Energy-Efficient FPGA Solutions for Large-Scale FFTs and Non-Uniform FFTs

A Software-Hardware Co-Design Approach for Radio Interferometry

Rubén Rodríguez Álvarez, Denisa Constantinescu, Miguel Peón Quirós, Adrien Devresse, Hamza Chouh, Shreyam Krishna, Etienne Orliac, David Atienza

EPFL - Embedded Systems Laboratory, EcoCloud, SCITAS

ruben.rodriguezalvarez@epfl.ch, denisa.constantinescu@epfl.ch

SEAMS Project

Sustainable & Energy Aware Methods for SKA

Scope

Energy-efficient computing with domain-specific accelerators
Multi-scale hardware-software co-design approach

Group Members

France

- **INSA Rennes**: Jean-F. Nezan, Mickaël Dardaillon, Hugo Miomandre, Jacques Morin
- **OCA**: Shan Mignot, Alain Miniussi, Chiara Ferrari, André Ferrari
- **OP**: Damien Gratadour

Switzerland

- **EcoCloud**: Miguel Peon Quiros, David Atienza
- **ESL**: Denisa Constantinescu, Rubén R. Álvarez, Basile Darne, David Atienza
- **SCITAS**: Adrien Devresse, Hamza Chouh, Etienne Orliac, Gilles Fourestey

Partners: EPFL, Laboratory of Astrophysics, MeerKAT, SKACH, SKAO

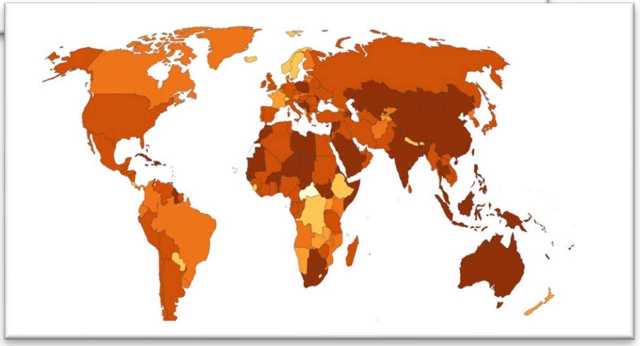
- Pipelines Profiling & Specification
- Single-Node SW/HW co-design
- Multi-Node Scale-up
- Integration and Testing
- On-Field Demonstrator

seams-project.com

Energy to completion

Electricity: Berlin's shock plan to adapt to the weather

Companies may soon have to adapt their production to the strength of the wind and the duration of sunshine, in order to relieve the electricity networks, put to the test by the intermittency of renewable energies. This is the option proposed by the Ministry of Economy and Climate in a note published in July. Enough to trigger the ire of the business world.

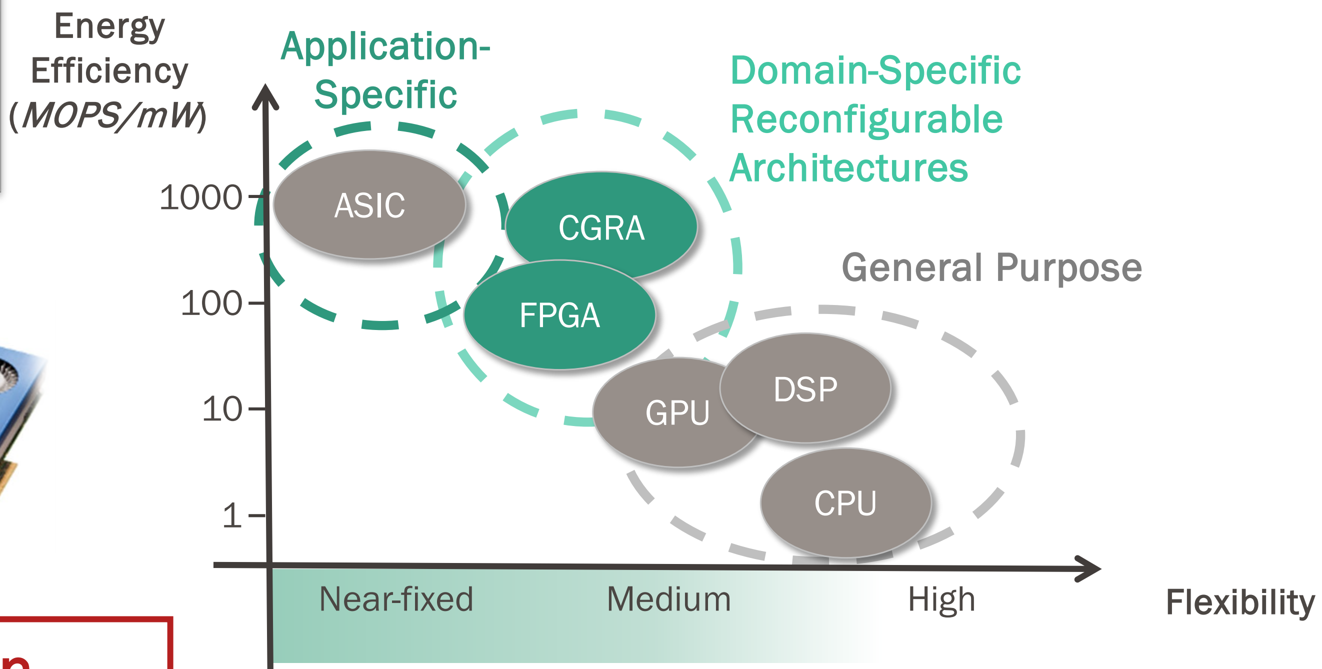


Global concern for energy consumption and lower carbon emission factor

SKA

- Exa-scale amounts of data
- Large computation -> Scalability in Multi-node
- Multi-kernel -> Domain-specific computation

Use domain-specific accelerators (GPUs, FPGAs, ASICs)



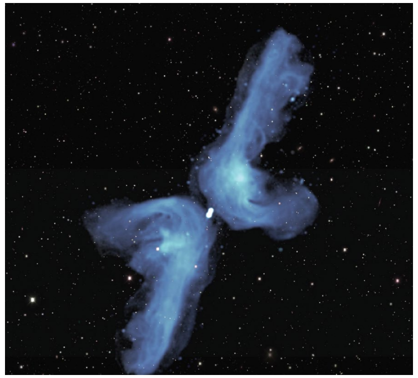
Adapted from *Kevin J M Martin. IPDPSW 2022*

Shift the focus to Energy-to-Completion

Roadmap

Visibility Domain | Image Domain

Correlated visibilities → Apply Gains → *Synthesis* → Image cleanning → *Analysis* → Apply Gains



*Non-unform points

Divide the operation in **Sub-Kernels**

FPGA Flow: Design → Validation → Profiling & Optimization → Design Space Exploration

