Rik Jongerius (IBM NL / ASTRON & IBM Center for Exascale Technology)
December 2012 / CALIM

ASTRON
Netherlands Institute for Radio Astronomy

IBM

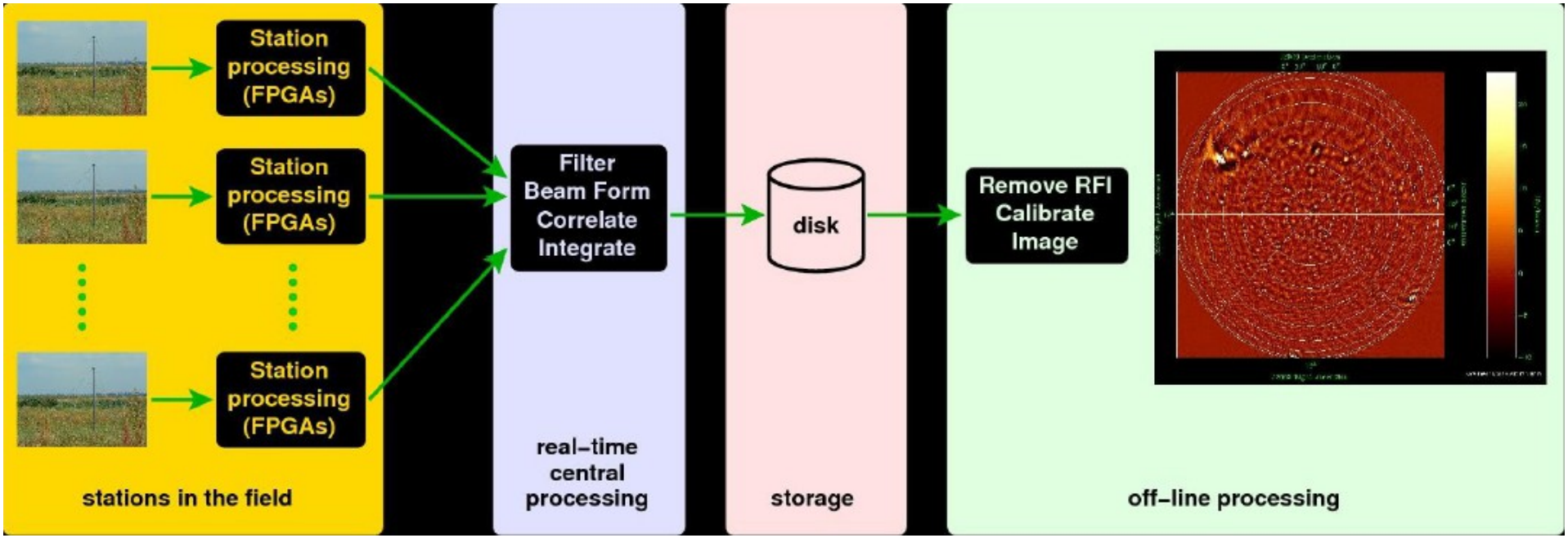
TU/e Technische Universiteit Eindhoven University of Technology

LOFAR Retrospective Analysis
Analyzing LOFAR station processing

Retrospective analysis goals

- Understand **compute and bandwidth distribution** for sky imaging
- How do these properties relate to the **required compute system**?
  - Architecture
  - System size
  - Energy usage
- Use LOFAR as a basis to determine models and scaling rules
- Apply the results of the LOFAR retrospective analysis to the SKA

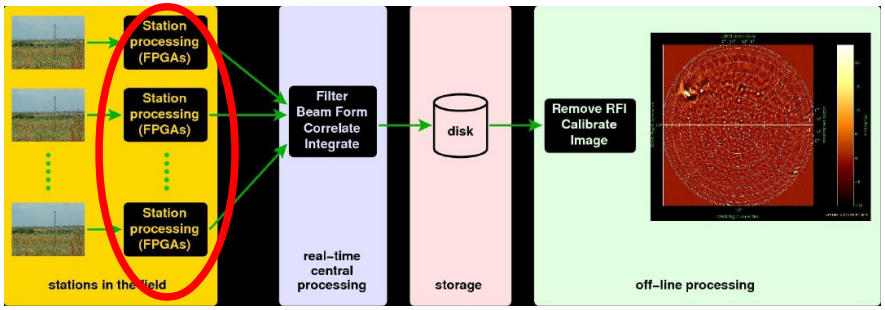
Content

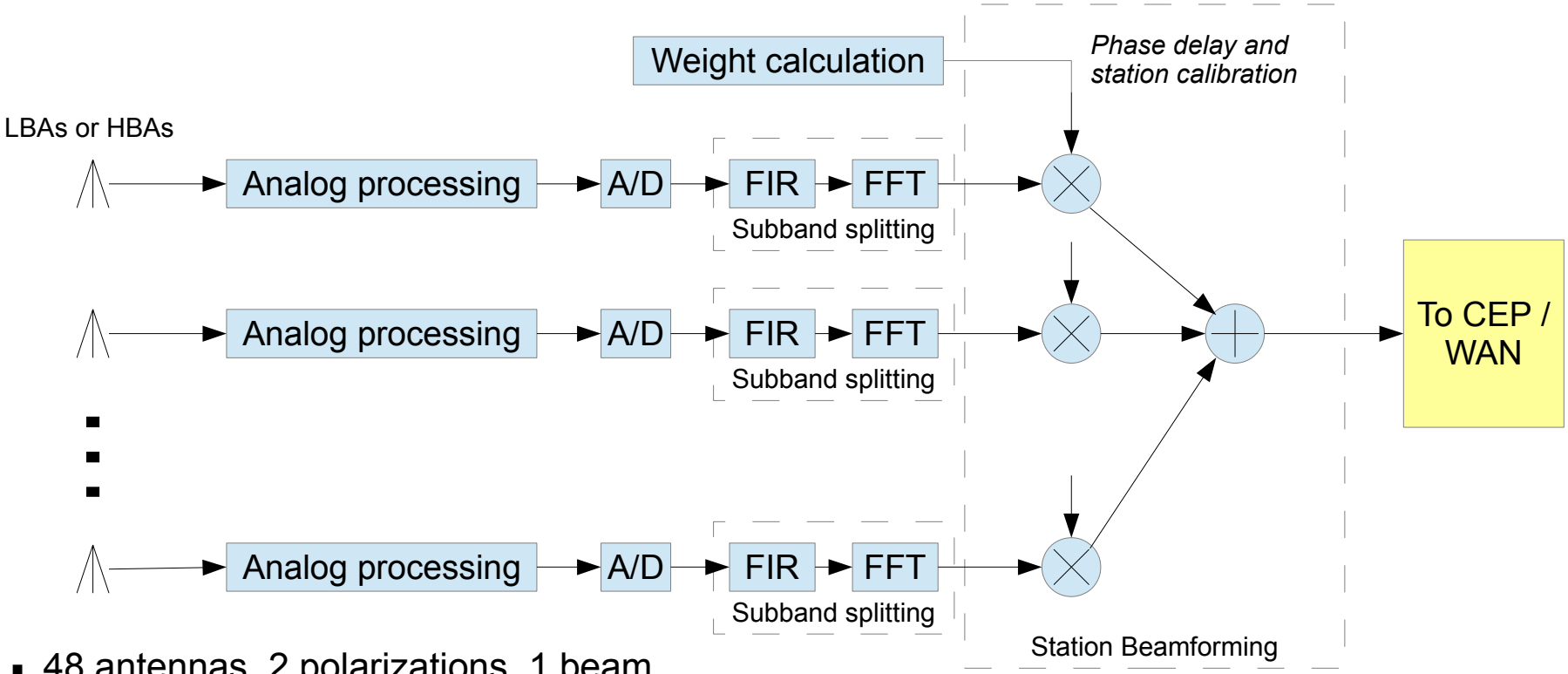


Source: John Romein

- Focus on station and central processing (correlator)
  - Basis to extend the analysis to post-correlator imaging
- Determine computational and bandwidth **scaling rules** for LOFAR digital signal path
- Analysis of **station processing on modern parallel platforms**
  - System size requirements
  - Energy consumption

