = \frac{\partial \sigma_{ij}}{\partial x_j}$$

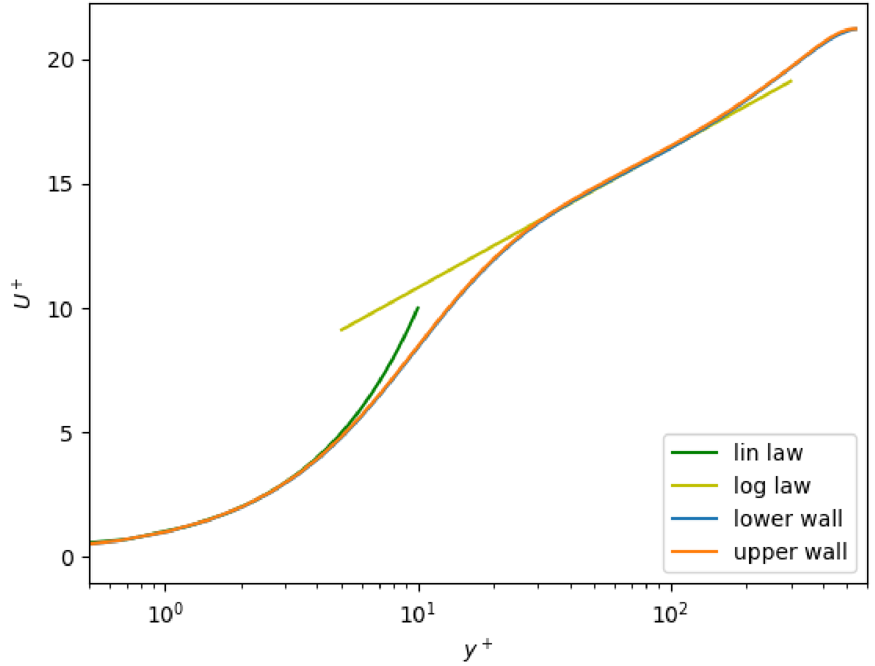
$$\frac{\partial \rho e}{\partial t} + \frac{\partial}{\partial x_j}(\rho h u_j) = \frac{\partial}{\partial x_j}(u_i \sigma_{ij} - q_j)$$

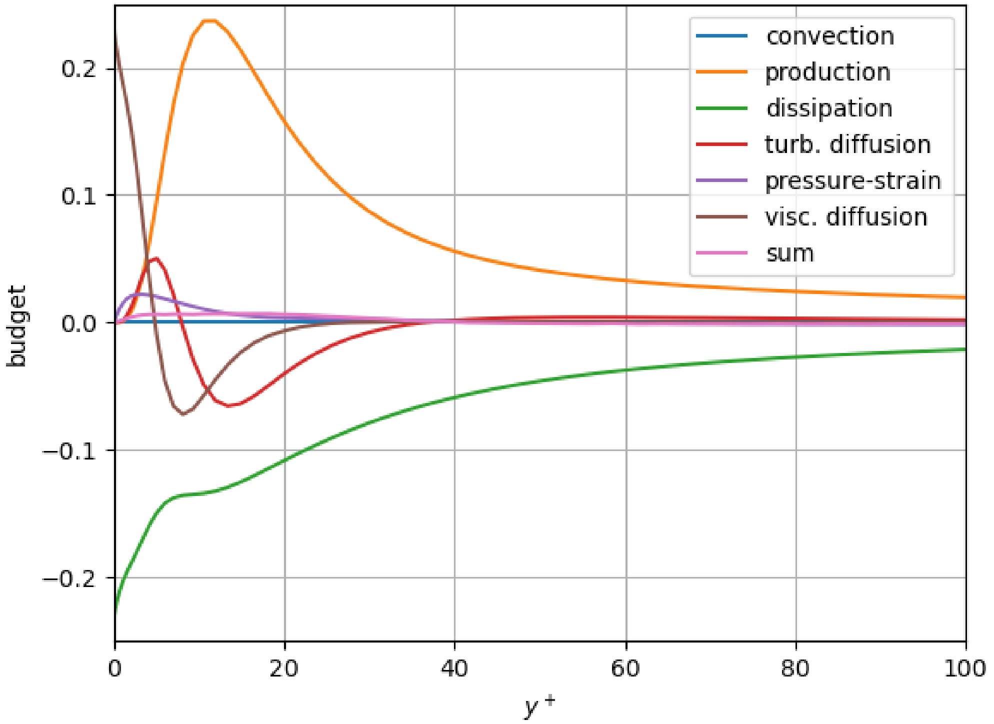
Compressible

Channel flow – SIMSON – Full single precision

- **Can we run in single precision? Including FFTs and all?**

Motivation: Jiménez et al. has state in single, although algebra in double. Previous TBL simulations in SIMSON have used single precision for MPI_ALLTOALL.

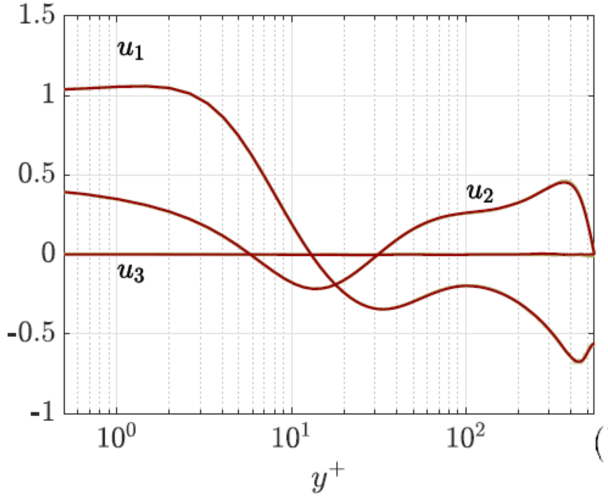




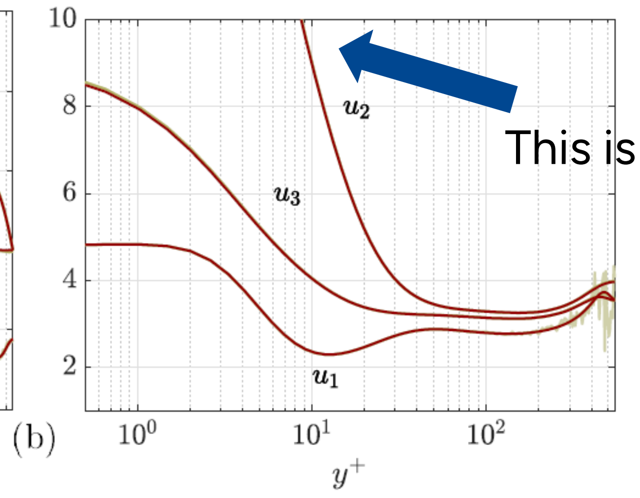
No difference. Somewhat surprising to us: We need to dig further: What quantity is sensitive? (vorticity, 2pt statistics, …). What is the influence of Re?

