Sustainable and Robust Linear Algebra Computations: A Brief Overview of European Efforts

Roman Iakymchuk

Sorbonne Université / Fraunhofer ITWM
roman.iakymchuk@sorbonne-universite.fr
roman.iakymchuk@itwm.fraunhofer.de

SIAM LA
May 20th, 2021

Sustainable and Robust Linear Algebra Computations at Exascale

MS56		
09:35AM	Roman Iakymchuk	Sustainable and Robust Linear Algebra Computations: A Brief Overview of European Efforts
09:55AM	Toshiyuki Imamura	Riken's Effort for Sustainable Linear Algebra Computations
10:15AM	Tiago Trevisan Jost	Fast Exploration of Variable Precision Linear Kernels
10:35AM	Fabienne Jézéquel	Precision Auto-Tuning and Control of Accuracy in High Performance Simulations

MS63		
11:15AM	Takeshi Ogita	Iterative Refinement and Verified Numerical Linear Algebra
11:35AM	Theo Mary	Multiple Word Arithmetic with GPU Tensor Cores: Theory and Practice
11:55AM	Takeshi Fukaya	Exploiting Lower Precision Computing in the GMRES(m) Method
12:15PM	Emmanuel Agullo	Variable Accuracy Storage through Lossy Compression Techniques in Numerical Linear Algebra

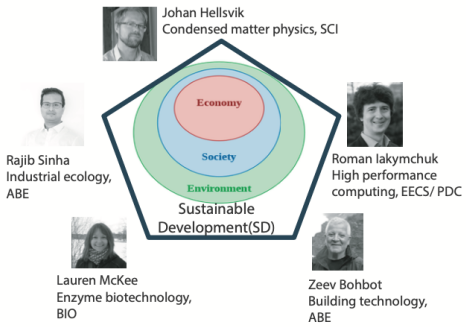
Outline

Sustainability is ...

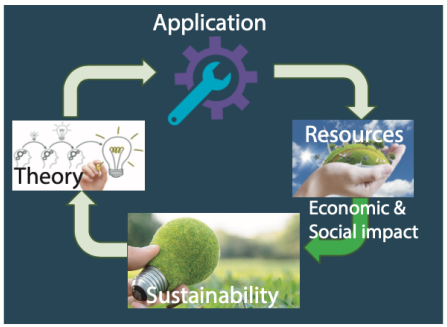
Incorporating Sustainable Development into the Design of a Student Project

Zeev Bohbot, Johan Hellsvik, Roman Iakymchuk, Lauren McKee, Rajib Sinha
KTH Royal Institute of Technology, Stockholm, Sweden

Who we are?



How is everything connected to Sustainability



KTH's goals for sustainability 2016-2020

KTH Environment Education
Basic knowledge about KTH's work with the environment and Sustainable Development (SD)

Education
Increase all employee's and student's knowledge
Integrate SD into all educational programs at all levels

Research
Increase research for SD
Increase the integration of SD in KTH's research base

How can we incorporate sustainable development in a student project?

Resources are needed to conduct tasks relating to the project. 'Resource Footprints are Good Proxies of Environmental Damage' (Steinmann et al 2017). Minimize the resource use of i) designing a project, ii) activities in the project, iii) component in the project, and iv) economic and social impacts

Sustainable development project framework
Study the current system (Systems approach)
Connect the sustainability issues (3 pillars of SD)
Apply IE perspective* to understand and solve the issues

*Waste as a resource, Industrial symbiosis, Closed loop/zero waste/circular economy, Material/energy flows, Life cycle thinking, Systems perspective

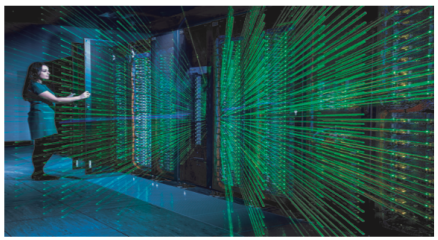
Energy-efficient architectures such as graphic processors (GPUs)

Sustainable algorithms

Some applications do not require 'full accuracy' answers: i) signal & video processing, ii) Monte Carlo simulation, iii) machine learning

Adaptive precision calculations
Mixed precision solvers with iterative refinement
Neural networks and deep learning

Application in Computer Science: High Performance Computing (HPC)



Reference

Steinmann, Z.J., Schipper, A.M., Hauck, M., Giljum, S., Wernet, G. and Huijbregts, M.A., 2017. Resource footprints are good proxies of environmental damage. Environmental science & technology, 51(11), pp.6360-6366.

