RADIOBLOCKS

Project ID: 101093934

# Assessment of the applicability of next-generation technology

| | |
|---|---|
| Deliverable: | D4.1 |
| Lead beneficiary: | Institute (Author: ASTRON) |
| Submission date: | 29 February 2024 |
| Dissemination level: | Public |

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101093934

**Abstract**

This document describes and motivates the technologies that we want to explore in the remainder of the RADIOBLOCKS project. The main technologies are NVIDIA GPUs paired with NVIDIA network interfaces, with 400 Gb/s Ethernet technology. Some future technologies are described as well, and will be considered for exploration, should they become available in the course of the project. We also motivate why we do not want to explore certain technologies. The selected technologies drive the choices for the compute cluster that will be purchased. The document was prepared by John Romein (ASTRON) and Mark Kettenis (JIVE), based on the input provided by the members of Work Package 4.

# Contents



RADIO BLOCKS

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101093934

# List of acronyms

| | |
|---|---|
| ACAP | *Adaptive Compute Acceleration Platform*; hybrid FPGA/vector-processor design by Xilinx |
| ADC | *Analog to Digital Converter* |
| ASIC | *Application-Specific Integrated Circuit*; Chip that is designed for some specific task |
| CBF | *Correlator / Beam Former*; hardware and software that combines the signals from multiple receivers |
| CPU | *Central Processing Unit*; general-purpose processor |
| CUDA | *Compute Unified Device Architecture*; programming environment for GPUs |
| DDR | *Double Data Rate*; volatile random-access memory |
| DMA | *Direct Memory Access*; access to main memory without CPU involvement |
| DPDK | *Data-Plane Development Kit*; toolkit for building applications that require high-speed network-packet processing |
| DPU | *Data Processing Unit*; processor in or close to the network |
| EPAC | *European Processor ACcelerator*; processor design by the EPI |
| EPI | *European Processor Initiative*; project that aims to design and build a new family of low-power processors |
| FFT | *Fast Fourier Transform* |
| FLOPs | *Floating-Point Operations per second* |
| FPGA | *Field-Programmable Gate Array*; configurable processor, typically used for real-time, streaming processing |
| GbE | *Gigabit/s Ethernet*; network standard. |
| GFLOPs | *Giga FLOPs*; $10^9$ FLOPs |
| GPU | *Graphics Processing Unit*; a highly efficient, parallel processor that was initially designed for graphics computations. and later for general-purpose processing |
| HBM | *High-Bandwidth Memory*; volatile random-access memory type that typically provides more memory bandwidth than DDR memory |
| HPC | *High-Performance Computing* |
| I/O | *Input/Output* |
| NIC | *Network Interface Controller*; hardware component that connects a computer to a network |
| OSFP | *Octal Small Format Pluggable*; network connector, somewhat bigger than QSFP |
| PCIe | *Peripheral Component Interconnect Express*; high-speed bus that connects the computer with devices like NICs and GPUs |
| RDMA | *Remote DMA; direct access to memory of other systems*, via the network |
| RFI | *Radio-Frequency Interference*; disturbance in the radio spectrum, created by an external source |
| QSFP | *Quad Small Form-factor Pluggable*; common network connector for high-speed interconnects |
| VHDL | *VHSIC Hardware Description Language*; language used to program FPGAs or to design ASICs. |

RADIO BLOCKS

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101093934

# 1 Introduction

This document describes and motivates the technologies that we want to explore in the remainder of the RADIOBLOCKS project. These choices are the result of a technology review meeting (Milestone 4.1) that we held in month 8. In that meeting, we discussed the contents of a "live" technology assessment document, that we worked on during the first eight months, and that describes all relevant technologies, including their advantages and disadvantages.

The main technologies of choice are NVIDIA GPUs and 400 Gb/s Ethernet technology. These were identified as promising technologies already before the submission of the RADIOBLOCKS proposal. We will motivate the choice for these technologies again, as well as motivate why we did not choose for certain technologies.

The choices drive the developments in the remainder of the project, but they are not carved in stone. Some technologies may be worth exploring later on in the project, but are either not available yet, or have shortcomings in their current implementations. And new technologies might be introduced that we are not even aware of — for example, the *tensor-core* technology was announced and introduced in as little as one year.

The cluster that we acquire, will consist of a mixture of general-purpose and more specialized hardware. The cluster will be primarily used to develop and demonstrate the common "radio blocks" and correlator applications that we develop in Work Package 4, but it will also contain the hardware that Work Package 5 needs for the development of their software. This also fosters collaboration between both work packages.

The remainder of this document is structured as follows. Section 2 describes the context of this work, and sketches the basic assumptions around a GPU correlator/beam former system. Section 3 provides a background on GPUs, describes available hardware, and substantiates the choices for particular GPU technologies. Section 4 describes the (Ethernet) network technology, and Section 5 describes the options to perform computations in the network interface. Section 6 briefly describes some FPGA technology, and Section 7 describes a few other accelerator alternatives. Section 8 describes the current state of the art in CPU technology, and Section 9 concludes.