Channel flow – SIMSON – Full single precision

- We looked at many things…



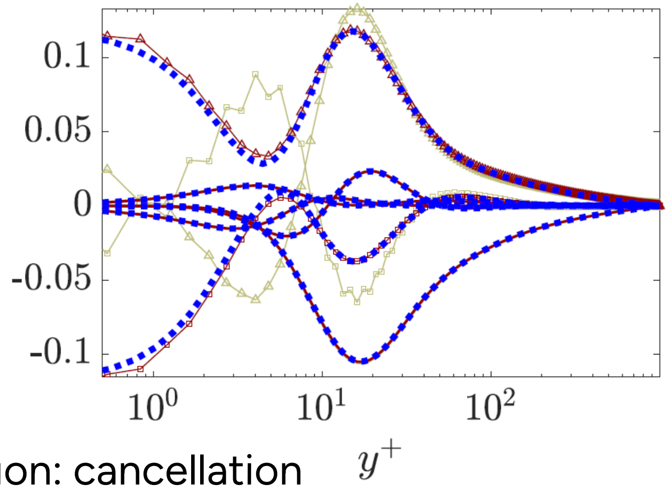
Skewness



Flatness

This is very sensitive!

- We found two issues:
  - The pressure
  - The mass-flux calculation: cancellation



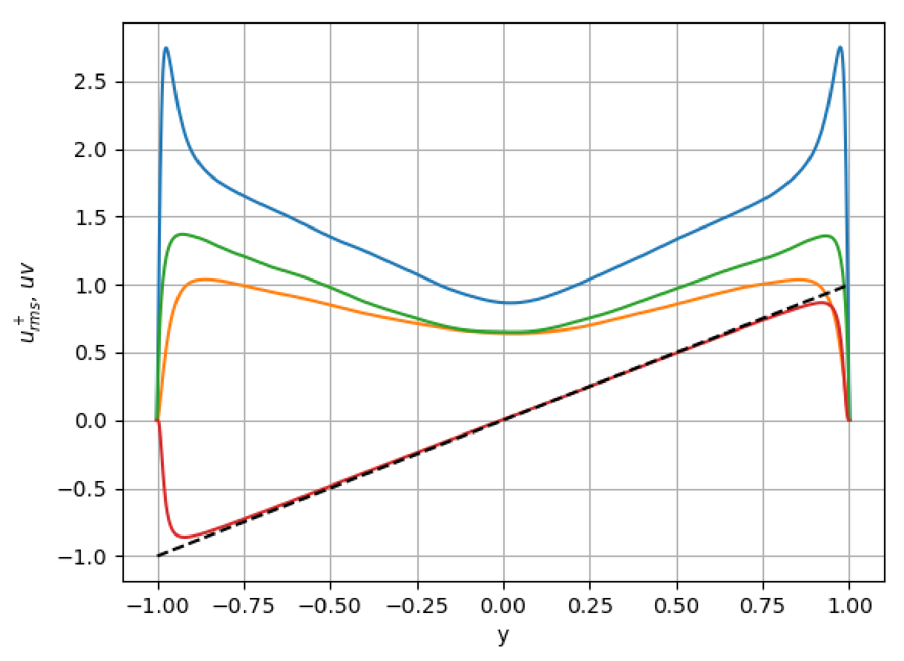
Budget for <u'v'>. The terms that are off are the pressure-transport terms.

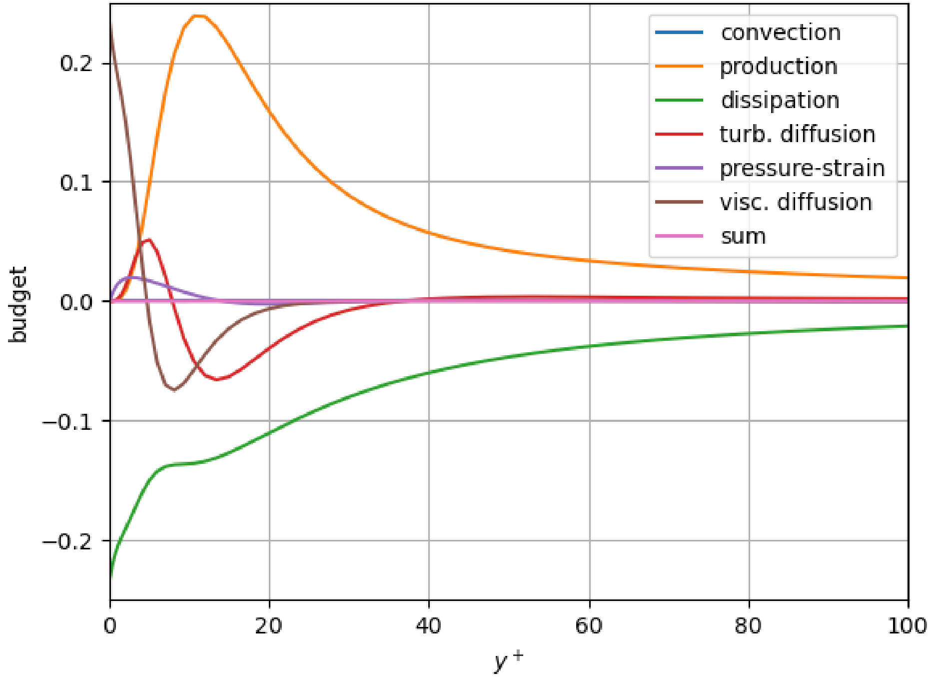
However, recomputing them (and only them) in FP64 fixed the problem (red)

Channel flow - SIMSON – State rounding

- **How low can we go with reducing the state (u,v,w,p) while keeping the calculations accurate (double precision)**

Motivation: exploit modern hardware more efficiently. But what terms?