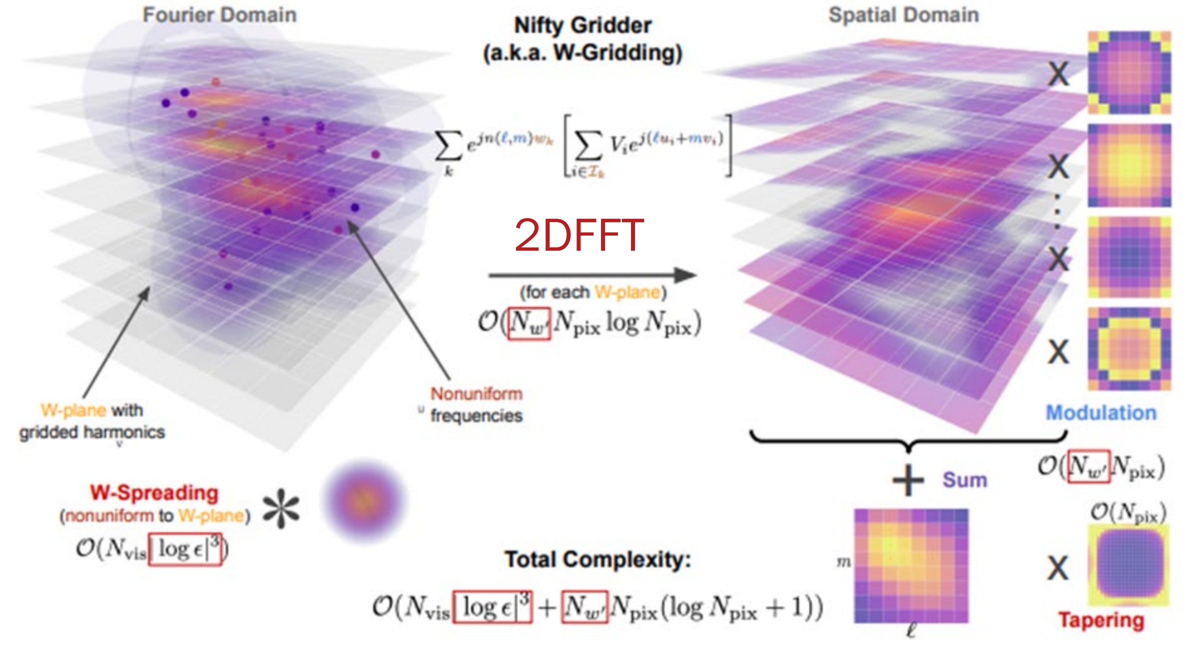
For each Sub-Kernel

Build a **Library** of accelerators

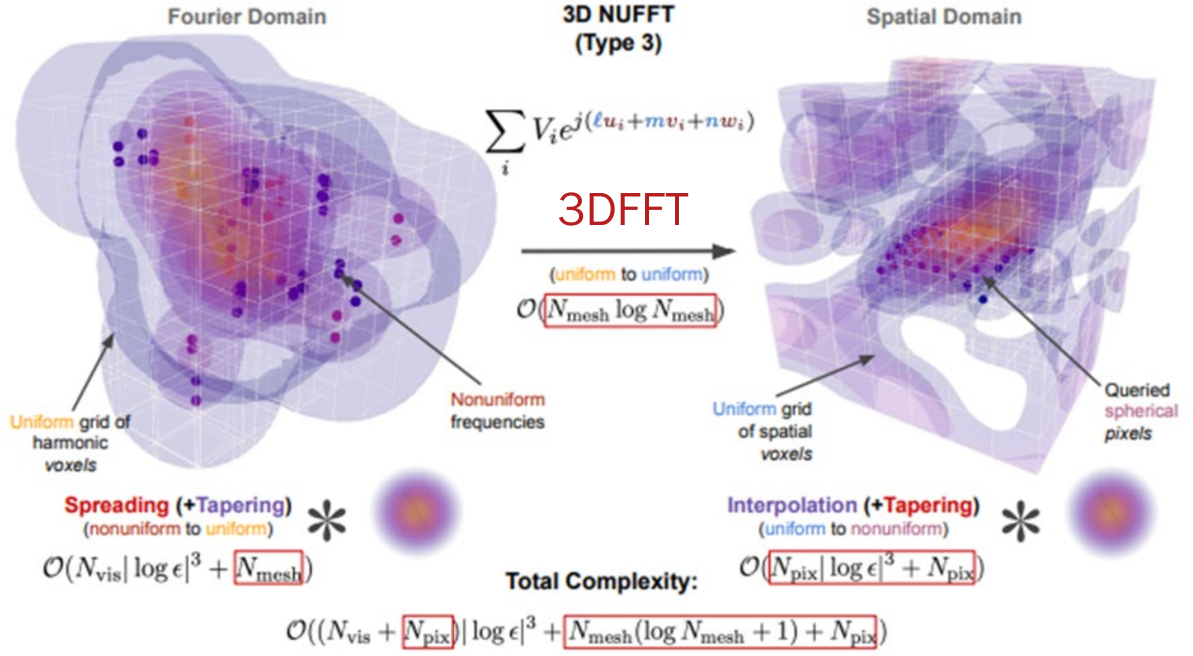
Scheduling → Design Space Exploration → Complete Execution

Synthesis (NUFFTs)

