Sparse-Days 2022

20-22 June 2022

Adaptive Precision Sparse Matrix–Vector Product

and its Application to Krylov Solvers

Roméo Molina

LIP6, Sorbonne Université

Service Online, Département Informatique, IJCLab

Joint work with

Stef Graillat, **Fabienne Jézéquel**, and **Theo Mary**

Today's floating-point landscape

		Bits Signif. (t)	Exp.	Range	$u = 2^{-t}$
bfloat16	B	8	8	$10^{\pm 38}$	4×10^{-3}
fp16	H	11	5	$10^{\pm 5}$	5×10^{-4}
fp32	S	24	8	$10^{\pm 38}$	6×10^{-8}
fp64	D	53	11	$10^{\pm 308}$	1×10^{-16}
fp128	Q	113	15	$10^{\pm 4932}$	1×10^{-34}

- Low precision increasingly supported by hardware

Today's floating-point landscape

		Bits Signif. (t)	Exp.	Range	$u = 2^{-t}$
bfloat16	B	8	8	$10^{\pm 38}$	4×10^{-3}
fp16	H	11	5	$10^{\pm 5}$	5×10^{-4}
fp32	S	24	8	$10^{\pm 38}$	6×10^{-8}
fp64	D	53	11	$10^{\pm 308}$	1×10^{-16}
fp128	Q	113	15	$10^{\pm 4932}$	1×10^{-34}

- Low precision increasingly supported by hardware
- **Great benefits:**
  - Reduced **storage**, data movement, and communications
  - Reduced **energy** consumption (5× with fp16, 9× with bfloat16)
  - Increased **speed** on emerging hardware (16× on A100 from fp32 to fp16/bfloat16)

Today's floating-point landscape

		Bits Signif. (t)	Exp.	Range	$u = 2^{-t}$
bfloat16	B	8	8	$10^{\pm 38}$	4×10^{-3}
fp16	H	11	5	$10^{\pm 5}$	5×10^{-4}
fp32	S	24	8	$10^{\pm 38}$	6×10^{-8}
fp64	D	53	11	$10^{\pm 308}$	1×10^{-16}
fp128	Q	113	15	$10^{\pm 4932}$	1×10^{-34}

- Low precision increasingly supported by hardware
- **Great benefits:**
  - Reduced **storage**, data movement, and communications
  - Reduced **energy** consumption (5× with fp16, 9× with bfloat16)
  - Increased **speed** on emerging hardware (16× on A100 from fp32 to fp16/bfloat16)
- **Some limitations too:**
  - Low accuracy (large u)
  - Narrow range

Adaptive precision algorithms

Mix several precisions in the same code with the goal of

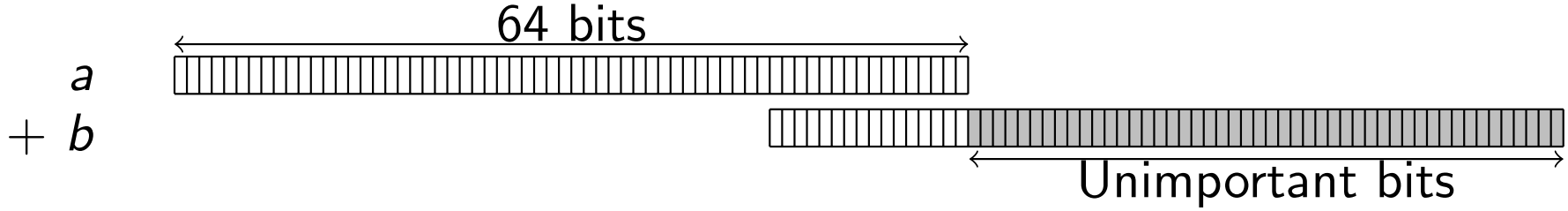
- Getting the performance benefits of low precisions
- While preserving the accuracy and stability of high precision

Adaptive precision algorithms

Mix several precisions in the same code with the goal of

- Getting the performance benefits of low precisions
- While preserving the accuracy and stability of high precision

Opportunity for mixed precision:
$\Rightarrow$ **all computations are not equally "important"!**



and small elements produce small errors :