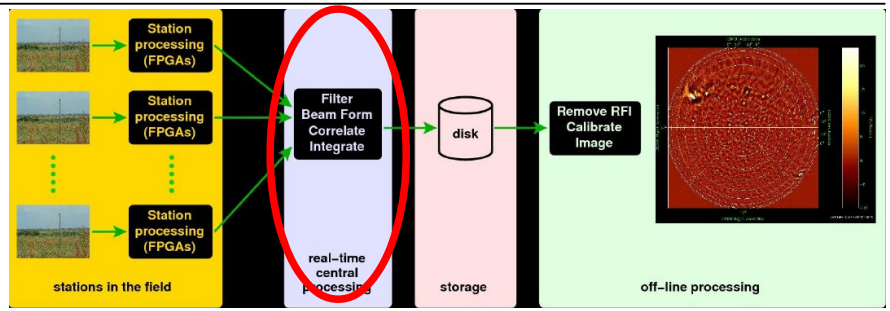
LOFAR station processing

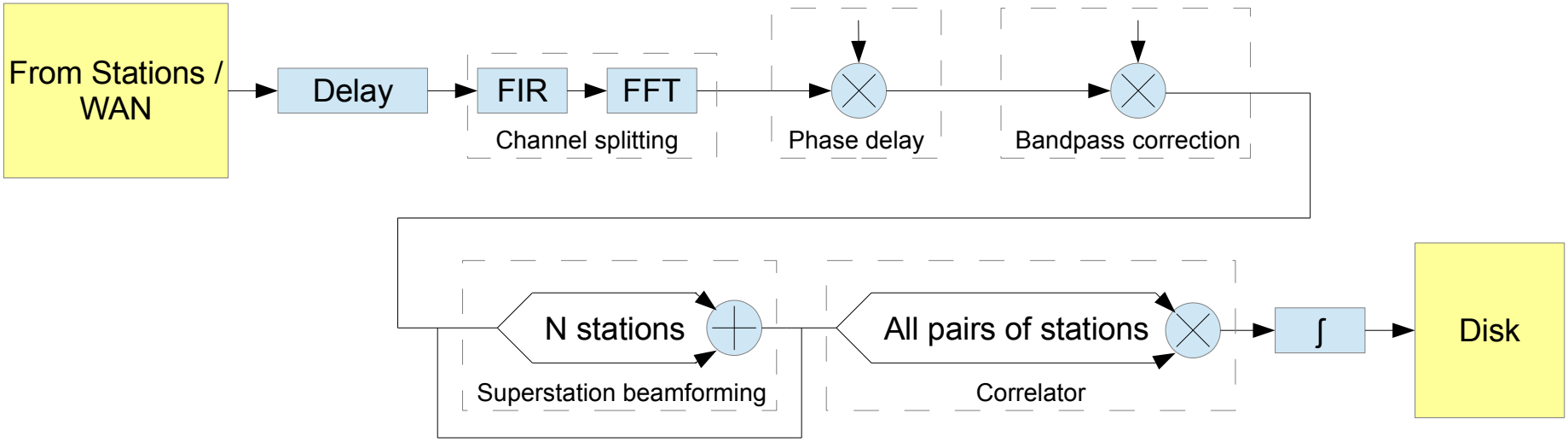




- 48 antennas, 2 polarizations, 1 beam
- 248/512 subbands beamformed

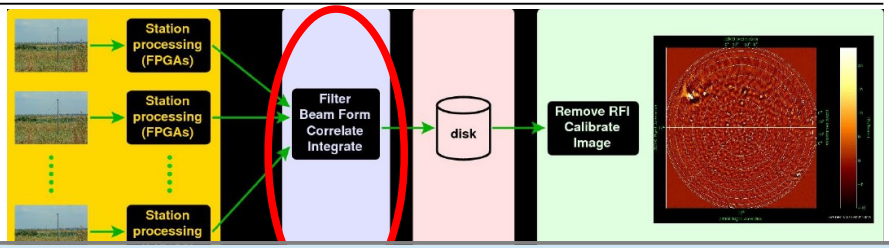
LOFAR central processing



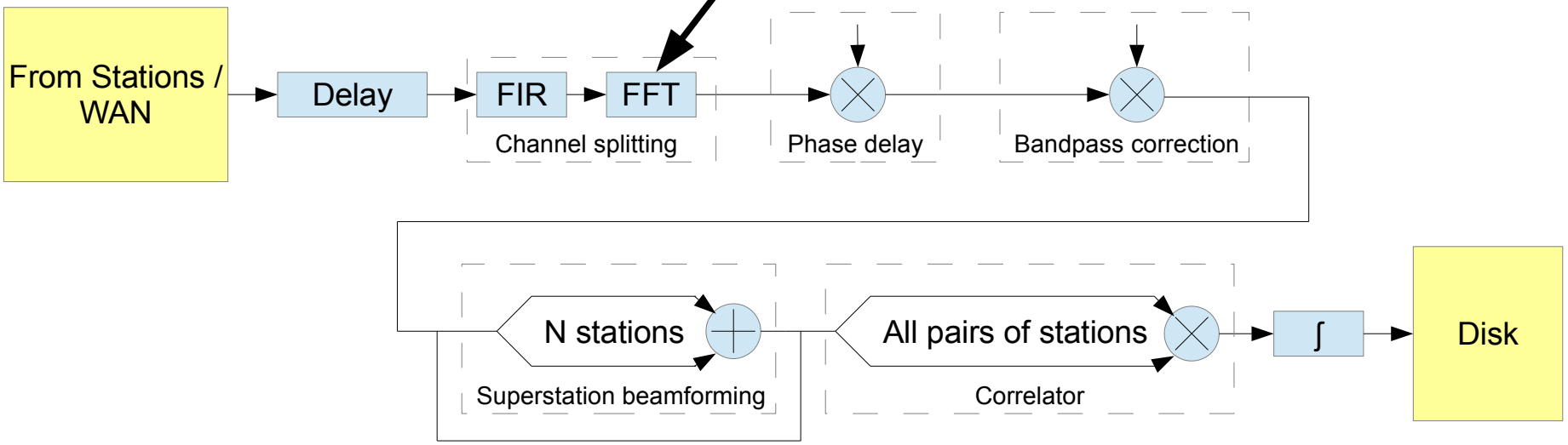


- 64 stations, 2 polarizations, 1 beam
- Channel splitting: 256 channels

LOFAR central processing

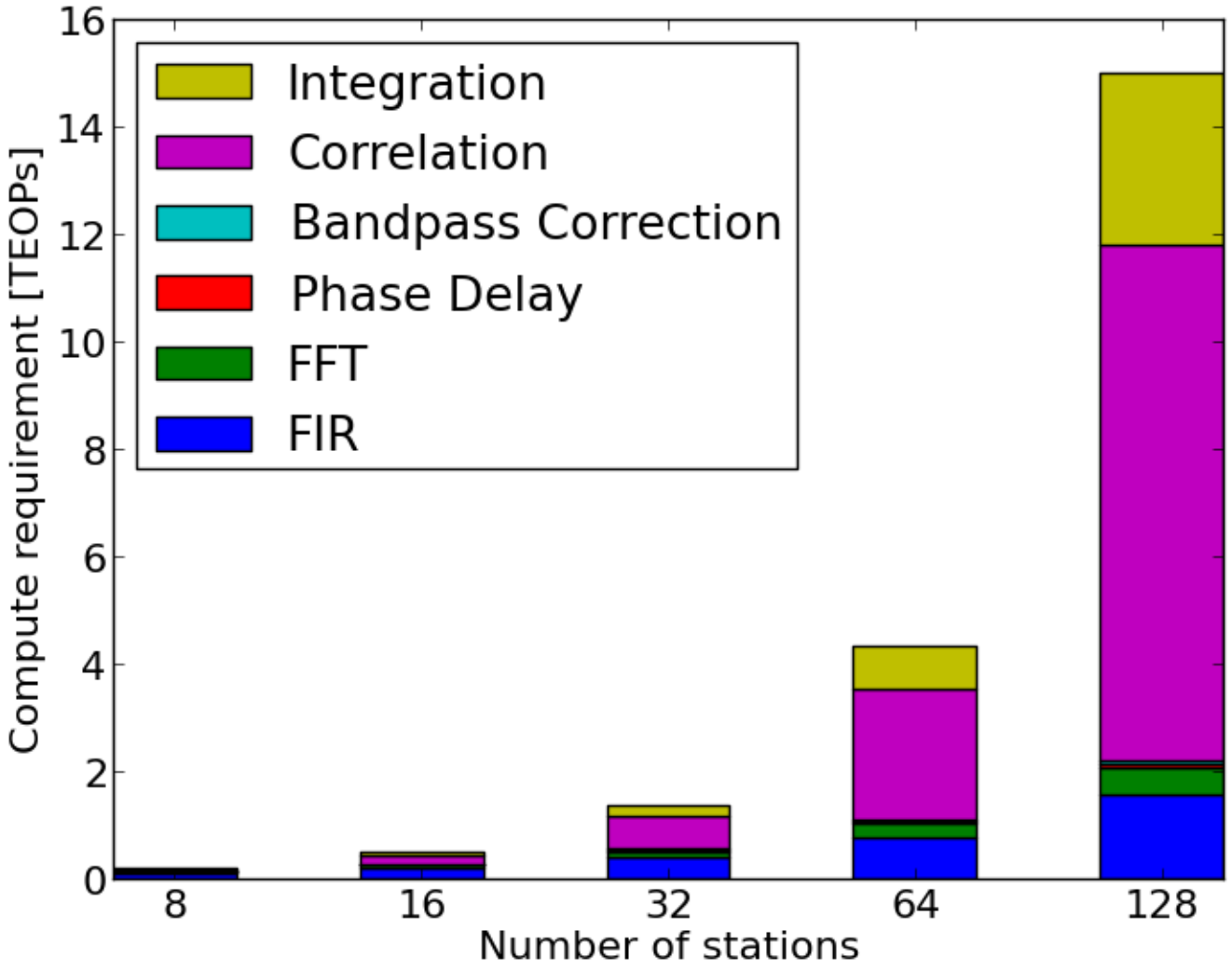


$$N_{ops} = N_{stations} N_{beams} N_{subbands} N_{polarizations} 5 N_{points} \log_2 (N_{points})$$



