and half Single Precision in Earth-System Models

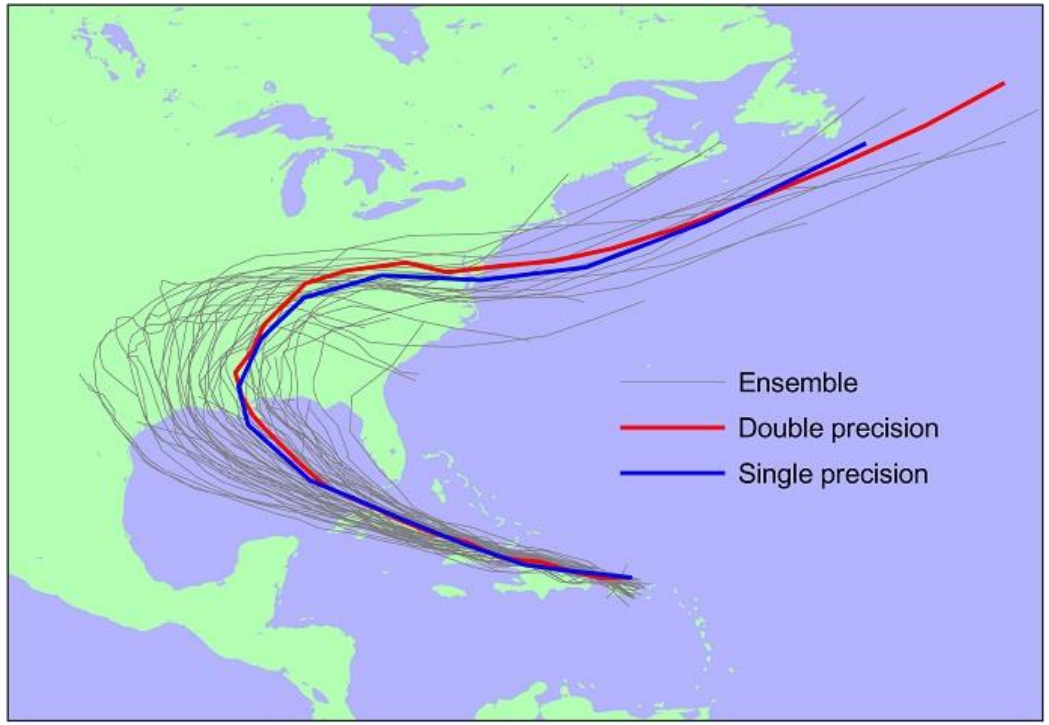
Sam Hatfield, Kristian Mogensen, Peter Dueben, Michail Diamantakis, Nils Wedi

samuel.hatfield@ecmwf.int

With thanks to Seiya Nishizawa and Hirofumi Tomita

ESiWACE2 has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823988.
This research used resources of the Oak Ridge Leadership Computing Facility (OLCF), which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725.