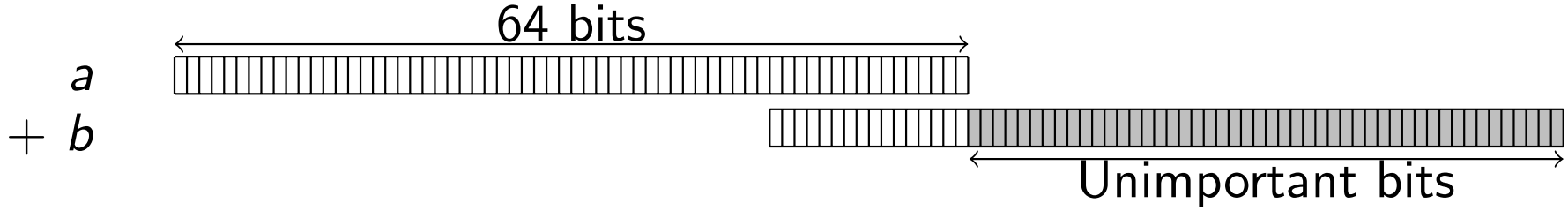
$$|\,\mathrm{fl}(a\,\mathrm{op}\,b) - a\,\mathrm{op}\,b| \leq u|a\,\mathrm{op}\,b|, \qquad \mathrm{op} \in \{+, -, *, \div\}$$

Adaptive precision algorithms

Mix several precisions in the same code with the goal of

- Getting the performance benefits of low precisions
- While preserving the accuracy and stability of high precision

Opportunity for mixed precision:
$\Rightarrow$ **all computations are not equally "important"!**



and small elements produce small errors :

$$|\,\mathrm{fl}(a \,\mathrm{op}\, b) - a \,\mathrm{op}\, b| \leq u|a \,\mathrm{op}\, b|, \qquad \mathrm{op} \in \{+, -, *, \div\}$$

$\Rightarrow$ We can adapt the precisions to the data at hand

Come see the poster to discover

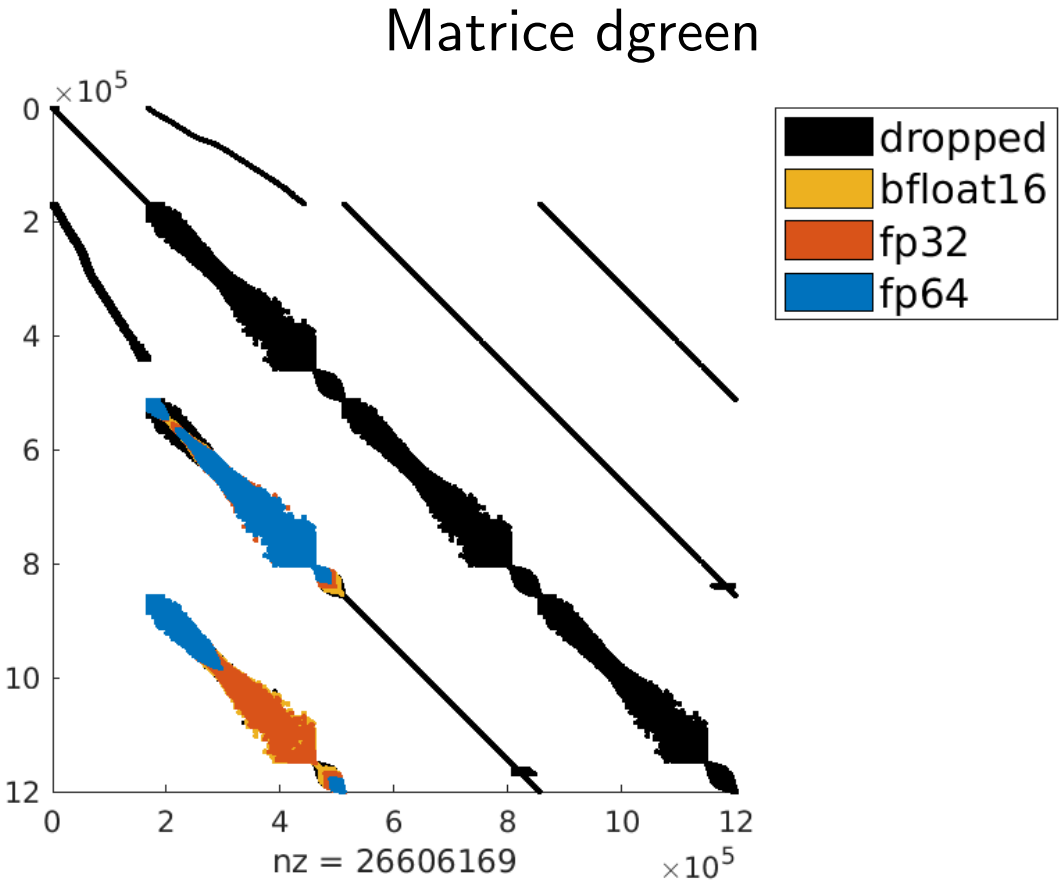
- **Adaptive precision SpMV**
  - Backward error analysis
  - Deduced algorithm
  - Experimental results

