NGT - Openlab ”Optimising Floating Point Precision” Workshop

Report of Contributions

https://indico.cern.ch/e/floatworkshop

Contribution ID: **1** Type: **not specified**

Problem statement from experiments and SFT

Contribution ID: **2** Type: **not specified**

Problem statement from Theory Department

Tuesday 1 July 2025 14:40 (10 minutes)

Presenter: FINKENRATH, Jacob

Contribution ID: **3** Type: **not specified**

Discussion - Questions

Tuesday 1 July 2025 14:50 (10 minutes)

Contribution ID: **4** Type: **not specified**

Floating Point Emulation in NVDIA Math Libraries

Tuesday 1 July 2025 15:30 (45 minutes)

The trends in computer architecture, primarily driven by AI-based applications (most recently, large language models), has led to a rapid increase in the reduced- and mixed-precision computing capabilities of GPUs. These processors demonstrate an outsized power-efficiency (FLOPS/watt) advantage over systems almost exclusively focused upon native single- and double-precision arithmetic. Thus, there is a great deal of motivation to leverage these capabilities, through the use of various mixed-precision algorithms and emulation techniques, to facilitate greater scientific computing throughput without sacrificing accuracy. We'll touch upon a number of these approaches and present real-world case studies that provide compelling evidence in support of this path to increasing the science per watt of supercomputers.

Author: RODRIGUEZ, Samuel (NVidia)

Presenter: RODRIGUEZ, Samuel (NVidia)

Contribution ID: **5** Type: **not specified**

Advancing AI with AMD: Open Source, Sovereign Innovation, and the Latest in CPU & GPU Performance

Tuesday 1 July 2025 16:15 (30 minutes)

Authors: ROSKOWETZ, Joerg; MAUDODI, Sayed

Presenters: ROSKOWETZ, Joerg; MAUDODI, Sayed

Contribution ID: **6** Type: **not specified**

Extended Precision in Convex Optimisation

Tuesday 1 July 2025 17:05 (20 minutes)

Semidefinite Programming is a matrix-form generalisation of linear programming, and is typically tackled using Interior Point Methods. These methods are of iterative nature and at each step, a matrix inversion needs to be performed. For small or sparse matrices, direct methods like sparse Cholesky factorisation are used. For dense matrices of larger size, like the ones that arise in convex relaxations of combinatorial problems, Krylov methods like Conjugate Gradient seem a better approach. We show how, as the dual-primal central trajectory approaches the feasible set and the tentative solution becomes rank-deficient, increasing the precision accelerates the convergence (in terms of number of CG iterations).

Presenter: HERRERA-MARTI, David (CEA France)

Contribution ID: **7** Type: **not specified**

Discussion - Questions

Tuesday 1 July 2025 17:25 (20 minutes)

Contribution ID: **8** Type: **not specified**

Using physics knowledge to improve numerical stability

Wednesday 2 July 2025 09:00 (30 minutes)

The numerically stable evaluation of scattering matrix elements near the infrared limit of gauge theories is of great importance for the success of collider physics experiments. We present a novel algorithm that utilizes double precision arithmetic and reaches higher precision than a naive quadruple precision implementation at smaller computational cost. The method is based on physics-driven modifications to propagators, vertices and external polarizations. [https://arxiv.org/abs/2406.07671]

Authors: E. Bothmann (speaker), J. M. Campbell, S. Höche, M. Knobbe

Presenter: BOTHMANN, Enrico (CERN)

Contribution ID: **9** Type: **not specified**

Double-double for virtual amplitude evaluation

Wednesday 2 July 2025 09:30 (30 minutes)

Two-loop virtual amplitudes are one of the key ingredients of an NNLO cross-section calculation. In this talk I would like to describe the precision requirements of evaluating such amplitudes via sector decomposition and quasi-Monte Carlo integration, and to report on satisfying them using double-double floating point number implementation within pySecDec on CPU and GPU.

Based on: 2402.03301, 2305.19768, and related work.

Presenter: MAGERYA, Vitaly (CERN)

Contribution ID: **10** Type: **not specified**

An overview of mixed precision strategies for scientific computing

Wednesday 2 July 2025 10:00 (30 minutes)

The increasing support of lower precision arithmetics in hardware provides new opportunities for high performance scientific computing. However, even though low precision arithmetics can provide significant speed, communication, and energy benefits, their use in scientific computing poses the challenge of preserving the accuracy and stability of the computation. To address this issue, a variety of mixed precision algorithms that combine low and high precisions have emerged. In this talk I will give an overview of mixed precision algorithms in numerical linear algebra, with a focus on recent advances to accelerate the solution of linear systems.

