Program

Monday

Time	Event
7:30-8:45	Breakfast
9:00-10:15	Participant introduction (Room N101)
10:15-10:45	Coffee and Tea
10:45-12:00	Participant introduction (Room N101)
12:15-13:00	Lunch
13:00-13:30	Participant introduction (Room N101)
13:45-14:45	Plenary talk: James Demmel
14:45-15:15	Discussions about breakout session logistics
15:15-15:45	Coffee and Cake
16:00-17:00	Breakout session focusing on an introduction to the topic
17:00-17:30	Breakout session structure debrief
18:00-	Dinner

Plenary talk

Speaker: James Demmel

Title: Portability with Low and Mixed Precision Arithmetic

Abstract: The large design space of different low precision arithmetics, and algorithms that use them, naturally leads to divergent answers, and the desire from users for reliable numerical guarantees. We describe 3 efforts in this direction, and ongoing challenges:

1. The P3109 Working Group is developing an IEEE Standard for Floating Point for Machine Learning. Originally conceived of as an 8-bit floating point standard, the advantages of using different formats for different ML problems has led it to grow to include many different numbers of mantissa bits, numbers

of exponent bits, rounding modes (including 3 kinds of stochastic rounding), signedness or not, and other features. We will give an update on the latest public design document, input from the public is welcome.

2. The availability of GEMM accelerators for some of these low precisions has motivated vendors to provide DGEMM implementations that use these accelerators to emulate FP64 arithmetic much faster, but potentially with compromises on accuracy. We have proposed a way to “Grade the BLAS”, i.e. a set of tests that can be used to say what error bound a BLAS implementation provides. Grades range from an “A” (for a conventional O(n^3) FP64 implementation) to a “C” (for a Strassen-like implementation). We describe how these grades influence the error analysis of common algorithms for solving Ax=b and least squares problems. One challenge is making these tests “ungamable”, so that a vendor cannot modify their proprietary (secret) BLAS implementation to pass our publicly available tests, without always guaranteeing the reported accuracy.
3. Another aspect of reliable numerics is “consistent” exception handling, i.e. dealing with Infs and NaNs. There are historical examples of plane and car crashes caused by mishandling exceptions. We propose a definition of “consistency” and describe ongoing efforts to provide a version of the BLAS, LAPACK and potentially other libraries that satisfy this definition.

Tuesday

Time	Event
7:30-8:45	Breakfast
9:00-9:15	Meet in the main lecture room for any announcements
9:15-10:30	Breakout session focusing on the state-of-the-art
10:30-11:00	Coffee and Tea
11:00-12:00	Breakout session focusing on the state-of-the-art
12:15-13:00	Lunch
13:15-14:15	Plenary talk: Theo Mary
14:15-15:15	Breakout session focusing on the state-of-the-art
15:15-15:45	Coffee and Cake
15:45-16:30	Breakout session focusing on the state-of-the-art
16:30-17:30	Meet in the main lecture room to summarise discussions
18:00-	Dinner

Plenary talk

Speaker: Theo Mary

Title: Beyond mixed precision: a guide to adaptive precision algorithms

Abstract:

This talk aims at providing a short guide to designing, developing, and analyzing adaptive precision algorithms. It is composed of two parts.

In the first part, I will explain the main principles of adaptive precision algorithms, and what makes them more powerful and expressive than traditional mixed precision ones. By the latter, I refer to algorithms that exploit a small number of precisions that are statically fixed in advance. In contrast, adaptive precision algorithms:

- can exploit a continuum of precisions, either in hardware when such a continuum is available (e.g., GPUs), or in software when such a continuum can be efficiently used (e.g., memory accessors);
- can deliver a continuum of accuracies, offering flexible performance--accuracy tradeoffs;

- dynamically adapt their behavior (choice of precisions) at runtime, depending on the data (both input and intermediate data arising during the computation);
- are provably accurate and robust by basing their strategy on rigorous error bounds derived analytically.

In the second part, I will illustrate these principles with various successful examples in linear algebra and beyond, including Krylov methods, preconditioners, matrix factorizations, matrix multiplication, low-rank approximations, randomized algorithms, tensors, neural networks, and nonlinear equations.

Wednesday

Time	Event
7:30-8:45	Breakfast
9:00-9:30	Meet in the main lecture room for any announcements plus group photo
9:30-10:30	Breakout session focusing on future directions
10:30-11:00	Coffee and Tea
11:00-12:00	Plenary talk: John Shalf
12:15-13:00	Lunch
13:15-17:30	Excursion
18:00-	Dinner

Plenary talk

Speaker: John Shalf

Title: A computer architects view of mixed and low precision arithmetic

Abstract: The end of Dennard scaling and the rapid growth in data rates from scientific instruments have fundamentally shifted the performance and energy balance of high-performance computing systems. Today, data movement—not arithmetic—dominates both energy consumption and time-to-solution. From a computer architect's perspective, this crisis calls for a rethinking of numerical precision, dataflow, and system organization rather than incremental improvements to traditional floating-point throughput. This talk examines low-precision and mixed-precision arithmetic as key architectural levers for restoring energy efficiency and performance scalability in future HPC systems. By tailoring numerical precision to algorithmic tolerance and application semantics, low-precision representations can dramatically reduce data movement, storage, and interconnect energy while enabling higher effective compute density. Mixed-precision approaches further allow systems to combine aggressive low-precision computation with selective high-precision refinement, achieving accuracy where it matters without paying its full cost everywhere.

Thursday

Time	Event
7:30-8:45	Breakfast
9:00-9:15	Meet in the main lecture room for any announcements
9:15-10:30	Breakout session focusing on future directions
10:30-11:00	Coffee and Tea
11:00-12:00	Breakout session focusing on future directions
12:15-13:00	Lunch
13:15-14:15	Plenary talk: Sherry Li
14:15-15:15	Breakout session focusing on future directions
15:15-15:45	Coffee and Cake
15:45-16:15	Breakout session focusing on future directions
16:15-17:30	Meet in the main lecture room to summarise discussions
18:00-	Dinner

Plenary talk

Speaker: Sherry Li

Title: Benefits and Obstacles of Mixed Precision Adoption in Exascale Math Libraries and Applications

Abstract:

In the exascale computing era GPUs have become mainstream computing engines. Given the significant speed advantage of lower precisions on GPUs, exploiting this hardware feature becomes important at all levels of the software stack. We will discuss recent mixed precision effort through the US Exascale Computing Project. The focus will be on the experience of the math libraries and
applications developers. We will present the usage scenarios, challenges and requirements from DOE's major math libraries that are using or will use reduced/mixed precision. We will also present the "wishlist" from the developers of what tools and vendor support could be useful to lower the barrier of adopting mixed precision in large software products.

Friday

Time	Event
7:30-8:45	Breakfast
9:00-9:15	Meet in the main lecture room for any announcements
9:15-10:30	Breakout session - Finalise discussions and tidy documentation
10:30-11:00	Coffee and Tea
11:00-12:00	Breakout session - Finalise discussions and tidy documentation
12:15-13:00	Lunch
13:00-	Workshop ends