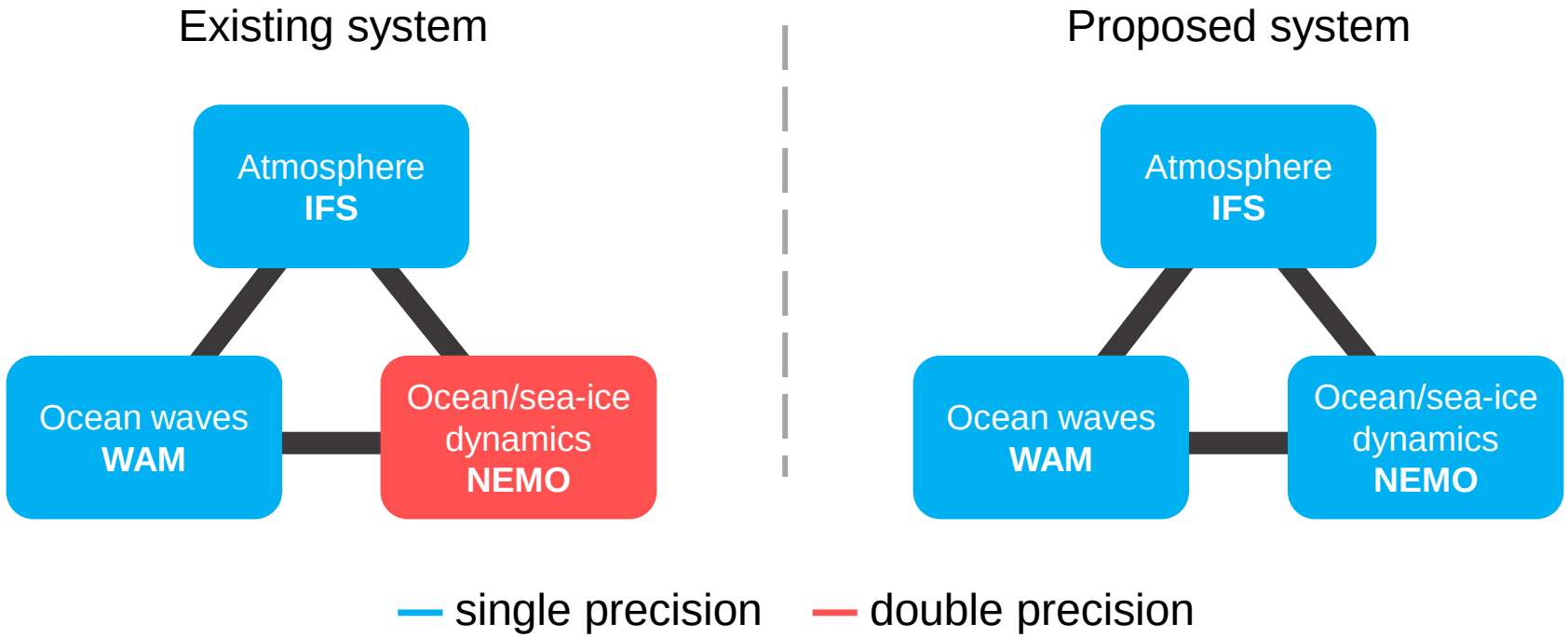
Single-precision weather forecasting at ECMWF



Hurricane Laura forecasts (22/08/2020)

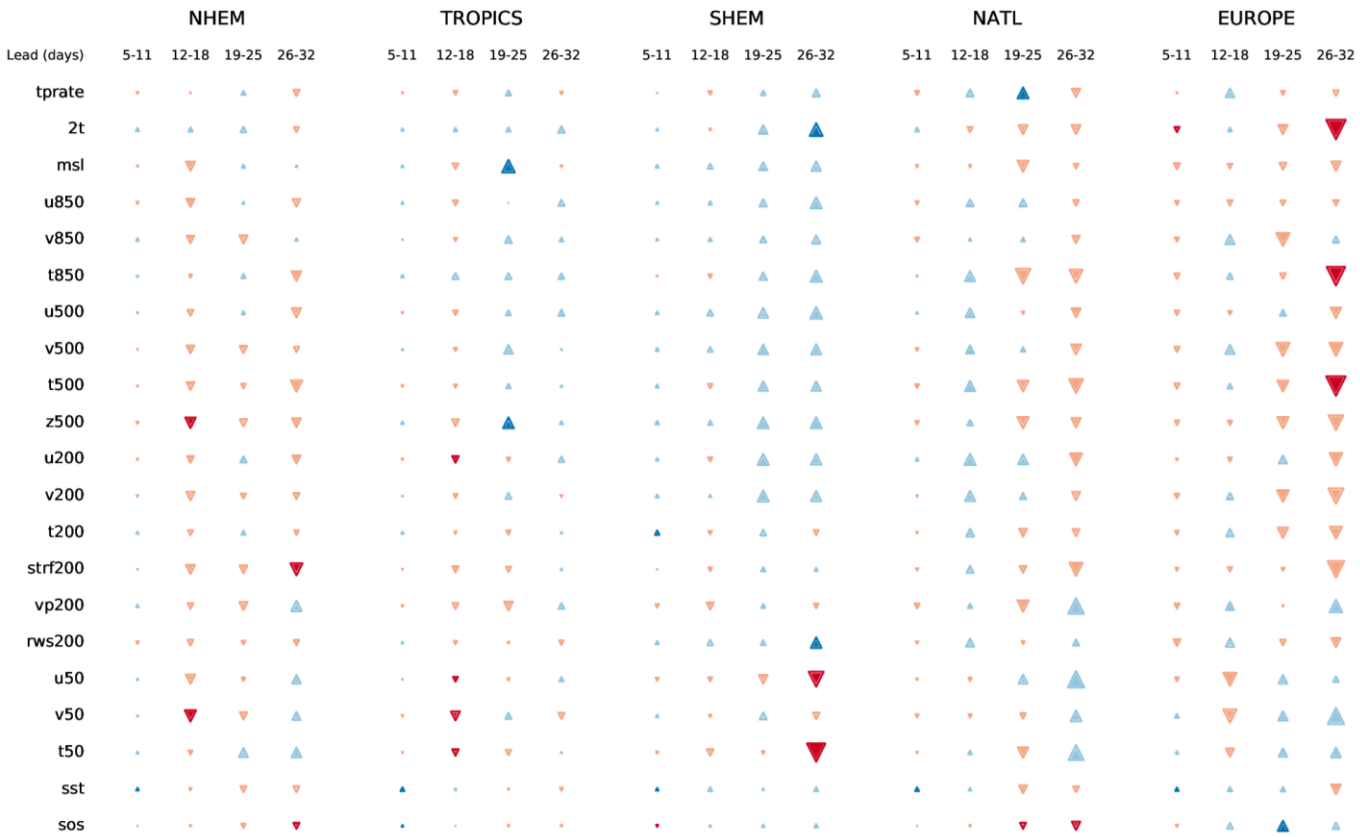
- Single precision now used in the ECMWF atmospheric model (IFS) for **operational weather prediction**
- Negligible impact on forecast scores, but **40% cost saving**
- Precision traded for higher vertical resolution → single precision allowed **more accurate forecasts for no extra cost**
- See Lang et al., *More Accuracy with Less Precision*, QJRMS (submitted)