# 2 Context

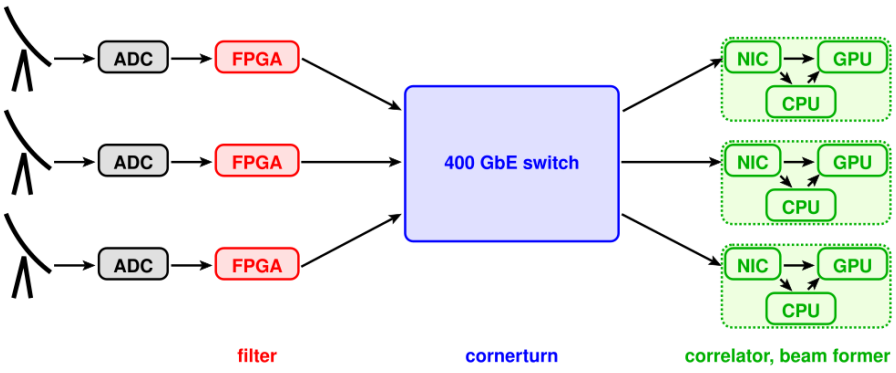


Figure 1: Data flow between the antennas and the correlator/beam former.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101093934

RADIOBLOCKS Work Package 4 covers the Correlator/Beam Former (CBF), and the data transport from the antenna digitizers to the CBF. Figure 1 depicts the data flow between the components. On the left, the antenna data are digitized by Analog-to-Digital Converters (ADCs). Typically, the FPGAs that read out the ADCs also filter (i.e., channelize) the data, so that the full frequency window that is observed, is split into independent frequency bands. This operation itself does not change the data rate, as time resolution is traded for frequency resolution — the amount of information in the data does not change, only the representation changes. However, frequency bands that are unusable (e.g., due to persistent Radio Frequency Interference (RFI)), may already be discarded here, reducing the data rate. A Fast Fourier Transform forms the base of filter, but the filter is often implemented as a PolyPhase Filter bank, a signal-processing technique that decreases signal leakage into neighboring frequency bands.

Filtering the data into independent frequency bands has an another important, practical, advantage: it provides a simple workload-distribution method for the CBF systems. On the right in Figure 1, the CBF machines are depicted. They combine the data from the different receivers, using the GPUs available in these systems. As the data has been filtered into different frequency bands, each CBF machine operates on its own, unique frequency band, independently of the other CBF systems.

Thus, near the antennas, each data stream contains all frequency bands of a single antenna, while the CBF machine need the data from all antennas, but only a single frequency band (or just a few bands). Hence, the data needs to be redistributed, in what is often called the *corner turn* or *transpose*. This is depicted in the middle of the figure. The data is thus transposed en route from the antenna digitizers to the CBF systems. This is the most efficient way to perform the corner turn, as it does not require extra hardware; the switch, a necessary component in the CBF cluster, performs that logical operation by design. Alternatively, the CBF systems can perform the corner turn internally on a second network, but this requires more network hardware, while the CBF systems no longer operate independently.

Figure 1 shows the core technologies that we need for Work Package 4: FPGAs, a network, and GPU systems. While Deliverable 4.2 explains the techniques to stream data efficiently from the digitizers into the CBF GPUs, this document evaluates the technologies that are required (or desirable) to build efficient correlators and beam formers.

# 3 GPUs

Graphics Processing Units (GPUs) owe their name to their original task: rendering pixels in video games. Their highly parallel execution units were much more efficient in performing large amounts of computations than general-purpose Central Processing Units (CPUs). Around 2010, NVIDIA introduced *CUDA* [32], a new ecosystem for using GPUs for general-purpose computing, which included a programming language, compiler, runtime system, libraries, drivers, and debugging and performance monitoring tools. The high revenues from the gaming market enabled rapid development of GPU technology.

GPUs have been adopted for radio-astronomical applications already from the early days, both for signal processing [12, 14, 16, 28, 43, 44] and for imaging tasks [42, 45, 46, 47]. Several instruments use GPUs to filter, correlate and/or beam form the signals from multiple receivers (e.g., LOFAR [14], CHIME [4, 20], and AARTFAAC [40]).

Around 2018, NVIDIA introduced *tensor core* technology in their GPUs [39], specifically designed to accelerate training and inference in deep-learning applications. The technology turns out to be a game changer for machine learning, which is rapidly evolving, and has a profound


This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101093934

societal impact on today's world. The same technology can be readily used for (some) signal-processing tasks like a correlator. The *Tensor-Core Correlator* [44] is a highly optimized library that correlates signals up to an order of magnitude faster and more energy efficient than regular GPU cores, hiding the nasty details of using tensor cores from the user. We expect the library to become the de facto standard GPU correlator library for radio-astronmical instruments in the years to come.

Essentially, the tensor-core correlator is our first "radio block", and GPU technology will play an important role in the correlator work package of the RADIOBLOCKS project. Therefore, we will discuss the various types of GPUs that are available today.

## 3.1 NVIDIA GPUs

NVIDIA is one of the largest and best-known manufacturers of GPUs. Below, we list three classes of GPU systems that we want to use in RADIOBLOCKS.

### 3.1.1 Discrete GPUs

Figure 2: An NVIDIA A100 GPU that can be inserted in the PCIe slot of a server machine.

The most common form of GPUs are *discrete GPUs*: devices as shown in Figure 2, that are inserted into a PCIe slot of a computer. The PCIe slot is used to communicate with the GPU, but also to provide power (up to 75 Watt). As nearly all available discrete GPU models use the PCIe gen4 x16 standard to communicate with the host CPU, the PCIe bandwidth is the same, regardless of whether it is an expensive and fast model, or a cheap and slower model.