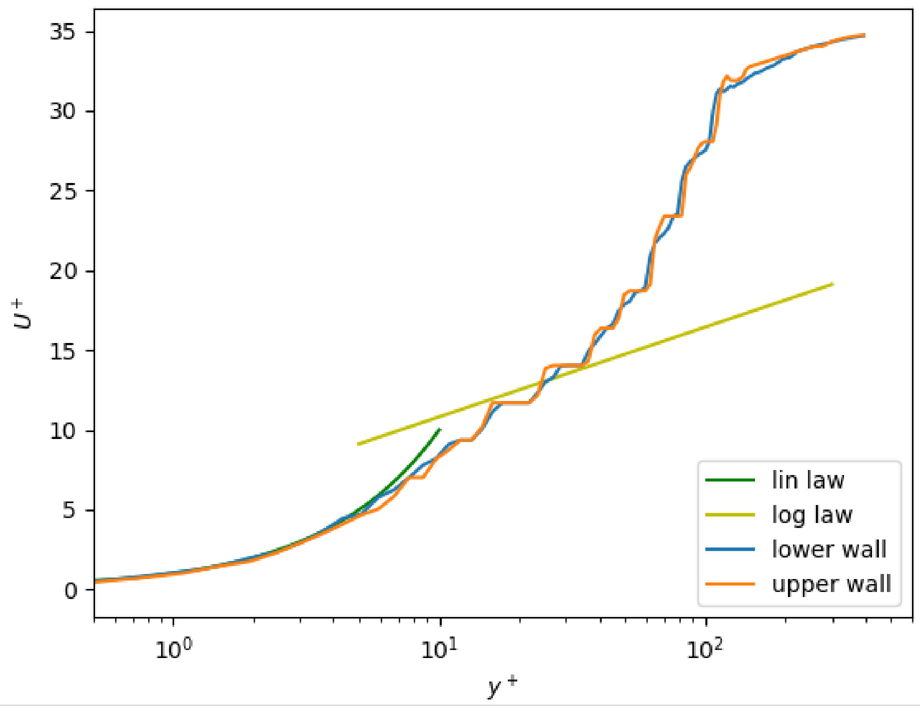


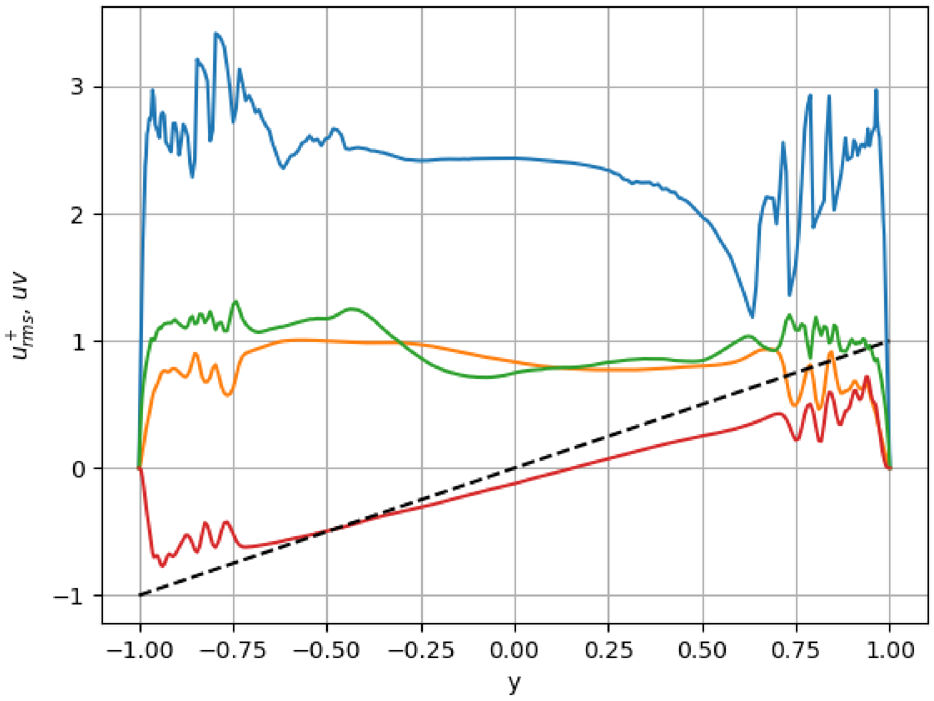
FP16 ("half-precision"): No noticeable difference

Channel flow - SIMSON – State rounding

- **How low can we go with reducing the state (u,v,w,p) while keeping the calculations accurate (double precision)**

Motivation: exploit modern hardware more efficiently. But what terms?