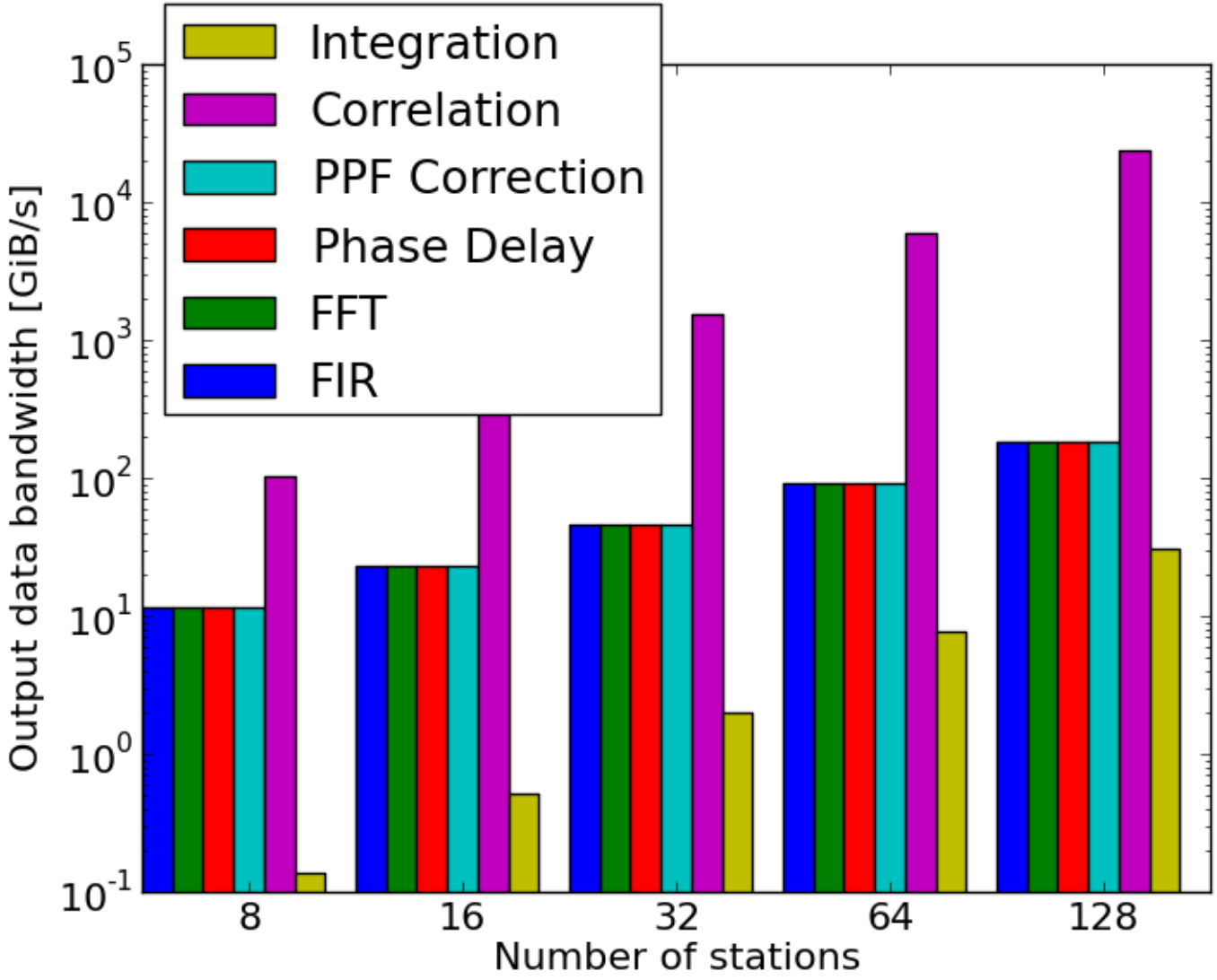
- 64 stations, 2 polarizations, 1 beam
- Channel splitting: 256 channels

Central processing compute scales quadratically

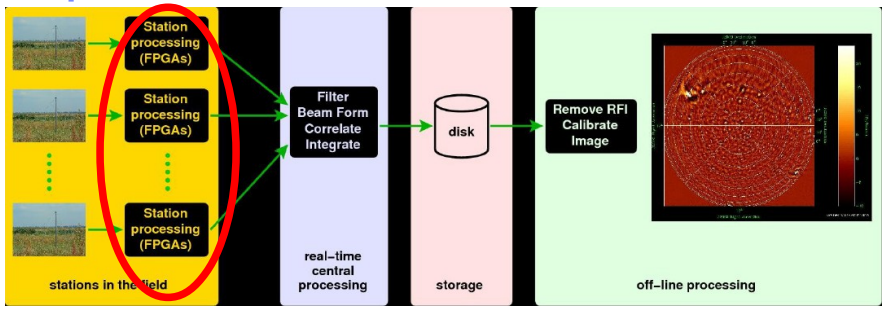


- TEOPS = Tera Effective Operations per second

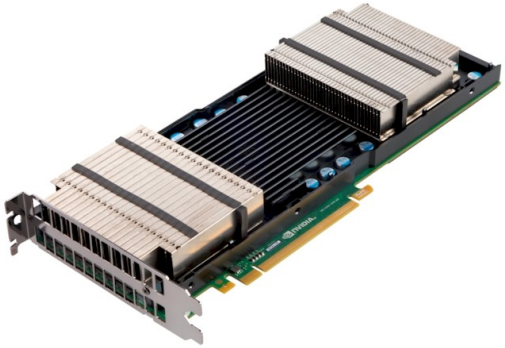
Central processing data reduction by integration

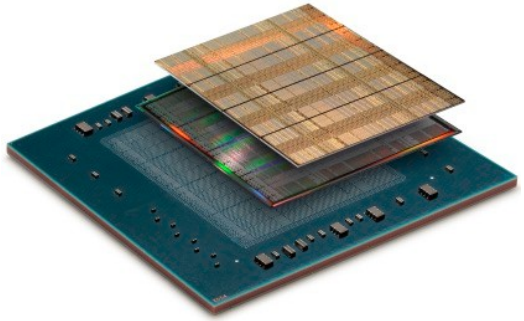


LOFAR station processing on parallel platforms



- Behavior of station processing on **modern, parallel platforms**





- Roofline analysis to extrapolate to full station size

- *How would we design LOFAR station processing, giving modern technologies?*

Fixed-point versus floating-point

- FPGA best performance with **fixed-point arithmetic**
- CPU and GPU best performance with **floating-point arithmetic**
  - Keep integer data path for bookkeeping (e.g. address calculations)

- Fixed-point implementation is more tedious
- 32-bit Floating-point arithmetic has **no impact on accuracy** of station processing

- ***Choose best arithmetic type for each platform***

Station processing: low computational intensity

Block	Parameter	Requirement (32-bit FP)
FIR	Compute	595.2 GFLOPs
	Antenna data in	71.5 GB/s
	FIR data out	71.5 GB/s
	Coefficients, delayed samples in/out	2.27 TB/s
	Computational intensity	0.25 ops/byte
FFT	Compute	480 GFLOPs
	FIR data in	71.5 GB/s
	FFT data out	71.5 GB/s
	Twiddle factors	357.5 GB/s
	Computational intensity	1.1 ops/byte
BF	Compute	74.2 GFLOPs
	FFT data in	34.6 GB/s
	BF data out	739 MB/s
	Calibration data	138.6 GB/s
	Computational intensity	0.43 ops/byte
Total	Compute	1.1 TFLOPs
	Total data bandwidth	3.0 TB/s
	Computational intensity	0.4 ops/byte

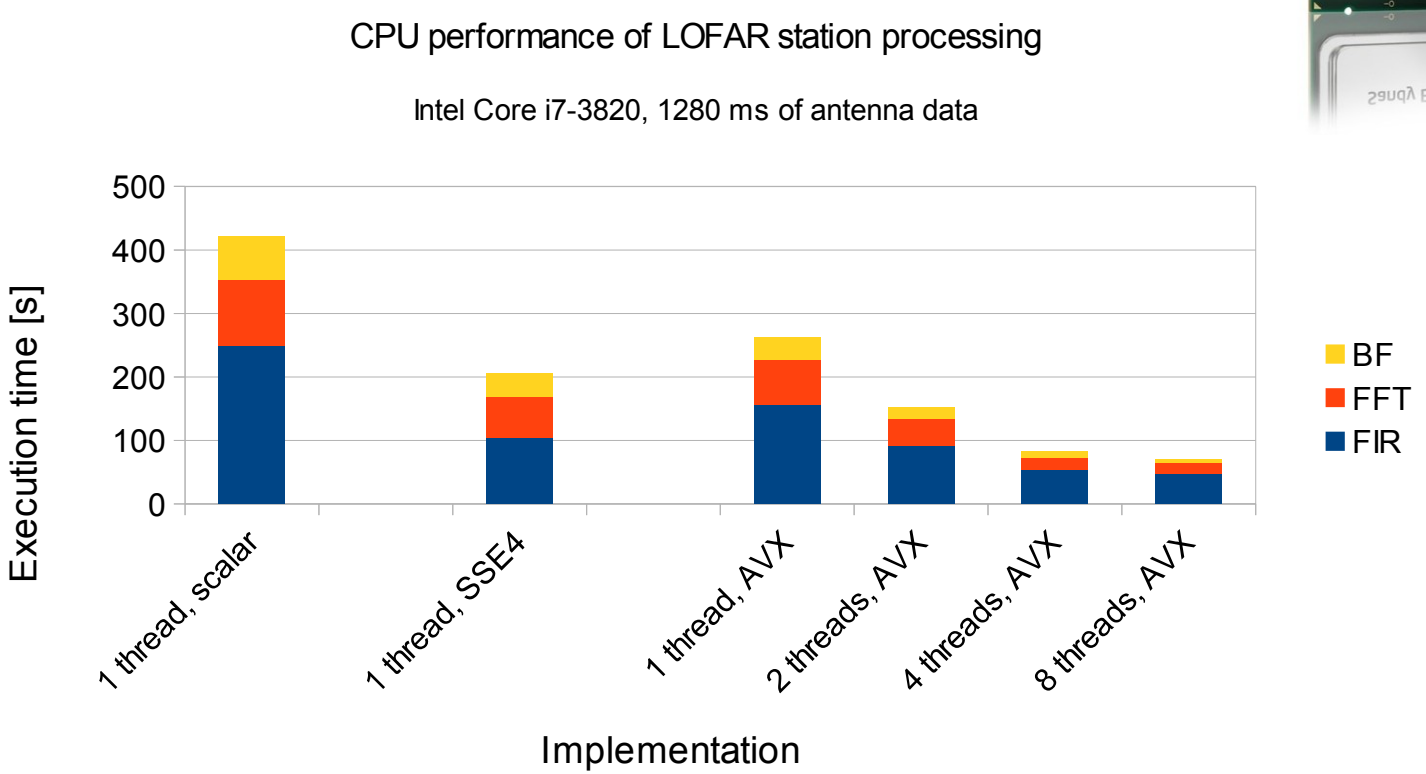
- Operational intensity: 59.9 ops/sample

GPU cannot reach peak performance for station processing

Type	Platform	Introduction	Peak performance	Off-chip bandwidth
CPU	Quad-core Intel i7-3820	2012	230.4 GEOPs	40 GB/s
			Quad-core, AVX	*40 lanes PCI-express 3.0*
GPU	Nvidia K10	2012	2 x 2288 GEOPs	16 GB/s
			Dual GK104	*16 lanes PCI-express 3.0*
FPGA	Xilinx Virtex-4 SX35	2004	76.8 GEOPs	11.2 GB/s
			192 DSP slices	*112 LVDS pairs*
FPGA	Xilinx Virtex-7 x690T	2010	1440 GEOPs	131 GB/s
			3600 DSP slices	*80 GTH transceivers*