Our (correlator) applications are typically limited by the PCIe bandwidth of a GPU, so it does not make sense to buy the fastest, most expensive, GPUs. Hence, for the cluster, we aim for relatively low-end (workstation-grade) GPUs. Higher-end GPUs are less cost effective and less energy efficient than lower-end GPUs for our use cases. The only model that supports PCIe gen 5 (with double the amount of bandwidth), is the NVIDIA H100, but this GPU is excessively expensive, and, still limited by PCIe bandwidth,

Apart from the PCIe-based discrete GPUs, there are some GPUs based on the SXM form factor. These provide high-bandwidth communication between all GPUs in a system through NVlink buses. Such systems are less interesting for correlator applications: GPU correlator systems can use multiple GPUs per system, but for a correlator, the GPUs need not communicate with other

RADIO BLOCKS

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101093934

GPUs, they only need to communicate with the CPU. Also, these high-end GPUs are too fast for the limited amount of external I/O bandwidth. As such systems are highly expensive, we do not consider them for the RADIOBLOCKS cluster.

Both the last generation of NVIDIA GPUs (called Ada) and the second but last generation (Ampere) are a viable choice. Ada-generation GPUs compute faster and more energy efficient than their Ampere-generation counterparts, but have, surprisingly, significantly lower memory bandwidth. The lower memory bandwidth is annoying, as both the filter and correlator GPU kernels rely on high memory bandwidth. Unfortunately, the unit prices of Ada-generation GPUs are much higher than that of Ampere-generation GPUs, so as long as Ampere-generation GPUs are still available, these may be a good choice.

We will make sure that the radio blocks that we will develop, support both the Ampere and Ada-generation GPUs. The next generation of GPUs, called Blackwell, is expected to be formally announced soon (although practical availability may be much later), and we intend to support them as well later on in the project.

### 3.1.2 NVIDIA Jetson AGX Orin

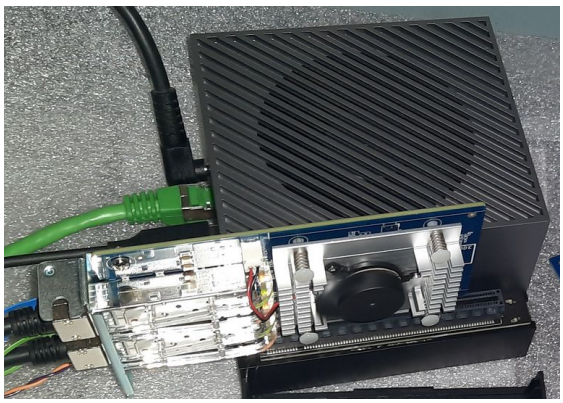

Figure 3: The Jetson AGX Orin is a compact 11cm x 11cm box. We use its expansion slot for a high-speed network interface.

Jetson GPUs [37] are compact embedded systems, meant for edge computing (see Figure 3). Unlike discrete GPUs, they have a tightly integrated CPU and GPU; they even share the same memory (up to 64 GB). Hence, there is no need to copy data between the CPU and GPU via the PCIe bus, eliminating an important bottleneck. However, their performance is much lower than that of most discrete GPUs, but so is their energy use, which is some tens of Watts for the entire system. The removal of the PCIe bottleneck and their low energy use make the Jetson an interesting platform. Essentially, even though this is based on an Ampere-generation GPU, it is by far the most energy-efficient system in our current cluster, using 40% less energy for the same workload compared to a contemporary server-based system with discrete GPUs.

Another reason why we want to explore these systems, is that they are designed for edge computing. We want to gain experience with this platform, to explore if we can use them for signal-processing tasks on GPUs in the field, near the antennas. For this, the platform should be capable of receiving streaming data at high speeds. Although NVIDIA sells these systems as 40 Gb/s Ethernet capable, our goal is to demonstrate receiving and processing (filtering and correlating) Ethernet packets at 100 Gb/s (the theoretical maximum that the PCIe bus of the


This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101093934

network interface can handle), which is quite a challenge on such a low-power device.

### 3.1.3 Grace Hopper Superchip

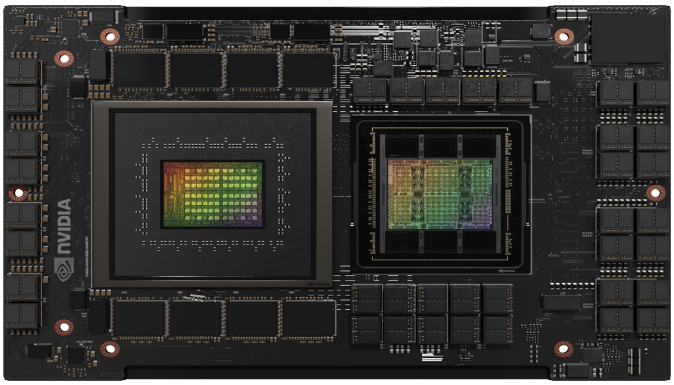

Figure 4: The GH200 Grace Hopper System-on-Module. The large chip on the left is the CPU, with LPDDR5 memory around it; the chip on the right is the Hopper GPU, with on-die HBM3 memory stacks.