Towards a fully single-precision Earth-system model



- What about other Earth-system components, e.g. the ocean model (NEMO)?
- Use of single precision in NEMO **nontrivial** (see talk by **Stella Paronuzzi**)

Verification of first fully single-precision coupled forecasts



dCRPSS skill-score chart for fully SP TCO199L91/eORCA1 ensemble hindcasts compared with operational system

- Forecast performance of **fully single-precision** coupled forecasts compared with existing system
- Blue: improvement
- Red: degradation
- **Mostly neutral**

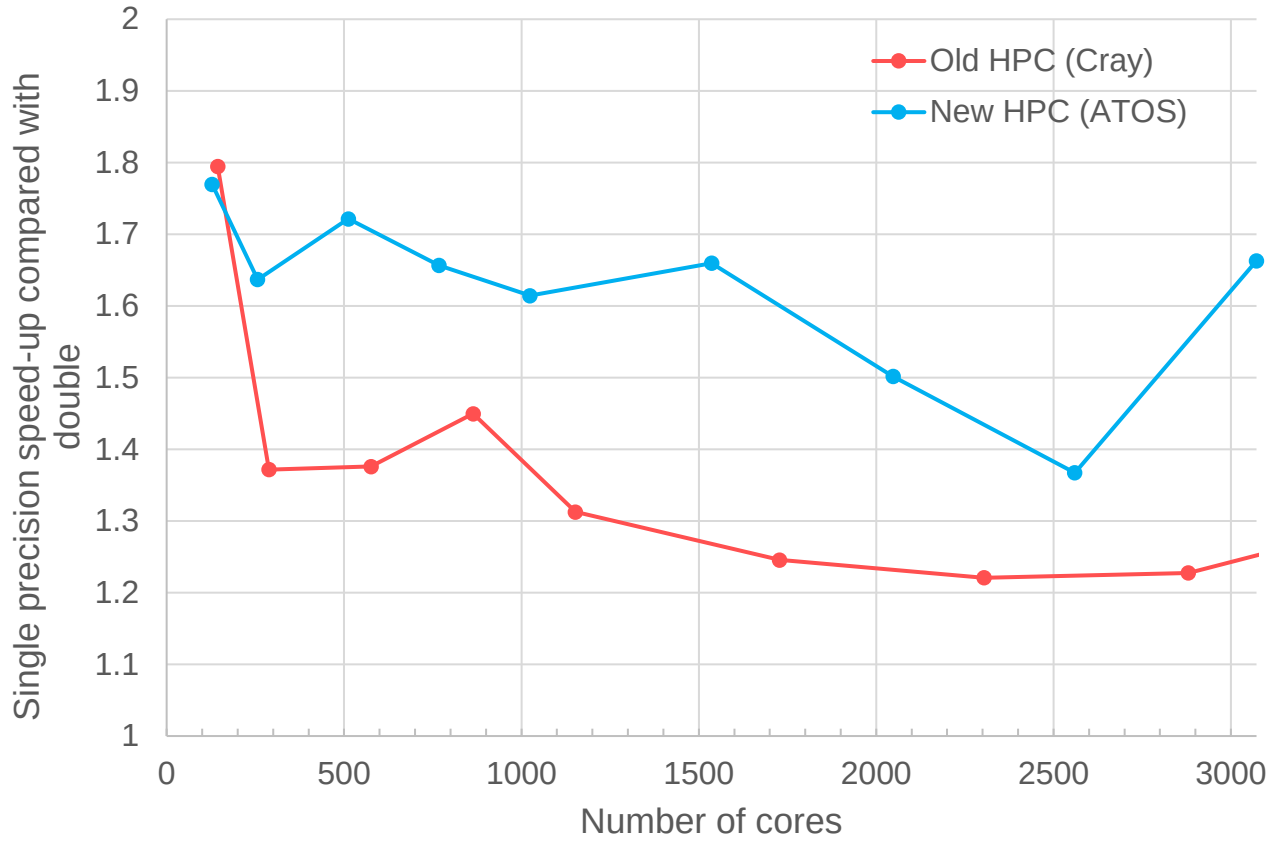
Significant increase