W-gridder



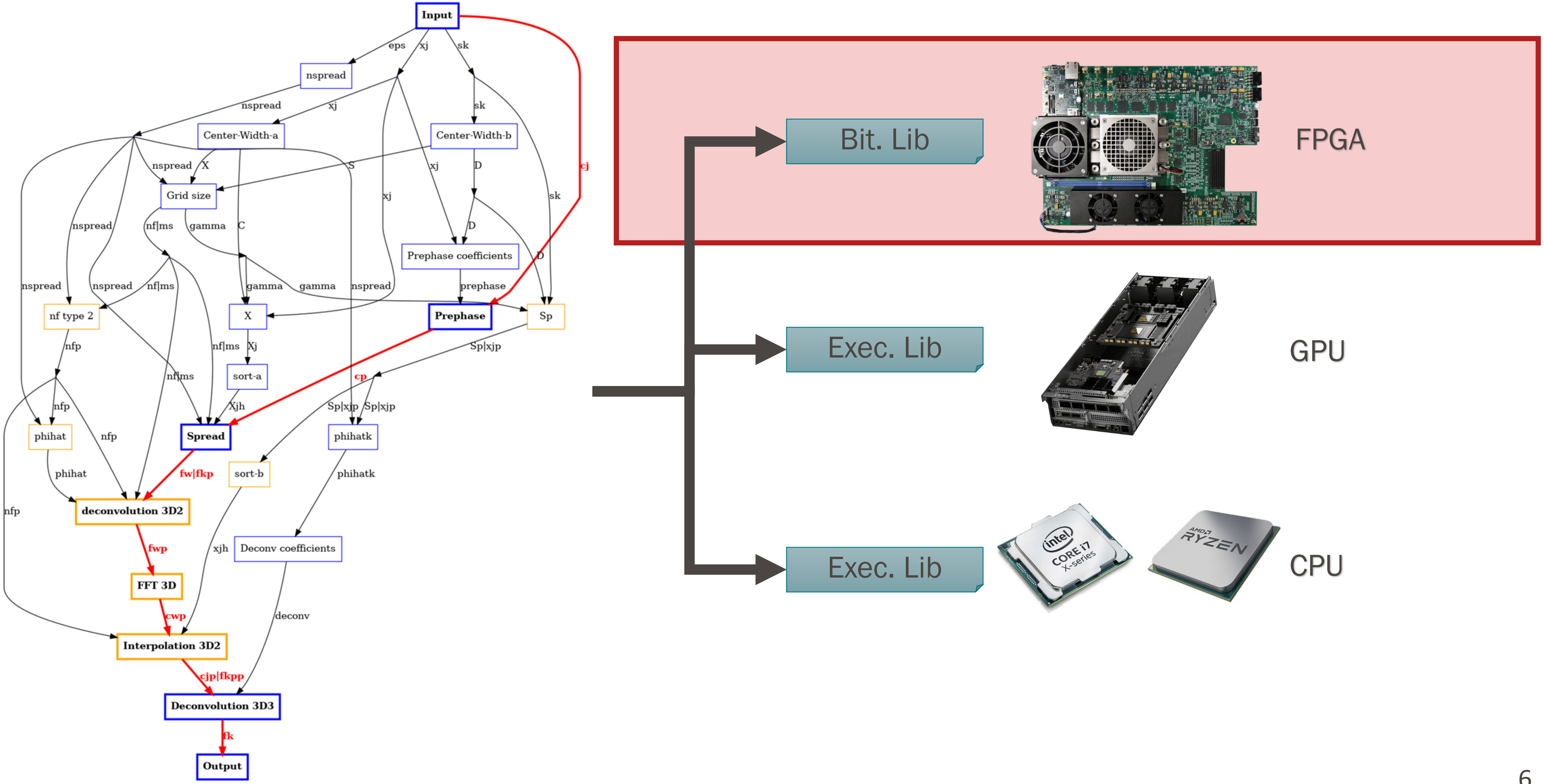
Finufft



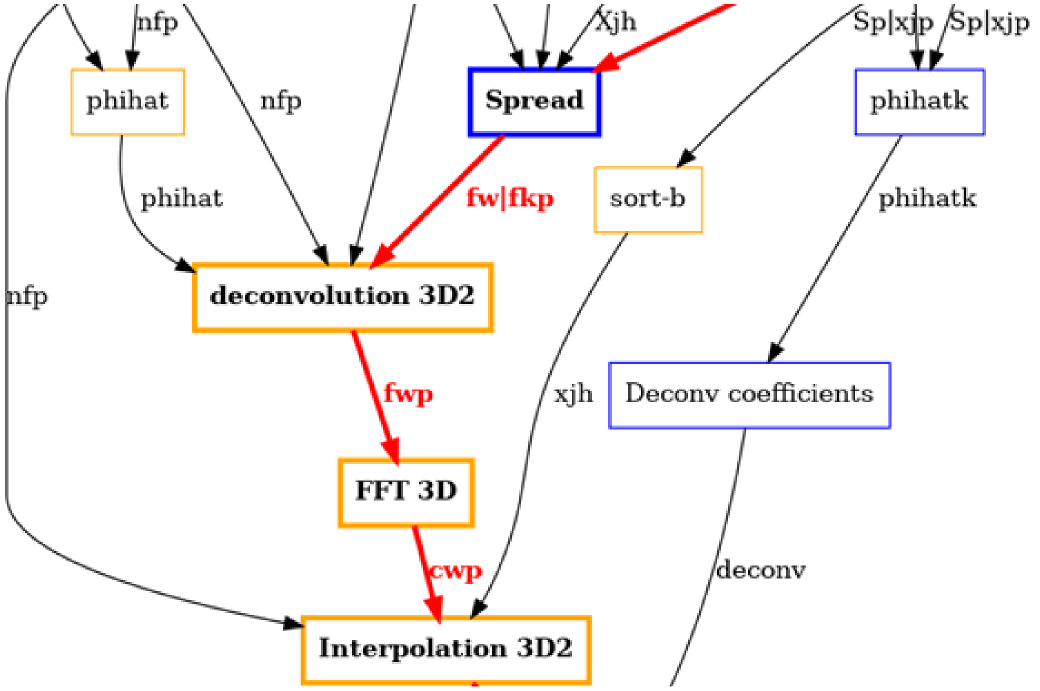
- Kashani et al. "HVOX: Scalable Interferometric Synthesis and Analysis of Spherical Sky Maps." (2023).
- Tolley et al. "BIPP: An efficient HPC implementation of the Bluebild algorithm for radio astronomy." (2023).
- Corda et al. "Near memory acceleration on high resolution radio astronomy imaging." MECO. IEEE, (2020).

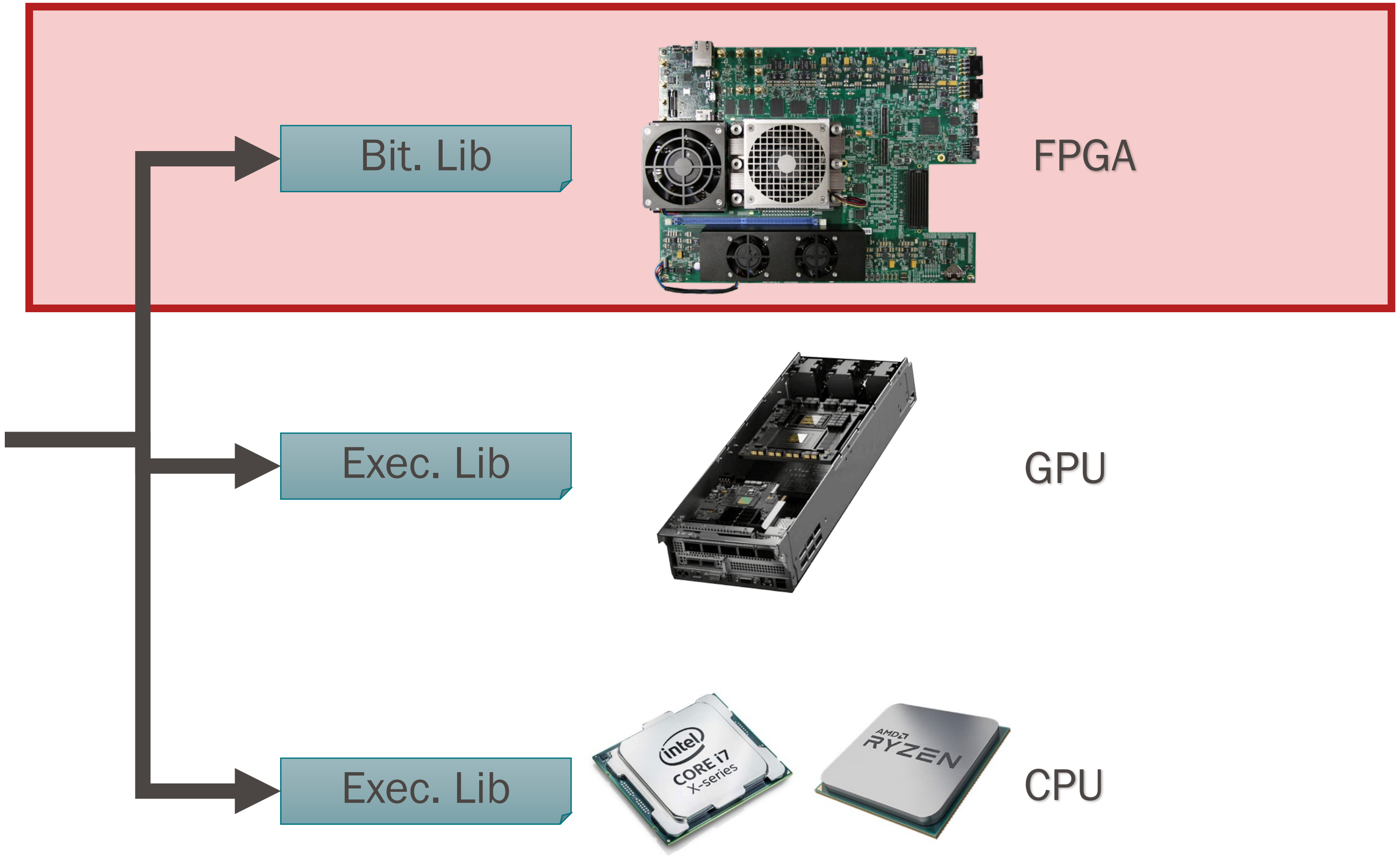
How do these algorithms map into an FPGA?