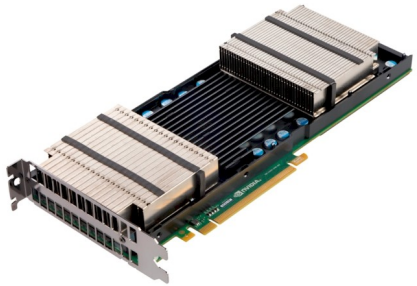
Type	Platform	Operational intensity	Required operational intensity	Max performance
CPU	Intel i7-3820	5.76 ops/byte	29.9 ops / byte (16 bit / sample)	100 %
GPU	Nvidia K10	286 ops/byte	29.9 ops / byte (16 bit / sample)	10.9 %
FPGA	Virtex-4 SX35	6.8 ops/byte	39.9 ops / byte (12 bit / sample)	100 %
FPGA	Virtex-7 x690T	11 ops/byte	39.9 ops / byte (12 bit / sample)	100 %

Single CPU performance 55.5x below real-time



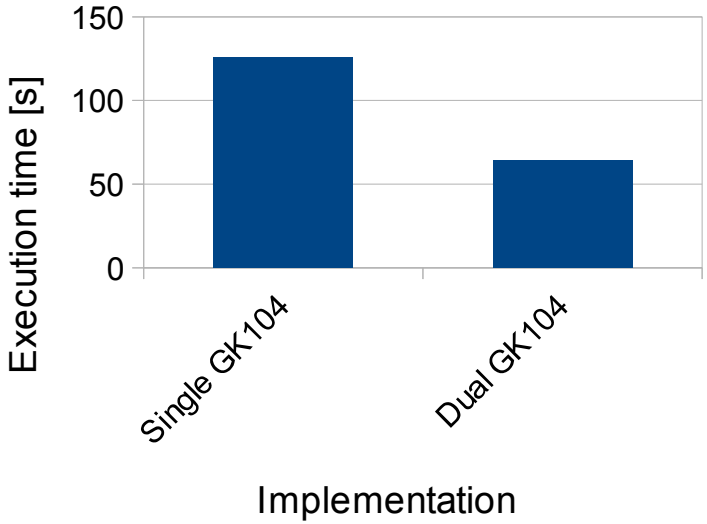
- Single-threaded AVX implementation worse compared to SSE4
  - *Different vectorization strategy required for scaling to multiple threads*

GPU performance 6.1x below real-time



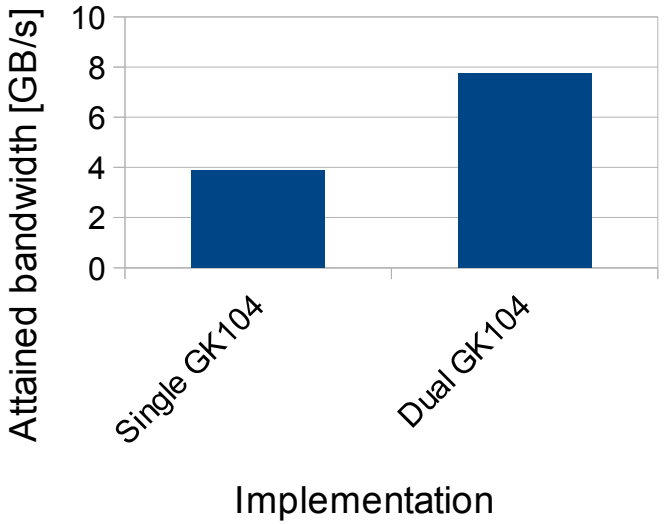
GPU performance of LOFAR station processing

Nvidia K10, 10240 ms of antenna data



Attained PCI-express bandwidth

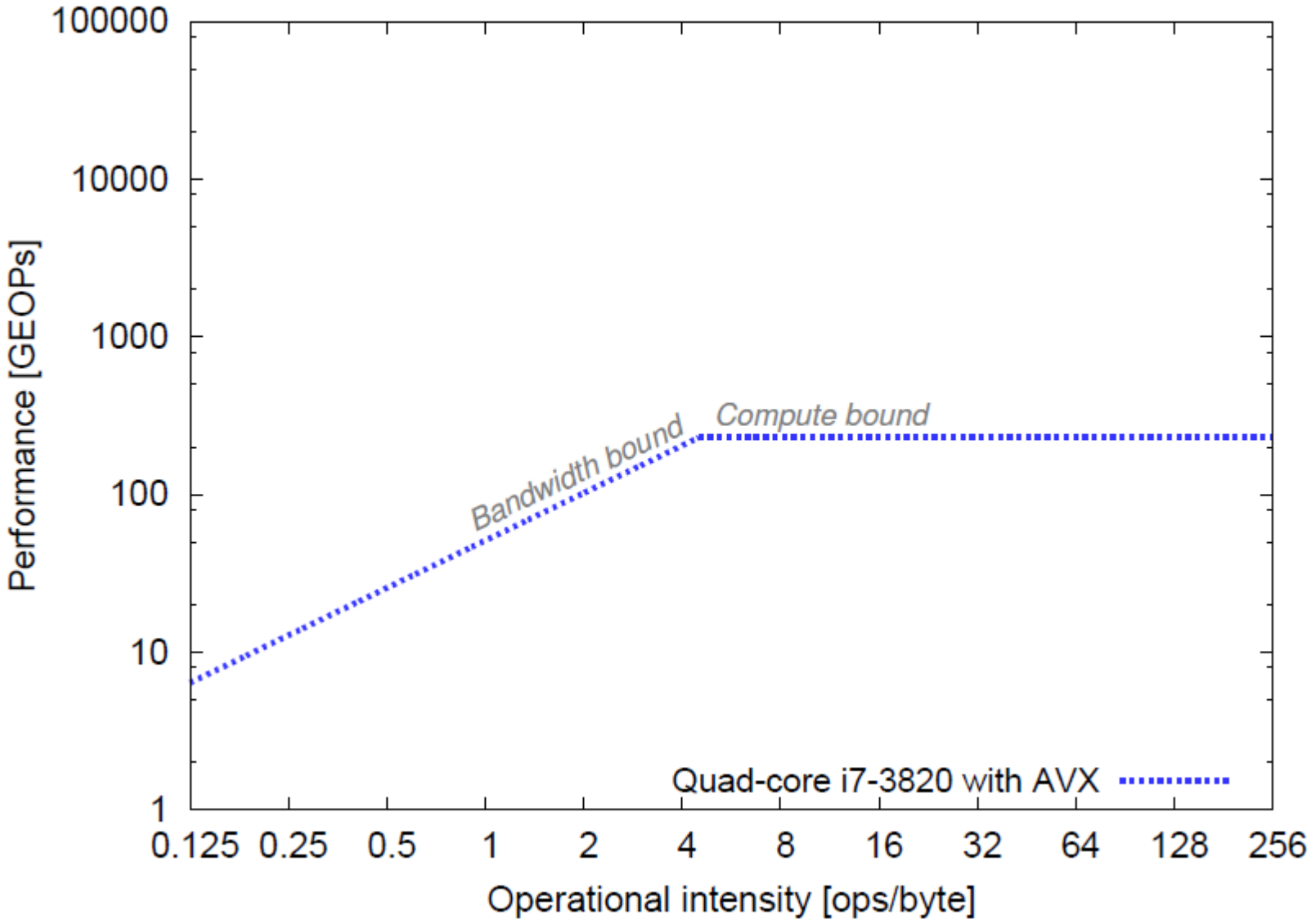
Nvidia K10



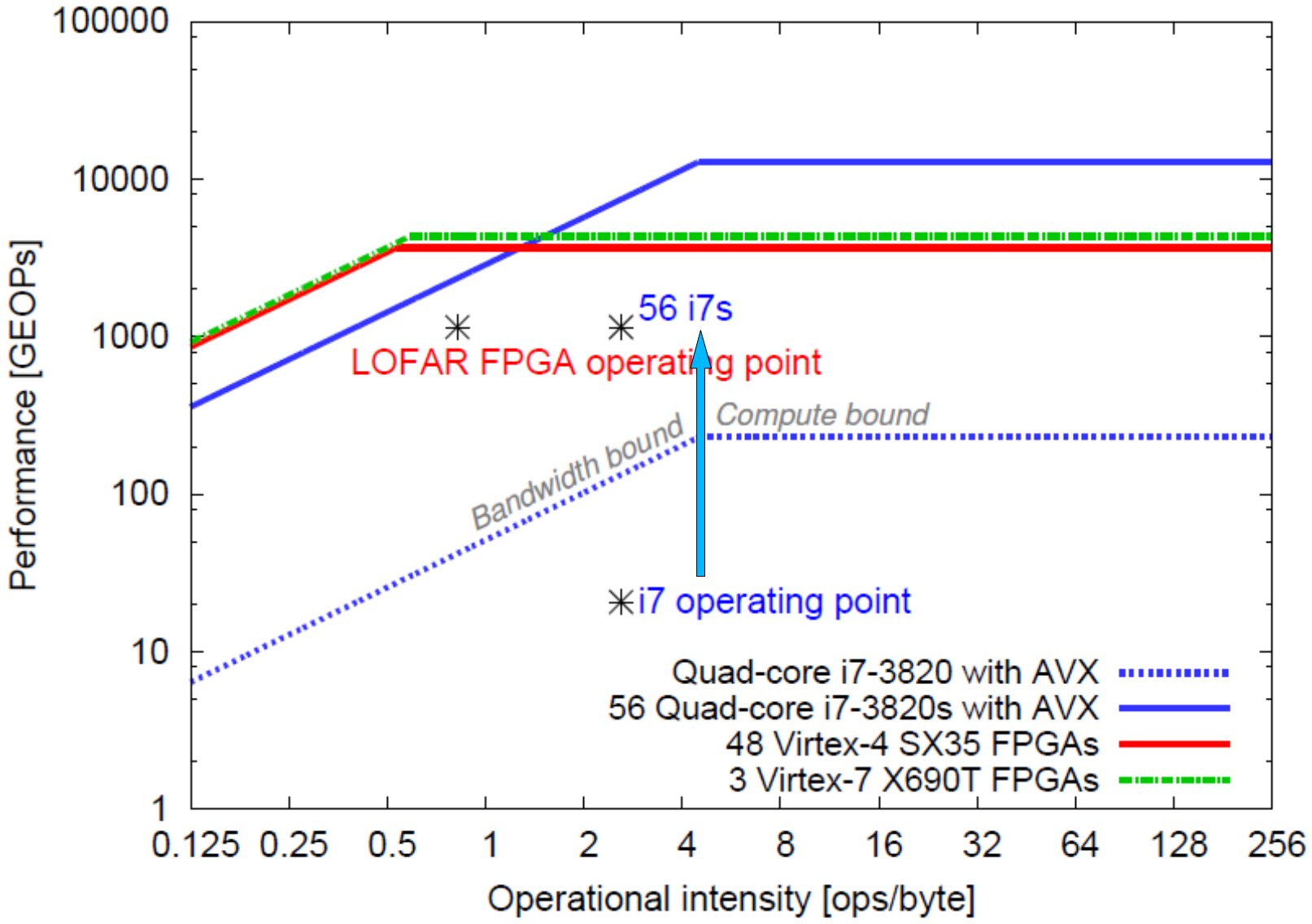
	GFLOPs	Reached % of peak performance	Theoretical max. % of peak performance
Single GK104	93.4	4.1 %	21.8 %
Dual GK104	183.2	4.0 %	10.9 %