Insignificant increase

Insignificant decrease

Significant decrease

Is NEMO actually faster with single precision?



- Yes, usually.
- Especially on our new supercomputer
- **Up to ~1.75× speed-up possible**
  - Similar to atmosphere
  - Depends on I/O solution

What about half precision?

- Half precision headline feature of top two machines:

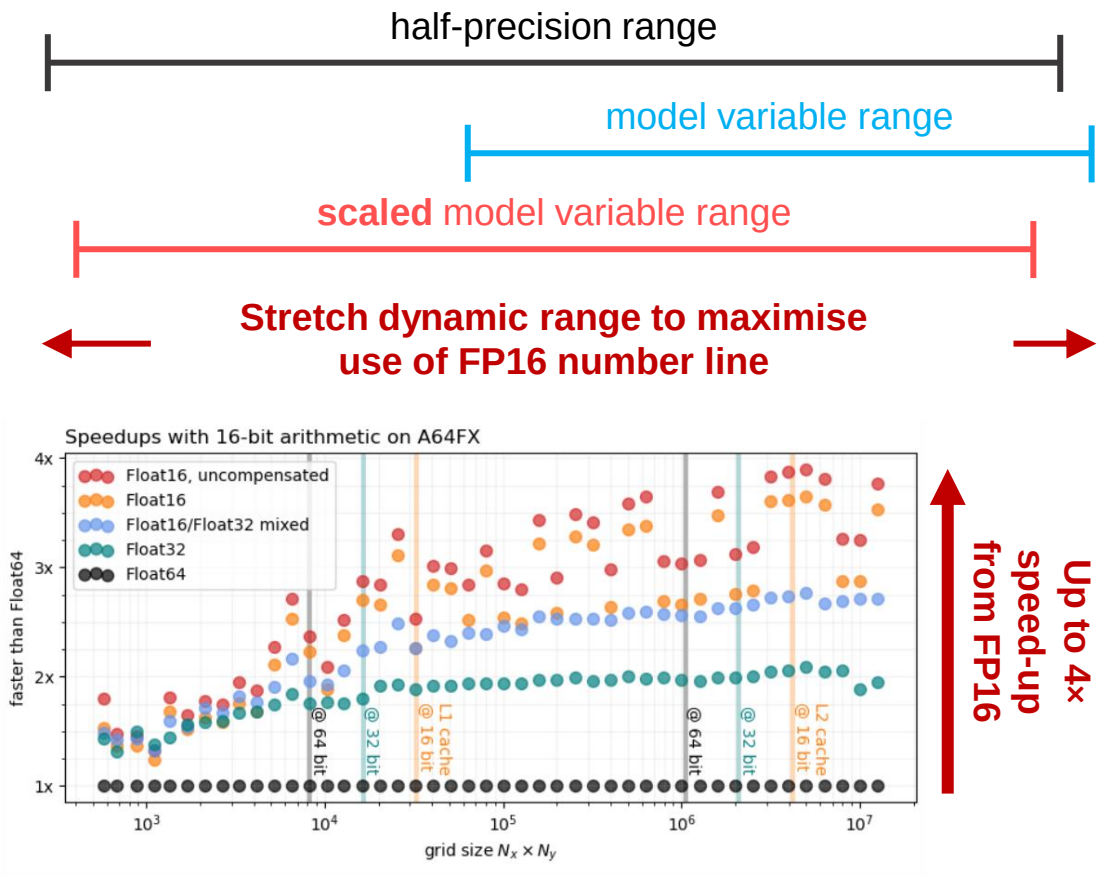
#1: Fugaku

- Fujitsu A64FX (ARM) CPUs allow half-precision operations
- ECMWF access gained through ECMWF/RIKEN memorandum-of-understanding
- (See talks by **Satoshi Matsuoka**, **Toshiyuki Shimizu** and **Yusuke Oishi**)