Finufft Synthesis



Finufft Synthesis





3D FFT takes 40%-90% of the computation

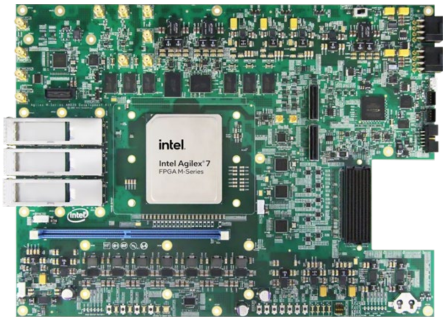
EPFL

FFT HW Design and Exploration

SEAMS
Sustainable & Energy Aware Methods for SKA

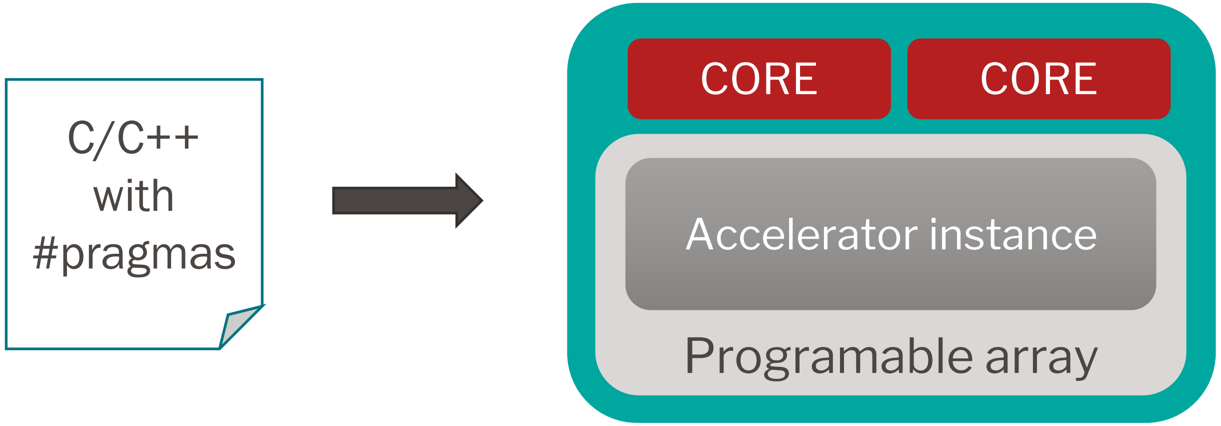
Characteristics	Agilex 7 M-Series Dev Kit	Alveo V80 Card
Internal memory	**370**Mb BRAM	**132**Mb BRAM + **541**Mb URAM
High Bandwith Memory (HBM2e)	**32**GB @ **1**TB/s	**32**GB @ **810**GB/s
Compute Elements	**3.9M LEs + 12.3K DSPs + 1.3M ALMs**	**2.6M LUTs + 10.8K DSPs**
Max Power (TDP)	**(2x) 240 Watts**	**190 Watts**
Global Memory (DDR4/5)	**64** GB	**32** GB
Comms	**16x** PCIe 5, **CXL**, GbE 116Gbps, **fiber optic**	**2x** PCIe 5
Technology	**7nm** Intel	**7nm** TSMC
Max Clock Freq	**500**MHz-**1**GHz	**600**MHz-**1**GHz

altera
An Intel Company



AMD
XILINX

High-Level Synthesis (HLS) for FPGAs



Characteristics:

- Mixed precision data types
- **Parallel, pipeline and serial**
- Resources constraints
- Code breakdown
- Highly parametrizable

Characteristics	HLS FPGA	CUDA GPU
Programming support	High	High
Productivity (design time)	Medium	High
Energy Efficiency	High	Low-Medium
Latency	Medium	Low
Scalability	High	High
Flexibility	High	Limited

We teach HLS and Co-Design; used it to accelerate **CNNs** and **genome alignment** applications