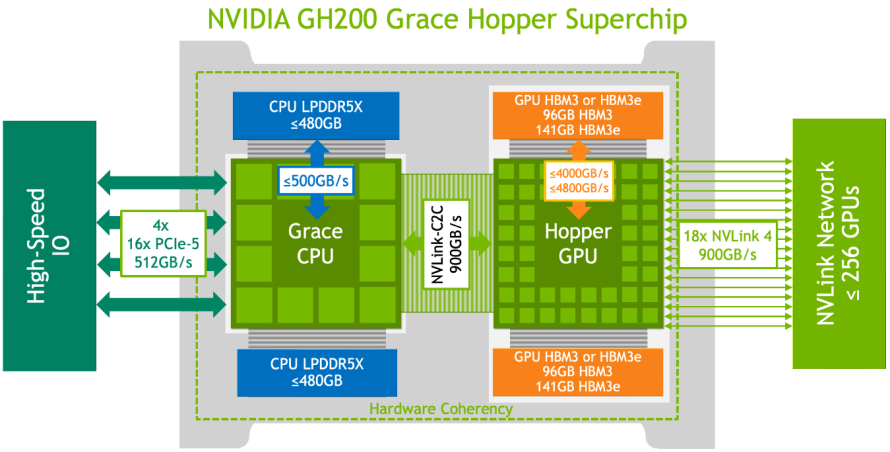


Figure 5: Schematic overview of the GH200 Grace Hopper System-on-Module. What really stands out, is enormous bandwidth between CPU and GPU.

Another highly innovative GPU technology that we definitely want to explore, is the NVIDIA *Grace Hopper Superchip* [36]. The CPU and GPU are much more tightly connected than in traditional server systems with discrete GPUs. The Grace Hopper is essentially a system on module (see Figure 4 and Figure 5), where a Grace CPU with 72 Neoverse 2 ARM cores with 480 GB of low-power DDR memory is connected by an NVlink bus to a Hopper GPU with 96 GB of High-Bandwidth Memory. Both the GPU and CPU can access each other's memories coherently (i.e., memory writes by one processor will be seen by the other processor, evicting stale cache entries if necessary). The bandwidth between CPU and GPU is 14 times higher than that of current-generation discrete GPUs that are connected by a PCIe gen4 bus. This is very useful, as the PCIe

RADIO BLOCKS

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101093934

bus is currently the limiting factor for practically all our radio-astronomical GPU applications.

Apart from the extremely high CPU-GPU interconnect, the GPU is by far the most powerful GPU that has every been produced. The peak 8-bit tensor-core performance is thirteen times higher than that of our current mostly-used GPU (the RTX A4000), and has nine times more memory bandwidth.

The Grace CPU also provides 64 PCIe gen5 lanes; in most systems, this allow connecting three 400 Gb/s Ethernet network interfaces, for a total of 1200 Gb/s of external communication. This should allow streaming external data at six times higher speeds into a GPU than what is possible with current-generation GPUs.

Our goals are to learn and to demonstrate how to fully exploit the exceptional compute power and I/O capabilities of this architecture. We will have to learn how to use new GPU hardware technologies and programming methods that have been introduced with this GPU generation, to harness its compute power. One Grace Hopper module should be powerful enough to replace the current 26-GPU LOFAR correlator *and* triple the input bandwidth to what is required for LOFAR 2.0, provided that we can successfully exploit its compute- and I/O capabilities in practice. We also want to demonstrate that this is a highly efficient architecture for wide-band instruments with extreme I/O requirements.

We ordered two Grace Hopper systems [41], to explore this highly interesting technology. They have been delivered recently, and our very first experiences confirm that these systems are exceptionally powerful.

## 3.2 AMD GPUs

AMD (formerly: ATI), like NVIDIA, has been a competitive GPU manufacturer for decades. NVIDIA was the first to introduce tensor cores though, and due to the enormous performance gain we obtained by computing correlations on tensor cores, there is a strong dependence on NVIDIA GPUs. Other vendors, like AMD, now also start introducing similar technologies in their latest generation of GPUs [6].

AMD GPUs are typically (much) cheaper than NVIDIA GPUs and their hardware is quite performant, but their programming environment is by far not as extensive as that of NVIDIA. In particular, NVIDIA introduced a new library, called cuFFTDx [33], which allows a much more efficient implementation of, for example, a PolyPhase Filter bank, because the FFT is embedded in a GPU kernel that can perform many more operations (e.g., FIR filtering, phase corrections, amplitude corrections) while reading and writing the data from and to GPU memory only once. The rocFFT library [9] for AMD GPUs is less efficient, as this requires reading and writing the data at least three times. Also, optimizing for AMD GPUs is typically more difficult than optimizing for NVIDIA GPUs.

Instruments like AARTFAAC and CHIME have been using AMD GPUs correlators successfully [4, 40]. We have been using AMD GPUs in the past, and want to continue doing so, to avoid a full dependence on a single vendor (NVIDIA).

### 3.2.1 Discrete GPUs

AMD has an interesting suite of workstation-grade GPUs (the W7000 series) [8], which offer the same quality as server-grade GPUs, but at a much lower price. AMD workstation-grade GPUs are also considerably cheaper than equally performant NVIDIA GPUs. We just ordered a recently introduced, mid-range GPU for experimentation and development, and to evaluate if this would be the right model for the regular GPU nodes in the RADIOBLOCKS cluster.

RADIO BLOCKS

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101093934

### 3.2.2 MI300A

The MI300A [7] is an interesting platform for future exploration. It will be AMD's answer to NVIDIA's Grace Hopper Superchip, with an even more tightly integrated set of CDNA3 GPU cores and 24 high-performance CPU cores on the same die. The exact specifications are not known yet, but as this system is designed for machine learning applications (and thus have powerful matrix multiplication units) this processor is likely a good match for our workloads.

An early version of an MI300A system [24] was showcased at the SuperComputing'23 conference. This particular model is of less interest to us, as this 4-socket system has "only" 1600 Gb/s of external I/O capabilities (which is more than Grace Hopper systems, but relatively low for a 4-socket system). Future systems with 1 or 2 sockets are likely to have a higher external I/O to compute ratio, which better matches our requirements.