Today's syllabus: resource awareness

X 30 =

- Sustainability is the **art of living well, within the ecological limits** of a finite planet. (Jackson, 2010)
- How can we incorporate sustainability into Scientific Computing, including HPC?

Sustainable HPC $\rightarrow$ Energy-efficient HPC

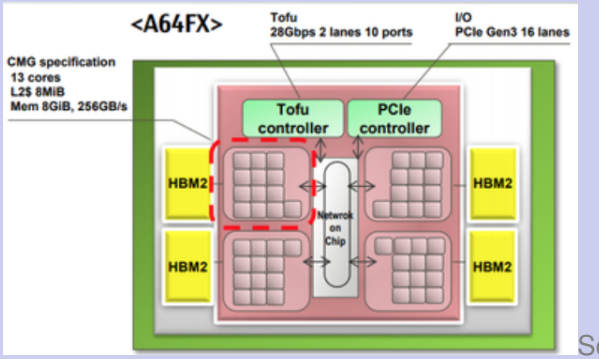
- **Energy-efficient architectures** such as graphic processors (GPUs) and FPGAs – Green HPC computing
- PDC@KTH extracts the produced heat to **warm up the main campus**
- CSCS at Switzerland proposes **'free cooling'** with the water from the lake of Lugano

Exascale computing and linear algebra

- Exascale computing is constrained by **power consumption**

→ Power-efficient hardware

- RIKEN's Fugaku w A64FX (FP64:FP32:FP16 = 1:2:4)
- EPI (ARM, FPGA, RISC-V)



Source: Fujitsu

- Linear algebra is known to be dominant by **double precision**

→ Energy-efficient algorithms

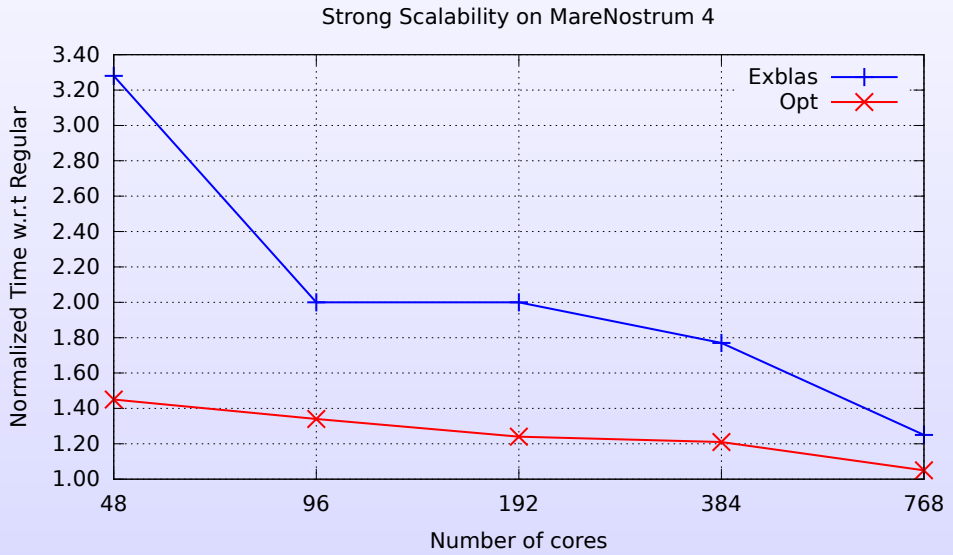
- **Mixed-precision with iterative refinement**
- Communication hiding or avoiding