HLS is a good fit for changing SW, portable HW, and design explorations

HLS Design Flow

SEAMS
Sustainable & Energy Aware Methods for SKA

C/C++

HLS compiler

Minutes

RTL

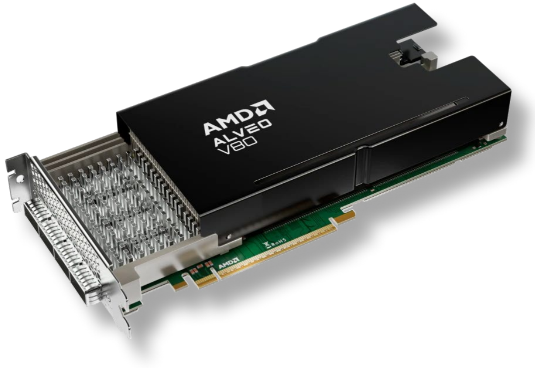
Synthesis tool

Hours

Executable

Bitstream

PCIe FPGA Card



HOST server

FPGA

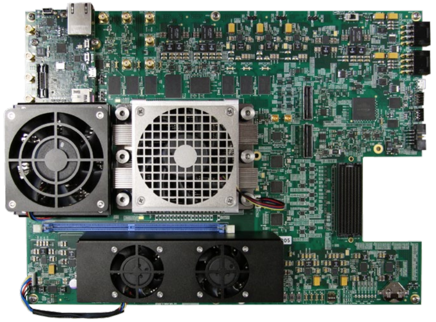
Control logic

Accelerator

MEM

POWER = HOST + FPGA

FPGA Standalone Board



FPGA

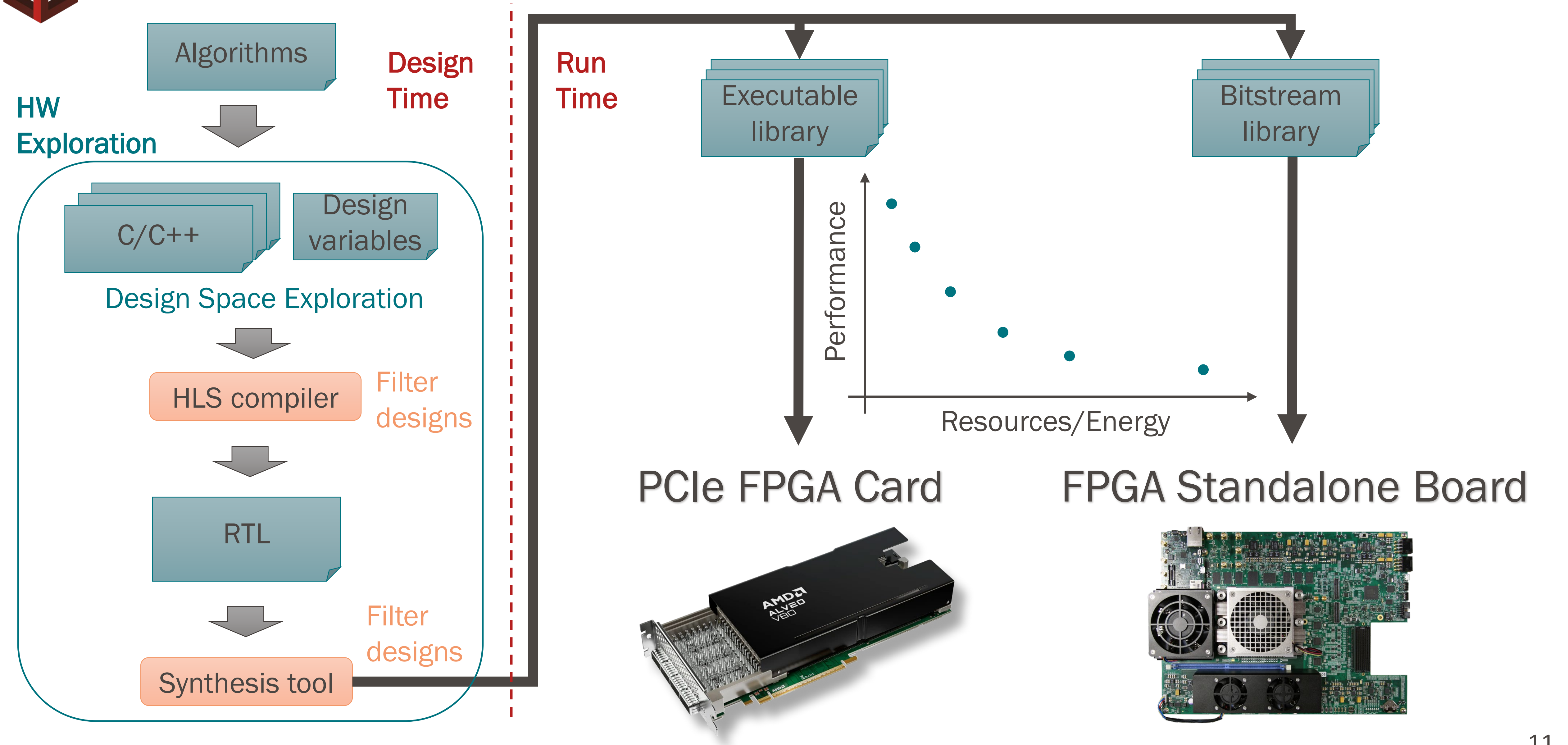