Come see the poster to discover

- **Adaptive precision SpMV**
  - Backward error analysis
  - Deduced algorithm
  - Experimental results

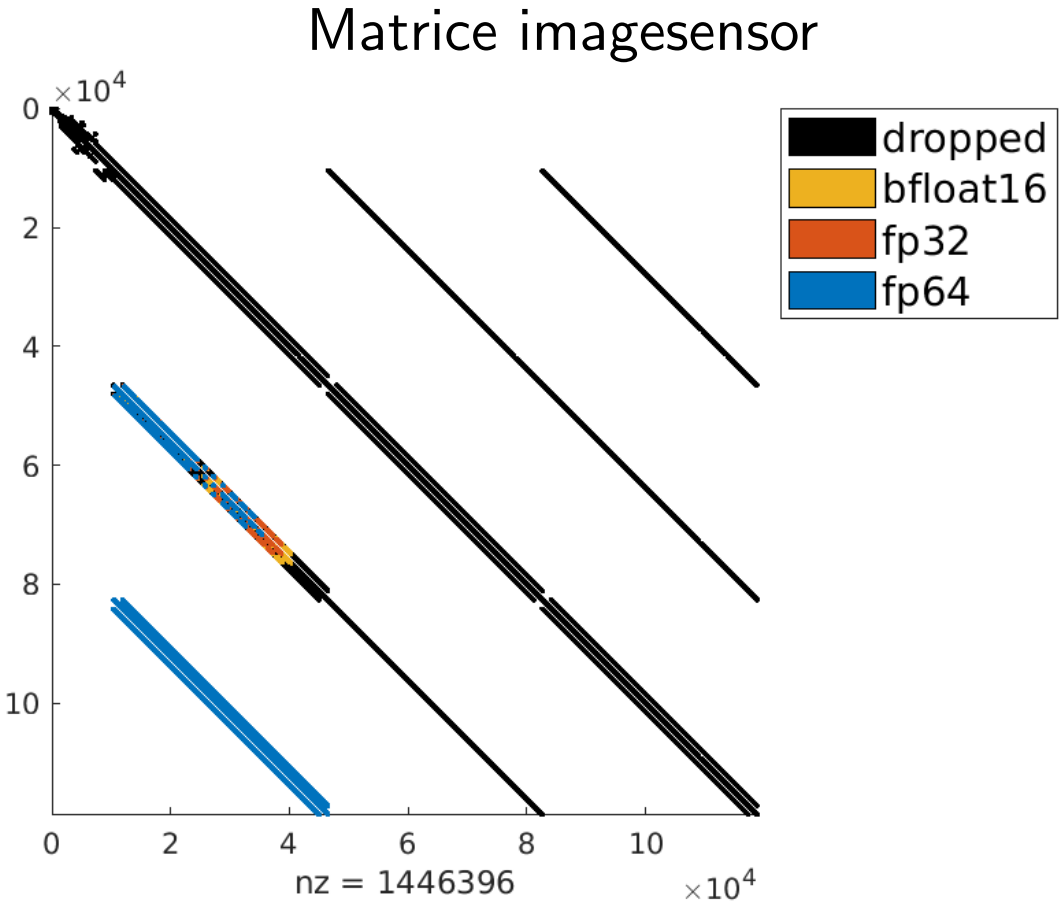
- **Application to Krylov solvers**
  - Importance of SpMV in GMRES and GMRES-IR
  - Convergence scheme