- 8.8x improvement over Intel Core i7-3820 (8 threads, AVX)

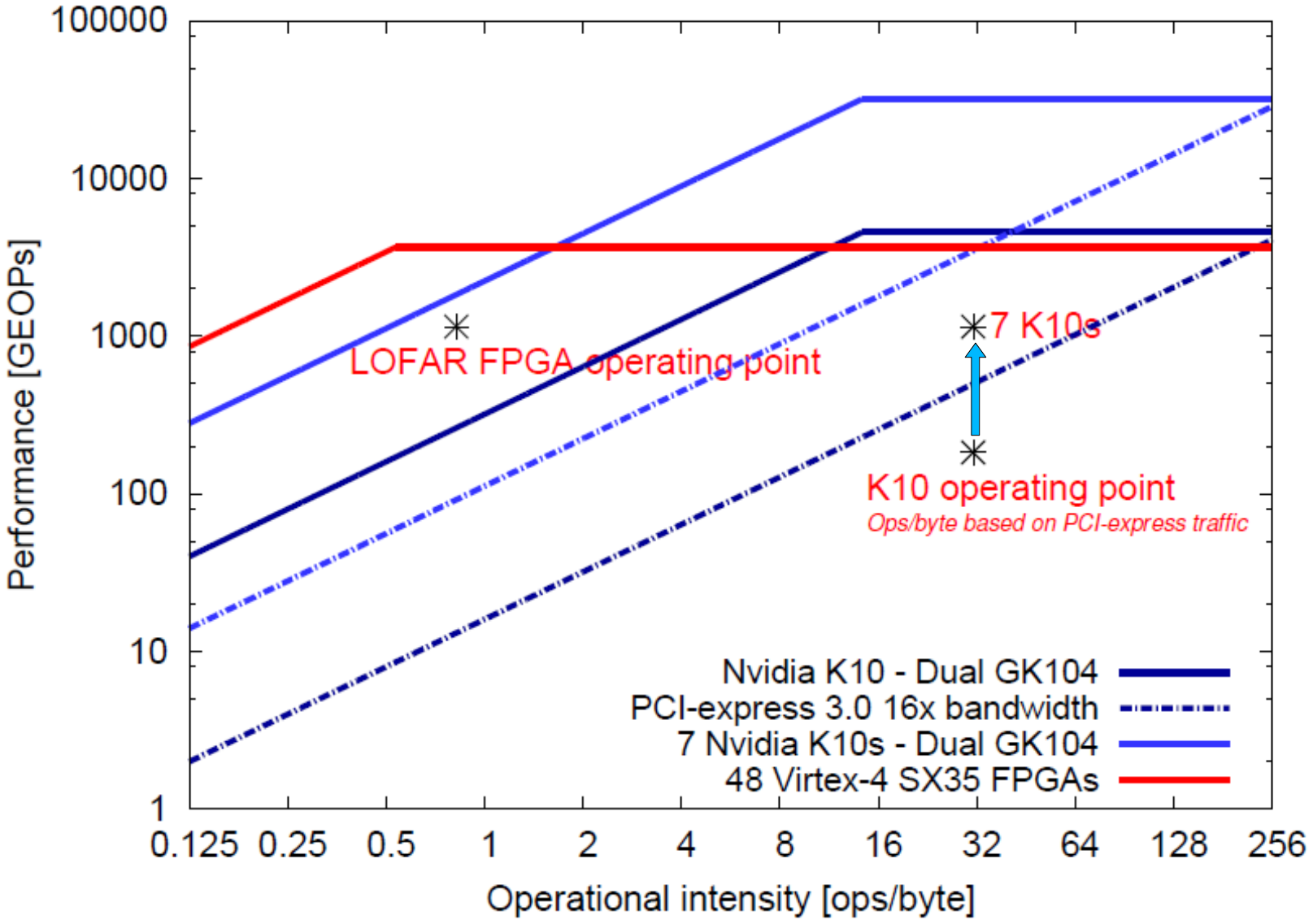
Roofline analysis



Roofline analysis: 3 Virtex-7s or 56 Intel Core i7s



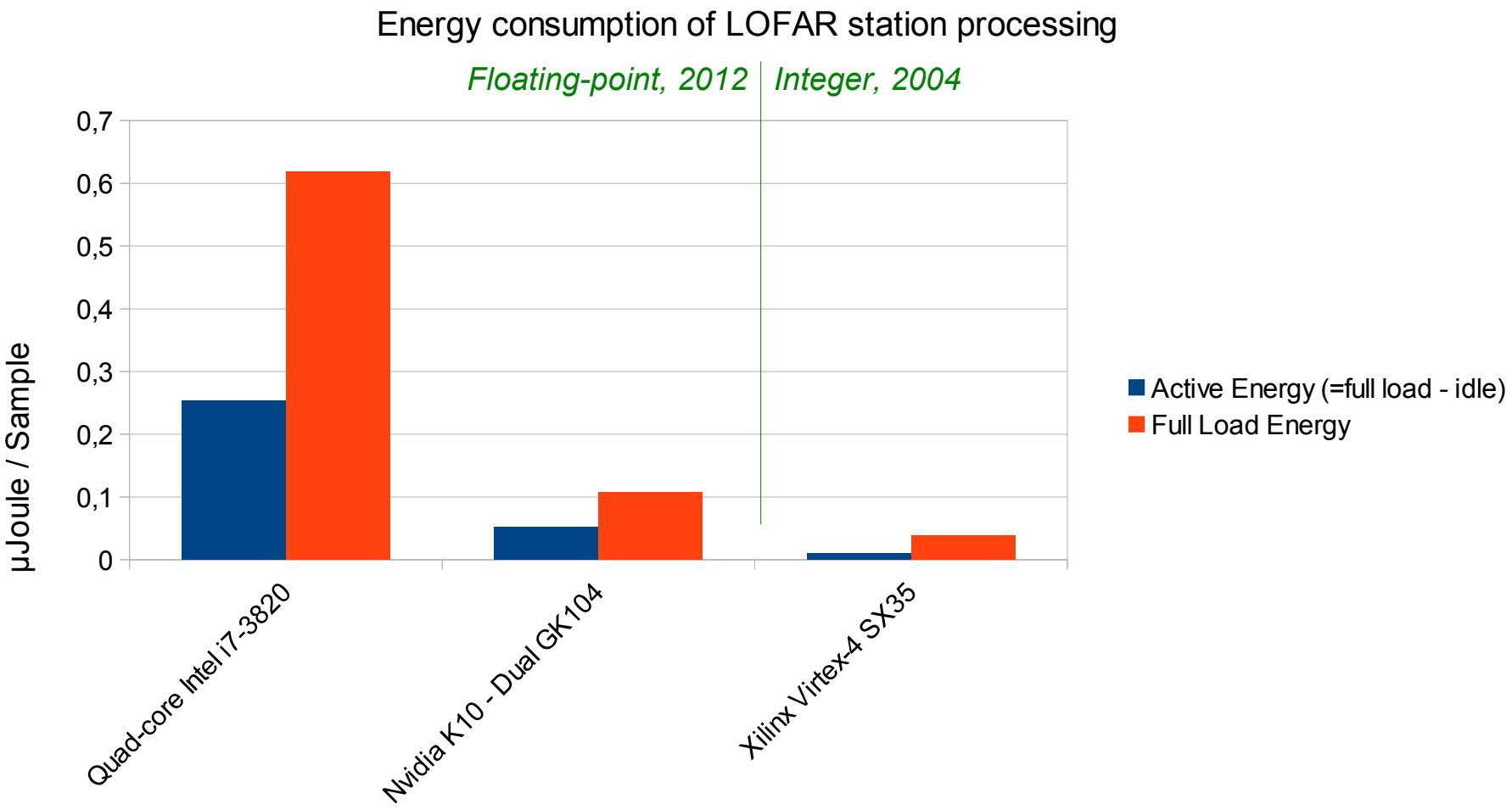
Roofline analysis: 7 Nvidia K10s



Power measurement setup

- Compare power consumption of complete system under full load
  - Favors FPGA implementation (no unused peripherals)
  - **Lower bound** on efficiency
- Compare power consumption of difference between full and idle load
  - Only compare active power
  - Does not include idle power of CPU/GPU/FPGAs
  - **Upper bound** on efficiency

FPGAs use least energy



Conclusions

- Algorithms in station processing have low operational intensity
  - Impossible to reach peak performance on certain platforms
- Streaming data model and low operational intensity favor FPGAs
  - Relatively high off-chip bandwidth
- For station processing use FPGAs
  - Twice as many GPUs as FPGAs
  - FPGAs are ~5x more power efficient compared to GPUs
    - 2004 FPGA platform vs 2012 GPU
    - Integer arithmetic on FPGAs vs floating-point

Future work

- Extend analysis to post-correlation calibration and imaging
  - Determine scaling rules
  - Map results onto hardware platforms
- Algorithms expected to have higher operational intensity
  - Possible to get closer to peak performance on GPUs
  - GPUs expected to be (more) competitive with FPGAs
- Include ASICs and DSPs as a potential low-energy hardware platform

Questions?