HOST controller

Accelerator

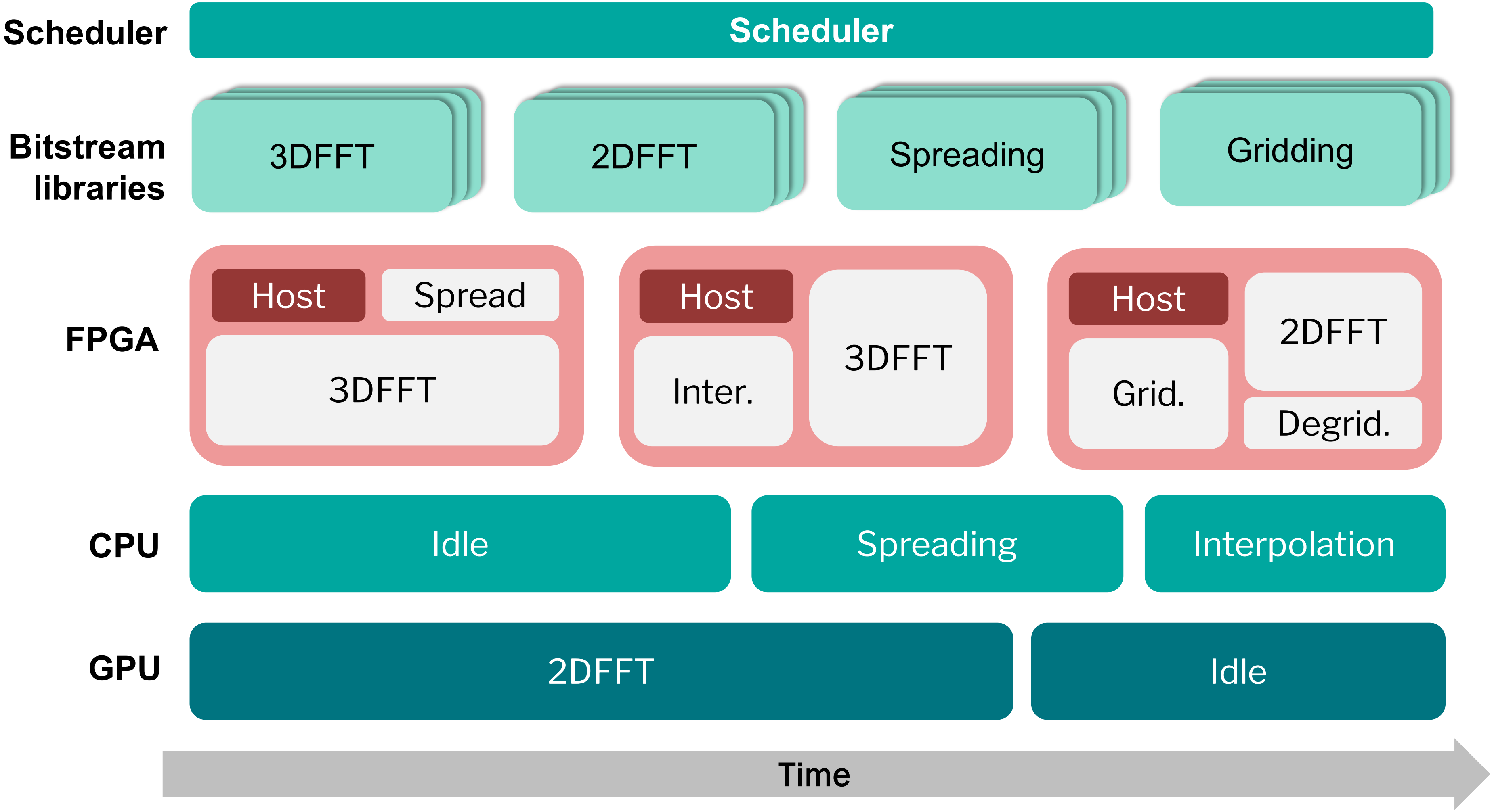
MEM

POWER = FPGA

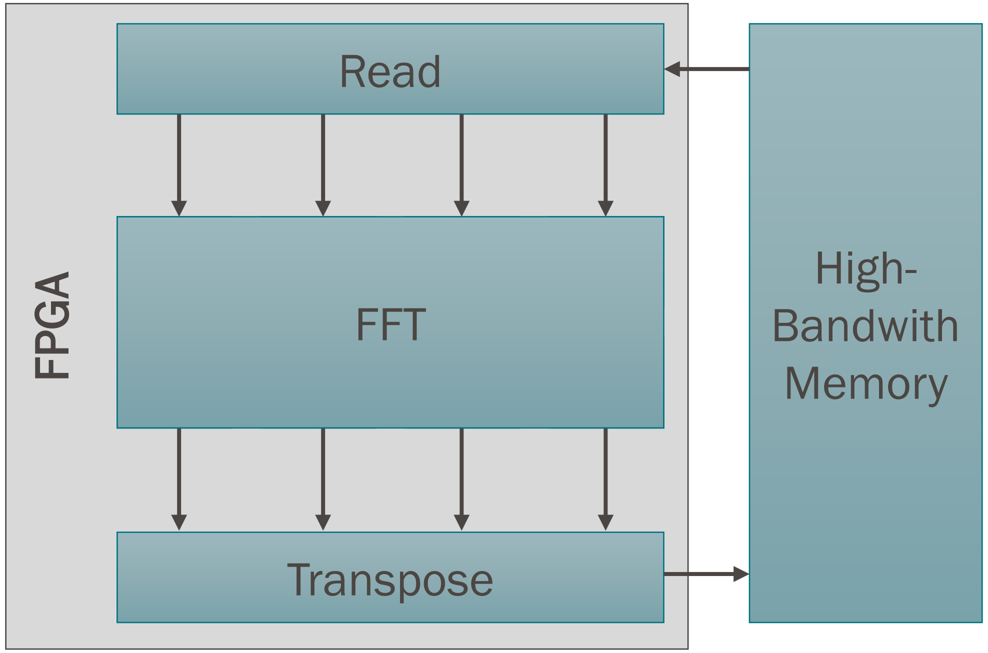
Design Space Exploration Design Flow

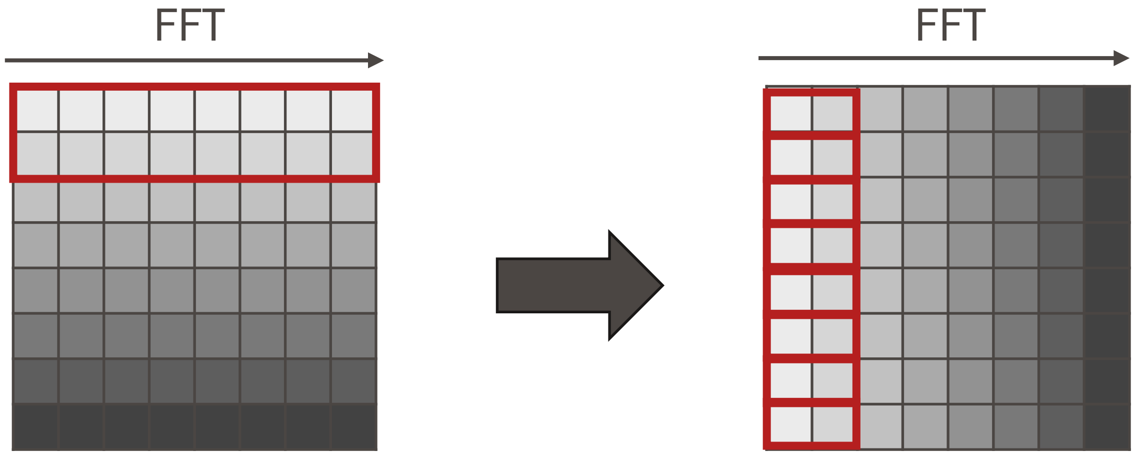


Run-time

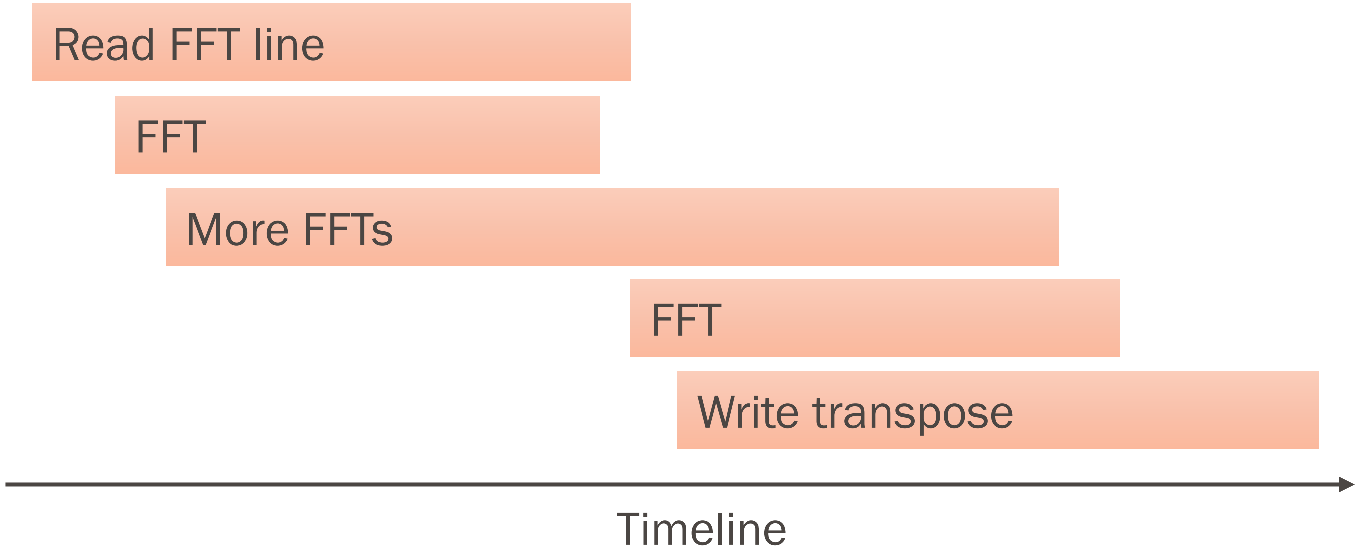

Scheduler
Scheduler
Bitstream libraries
3DFFT
2DFFT
Spreading
Gridding
FPGA
Host
Spread
3DFFT
Host
Inter.
3DFFT
Host
Grid.
2DFFT
Degrid.
CPU
Idle
Spreading
Interpolation
GPU
2DFFT
Idle
Time