Robustness of Algorithms

- **Robustness: accuracy and reproducibility**
- FP ops are non-associative : $(-1+1)+2^{-53} \neq -1+(1+2^{-53})$
- Non-reproducibility in PCG: `dot`, `axpy`, and `spmv`

→ Solution : ExBLAS (ParCo15, NRE15, **JCAM20**, IJHPCA20)

3D Poisson equation with 27 stencil points and $tol = 10^{-8}$

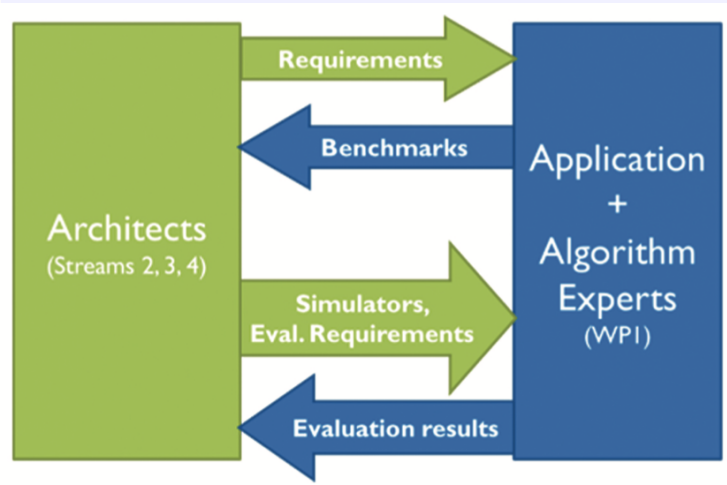


Iteration	Residual			
	MPFR	*Original* 1 proc	*Original* 48 procs	*Exblas & Opt*
0	0x1.19f179eb7f032p+49	0x1.19f179eb7f03**3**p+49	0x1.19f179eb7f03**3**p+49	0x1.19f179eb7f032p+49
2	0x1.f86089ece9f75p+38	0x1.f86089**f08810d**p+38	0x1.f86089e**d07a76**p+38	0x1.f86089ece9f75p+38
9	0x1.fc59a29d329ffp+28	0x1.fc59a29d**1b6a**p+28	0x1.fc59a29d**2e989**p+28	0x1.fc59a29d329ffp+28
10	0x1.74f5ccc211471p+22	0x1.74f5cc**b8203ad**p+22	0x1.74f5ccc**1fafef**p+22	0x1.74f5ccc211471p+22
...	...	...	...	...
40	0x1.7031058eb2e3ep-19	0x1.703105**aea0e8a**p-19	0x1.7031058e**8ff5a**p-19	0x1.7031058eb2e3ep-19
42	0x1.4828f76bd68afp-23	0x1.4828f**6fabbf2a**p-23	0x1.4828f76b**b9038**p-23	0x1.4828f76bd68afp-23
45	0x1.8646260a70678p-26	0x1.864626**01300d2**p-26	0x1.8646260a7**1301**p-26	0x1.8646260a70678p-26
47	0x1.13fa97e2419c7p-33	0x1.13fa9**8038c44e**p-33	0x1.13fa97e**54e903**p-33	0x1.13fa97e2419c7p-33

Table 3: Accuracy and reproducibility comparison on the intermediate and final residual against MPFR for a matrix with condition number of 10^{12}. The matrix is generated following the procedure from Section 5.1 with n=4,019,679 (159^3).