The showcased system had 512 GB of HBM3 memory, but no DDR memory, so the amount of memory for such a powerful system is limited, but the memory is fast. The MI300X will have even more GPU power than the MI300A, but has no integrated CPU, which makes it less interesting for our use cases.

## 3.3 Intel GPUs

Intel has a broad range of GPU accelerators, named $X^e$ [26], ranging from the integrated graphics in laptop processors to the massive Ponte Vecchio systems built for data centers. They are programmed through OneAPI, a unified programming environment for CPUs, GPUs, and FPGAs.

Even though the $X^e$ architecture is a major leap forward from previous GPU generations and there is support for limited-precision matrix multiplications, and it is one of the very few accelerators that supports PCIe gen 5, we currently do not consider Intel GPUs, as they seem to not reach the same performance and energy efficiency as the latest NVIDIA or AMD GPUs. At this point in time, we do not think that the effort to port the RADIOBLOCKS GPU libraries to OneAPI (which may be considerable) is worth it.

# 4 Network switches and interfaces

High-speed I/O is a major topic of this work package. Thanks to innovations like tensor cores, GPU correlators are almost two orders of magnitude faster than a decade ago. The I/O requirements scale proportionally, but I/O technology did not improve at a similar rate, and what was once a compute-bound problem is now an I/O-bound problem.

As radio-astronomical instruments typically use FPGAs for digitization, filtering near the receivers, and stream data to centrally located GPU systems that may be far away, there is only one obvious data transport technology: Ethernet. Ethernet is well supported by both FPGAs and CPU/GPU systems, works over any distance, can be switched, and does not come with proprietary restrictions (unlike PCIe, NVlink, InfiniBand, and OmniPath).

In the RADIOBLOCKS project, we want to use 400 Gb/s Ethernet (400 GbE) technology. This matches the speed of PCIe gen5 systems; such systems are available right now. Whereas on the computer side 400 GbE will be used, on the switch side 800 GbE is necessary. This is because of the following: a switch port has 8 lanes of 112 Gb/s, while a 400 GbE network interface (NIC) port has 4 lanes of 112 Gb/s, and through the use of (Direct-Attach Copper) breakout cables, one switch port connects to two NIC ports. If switch ports would have had 4 lanes of 112 Gb/s, or if the NIC ports would have used 8 lanes of 56 Gb/s, we could have used 400 GbE switches (which are amply available), but all 400 GbE switches have 8 lanes of 56 Gb/s, while there are no 8-lane

RADIO BLOCKS

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101093934

NICs. The first 400 GbE NICs, 800 GbE switches, and matching cables have been announced, but are still difficult to obtain.

Unfortunately, the market introduction of 400 GbE can best be described as chaotic. NVIDIA, currently the only vendor of 400 GbE network interfaces, introduced a new connector type, called OSFP. This connector is slightly larger than the commonly used QSFP connector, to allow optical transceivers that can be inserted in the switch or NIC ports to dissipate more power. However, there are at least three different connector types, all called OSFP, which are incompatible with each other and have different cooling solutions.

We prefer to avoid the use of OSFP. This is largely due to it being incompatible with our current QSFP-based network that connects our FPGAs and GPUs, other than via using expensive optical converters instead of relatively cheap copper cables.

The only 400 GbE NIC with a QSFP112 connector is also from NVIDIA [38]. They are available for purchase right now, but only as engineering samples. This means that there is no support for these devices, and others reported that firmware updates do not work on them [17]. Hence, we have little confidence that this is the right solution for the new cluster.

Chelsio announced the availability of their new Terminator-7 ASICs [15] that support 400 GbE, but network cards based on these ASICs have not been announced yet. Also, it is unlikely that we can use the Data Plane Development Kit (DPDK) [19] with such network interfaces to receive network packets directly in GPU memory, as we currently do with NVIDIA network interfaces.

With respect to 800 GbE network switches, there is also little to choose from. The only available 800 GbE NVIDIA switch has OSFP ports, which requires expensive optical cables to connect to QSFP112 network interfaces. QSFP-based switches are not expected soon. Arista announced a 800 GbE switch [10] with the right port type (QSFP-DD800, which is compatible with QSFP112), but this switch is based on an older, low-end switch ASIC: the Tomahawk 4, and we do not know if this switch chip is capable of properly handling mixed signal speeds (earlier versions of the Tomahawk ASICs were not). Edge-Core announced a new switch [21], based on the Tomahawk 5 ASIC. This switch has more switch ports than we need, but may still be a cost-effective solution. Finally, Nokia announced 800 GbE core routers [29], but they provide all kinds of routing facilities (with an associated price tag) that we do not need, and the switch ports do not have all-to-all full-speed connectivity. We are not aware of any other potential solution.

We will monitor the market developments for a few more months. If 400 GbE NICs, 800 GbE switches, and QSFP112 cabling remain problematic, the fallback solution would be using 200 GbE NICs, 400 GbE switches, and QSFP56 cabling, which are readily available. The disadvantage would be that we need twice the amount of cabling and NIC ports to achieve the desired data rates, and that we need (expensive) optical converters between the FPGA QSFP112 network ports and the switch.

# 5 DPUs

Data Processing Units are essentially “smart” network interfaces that can be programmed to perform some operations on incoming and/or outgoing data. For example, such smart NICs can contain hardware to encrypt or decrypt data. More interesting for our use cases, would be to perform some signal-processing tasks like filtering or RFI detection. However, such tasks can consume a considerable amount of processing power.