FFT HW Design and Exploration



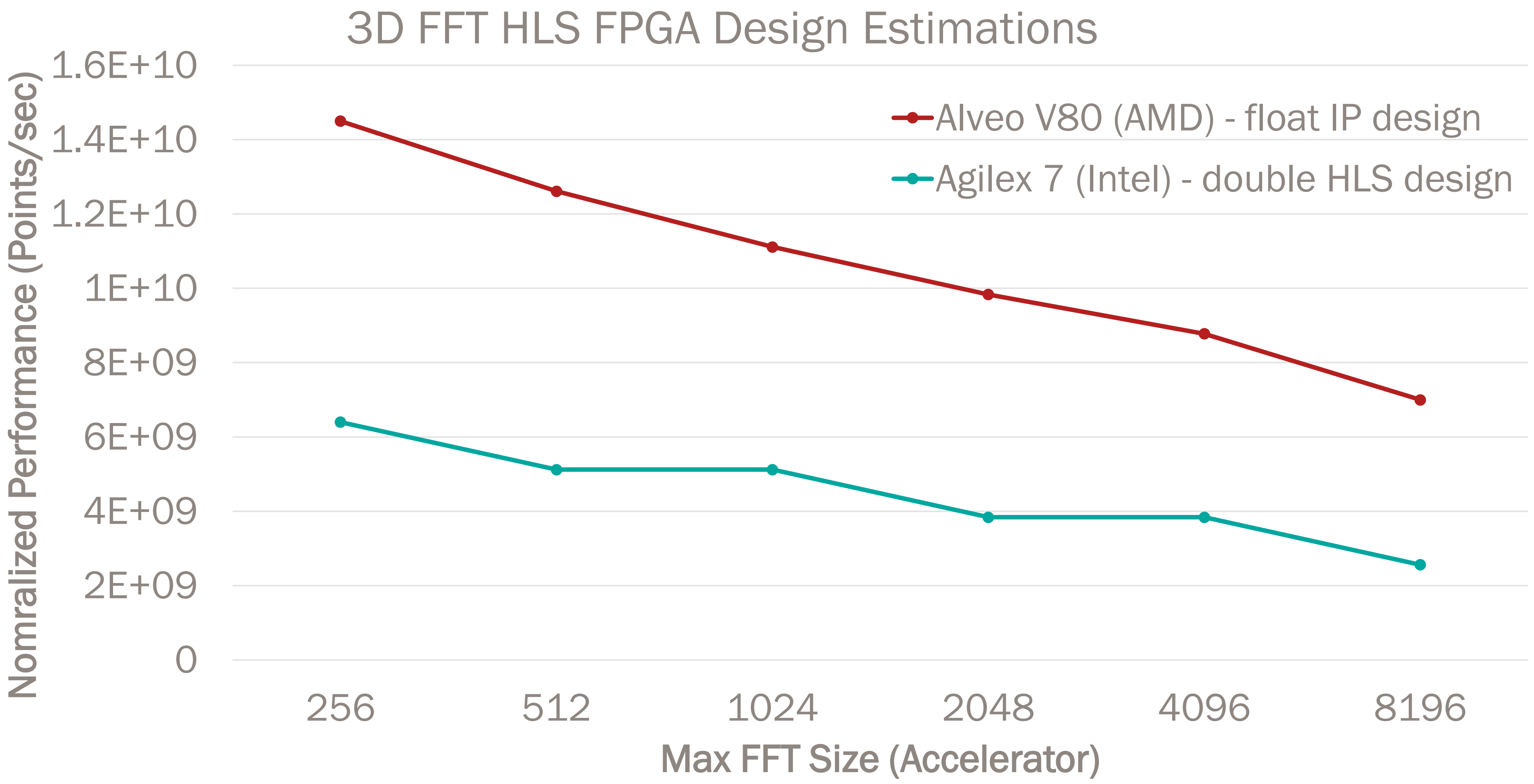


Consecutive transfers to memory takes less time and energy



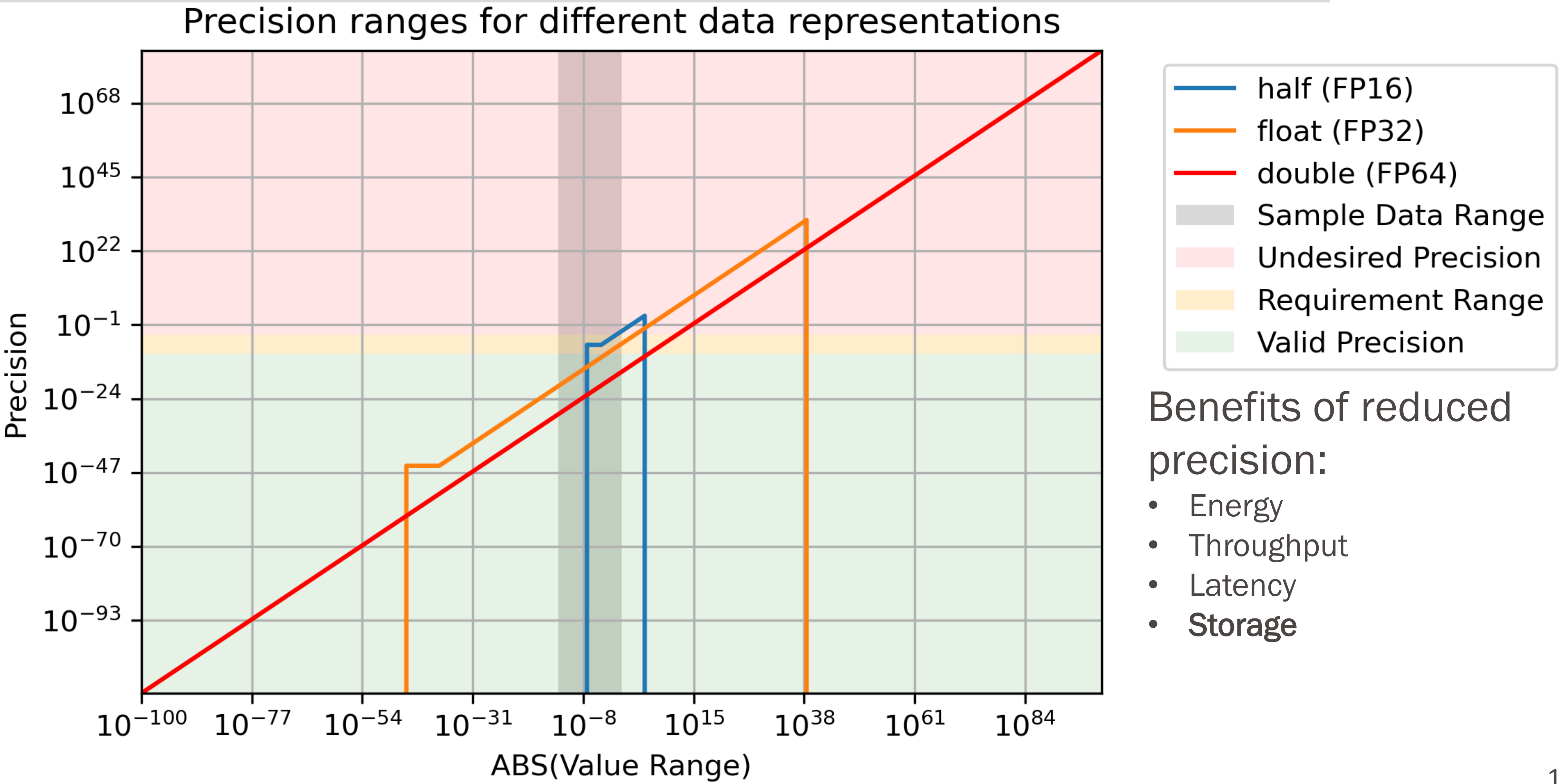
Design Variables
parallel FFTs
Transpose buffer
FFT stages
Data format
FFT Max FFT size

3D FFT HW Preliminary Results



Precision in for FINUFFT Synthesis (BIPP)

Sample data extracted from bipp execution, simulated with OSKAR for SKA-Low configuration

