Mixed-Precision Arithmetic in Earth-System Modelling

Sam Hatfield, Peter Düben, Kristian Mogensen, Nils Wedi

samuel.hatfield@ecmwf.int
European Centre for Medium-Range Weather Forecasts
The quiet revolution
Tomorrow’s workhorse supercomputers

Top500 #1 Fugaku, RIKEN R-CCS

Top500 #2 Summit, © ORNL and Carlos Jones
Outline

Part 1
   Floating-point numbers
   Why reduce precision?

Part 2
   Case studies
   Single-precision at ECMWF
Floating-point numbers
Real numbers on computers

Numerical models use **real number arithmetic**
We therefore need a way to map a number from the real number line:

\[
\begin{align*}
\mathbb{R} & \quad \pi \\
-\infty & \quad \cdots & \quad \cdots & \quad \infty
\end{align*}
\]

\[
\begin{align*}
-4 & \quad -3 & \quad -2 & \quad -1 & \quad 0 & \quad 1 & \quad 2 & \quad 3 & \quad 4
\end{align*}
\]

to a finite bitstring:

\[
\pi \approx 010000000000010010010000011111111011 \\
0101010001000100001011010100011000
\]

64 bits

so that we can do that arithmetic on a computer

On a finite computer this is **inherently imperfect**
The obvious way: fixed-point numbers

We can create a crude real number format from an integer simply by placing a “binary point” somewhere, e.g.

10110110 = 182 8 bit integer
10110.110 = 22.75 8 bit fixed-point number (binary point at 5th place)

- Advantages
  - We can reuse the integer arithmetic chip (fast)

- Disadvantages
  - Very low precision (depending on the position on the number line)
  - Limited range
A better way: floating-point numbers

Instead we use **floating-point** numbers:

\[ x = \text{fixed-point number} \times 2^{\text{integer bias}} \]

significand

exponent

e.g. for a 64 bit “float”

\[ \pi = 1.5707963267948966 \times 2^{1024 - 1023} \]
Floating-point standard

Three formats according to IEEE:

- **Double**:
  - Sign: 1 bit
  - Exponent: 52 bits
  - Significand: 11 bits
  - Exponent value: $1010101000100010001010100011000$
  - Value: $3.141592653589793$

- **Single**:
  - Sign: 1 bit
  - Exponent: 23 bits
  - Significand: 8 bits
  - Exponent value: $1001001000011111011$
  - Value: $3.1415927$

- **Half**:
  - Sign: 1 bit
  - Exponent: 10 bits
  - Significand: 5 bits
  - Exponent value: $1001001000$
  - Value: $3.14$
Floating-point properties

You should keep in mind:

- **Machine epsilon**: smallest number that can be added to 1 to produce a different number (Fortran: `EPSILON`)
- **Smallest (non-subnormal) number**: smallest representable number in the “normal” range (Fortran: `TINY`)
- **Largest number**: largest representable number (Fortran: `HUGE`)

<table>
<thead>
<tr>
<th>Precision</th>
<th>double</th>
<th>single</th>
<th>half</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPSILON(x)</td>
<td>2.220446049250313 \times 10^{-16}</td>
<td>1.1920929 \times 10^{-7}</td>
<td>0.000977</td>
</tr>
<tr>
<td>TINY(x)</td>
<td>2.2250738585072014 \times 10^{-308}</td>
<td>1.1754944 \times 10^{-38}</td>
<td>0.00006104</td>
</tr>
<tr>
<td>HUGE(x)</td>
<td>1.7976931348623157 \times 10^{308}</td>
<td>3.4028235 \times 10^{38}</td>
<td>65504 !</td>
</tr>
</tbody>
</table>
Digression: alternatives to floats

Could “posits” replace floats?

[Klöwer et al., 2020]
Why reduce precision?
Initial and model error

Lorenz ‘63 example
A communication/memory-bound world

Basically all models are now **memory-bound**

Reducing precision effectively **increases the communication bandwidth**, thereby accelerating models
Case studies
Emulating reduced-precision computations

Most hardware only supports **double** and **single**-precision arithmetic. How do we assess feasibility of reducing precision without having to port to GPUs etc.? **Software emulation!**

![Binary representation of numbers with different precisions](image)

Downside: we can only estimate computational cost savings

[Dawson and Düben, 2017]
Single-precision in a realistic atmospheric model

Compare double-precision with single-precision in OpenIFS

\[
\text{diff}(\text{single}, \text{double}) \approx \text{std}(\text{ensemble}) < \text{diff}(\text{double member}, \text{double mean})
\]

[Düben and Palmer, 2014]
Scale-selective precision (1)

High-precision for large scales, low-precision for small scales?