Backup slides

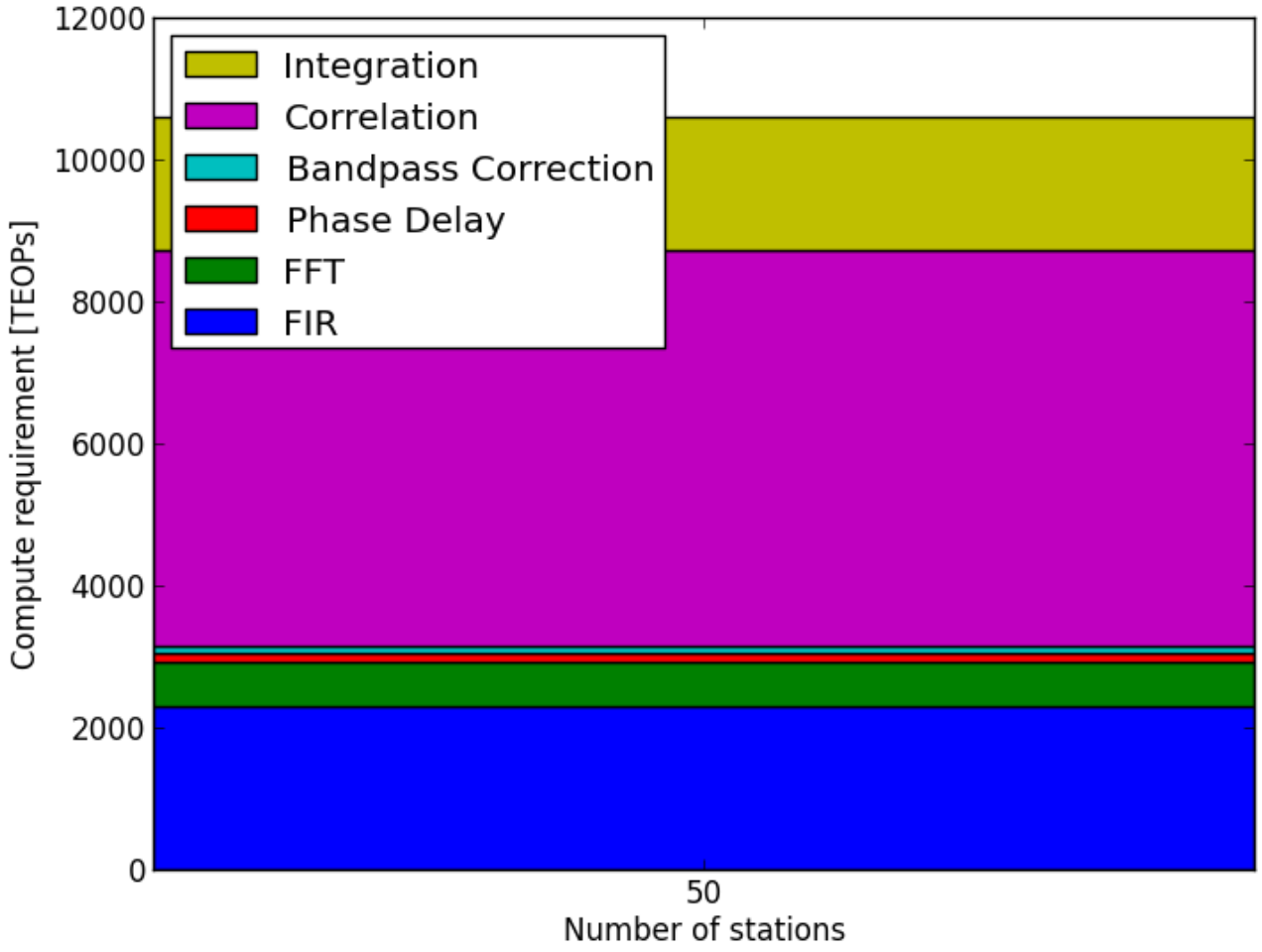
SKA phase 1 – Memo 130 [I]

- AA-Low
- 11.200 antennas/station, 50 stations
- 3040 subbands, 125 channels, 480 beams
- 380 MHz bandwidth

- Preliminary results

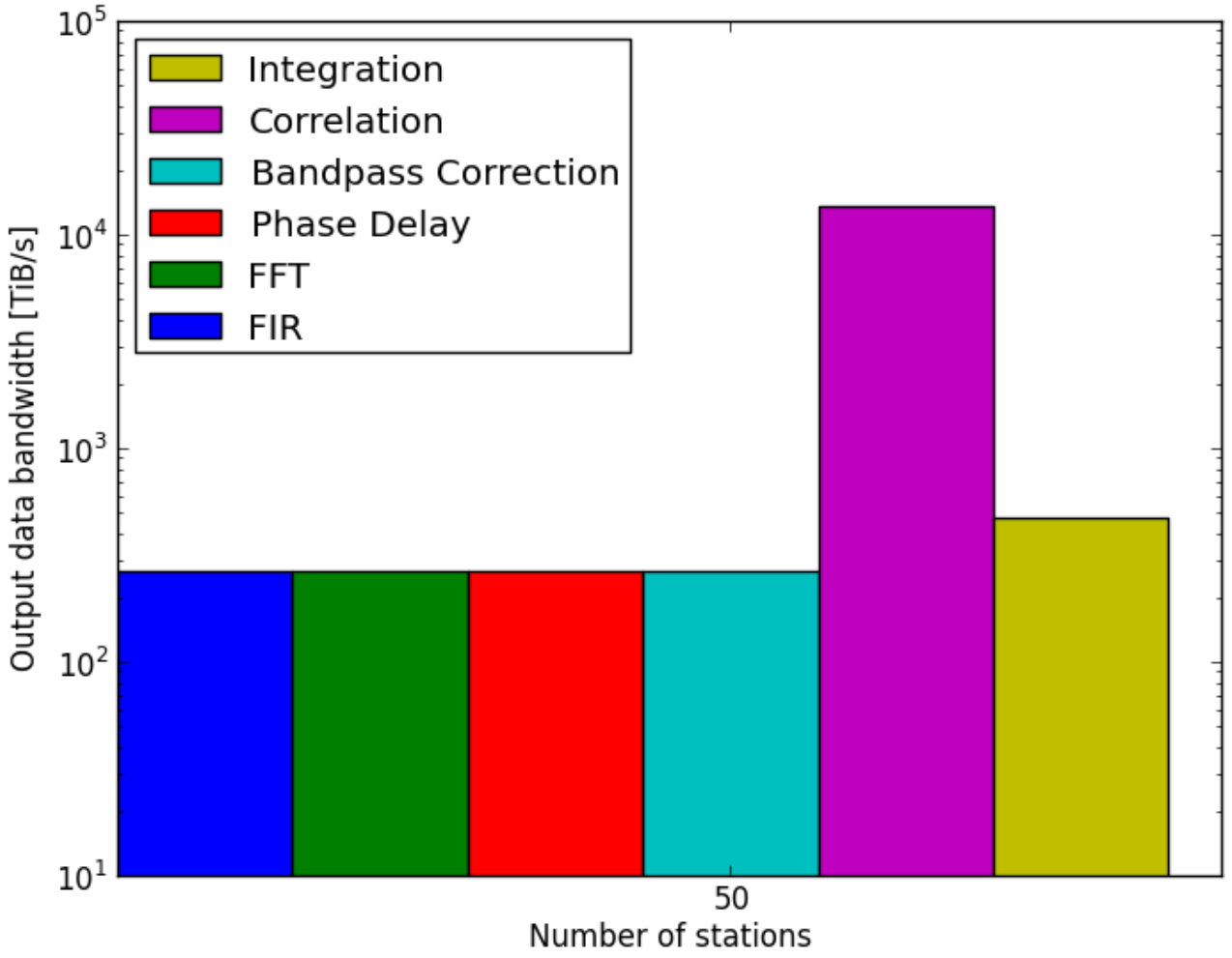
SKA phase 1 – Memo 130 [II]

- Central processor: compute (preliminary results)