Example of mixed-precision SpMV gains



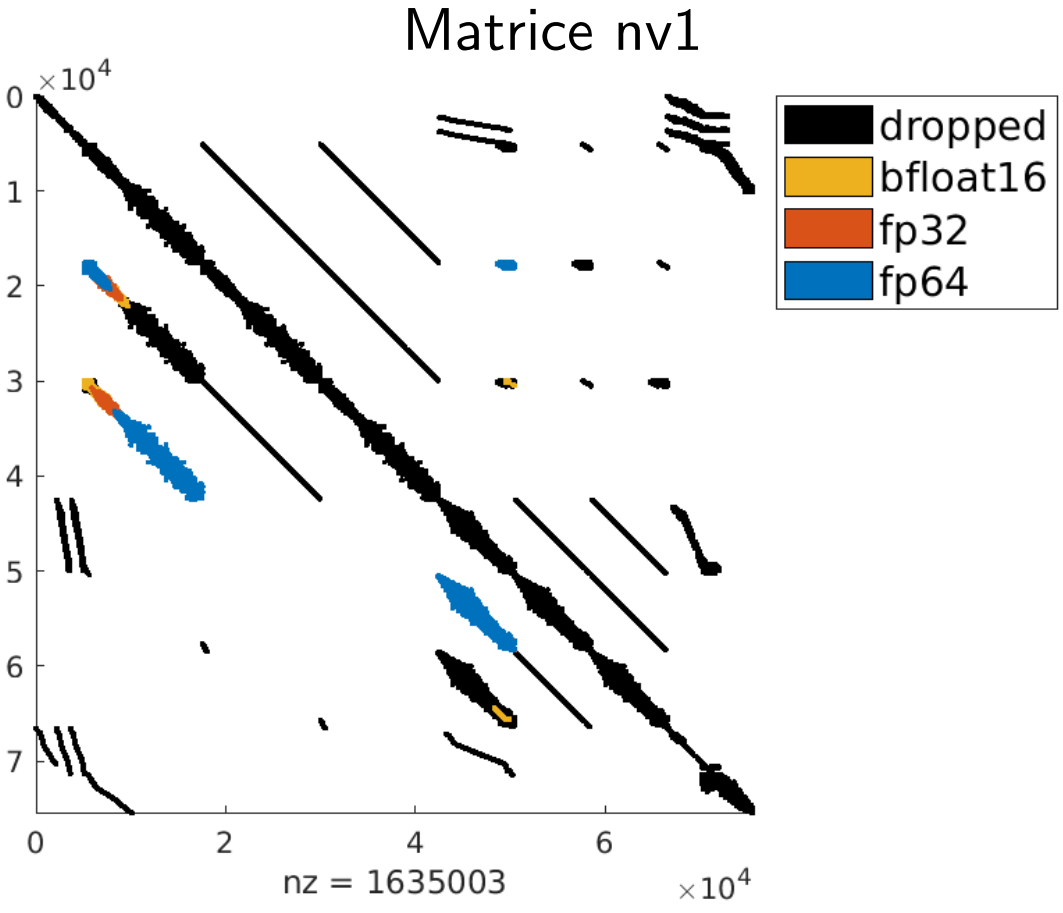
For some matrices, many elements can be dropped that leads to major gains.

Example of mixed-precision SpMV gains



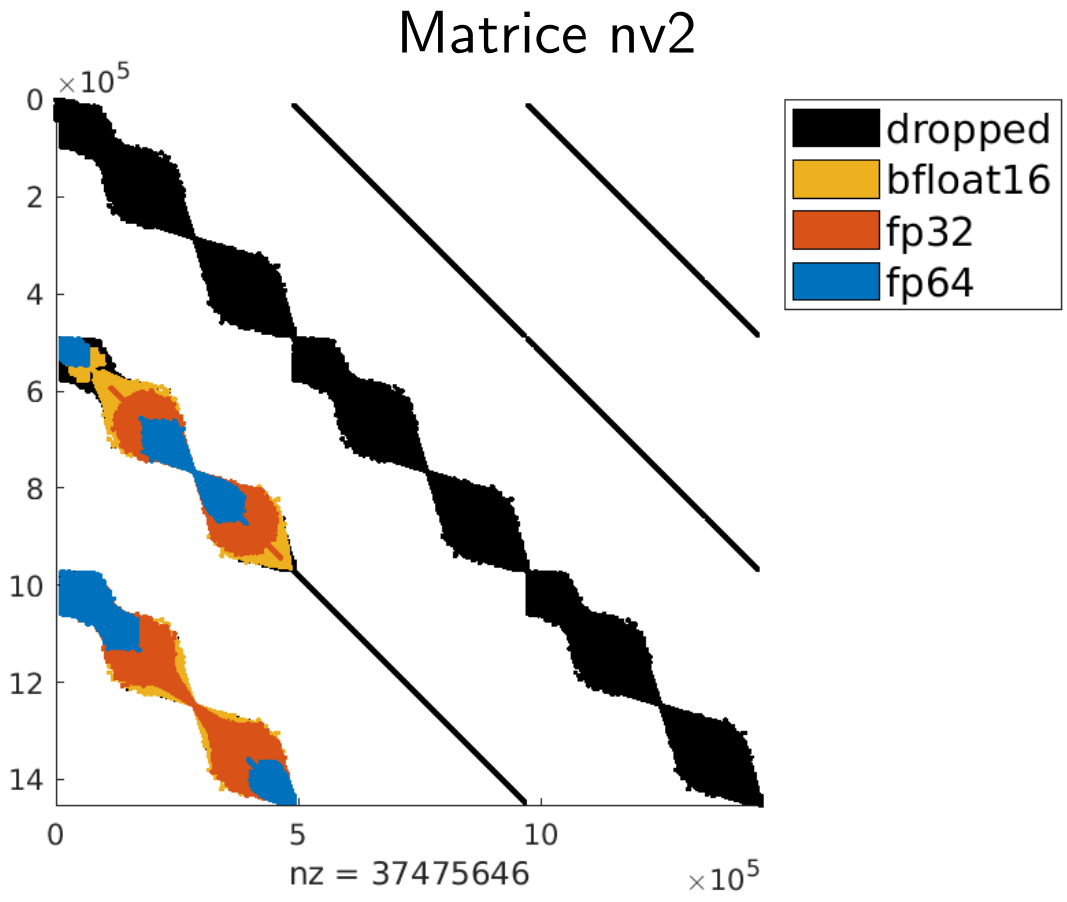
For some matrices, many elements can be dropped that leads to major gains.