Presenter: MARY, Theo (Computer Lab of Paris 6 (Lip6))

Contribution ID: **11** Type: **not specified**

TNL: Numerical Library for Modern Parallel Architectures

Wednesday 2 July 2025 12:05 (15 minutes)

TNL (www.tnl-project.org) is a collection of building blocks that facilitate the development of efficient numerical solvers and HPC algorithms. It is implemented in C++ using modern programming paradigms in order to provide a flexible and user-friendly interface similar to, for example, the C++ Standard Template Library. TNL provides native support for modern hardware architectures such as multicore CPUs, GPUs, and distributed systems, which can be managed via a unified interface. In our presentation, we will demonstrate the main features of the library together with efficiency of the implemented algorithms and data structures.

Presenter: OBERHUBER, Thomas (Czech Technical University in Prague)

Contribution ID: **12** Type: **not specified**

VXP: Extended Precision Accelerator

Tuesday 1 July 2025 16:45 (20 minutes)

in this talk, we propose a RISC-V-based accelerator aimed at extended precision computing for scientific computing applications. Furthermore, we show how it can help improving convergence of iterative solvers in real use cases. Lastly, we present details about our hardware implementations and results obtained on real silicon prototypes.

Presenter: GUTHMULLER, Eric (CEA France)

Contribution ID: **13** Type: **not specified**

Discussion - Questions

Wednesday 2 July 2025 12:35 (30 minutes)

Contribution ID: **14** Type: **not specified**

Floating-Point Error Estimation Using Automatic Differentiation

Wednesday 2 July 2025 14:20 (30 minutes)

Floating-point errors highlight the inherent limitations of finite-precision computing, and if left unaddressed, they can lead to severe consequences. In high-precision applications, accurately quantifying these uncertainties is essential. Various approaches have been explored to tackle floating-point errors, including increasing numerical precision, employing compensation algorithms, and applying both statistical and non-statistical estimation techniques. One widely used method for dynamic error estimation is Automatic Differentiation (AD). However, current AD-based tools often require manual code annotations or modifications. Additionally, AD tools based on operator overloading typically necessitate repeated gradient computations across different inputs and inherit the inefficiencies of the operator overloading approach.

In this work, we introduce a customizable approach for leveraging AD to automatically generate source code that estimates floating-point uncertainties in C/C++ applications using Clad. Our framework, CHEF-FP supports automatic error annotation and allows integration with user-defined error models. We also share our progress in extending this approach to GPU-based applications.

Presenter: VASILEV, Vassil (Princeton University)

Contribution ID: **15** Type: **not specified**

Precision auto-tuning and control of accuracy in high performance simulations

Wednesday 2 July 2025 14:50 (30 minutes)

In the context of high performance computing, new architectures, becoming more and more parallel, offer higher floating-point computing power. Thus, the size of the problems considered (and with it, the number of operations) increases, becoming a possible cause for increased uncertainty. As such, estimating the reliability of a result at a reasonable cost is of major importance for numerical software. In this talk we present an overview of different approaches for accuracy analysis (guaranteed or probabilistic ones) and the related software. We also describe methods to improve the results accuracy. We present the principles of Discrete Stochastic Arithmetic (DSA) that enables one to estimate rounding errors in simulation codes. DSA can be used to control the accuracy of programs in half, single, or double precision via the CADNA library, and also in arbitrary precision via the SAM library. Thanks to DSA, the accuracy estimation and the detection of numerical instabilities can be performed in parallel codes on CPU and on GPU. Most numerical simulations are performed in double precision, and this can be costly in terms of computing time, memory transfer and energy consumption. We present tools for floating-point auto-tuning that aim at reducing the numerical formats used in simulation programs.

Presenter: JÉZÉQUEL, Fabienne (LIP6, Sorbonne Université)

Contribution ID: **16** Type: **not specified**

Emulating Matrix Multiplication Using Mixed-Precision Computation

Wednesday 2 July 2025 16:20 (30 minutes)

This talk introduces a method for emulating matrix multiplication through mixed-precision computation. As exemplified by the Matrix Engine on GPUs, low-precision arithmetic can be performed significantly faster than conventional FP32 or FP64 operations. We present Ozaki Scheme I and II, which leverage low-precision arithmetic to achieve accuracy comparable to standard FP64, and discuss their numerical performance.