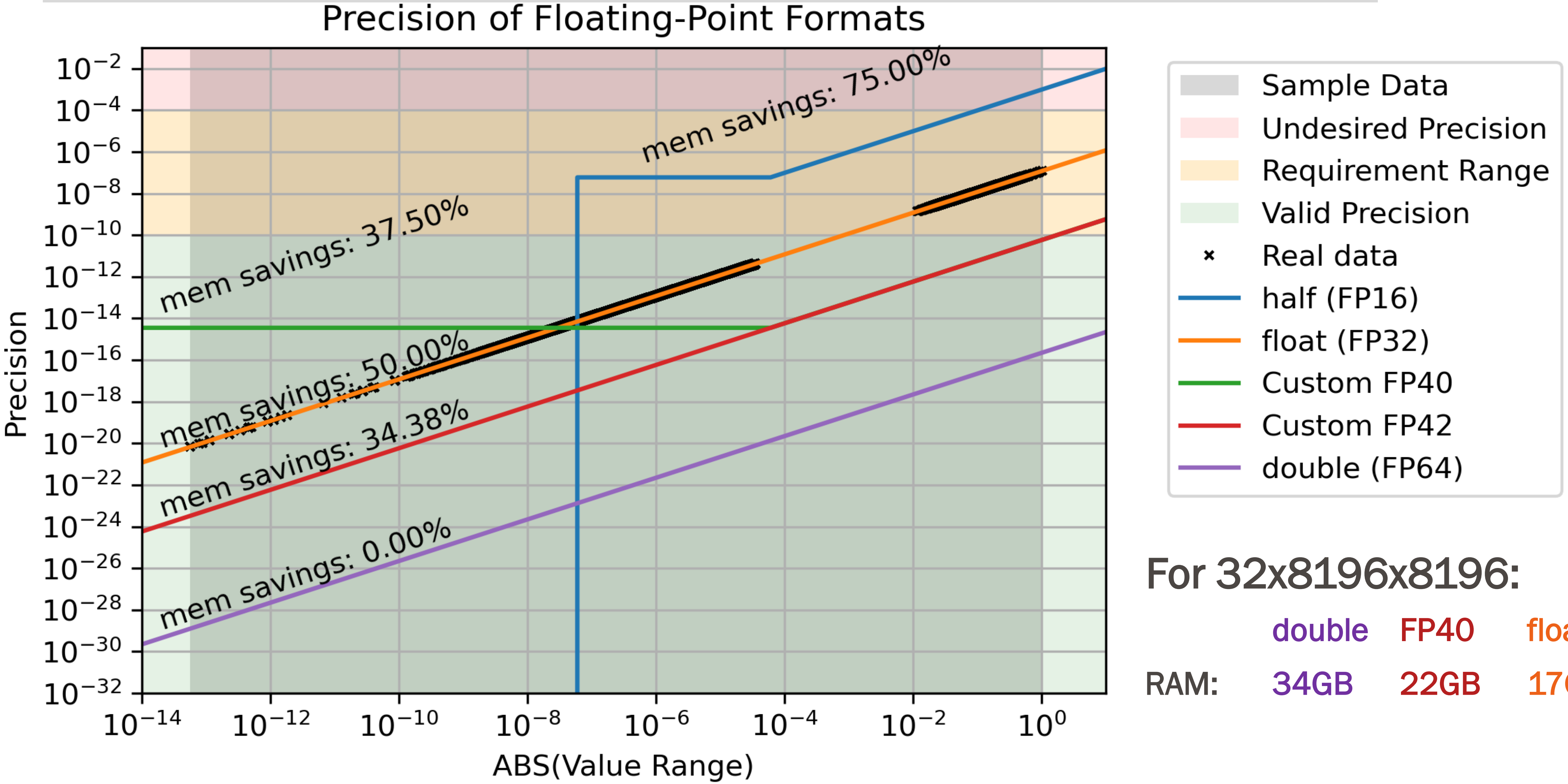


Benefits of reduced precision:

- Energy
- Throughput
- Latency
- **Storage**

Precision in for FINUFFT Synthesis (BIPP)

Sample data extracted from bipp execution, simulated with OSKAR for SKA-Low configuration

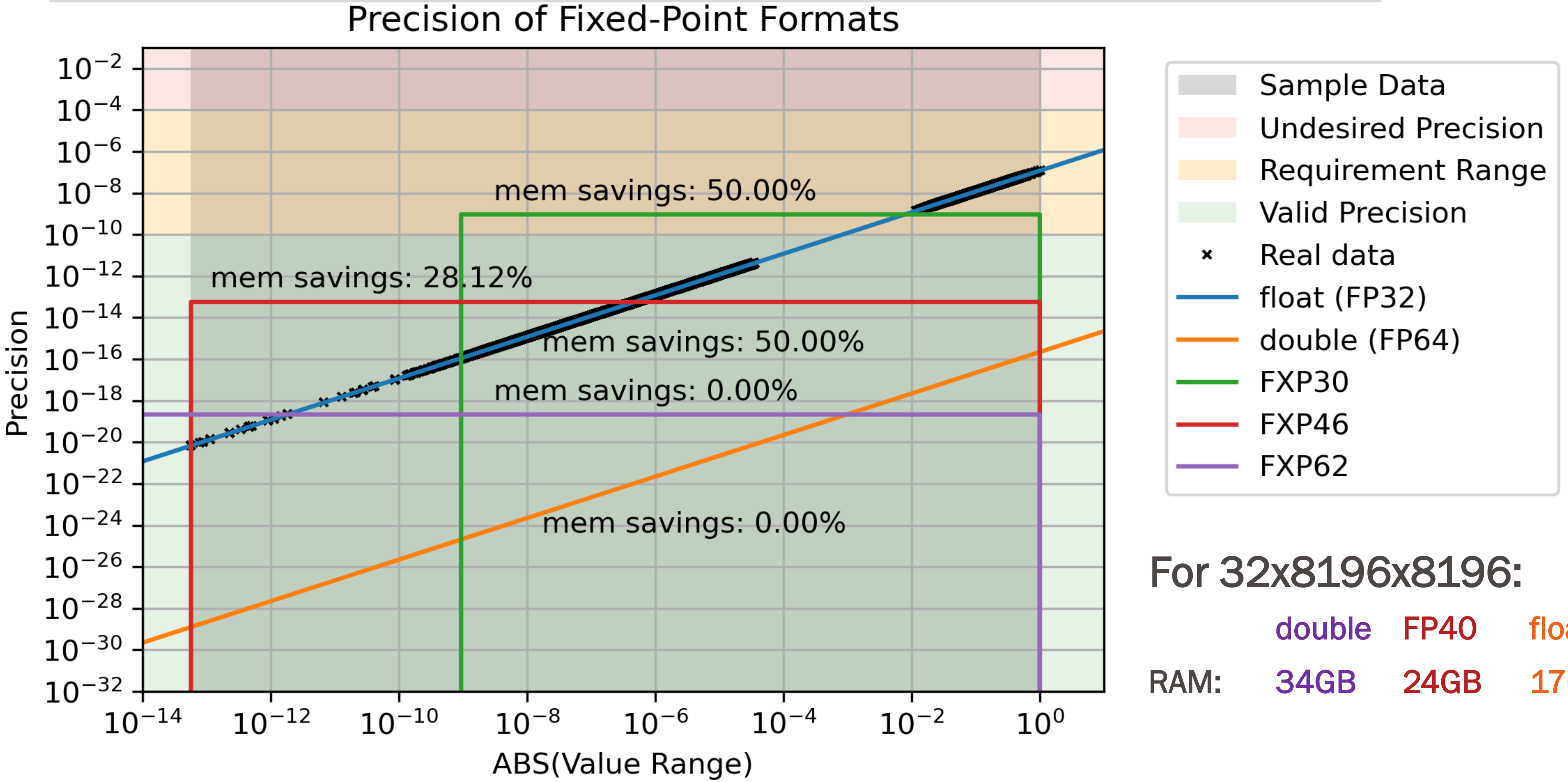


For 32x8196x8196:

	double	FP40	float
RAM:	34GB	22GB	17GB

Precision in for FINUFFT Synthesis (BIPP)

Sample data extracted from bipp execution, simulated with OSKAR for SKA-Low configuration



For 32x8196x8196:

	double	FP40	float
RAM:	34GB	24GB	17GB

Conclusion & Follow Up

Done

- Deploy flexible algorithms using FPGAs
- Accelerate kernels with an FPGA
- Explore the Design Space
- Share resources among different kernels

Ongoing Exploration

- FPGAs improve the energy consumption
- FPGAs match (even increase) the performance of GPUs
- Custom precision data formats are beneficial
- Solve memory-bounded workloads in FPGAs
- Reconfigure the FPGA at run-time for dynamic workloads

Inputs Needed

- Other algorithms to accelerate (i.e. ML)
- Dynamic range of real data (at different stages)
- Precision & latency requirements for different use cases
- Precision metrics (i.e. SNR)
- Scalability of algorithms

Thank you!

Ruben

EPFL - Embedded Systems Laboratory
ruben.rodriguezalvarez@epfl.ch
denisa.constantinescu@epfl.ch

Backup Slides

Types of NUFFTs algorithms

Method	Spread	FFT	Interpolation
$NUFFT_1$	$N_{vis} \lvert\log \epsilon\rvert^2$	$N_{pix} \log N_{pix}$	N_{pix}
W-gridding	$N_{vis} \lvert\log \epsilon\rvert^3$	$N_{w'} N_{pix} \log N_{pix}$	$N_{w'} N_{pix}$
$NUFFT_3$	$N_{vis} \lvert\log \epsilon\rvert^3 + N_{mesh}$	$N_{mesh} \log N_{mesh}$	$N_{pix} \lvert\log \epsilon\rvert^3 + N_{pix}$

Kashani, Sepand, et al. "HVOX: Scalable Interferometric Synthesis and Analysis of Spherical Sky Maps." *arXiv preprint arXiv:2306.06007* (2023).