European Processor Initiative

General Overview



Source: EPI

- 26+1 partners, including Fraunhofer ITWM
- ARM-based chips production by SiPearl
- **EPAC** – first version of RISC-V EPI accelerator architecture
- Software level: support of RISC-V vector intrinsics and auto parallelization of C/C++ codes
- Automotive high-performance computing PoC, e.g. ADAS functionality

European Processor Initiative

VRP and STX

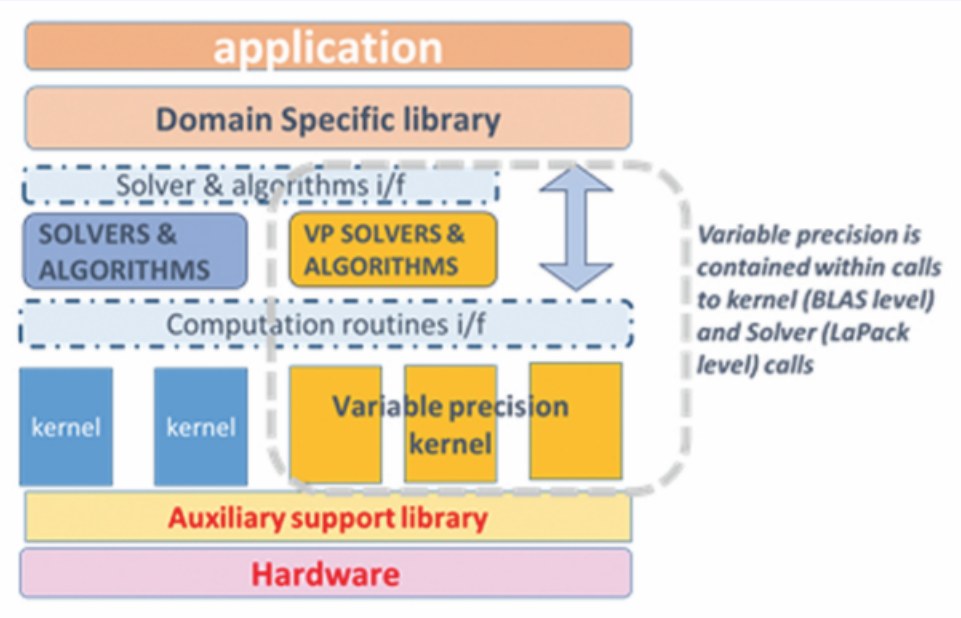
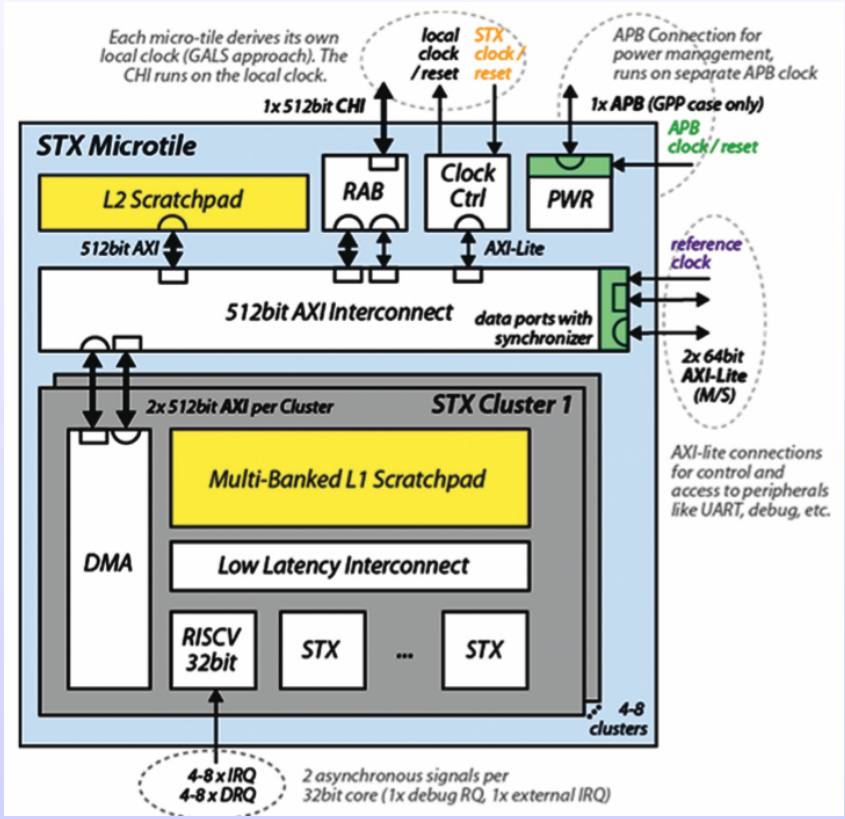