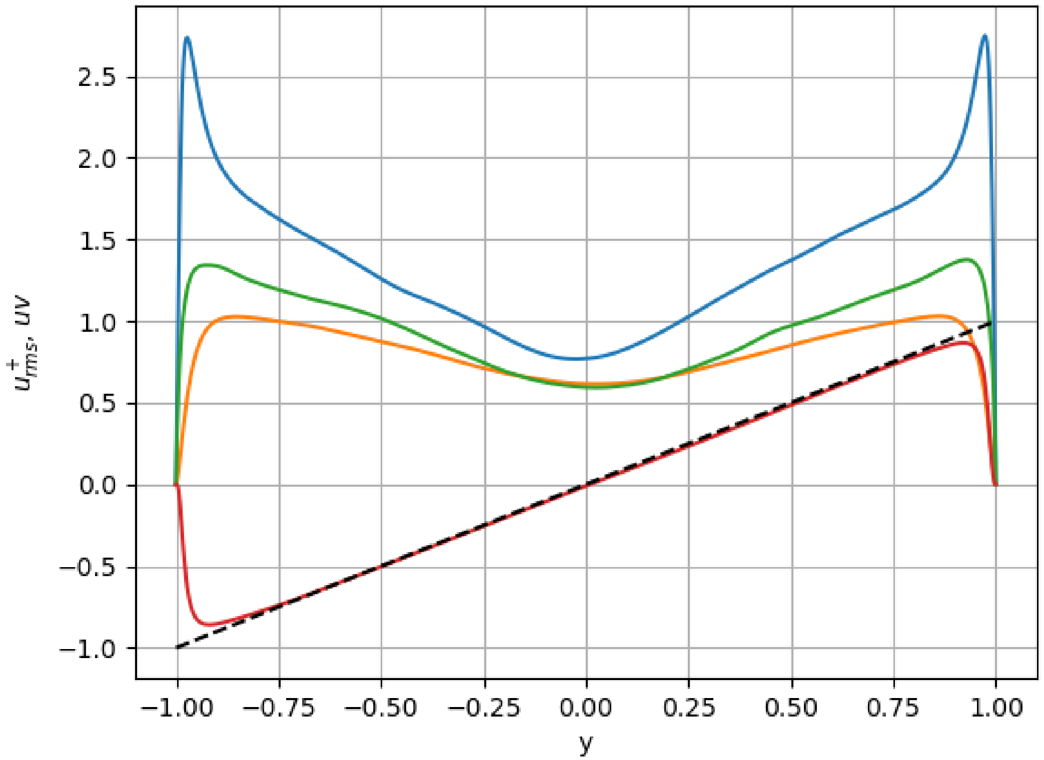


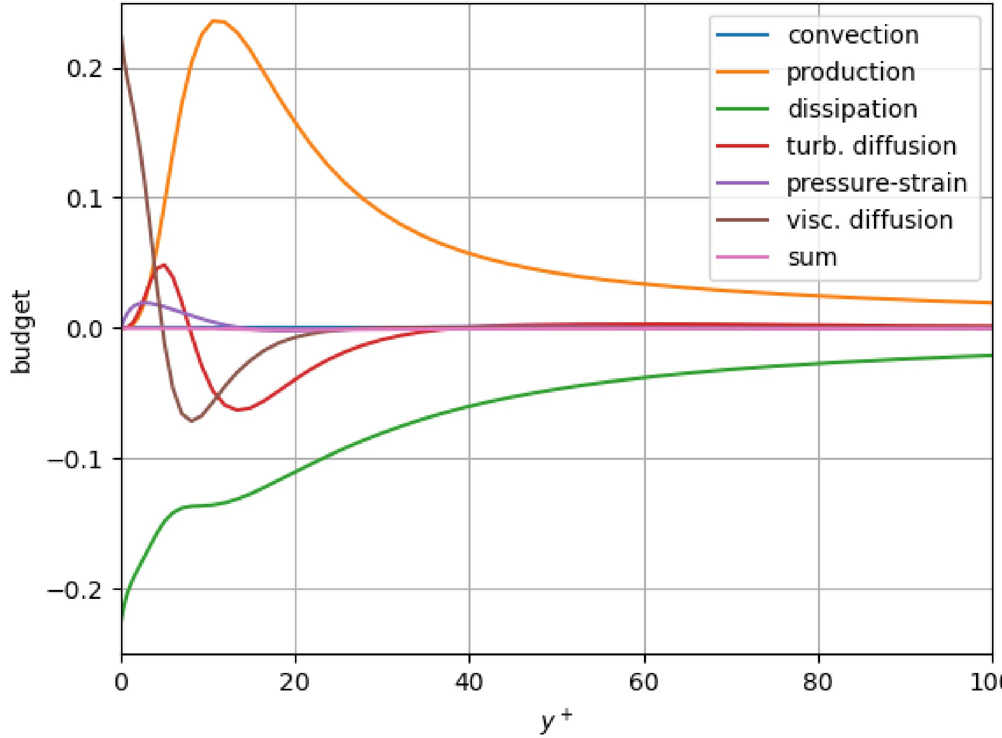
FP08 ("quarter-precision" e5m2): now things get bad, but keep in mind, this is 3 counting digits...

Channel flow – SIMSON – Convective term

- **How low can we go with reducing the convection operator?**

Most costly operation in SIMSON. Term exists in all codes, regardless of formulation.

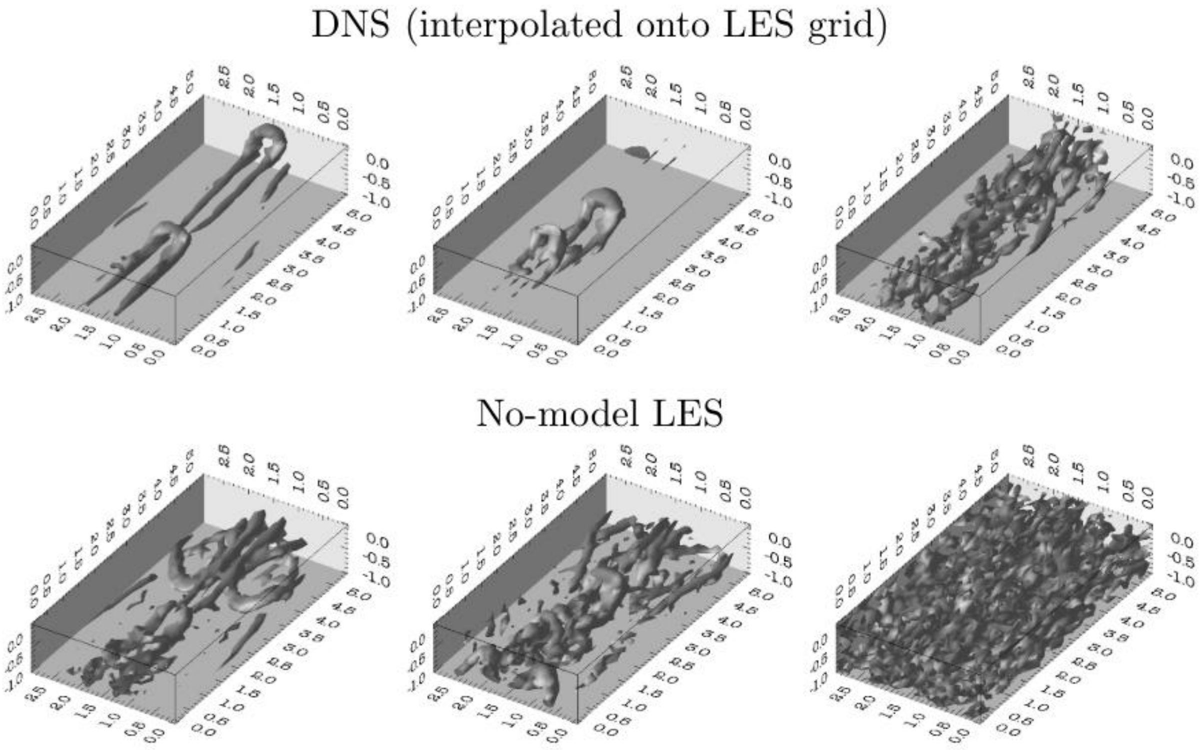




FP08 ("quarter-precision"): if restricted to convective operator, things are ok.

K-type Transition

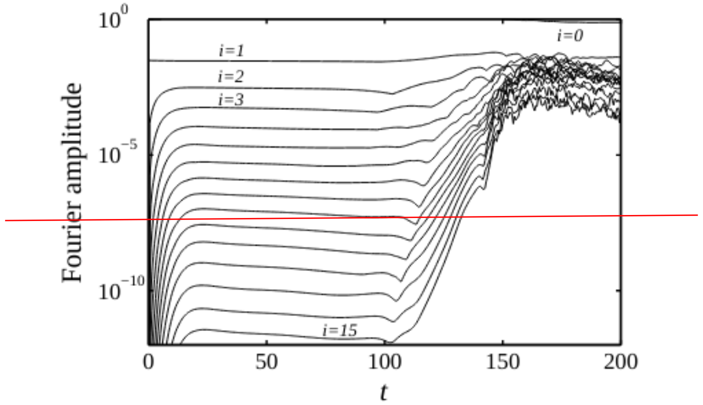
- What about non-turbulent flows?
  Question: Does low precision introduce additional disturbances?
- Transitional flow: K-type transition (going back to Nishioka 1975, Kleiser 1980, Schlatter 2005)
- Rel. low deterministic disturbances (3% for 2D wave, 0.1% for 3D wave)
- "low" Reynolds number (Reb=3333) typical for transition
- Motivation: LES (i.e. low resolution) is problematic (Schlatter 2005)
- Is precision leading to similar effects?
- Note that resolved near-wall turbulence is very similar to repeated transition (Schoppa and Hussain 2001)