This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101093934

## 5.1 NVIDIA BlueField 3

NVIDIA BlueField is probably the best-known DPU, and the BlueField 3 has been introduced recently [30]. It has a 400 GbE NIC that can both be used from the host CPU and from a 16-core ARM CPU that is integrated in the NIC. This 16-core ARM CPU runs its own (Linux) operating system, and can be programmed through the DOCA toolset [34].

We considered BlueField, but for the tasks we have in mind (for example filtering, or data-quality inspection), the built-in processor is too slow, and the internal memory bandwidth is too limited. For the DPU technology to be viable for our use cases, these systems need to be computationally more powerful.

## 5.2 NVIDIA Converged accelerators

A very interesting concept is the *converged accelerator* [31], which is basically a DPU and GPU combined on a single board. The DPU is responsible for handling the network interface, and the GPU provides the processing power to perform the computationally intensive tasks. The DPU and GPU are connected through a dedicated PCIe switch, that provides a high-speed interface between the DPU, the GPU, and the host system. Network packets can be directly transferred between the DPU and GPU (at full PCIe bus speed) such that incoming packets with digitized antenna samples can be directly processed by the GPU; the packet data remains on the card and is not transferred to the host system. One can even perform *all* processing on the DPU and the GPU; the host system then only provides electrical power to the converged accelerator card. This would provide a highly energy-efficient solution, but the limited amount of GPU memory allows incoming data to be buffered for no more than some hundreds of milliseconds, making it more difficult to meet the real-time requirements of a correlator application.

Alongside with the introduction of the Hopper GPU, NVIDIA announced a converged H100 CNX accelerator that contained a Hopper GPU and 400 GbE on the same card. Unfortunately, these were not only very expensive, they have been silently canceled. Other converged accelerator models are limited to 200 Gb/s, making them less attractive for this study.

We will follow this technology, but at this point in time, converged accelerators are not cost efficient.

## 5.3 Other DPUs

There are a number of other DPUs on the market, such as the Octeon 10 from Marvell, Pensando (now part of AMD), and the MPPA DPUs by Kalray. However, these are limited to 100 GbE, or at best 200 GbE, thus we will not pursue DPUs from this generation.

# 6 FPGAs

FPGAs are reconfigurable processors that can perform, amongst others, signal-processing operations on streaming data, in real time. Unlike GPUs, they typically have programmable transceivers that can be used to read Analog-to-Digital converters, hence FPGAs are often used near the antennas where the data are digitized, filtered, and packetized.

FPGAs are also suitable as processor for correlators. However, they lack the flexibility of GPUs, and programming them is more difficult, time consuming, and error prone [47], while talented FPGA programmers are even more scarce than talented GPU programmers. Also, they


This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101093934

used to be much more energy efficient than CPUs and GPUs, but this advantage seems to have disappeared. Even though a direct comparison between a functionally equivalent GPU and FPGA correlator would be highly interesting, we currently do not plan for this, due to the large amount of programming effort to implement the FPGA correlator firmware.

We do, however, plan to include a few FPGAs in the RADIOBLOCKS cluster that mimic the behavior of the antenna digitizers. These FPGAs will be used as real-time packet generators for the GPU correlators that we will develop. This will help us to investigate and develop the high-speed data transfer methods between the antenna digitizers and the correlator, which are based on Remote Direct Memory Access (RDMA) and DPDK, respectively (these are extensively described in Deliverable 4.2). Hence, in the following subsections, we will describe a few FPGA options.

Some PCIe cards with FPGAs can also be used as a "smart NIC", instead of a regular high-speed Ethernet network interface. They could be used in correlator machines to receive Ethernet packets, with little receive overhead. However, for this purpose, we prefer the use of regular network interfaces, as they are cheaper, much easier to use, and consume far less power than FPGA-based smart NICs.

## 6.1 Intel Agilex

Agilex [25] is the name of the latest generation of Intel (formerly: Altera) FPGAs. There are different types of Agilex FPGAs; for us the Agilex-I (or Agilex-M) series are the most interesting ones, as they come with 400 GbE interfaces.

FPGA development boards often come in one of two flavors: PCIe-based cards with a form factor that is similar to discrete GPUs and can be inserted in the PCIe slot of a server machine, and stand-alone developments that use built-in ARM CPU cores as a "host" system. We chose for a stand-alone development kit from iWave, based on its capabilities, price, and availability.

Later on in the project, we may consider some PCIe boards from, for example, Bittware (such as the IA-860 [13]). The advantage of these boards is that the FPGA is programmable in a high-level programming language, OneAPI (basically, a further development of OpenCL and SYCL). This reduces the programming effort, compared to the use of traditional Hardware Description Languages like Verilog or VHDL. In two earlier projects, we evaluated a predecessor of the OneAPI toolkit, and were quite positive about it. However, the 400 GbE network interfaces are not supported by OneAPI. We could add the support (like we did before for 40 GbE), but the benefit of a reduced programming effort then no longer applies (unless the development effort would be amortized over multiple OneAPI applications that would profit from the 400 GbE support).

## 6.2 AMD ACAP

AMD (formerly: Xilinx) has come with an interesting, new concept: the Adaptive Compute Acceleration Platform (ACAP) [5]. Basically, it combines traditional FPGA logic with programmable vector processors, scalar Digital Signal Processors, regular ARM CPU cores, and a Network-on-Chip that connects all these components, all in a single System-on-Chip (see Figure 6). The idea is that vector processors take over the compute tasks from the programmable logic, as the vector processors compute much faster and energy efficient than FPGA logic.