[Chantry et al., 2018]
Scale-selective precision (2)

Hurricane Irma core position

Average precision of "scale-selective" (across all wavenumbers): 8.6 significand bits
Reduced-precision Legendre transforms (1)

Profile of IFS at TCo7999 (∼1.5 km)

- Target the spectral transforms to accelerate high-resolution IFS simulations
- GPUs allow half-precision or Tensor Core matrix multiplications

Physical parametrizations
Dynamics
Semi-implicit calculations
Spectral transforms
Reduced-precision Legendre transforms (2)

The largest half-precision number is **65504**
So we must **rescale** variables to avoid overflows

![Variable range diagram with overflow indicator]
Reduced-precision Legendre transforms (2)

The largest half-precision number is \textbf{65504}

So we must \textbf{rescale} variables to avoid overflows

![Diagram showing the process of rescaling and transforming variables.]

- Rescale
- Variable range
- Legendre transform
- Unrescale

0 \quad 65504
Half-precision rounding errors also cause problems for the **largest scale** calculations.

So keep them at double-precision.
Reduced-precision Legendre transforms (4)

- “Model uncertainty” = double-precision with SPPT
- error(half-precision) < model uncertainty if we protect the first $c$ (in this case 10) modes
Reduced-precision Legendre transforms (5)

Ensemble forecasts (25 km)

Deterministic forecasts (9 km)

[Hatfield et al., 2019]
Single-precision at ECMWF
Single-precision ocean and atmosphere

ECMWF’s philosophy on single-precision:

- Same source code for double and single-precision
- Upgrade specific parts to double-precision where necessary
- Roadmap: **fully single-precision coupled forecasts**

```plaintext
REAL(wp) :: x
```

- Single-precision binary
- Double-precision binary
Single-precision in the atmosphere (development issues)

Promote zeroth mode to double-precision:

Change in mean sea-level-pressure error SP vs. DP

EUROPEAN CENTRE FOR MEDIUM-RANGE WEATHER FORECASTS

Sam Hatfield
Single-precision in the atmosphere (development issues)

Promote zeroth mode to double-precision:

Change in mean sea-level-pressure error SP vs. DP
Single-precision in the atmosphere (status)

- Single-precision atmosphere operational default for HRES and ENS (not DA) from Cy47r2 onwards (∼ January 2021)
- \(1.7 \times\) times speed-up compared with double-precision
- Allows free upgrade from 91 to 137 levels → improvement in forecast skill
Next: single-precision in the ocean

- NEMO now coupled to IFS in **all forecast products**
- 20% of total cost of EPS but **60% cost of seasonal system**
- Develop **single-precision** capability in NEMO, building on work at BSC [Tintó Prims et al., 2019]
- Single-precision in sea-ice and icebergs novel

Example bug: orphaned icebergs

Affects mostly single-precision **but also double-precision**
Single-precision in the ocean (status)

- Single-precision NEMO stable at 0.25° resolution with sea-ice and icebergs (operational configuration)
- Speed-up: **up to 1.7×** (depends on I/O)

Change in sea-surface temperature RMSE single vs. double-precision
Single-precision in the ocean (status)

Northern Hemisphere sea-ice concentration bias
Supercomputing is undergoing a **paradigm shift**

**Precision** will become an additional “knob” for **cost/accuracy** trade-off

Earth-System modelling is an ideal application for exploring precision:
- We have **model error** and **initial error**
- Our models are **memory/communication-bound**

Single-precision will be used operationally in the IFS **starting next year**

Single-precision ocean and coupled modelling under testing
References

Scale-Selective Precision for Weather and Climate Forecasting.

Rpe v5: An emulator for reduced floating-point precision in large numerical simulations.

Benchmark Tests for Numerical Weather Forecasts on Inexact Hardware.

Accelerating High-Resolution Weather Models with Deep-Learning Hardware.

Choosing the optimal numerical precision for data assimilation in the presence of model error.
*Journal of Advances in Modeling Earth Systems*.

Weather and climate models in 16-bit arithmetics : Number formats , error mitigation and scope.

How to use mixed precision in Ocean Models.
*Geoscientific Model Development Discussions*, pages 1–21.