#2: Summit

- NVIDIA V100 GPUs allow half-precision and some mixed half/single-precision operations (TensorCore)
- ECMWF access gained through INCITE project
- (See talk by **Andreas Mueller**)

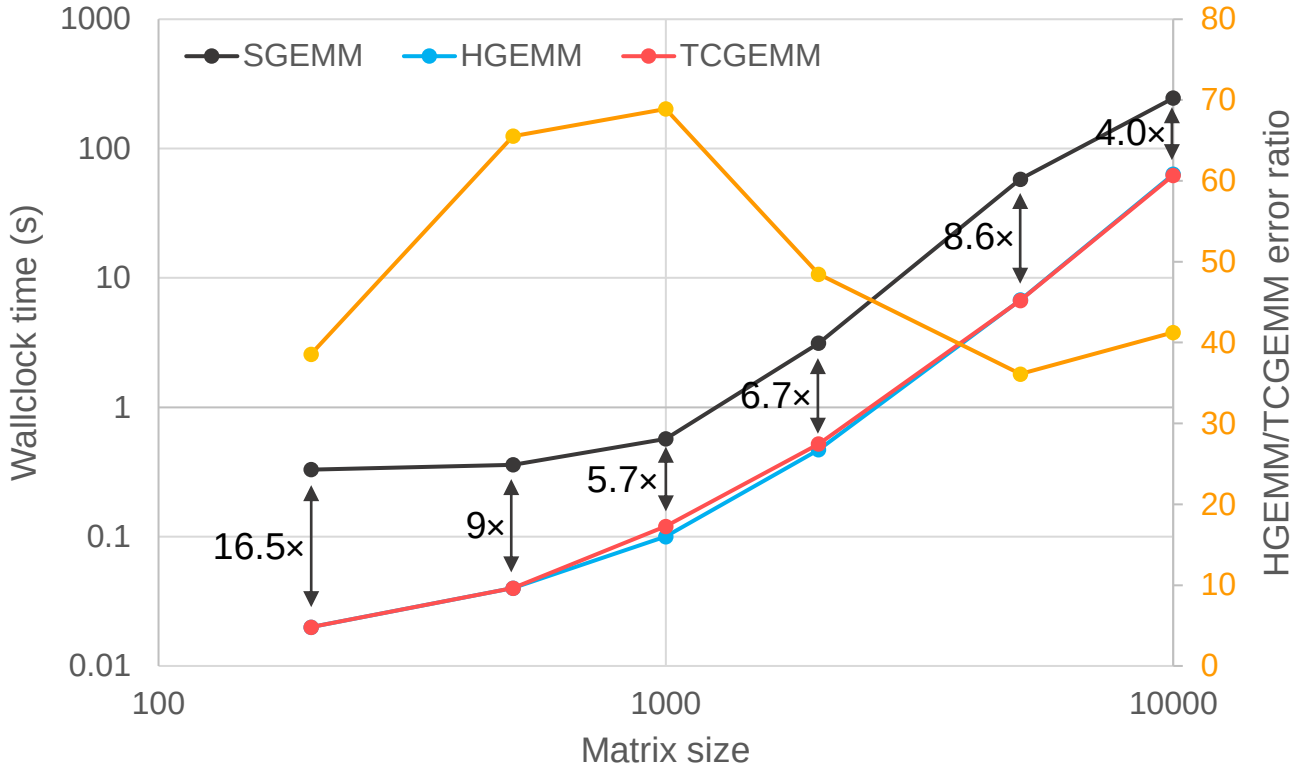
Squeezing/stretching models into half precision



Kloewer et al., https://doi.org/10.1002/essoar.10507472.2

- Half-precision has low dynamic range
- Variables must be **rescaled** to fit
- Cf. work by Milan Kloewer
- Successful **model-wide** use of half precision by rescaling variables
- Up to **~4× speed-up of shallow water model on A64FX** (Isambard 2)
- Probably not possible with complex models *throughout*