Presenter: OZAKI, Katsuhisa (Shibaura Institute of Technology)

Contribution ID: **17** Type: **not specified**

Discussion - Questions

Wednesday 2 July 2025 10:50 (15 minutes)

Contribution ID: **18** Type: **not specified**

Mixed precision ab initio tensor network state methods adapted for NVIDIA. Blackwell technology via emulated FP64 arithmetic

Wednesday 2 July 2025 11:35 (30 minutes)

An overview of recent advances in tensor network state (TNS) methods are presented that have the potential to broaden their scope of application radically for strongly correlated quantum many body systems. Novel mathematical models for hybrid multiNode-multiGPU parallelization on high-performance computing (HPC) infrastructures will be discussed. Scaling analysis on NVIDIA DGX-A100 and DXG-H100 platforms reaching quarter petaflops performance on a single node will also be presented. Finally, we discuss cutting edge performance results via mixed precision spin adapted ab initio Density Matrix Renormalization Group (DMRG) electronic structure calculations utilizing the Ozaki scheme for emulating FP64 arithmetic using 8-bit integer logic. By approximating the underlying matrix and tensor algebra via finite number of INT8 slices we demonstrate for chemical benchmark systems that chemical accuracy can be reached even with mixed precision arithmetic. We also show that due to its variational nature, DMRG provides an ideal tool to benchmark accuracy domains and performance of new hardware developments and related numerical libraries. Detailed numerical error analysis and performance assessment are presented also for subcomponents of the DMRG algebra by interpolating systematically between double and single precision. Our analysis paves the way for utilization of state-of-the-art Blackwell technology in tree-like tensor network state calculations opening new research directions in material sciences and beyond.

Presenter: LEGEZA, Ors (Wigner Research Centre for Physics, Hungary)

Contribution ID: **19** Type: **not specified**

Experiences with CADNA and the Madgraph5 Event Generator

Wednesday 2 July 2025 15:20 (30 minutes)

This talk presents a summer student project that explored the numerical stability of MadGraph5 using CADNA. It focuses on how CADNA's warning system and its ability to quantify floating-point precision were used to assess whether MadGraph5 can operate reliably with single-precision floating-point numbers.

Author: HAGEBOECK, Stephan (CERN)

Presenter: HAGEBOECK, Stephan (CERN)

Contribution ID: **20** Type: **not specified**

Discussion - Questions

Wednesday 2 July 2025 16:50 (20 minutes)

Contribution ID: **21** Type: **not specified**

Closing Session

Contribution ID: **22** Type: **not specified**

Closing Session

Wednesday 2 July 2025 17:10 (10 minutes)

Contribution ID: **23** Type: **not specified**

Float32 Expansions –A Possible Answer for Scientific Computing in the Era of AI-Driven GPU Development

Wednesday 2 July 2025 12:20 (15 minutes)

In recent years, the emergence of large language models has led GPU vendors to prioritize performance improvements for lower-precision arithmetic, often at the expense of continued development for Float64. Meanwhile, scientific computing has increasingly relied on GPGPU acceleration, where double precision is still essential. Multi-word expansions for single-precision floating point numbers may offer a viable alternative—providing comparable or even superior precision while achieving better performance than native double precision. In this talk, we will present results using a CUDA-enabled, templated, and ported version of the QD library within the TNL framework, applied to existing numerical algorithms.

Presenter: STLOUKAL, František (Czech Technical University in Prague)

Contribution ID: **24** Type: **not specified**

Problem statement from experiments and SFT

Tuesday 1 July 2025 14:10 (30 minutes)

Author: INNOCENTE, Vincenzo (CERN)

Presenter: INNOCENTE, Vincenzo (CERN)

Contribution ID: **25** Type: **not specified**

Welcome Session

Tuesday 1 July 2025 14:00 (10 minutes)

Author: ROISER, Stefan (CERN)

Presenter: ROISER, Stefan (CERN)

Contribution ID: **26** Type: **not specified**

Adaptive Floating-Point Quantization for Efficient Neural Networks

Wednesday 2 July 2025 10:30 (20 minutes)

The rapid growth of deep learning models, particularly Large Language Models (LLMs), which have increased their parameter counts nearly tenfold annually since 2018, has intensified the need for more efficient, power-aware deployment strategies. Quantization is a widely adopted technique for reducing the computational and memory footprint of neural networks by lowering numerical precision.

This work investigates a floating-point quantization approach to adaptively reduce bitwidths for weights and activations while preserving model accuracy. A quantization-oriented methodology is presented, which analyzes the distribution of tensor values to guide the design of custom floating-point formats. Experimental results on Recurrent Neural Networks demonstrate that this approach achieves an average 3.5× reduction in bit usage, with only a 0.5% drop in top-1 accuracy, using quantization-aware training (QAT).

Building on this work, a follow-up contribution extended the AMD/Xilinx deployment flow by enabling support for arbitrary floating-point in the Quantized Neural Network format QONNX, complementing the existing support in the QAT library Brevitas and completing the quantization path toward hardware acceleration with the AMD FPGA NN library FINN.

Presenter: GHIELMETTI, Nicolo (CERN)