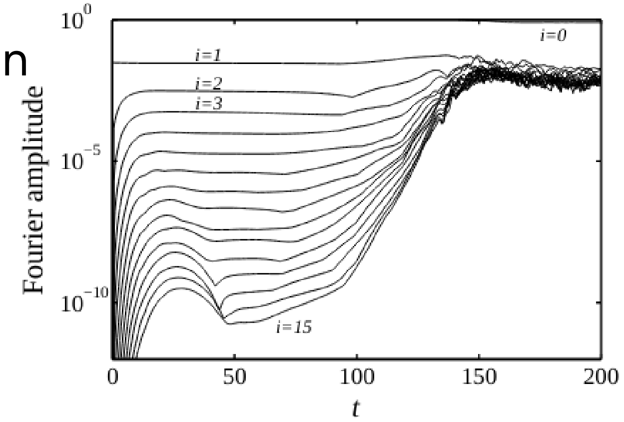
K-type Transition

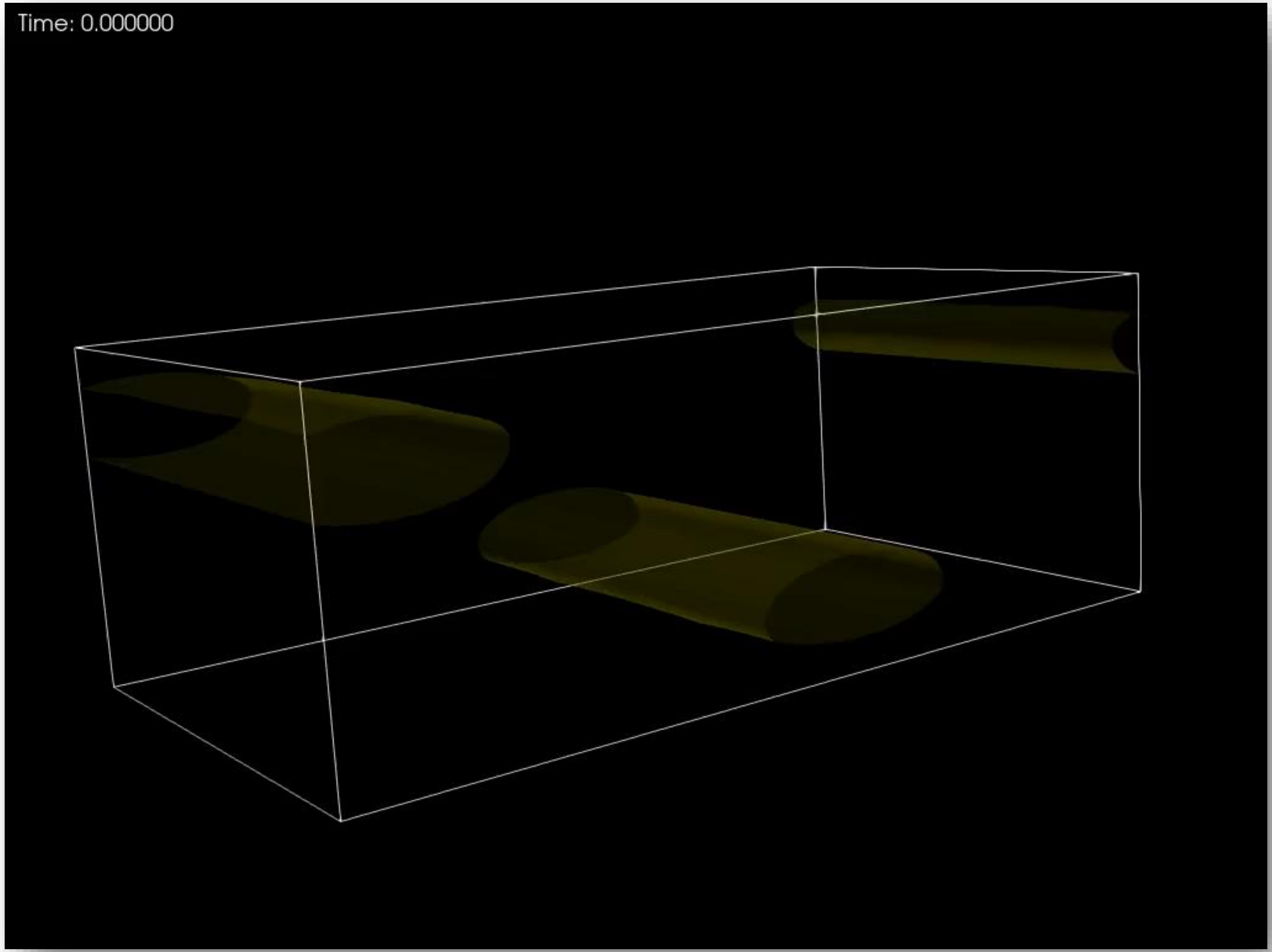
- Flow case in Neko (~120^3, double precision)
- Original case by Gilbert and Kleiser (1986)
  Our setup according to Schlatter 2005.