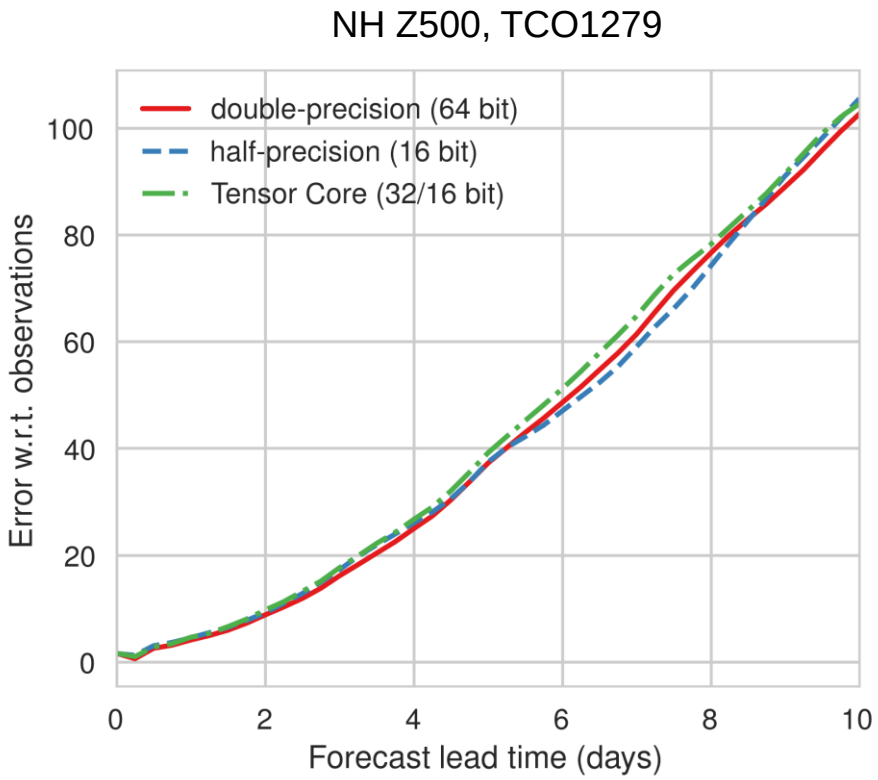
Targeted acceleration of model kernels with half precision



GEMM benchmark on NVIDIA V100
"TC" = TensorCore (mixed half/single)

- Half precision promises up to **order of magnitude** acceleration of certain kernels, **especially GEMM**
- Mixed half/single-precision algorithms (e.g. NVIDIA TensorCore) allow **half-precision speed** with **single-precision accuracy**

Where can we use half precision at ECMWF?



Skill of forecasts using **half-precision Legendre transforms** compared with **double precision**

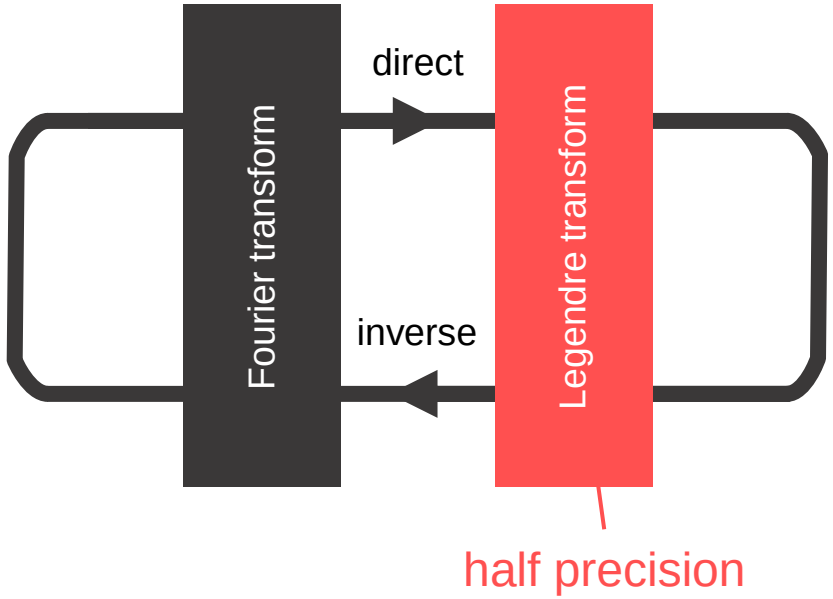
Hatfield et al. 2019, https://doi.org/10.1145/3324989.3325711