Example of mixed-precision SpMV gains



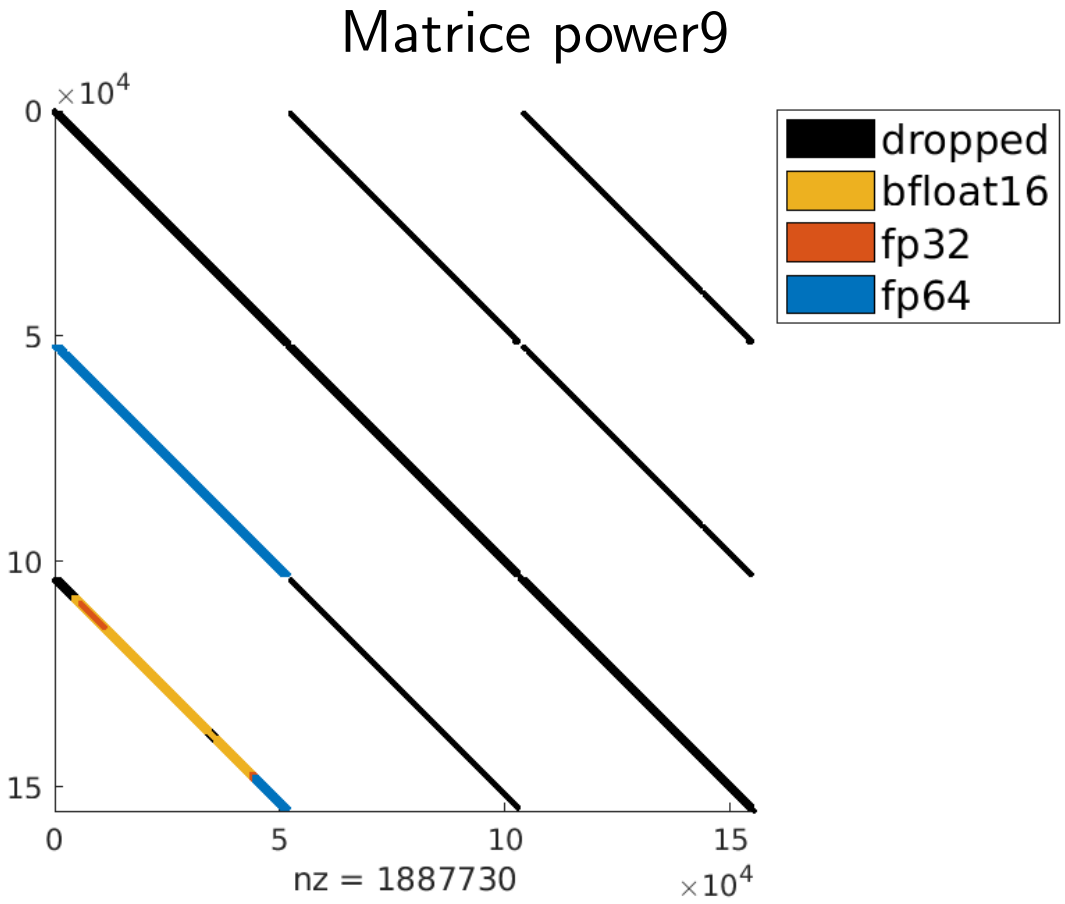
For some matrices, many elements can be dropped that leads to major gains.

Example of mixed-precision SpMV gains



For some matrices, many elements can be dropped that leads to major gains.

Example of mixed-precision SpMV gains



For some matrices, many elements can be dropped that leads to major gains.