We started exploring the ACAP concept, by experimenting with the different parts of the processor. In collaboration with partner universities, we have been able to implement a PolyPhase Filter bank on the vector units of the ACAP with similar properties as the filter in the LOFAR 2.0


This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101093934

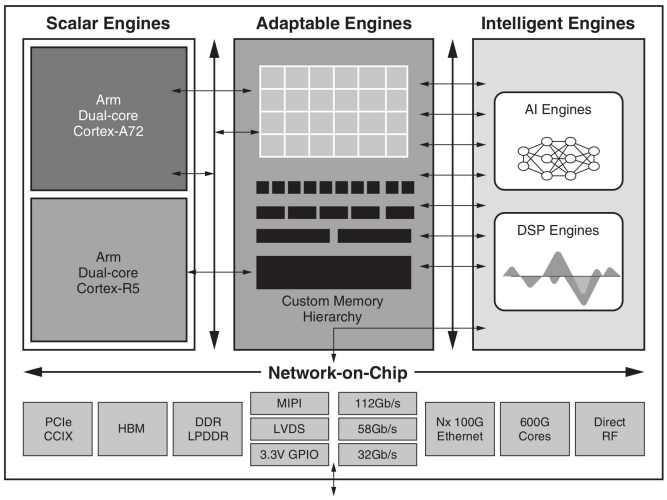


Figure 6: ACAP schematics.

system. The achieved performance is excellent. The vector units are easier to program than the FPGA logic, but more complex than a CPU or GPU. The software ecosystem is less mature compared to that of a CPU or GPU, but still improving. Based on further experience that we obtain in the near future, we may or may not proceed with this architecture in the course of this project.

# 7 Other accelerators

## 7.1 Intel Habana Gaudi2

The Intel Habana Gaudi2 [27] processor is developed to accelerate deep-learning tasks, and is highly efficient in performing limited-precision matrix multiplications (like tensor cores). Moreover, it has twenty-four 100-GbE interfaces, an enormous amount of Ethernet connectivity. The combination of the above forms a promising architecture for correlators. Unfortunately, these accelerators are only available in 8-socket systems, where twenty-one of the twenty-four Ethernet links are used to (directly) connect to each of the seven other accelerators, leaving only three 100-GbE links per chip for external communication. Our correlator applications do not need this internal connectivity, as we assume that the data already underwent a "corner turn" transpose on a separate Ethernet switch. It would have been an interesting platform if all twenty-four links were exposed for external communication, but three links per chip is insufficient.

## 7.2 European Processor Initiative EPAC

In addition to a general purpose processor that uses the ARMv8 architecture (see below) the European Processor Initiative is also developing an accelerator, called EPAC, that is based on the RISC-V instruction set [22]. The EPAC integrates three different accelerator tiles: A RISC-V vector tile (VTILE), a Deep Learning and Stencil accelerators (STX) tile, and a variable floating point precision core (VRP) tile.

RADIO BLOCKS

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101093934

The RISC-V vector tile is based on a draft version of the RISC-V "V" vector instruction set extension. Since significant changes have been made in the final version of this instruction set extension, the long-term future of software support for such a tile is uncertain. And since this is effectively just a CPU instruction set extension, it is unclear what benefits this tile offers over using the standard vector instructions offered by any modern CPU.

The variable floating-point precision core is targeted at supporting floating-point calculations at higher than double precision. Most algorithms used in radio astronomy have no need for such precision.

The STX tiles are probably the most interesting EPAC accelerators for our application. These units claim to provide 64 GFLOPS of FP64 performance per EPAC at 5-10x the energy efficiency of a normal CPU vector instruction unit. This tile also includes units that are specifically targeted at optimizing FFTs. Since FFTs are an important part of many of the algorithms in radio astronomy, it would be interesting to evaluate this architecture. Unfortunately EPAC appears to be still under development.

# 8 CPUs

In the face of competition for accelerators such as GPUs, CPU development continues as well. Over the past years several CPU architectures have seen instruction set architecture extensions targeted at AI workloads for example. We only consider the most recent generations of server-class (or workstation-class) CPUs that are currently available.

## 8.1 4th and 5th Generation Intel Xeon Scalable Processors (Sapphire Rapids, Emerald Rapids)

The differences between Sapphire Rapids [1] and Emerald Rapids [3] are fairly small. Sapphire Rapids CPUs use cores with Golden Cove cores and have the following characteristics:

- 80x PCIe 5.0
- CXL 1.1
- 8–60 cores
- 8 DDR5 memory channels (up to DDR5-4800)
- 2-, 4- and 8-socket variants

Emerald Rapids CPUs use Raptor Cove cores, which is a refresh of the Golden Cove microarchitecture that does not add significant features. These CPUs are available with up to 64 cores, higher memory bandwidth (up to DDR5-5600) and slightly higher base clock frequencies (but lower turbo frequencies).

These server-class CPUs from Intel are available with several built-in accelerators. The accelerators that are relevant for HPC/AI workloads are:

- Advanced Matrix Extension (AMX): An instruction set extension that provides 2D register files (tiles) and a Tile Matrix Multiplication instruction (TMUL). This operation is very similar to the matrix multiplication operations offered by the tensor cores of the NVIDIA GPUs. Currently these units support INT8 and BF16 data types, at 2048 INT8 or 1024 BF16 operations per cycle (per core). The AMX instruction set architecture allows for additional data types that may be supported by future generations. Use of this accelerator does not require an additional license.