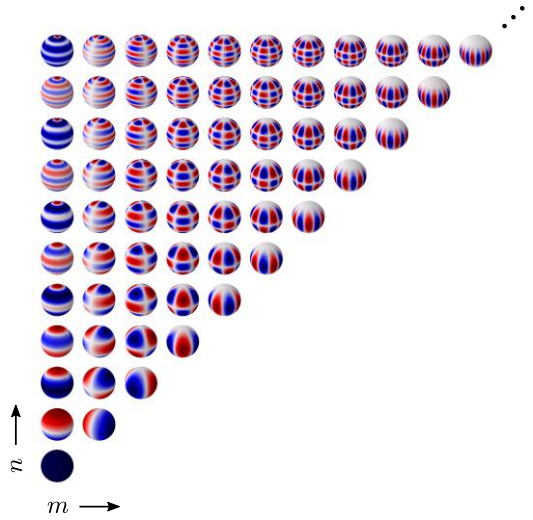
- **Legendre transforms** of the IFS a good target for half precision
  - Bottleneck at high resolution
  - Compact code
  - Algorithmically simple → series of GEMMs
- Preliminary software emulation studies (Hatfield et al. 2019):
  - Half precision can be used in Legendre transforms even up to TCO1279 (9 km globally) resolution
  - Necessary to **rescale** inputs/outputs, as before