Full DNS



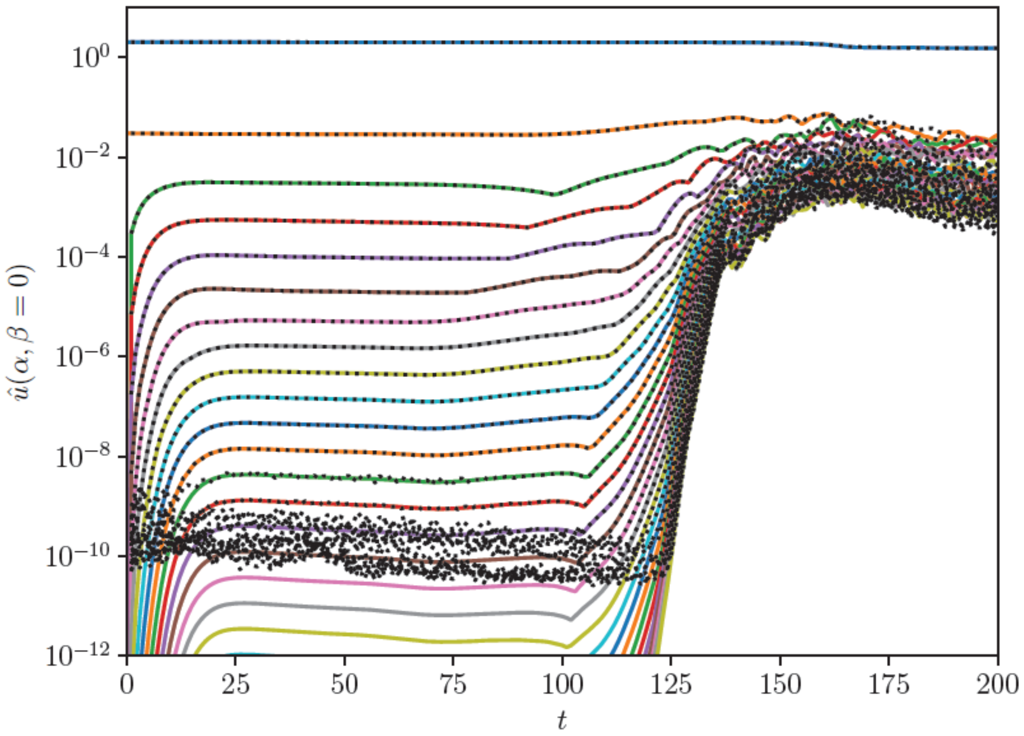
Low resolution DNS



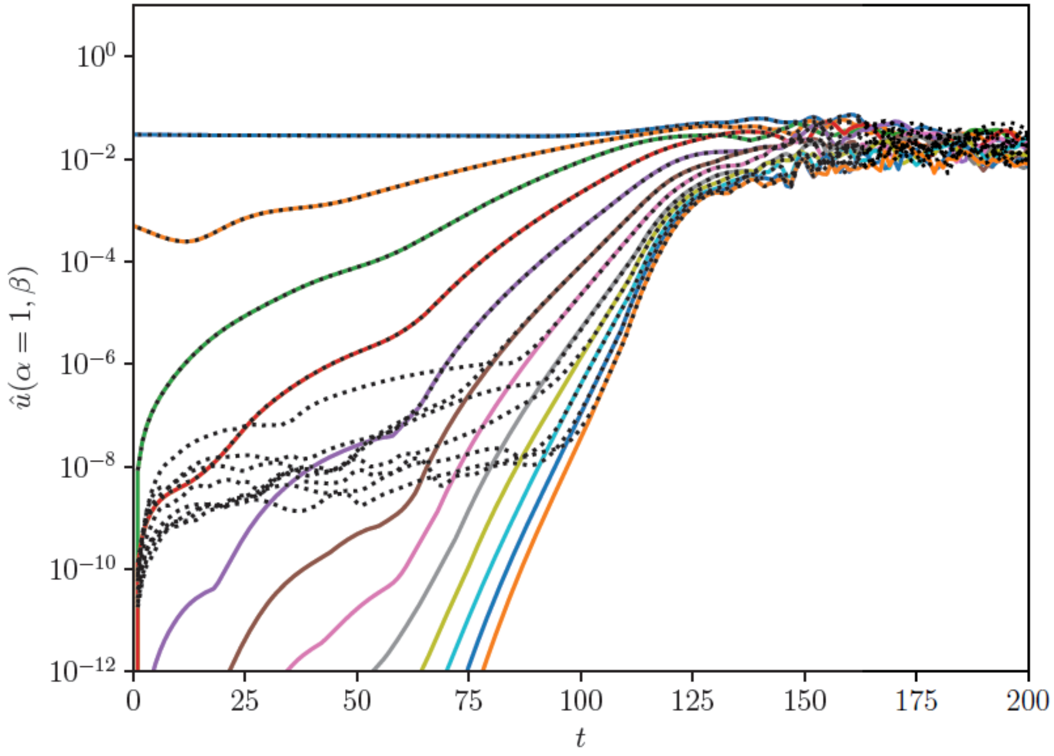


K-type Transition

- Double vs single precision (Simson and Neko)



Spanwise-constant modes

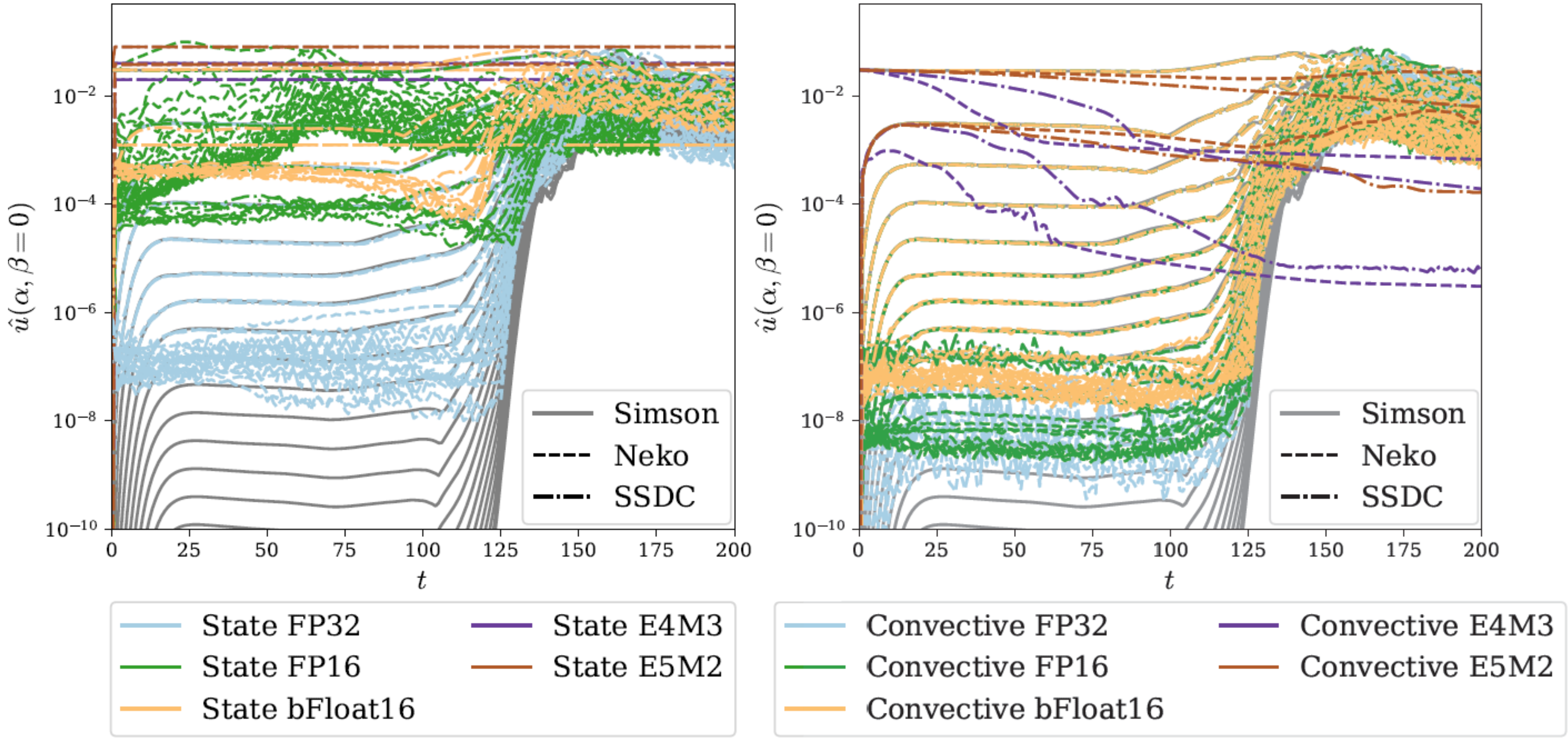


Spanwise modes

No difference between single and double!

