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 12:00 (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