Fig: Layered programming model

- for large ill-conditioned systems
- "when the standard precision unit cannot reach the expected accuracy, the variable precision unit takes the relay"
- zero-copy from GPP to VRP

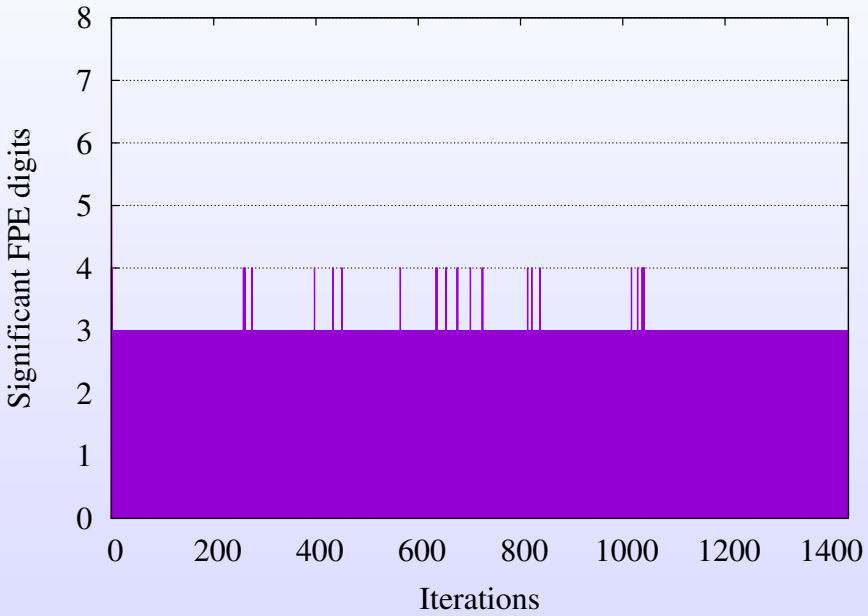


- Stencil/ tensor accelerator
- Energy efficiency
- Posit-based ML & DNN Acceleration

Source: EPI

Possibilities to exploit VRP

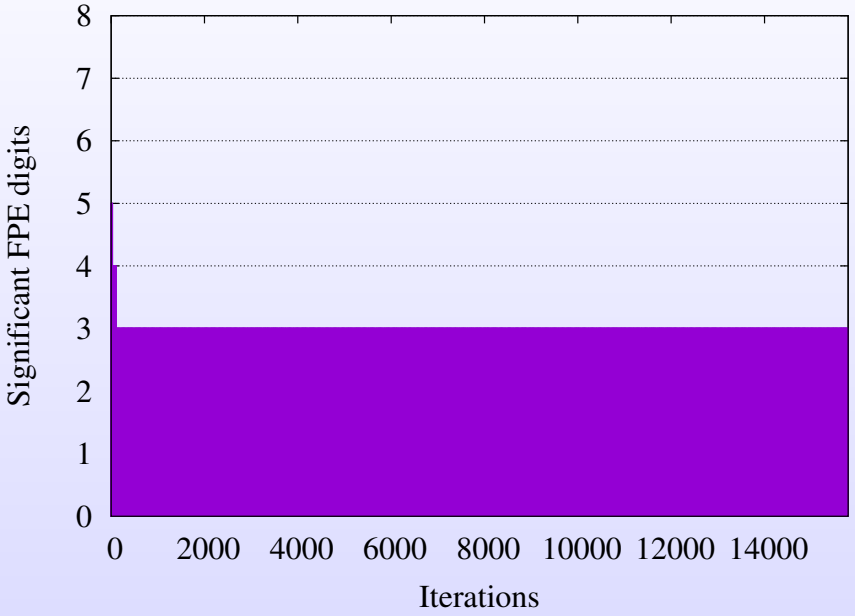
Required precision in PCG to keep every bit



msc01050 (Boeing)