K-type Transition

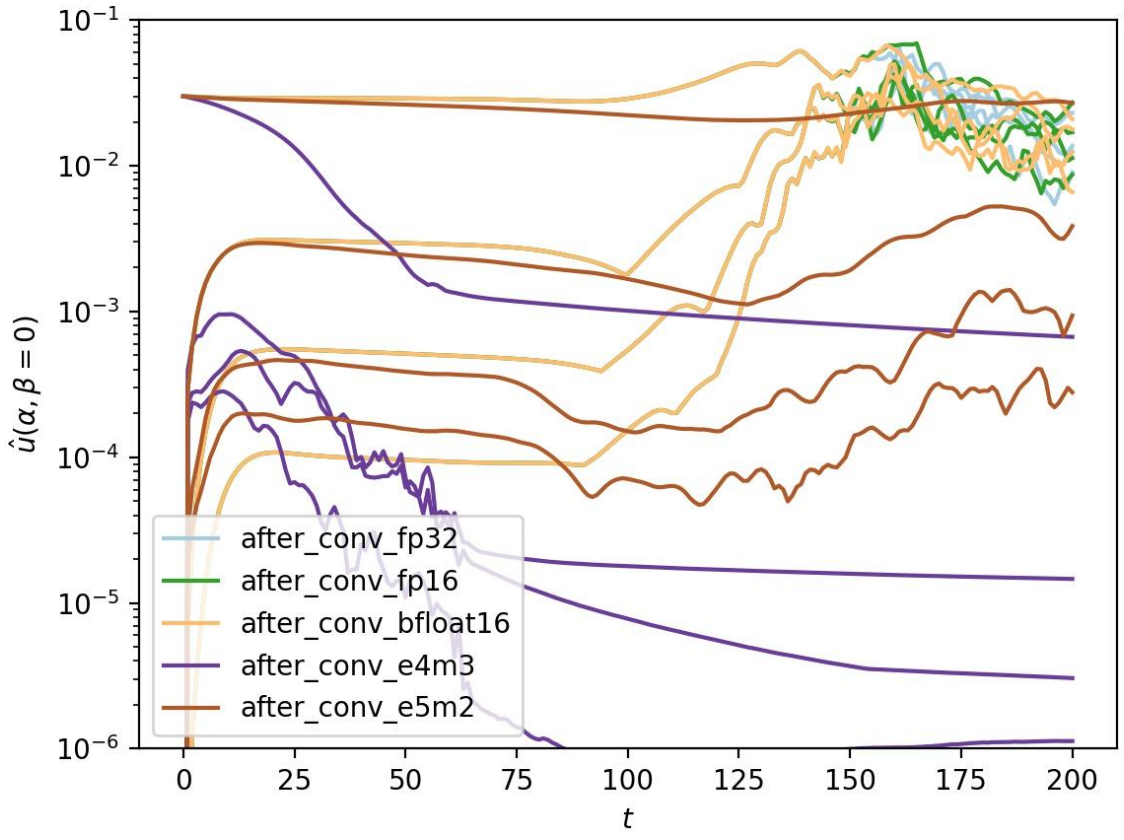
- Let's go to even "worse" precisions



Rounding of convective term works down to bfloat16!

K-type Transition – Neko

- **Growth of beta=0 mode**
  - Verified against SIMSON and Schlatter 2004
- **Single vs double**
- **Round convective term**
- **Geometry is not deformed**
  - Cancellation of error sources
- **Flow separation?**
  - Curved geometry
  - LES/Implicit LES

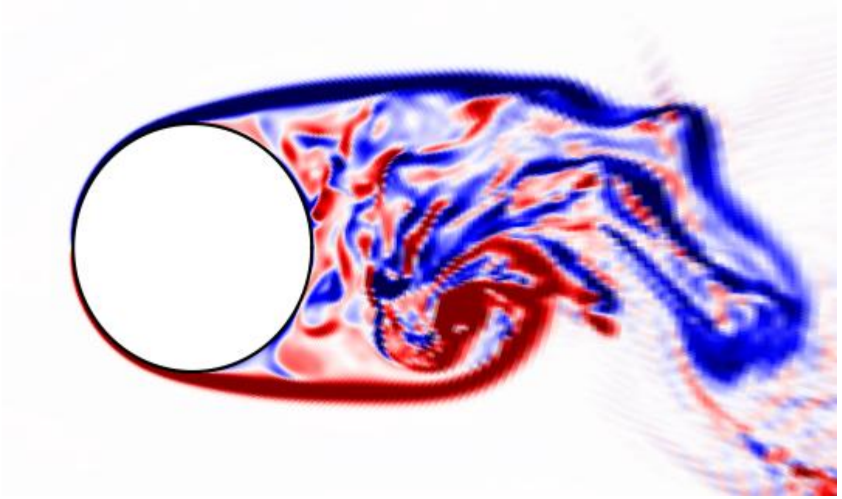


Cylinder at $Re_D = 3900$

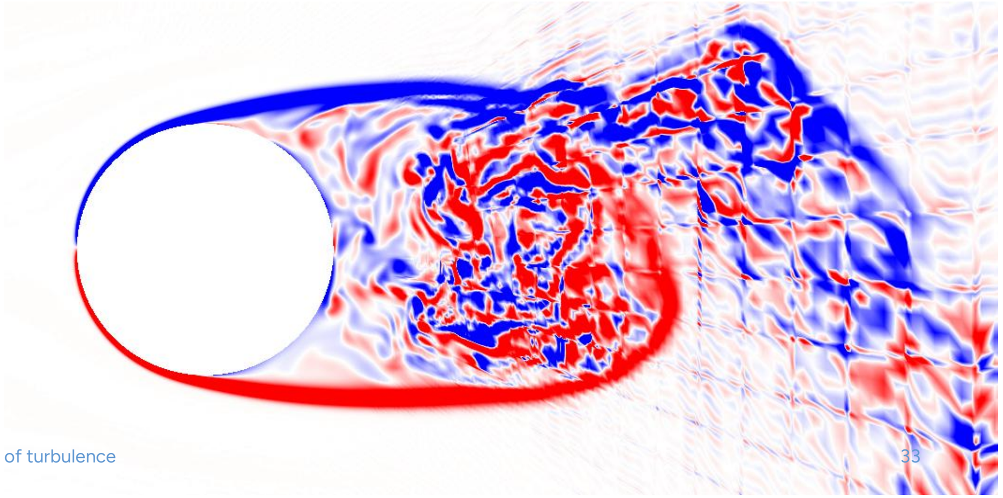
- **Common high-fidelity CFD benchmark**
  - Goes back to Kravchenko and Moin 2000
- **512x512x128 grid points (cylindrical coords)**
- **Stable thin shear layer**
  - Sensitive to numerical error
- **Seen clearly when plotting vorticity**
- **Near-wall quantities?**