Half-precision Legendre transforms on Fugaku

Spectral transform benchmark

grid point space

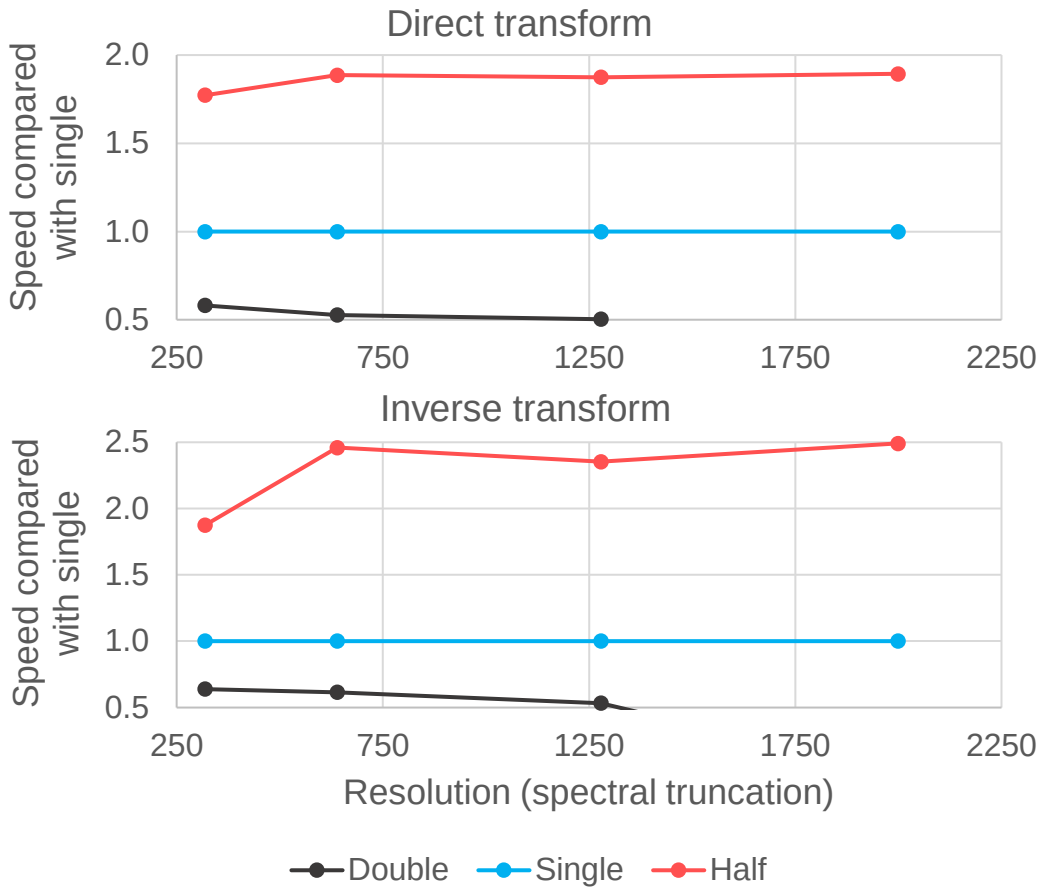




spectral space

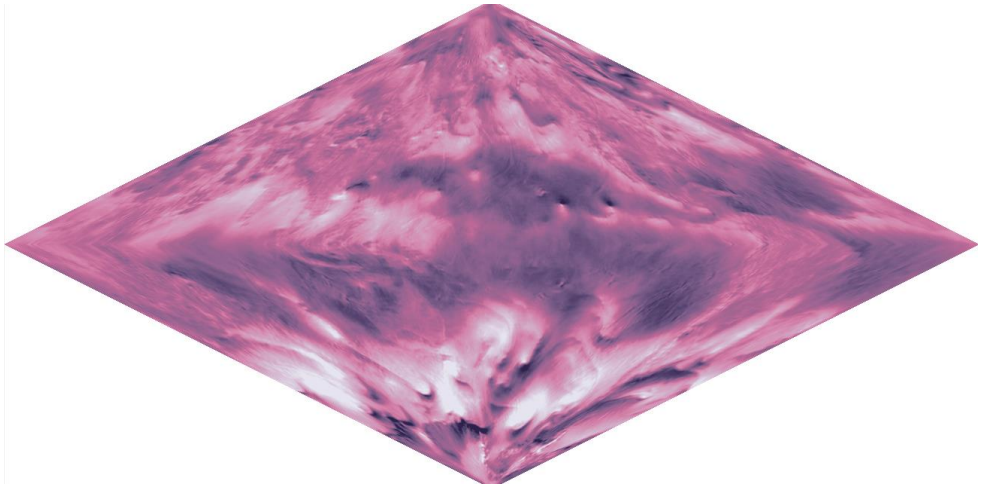
Repeat many times

Another 2× speed-up from half precision?

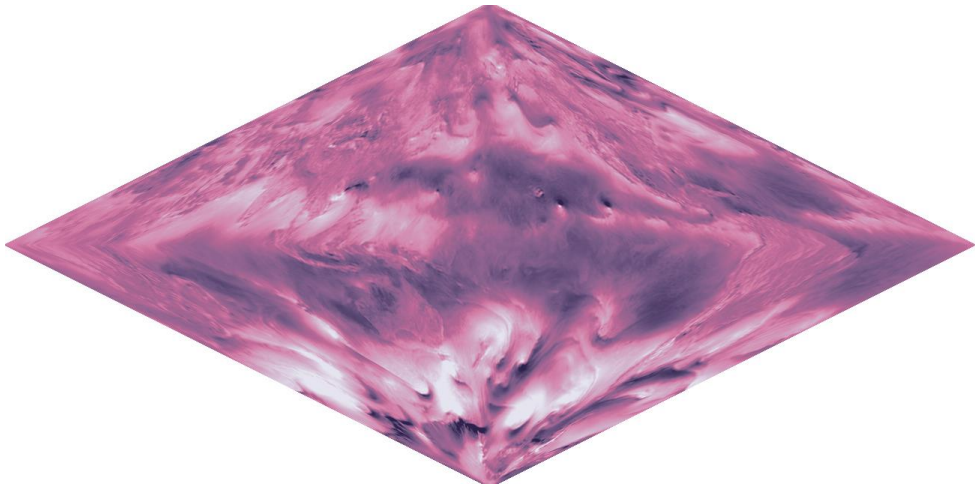


- **Pretty much.**
- Speeds for individual GEMM calls of Legendre transform at TCO1999 resolution, compared with single precision:
  - Inverse (LEINV) antisymmetric (1): 1.29×
  - Inverse (LEINV) symmetric (2): 1.86×
  - Direct (LEDIR) antisymmetric (1): **2.48×**
  - Direct (LEDIR) symmetric (2): **2.49×**
- **However**, transposing one or more matrices kills half-precision speed-up

Spectral transform tests on Fugaku



Half precision



Single precision

Zonal wind field (cubic octahedral grid) after 10 direct-inverse spectral transforms

- Differences of wind between single precision of around 1 m/s after 10 direct-inverse transforms
- Comprehensive meteorological evaluation awaits porting of IFS to Fugaku (in progress)

Conclusion

- Single precision **gives ~1.7× speed-up of weather simulation**
  - Already operational in the atmosphere
  - Under development in the ocean
- **Targeted use of half precision** may provide further acceleration
- But, we must be careful
  - Half precision has low range → rescaling of model variables
- Experiments into spectral transform underway at ECMWF on:
  - Fugaku — native support on CPUs
  - Summit — mixed single/half support on GPUs (TensorCore)