$NNZ = 26,198$

$cond(A) = 9.0e+15$

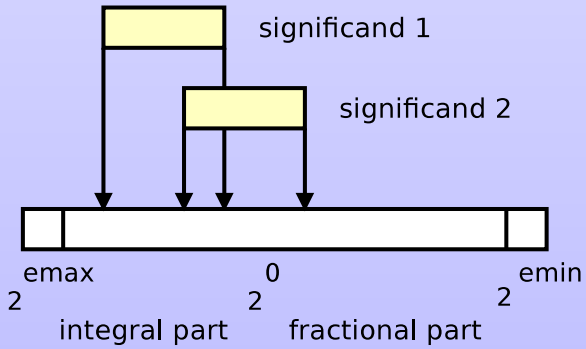


gyro_k (Oberwolfach)

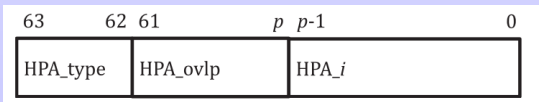
$NNZ = 1,021,159$

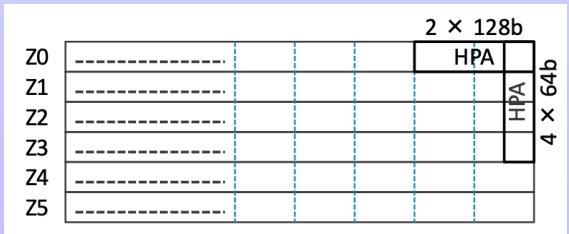
$cond(A) = 1.1e+09$

Kulisch accumulator



ARM High-Precision Anchored





Source: ARM, IEEE ToC18

Textarossa

- Started 01/04/2021
- 11 partners, including Fraunhofer ITWM
- Enable mixed-precision computing through libraries and compilers
- HW and compiler support of posit
- Fast task scheduling and low-latency intra/inter-node communication

ImPreNum (FR, 2018-23)

Improving Predictability of Numerical Computations

Objectives

- accuracy as a first-class concern in numerical computing
- add accuracy considerations to the cost/performance trade-offs

$\rightarrow$ from hardware to languages, compilers, and numerical libraries like BLAS and LAPACK