FP64 LES with PadeLibs, vorticity

Incoming flow direction

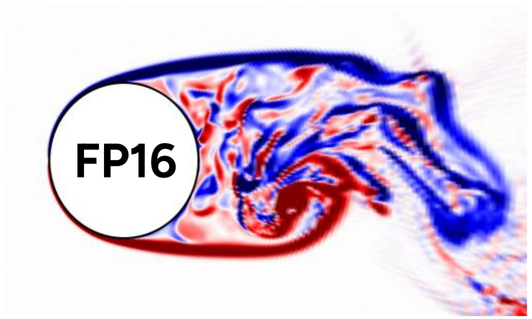


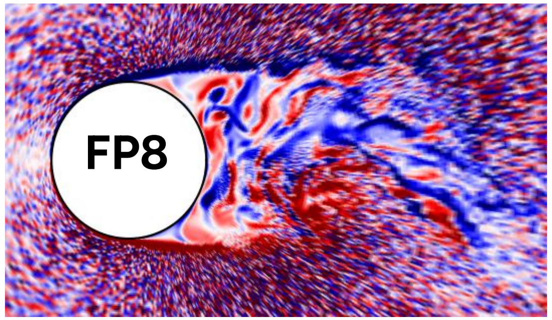
Underresolved simulation with Neko

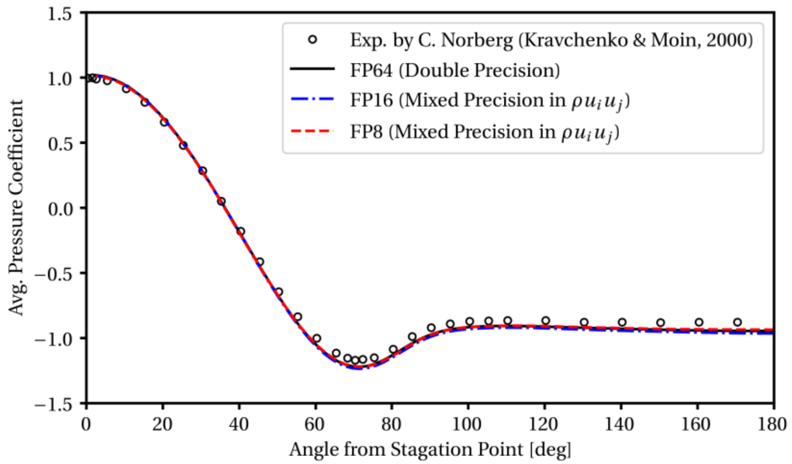


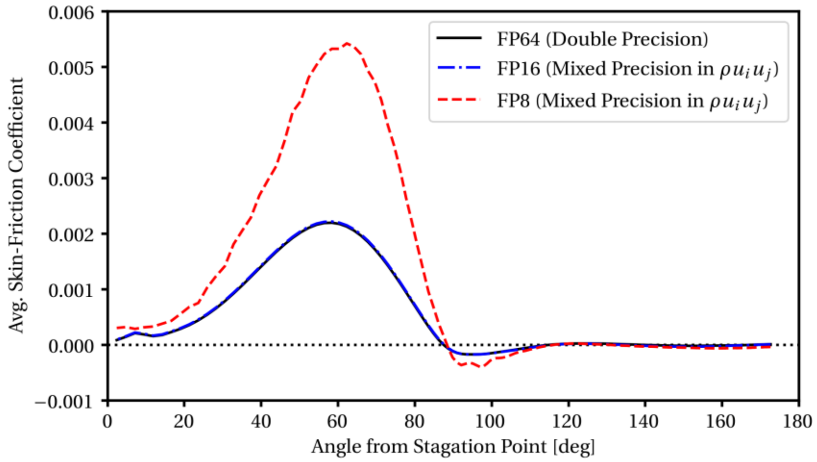
Cylinder at Re_D=3900

- **PadeLibs – Wall quantities**
  - Perturbing convective term
  - (Neko – Single vs double, no observable difference)
    - Time convergence and precision
- **Pressure coefficient and skin-friction coefficient**
  - 8bit significantly alters simulation
  - Separation point is rather stable
- **Instantaneous vorticity**









Performance measurements

- Not the focus, but some interesting results

GPU	β	π_{FP64}	π_{FP32}	FP32 vs FP64
RTX4080	0.72 TB/s	0.76 TFlop/s	48.7 TFlop/s	$2 - 2.5\times$
A40	0.7	0.58	37.4	$2 - 2.5\times$
A100	1.56	9.7	19.4	$1.5 - 2\times$

Table 6: Bandwidth to global memory (DRAM or HBM) β in TB/s, and performance in TFlop/s

- For Neko (GPU): RTX4080 in FP32 matches more or less the A100 in FP64
  → cheaper consumer-grade cards may be used to prototype

- Simson (CPU): Reduction to about 55% when going from double to single.
  → even larger reduction if data transfer could be reduced to half precision.

Channel flow – Remarkably robust

- **Similar trend for all codes**
  - Exact precision when things break differs, but single precision (FP32) without problems
  - FP16 for convection, no difference for any code

- **Large potential for canonical turbulent wall-bounded flows**
  - Speed up DNS, low-fidelity simulations (transients, inflow, ensembles). Potentially also for LES…

- **Statistics not a particularly sensitive measure**
  - Lower precision covers phase-space

- **What about higher Re?**
  - We tested up to Re_τ=2000 in large domains with full single precision, no difference.
  - It must break down at some point (high Re, size etc.) but nothing we have observed.

- **Laminar-Turbulent Transition?** Even that did not show sensitivity.

Summary

- All simulation results (turbulent channel, transition to turbulence, separated flow) remarkably robust with regards to precision
  - Different ways of rounding (state, term or noise)
- Accurate flow simulations with low precision seem to be within reach
  - Large performance/price opportunities on current and upcoming computer architectures
- **For the solvers and cases that we considered**

- **Going forward:**
  - Increasingly utilize lower precision in our simulations (so far only single in hardware)
    - Reduce time to solution, number of core hours and storage
  - Connection to residuals and tolerances? Can we reduce the time to solve for pressure?
  - Carefully evaluate uncertainty of simulations across all solvers and cases
    - Generalizability
    - Perturbing simulation != Actually running in lower precision