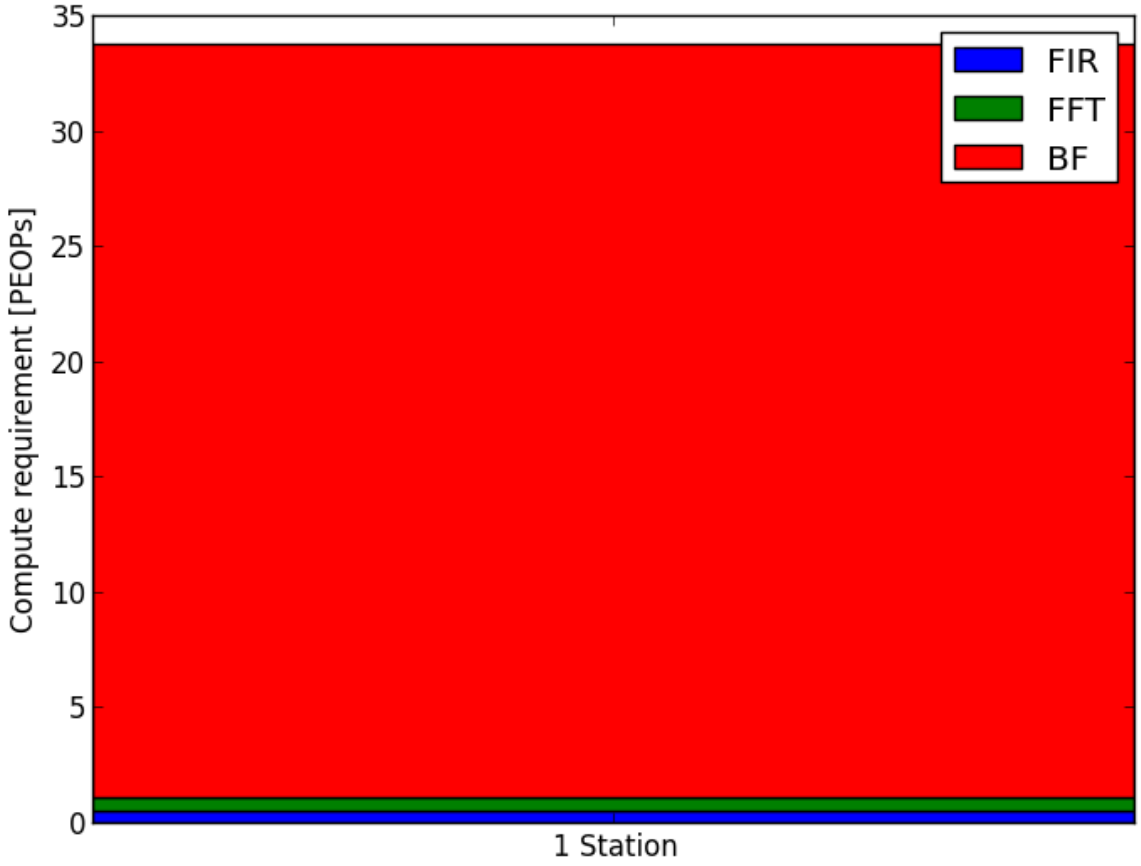
SKA phase 1 – Memo 130 [III]

- Central processor: data rates (preliminary results)



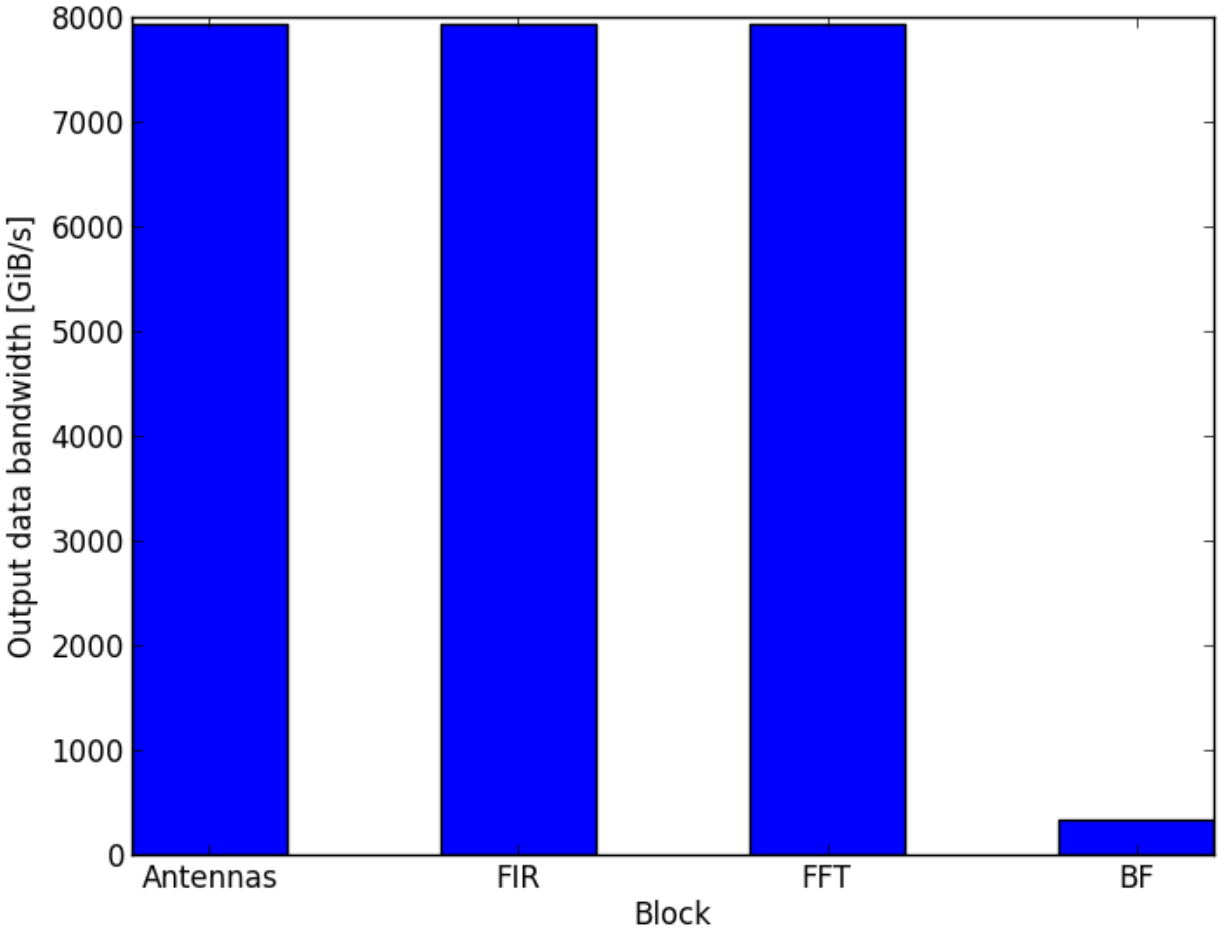
SKA phase 1 – Memo 130 [IV]

- Station processing: compute (preliminary results)



SKA phase 1 – Memo 130 [V]

- Station processing: data rates (preliminary results)



Platform selection — CPU

- Quad-core Intel Core i7-3820 @ 3.6 GHz
  - Launched Q1 2012
  - AVX extensions (256 bit- or 8 float-wide SIMD)
- Theoretical peak performance (quad-core , AVX)
  - 230.4 GFLOPs

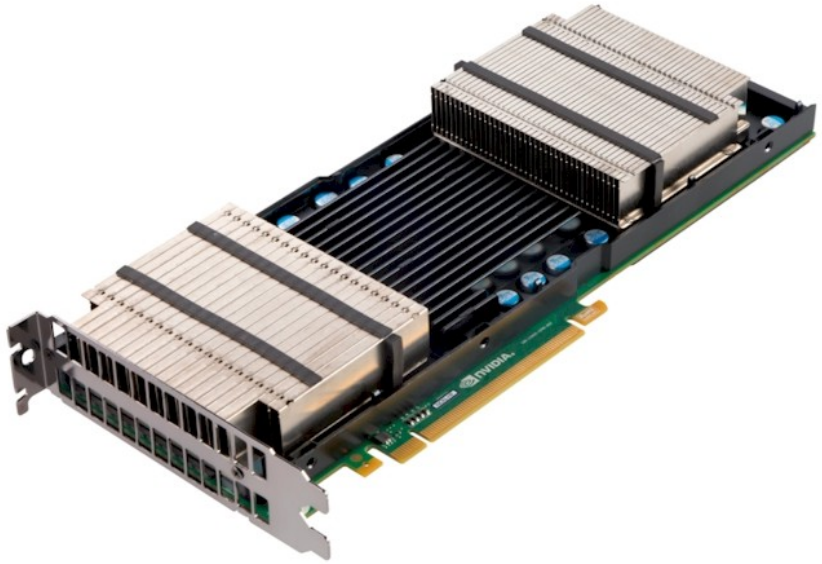
	DDR3	PCI Express 3.0 40 lanes
Theoretical peak BW	51.2 GB/s	40 GB/s
Operational intensity	4.5 ops/byte	5.76 ops/byte

- For **32 bit / sample**: 15.0 ops/byte, BW sufficient to sustain peak performance

Source: ark.intel.com

Platform selection — GPU

- Nvidia Tesla K10
  - Launched May 2012
  - Dual Kepler GK104 GPU



- Theoretical peak performance
  - 2288 GFLOPs per GPU

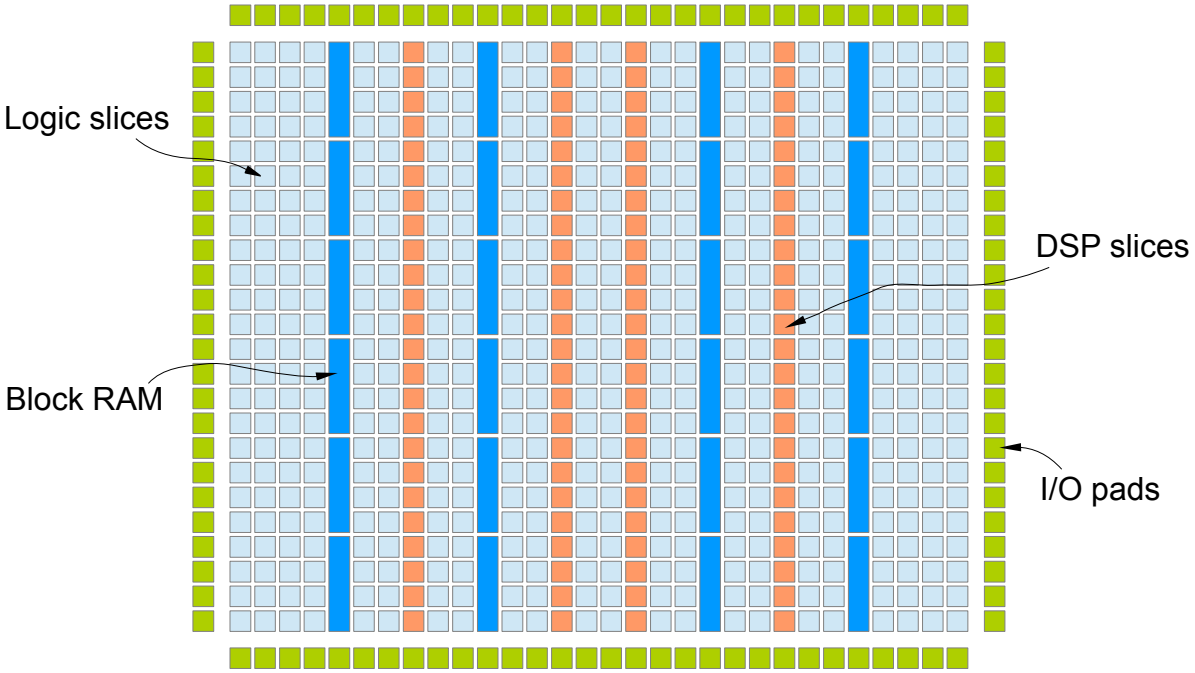
	GDDR5	**PCI Express 3.0 16 lanes**
Theoretical peak BW	160 GB/s	16 GB/s
Operational intensity	28.6 ops/byte	286 ops/byte