Key points

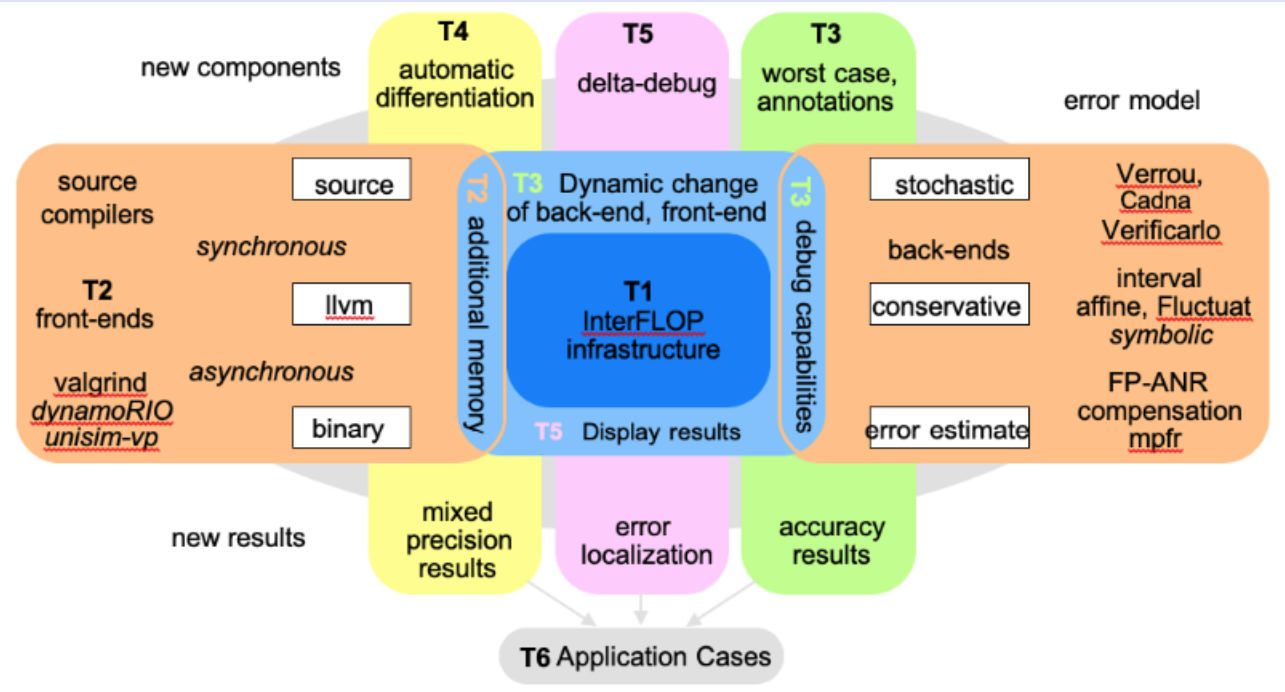
- modified RISC-V processor with a variable-precision accelerator, extensions to the C language and compiler support
- enabling precision control at the lower levels of the computing stack
- formulate accuracy requirements $\rightarrow$ to exploit this precision control

- More on ImPreNum at 10:15AM by Tiago Trevisan Jost

InterFlop (FR, started 02/2021)

Objectives (7 partners w Sorbonne)

- combine a set of error analyzes that cover a large number of possible inputs
- propose new floating-point formats
- improve precision auto-tuning
- mixed precision and probabilistic error analysis
- provide original solutions to visualize and interpret results



Minimal Precision Computing (FR-JP)

- A ***minimal-precision computing*** scheme/system involving both hardware & software stack, combining three key components:

Build upon available technologies (most of them are in-house)

1. **Precision-tuning based on a numerical validation method**
   - Validation libraries with stochastic arithmetic: CADNA & SAM (Sorbonne U.)
   - Precision-tuner with CADNA: PROMISE (Sorbonne U.)
2. **Arbitrary (high-/mixed-) precision arithmetic / numerical libraries**
   - Arithmetic library: MPFR (GNU), QD (Hida et al.)
   - Arbitrary-precision BLAS/LAPACK: MPLAPACK (Nakata)
   - Accurate BLAS: ExBLAS (Sorbonne U.), OzBLAS (TWCU/RIKEN), etc.
3. **FPGA as arbitrary-precision hardware**
   - FPGA-GPU-CPU system: “Cygnus” (U. Tsukuba)
   - HLS compilers: SPGen (RIKEN), Nymble (TU Darmstadt/RIKEN)

Thank you for your attention!

Sustainable and Robust Linear Algebra Computations at Exascale

MS56		
09:35AM	Roman Iakymchuk	Sustainable and Robust Linear Algebra Computations: A Brief Overview of European Efforts
09:55AM	Toshiyuki Imamura	Riken's Effort for Sustainable Linear Algebra Computations
10:15AM	Tiago Trevisan Jost	Fast Exploration of Variable Precision Linear Kernels
10:35AM	Fabienne Jézéquel	Precision Auto-Tuning and Control of Accuracy in High Performance Simulations

MS63		
11:15AM	Takeshi Ogita	Iterative Refinement and Verified Numerical Linear Algebra
11:35AM	Theo Mary	Multiple Word Arithmetic with GPU Tensor Cores: Theory and Practice
11:55AM	Takeshi Fukaya	Exploiting Lower Precision Computing in the GMRES(m) Method
12:15PM	Emmanuel Agullo	Variable Accuracy Storage through Lossy Compression Techniques in Numerical Linear Algebra