- For **16 bit / sample** over PCI-express:
  - 29.9 ops/byte for station processing: max. 10.9 % of peak performance

Source: www.nvidia.com

Platform selection — FPGA [I]

- Fixed-point performance: **giga effective operations per second** (GEOPs)
- Peak performance based on available DSP slices
- Peak memory bandwidth based on available block RAMs
  - Intermediate data fits in block RAMs for FPGA platforms used in the analysis



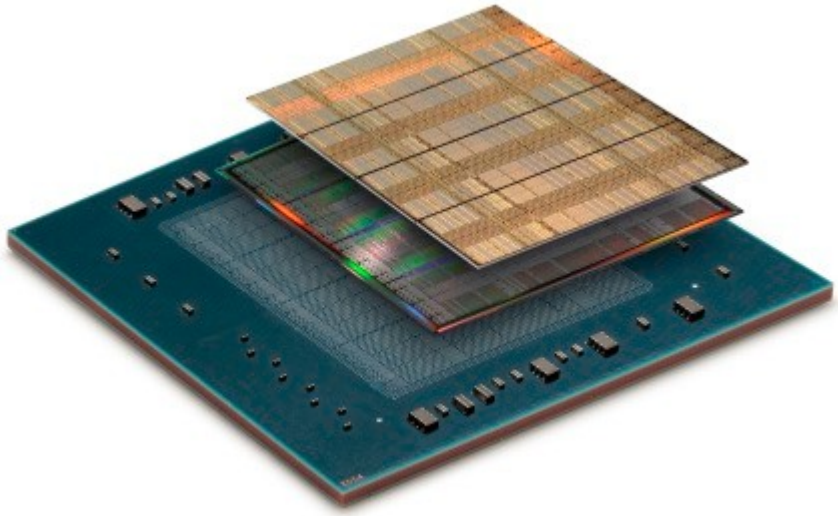
Platform selection — FPGA [II]

- In the field: Xilinx Virtex-4 SX35
  - Launched 2004
  - 192 DSP slices
  - 192 block RAMS
- Logic clocked @ 200 MHz
- Theoretical peak performance (DSP slices):
  - 76.8 GEOPs
- Theoretical peak memory bandwidth (block RAM):
  - 143 GB/s

	112 LVDS pairs
Theoretical peak BW	11.2 GB/s
Operational intensity	6.8 ops/byte

- For **12 bit / sample:** 39.9 ops/byte, BW sufficient

Platform selection — FPGA [III]



- Xilinx Virtex-7 x690T
  - Launched June 2010
  - 3600 DSP slices
  - 2940 block RAMS
- Assume logic clocked @ 200 MHz
- Theoretical peak performance (DSP slices):
  - 1440 GEOPs
- Theoretical peak memory bandwidth (block RAM):
  - 2352 TB/s

	80 GTH transceivers
Theoretical peak BW	131 GB/s
Operational intensity	11 ops/byte

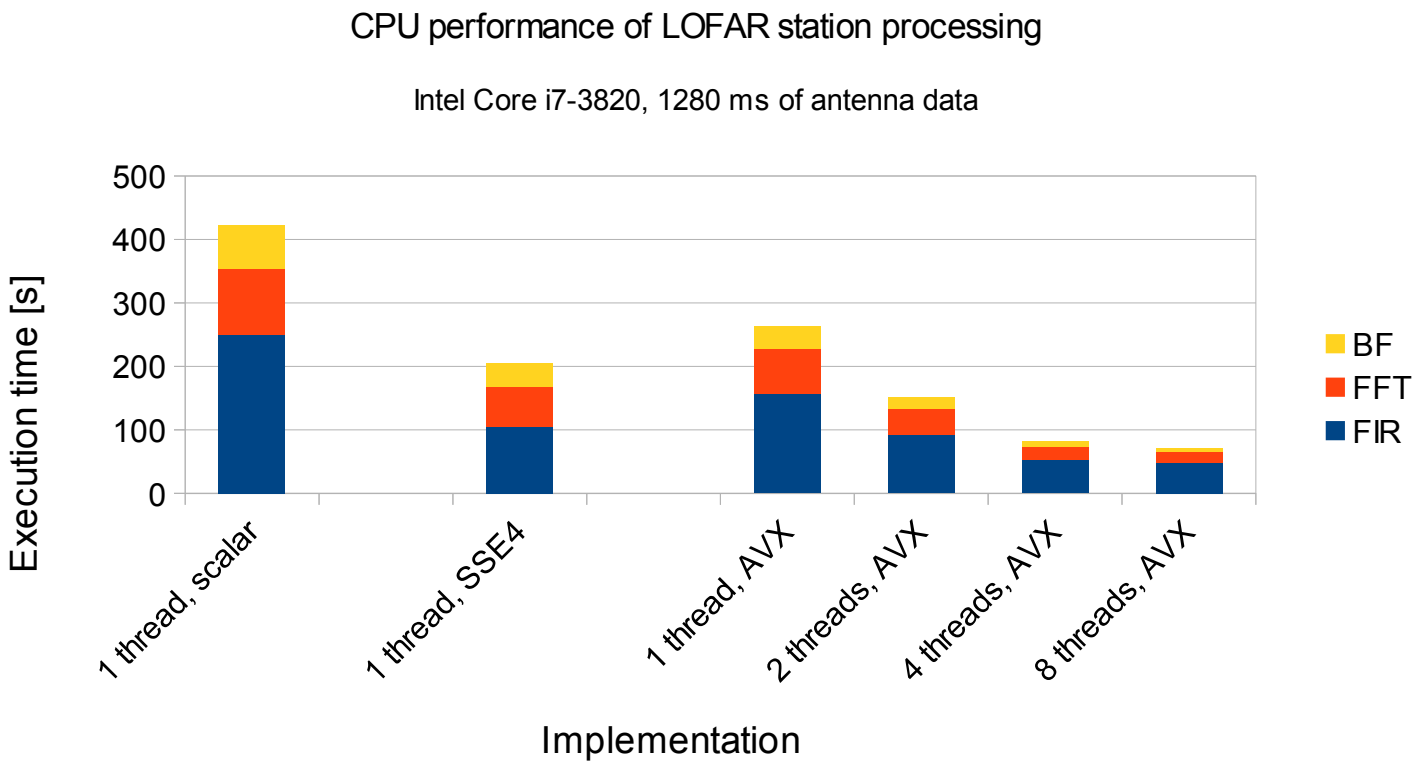
- For **12 bit / sample:** 39.9 ops/byte, BW sufficient

LOFAR station processing requirements

Block	Parameter	Requirement (32-bit FP)	*Requirement (FPGA)*
FIR	Compute	595.2 GFLOPs	*595.2 GEOPs*
	Antenna data in	71.5 GB/s	*26.8 GB/s*
	FIR data out	71.5 GB/s	*44.7 GB/s*
	Coefficients, delayed samples in/out	2.27 TB/s	*992 GB/s*
	Computational intensity	0.25 ops/byte	*0.56 ops/byte*
FFT	Compute	480 GFLOPs	*480 GEOPs*
	FIR data in	71.5 GB/s	*44.7 GB/s*
	FFT data out	71.5 GB/s	*40.2 GB/s*
	Twiddle factors	357.5 GB/s	*178.8 GB/s*
	Computational intensity	1.1 ops/byte	*1.8 ops/byte*
BF	Compute	74.2 GFLOPs	*74.2 GEOPs*
	FFT data in	34.6 GB/s	*19.4 GB/s*
	BF data out	739 MB/s	*369.5 MB/s*
	Calibration data	138.6 GB/s	*69.3 GB/s*
	Computational intensity	0.43 ops/byte	*0.83 ops/byte*
Total	Compute	1.1 TFLOPs	*1.1 TEOPs*
	Total data bandwidth	3.0 TB/s	*1.4 TB/s*
	Computational intensity	0.4 ops/byte	*0.8 ops/byte*

- Operational intensity: 59.9 ops/sample

Single CPU performance 55.5x below real-time



Performance compared to platform peak:

	GFLOPs	% of peak performance
1 thread, scalar	3.5	45.9 %
8 threads, AVX	20.7	9.0 %

- Single-threaded AVX implementation worse compared to SSE4
  - *Different vectorization strategy required for scaling to multiple threads*

Roofline analysis

- Real-time station processing is done on 48 Xilinx Virtex-4 SX35 FPGAs
- Based on the roofline model, modern HW platforms require at least:
  - 3 Xilinx Virtex-7 x690T FPGAs
  - 56 Intel Core i7-3820 CPUs
  - 7 Nvidia Tesla K10 GPUs
- Sustaining full PCI-express 3.0 bandwidth, would require 4 Nvidia Tesla K10s

Available power measurements

- CPU and GPU: Measure power delivered by PDU
  - Full digital system measurements: both system idle and under load
  - Downside: consumption of unused peripherals
- GPU: Tesla's power measurement features
  - Only GPU power consumption
  - Downside: low time resolution
  - Downside: no host power consumption
- In-the-field FPGAs: Measure power consumption of digital boards
  - Full digital system measurement with little additional devices (idle and load)