This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101093934

- Deep Learning Boost (DL Boost): This instruction set extension provides a new VNNI instruction that optimizes matrix multiplications when using AVX-512. Use of this extension does not require an additional license.
- Data Streaming Accelerator (DSA): A DMA engine that can be used to offload copying data in memory. Based on Intel's documentation it seems that a single instance of this accelerator is available on all models, but on some models additional instances can be unlocked by purchasing a license.

We are interested in exploring the use of AMX, but the lack of support for the FP16 and complex data types in the current AMX implementation may be problematic for some of the targeted applications (FP16 and complex data types are expected to be implemented in future generation Intel CPUs). However since the total number of operations per cycle is an order of magnitude smaller than what is offered by GPUs this will have low priority. As we expect systems with these CPUs to be widely available we do not see the need to include these machines in our cluster.

## 8.2 AMD EPYC Server processors (Genoa, Bergamo, Siena)

AMD's current offerings for server-class CPUs [2] are based on the Zen-4 microarchitecture. Genoa CPUs have Zen-4 cores and the following characteristics:

- 128x PCIe 5.0
- 16–96 cores
- 12 DDR5 memory channels (up to DDR5-4800)
- Single- and dual-socket variants

Bergamo and Sienna CPUs have Zen-4c cores. These are functionally equivalent to Zen-4 cores but as a result of the space-optimized design these cores will run at somewhat lower clock rates. Bergamo CPUs are available with up to 128 cores. Bergamo CPUs are only faster than Genoa CPUs for workloads that are very well parallelized. Sienna CPUs have only 6 DDR5 memory channels and up to 64 cores and can only be used in single-socket systems. For our applications, these CPUs are less interesting than the Genoa CPUs: GPU correlators typically do not need that many CPU cores, and the calibration and imaging pipelines are not fully parallelized, so the sequential parts of the processing pipelines would suffer from the lower clock speeds.

These AMD CPUs do not offer the AMX instruction set extension that modern Intel CPUs provide and do not implement all of the AVX-512 vector instructions. In particular they lack the BF16 and FP16 extensions that provide support for half-precision floating-point that would be of interest for a CPU correlator. Also AMD's AVX-512 implementation does not truly implement 512-bit vectors but instead uses existing 256-bit units so the theoretical performance is expected to be half of what Intel CPUs provide at the same clock speed.

While the per-core theoretical performance of AMD CPUs is lower than that of Intel CPUs, the price per core of the AMD CPUs is significantly lower. This actually makes the price/performance ratio of AMD CPUs more attractive than the Intel CPUs.

Also worth mentioning are the AMD Threadripper PRO processors. These are almost identical to the Genoa CPUs, but run at much higher (about 35%) clock rates. On the downside, they have only 8 memory channels instead of the 12 memory channels for Genoa CPUs. Threadripper PRO processors target the workstation market, but the first Threadripper PRO systems with a Board Management Controller (for remote control), registered ECC memory (server-class self-correcting memory), and rack rails have just been announced [11]. Such systems have all


This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101093934

properties of server-grade systems, but provide much higher performance. We seriously consider acquiring some of these systems for the RADIOBLOCKS cluster, as these contain the fastest (x86_64) processors available today [18], are cost effective, and can be equipped with discrete PCIe gen5 GPUs and network interfaces.

## 8.3 NVIDIA Grace

The Grace part of the Grace-Hopper superchip is a powerful CPU in its own right. And it is available in a Grace CPU superchip [35], where the Hopper GPU chip is replaced by an additional Grace CPU chip, as well. Grace uses Arm Neoverse-V2 cores, which is ARM's most recent core targeted at HPC workloads. The Grace CPUs have the following characteristics:

- 128x PCIe 5.0
- 72 or 144 cores
- 32 LPDDR5X memory channels

These processors integrate Arm's Scalable Vector Extensions (both SVE and SVE2 are supported) with 128-bit vectors. This is four times smaller than the 512-bit vectors offered by AVX-512, but Grace can issue four SVE instructions per clock instead of just two on typical AVX-512 implementation. And SVE includes support for half-precision floating point which is something that not all AVX-512 implementations provide. AVX-512 has a reputation for being rather power hungry so Grace may still perform favourably in terms of FLOPs per Watt.

We will be able to evaluate the performance of these cores on the Grace-Hopper systems. Therefore there is little benefit in acquiring a Grace-Grace system.

## 8.4 European Processor Initiative Rhea processor

The first general purpose EPI processor, called Rhea, uses ARM Neoverse-V1 cores [23]. This is the predecessor of the Neoverse-V2 that is used by Grace. The Rhea CPUs are still under development so some of its characteristics are still unknown.

- PCIe 5.0
- 64 cores
- HBM2E and/or DDR5 memory

These processors also provide SVE (but not SVE2), but this time with 256-bit vectors. They can only issue two SVE instructions per clock, so the total FLOPS should be comparable to what Grace provides. The EPI processor is supposed to integrate an accelerator tile that provides a subset of the EPAC (see above), but details about what functionality will be provided do not seem to be available.

# 9 Conclusion

In this document, we described and substantiated the choices for the technologies that we want to explore and demonstrate in the RADIOBLOCKS project. GPUs play an important role: especially the Grace Hopper Superchip is a highly innovative System-on-Module that provides an order of magnitude more CPU-GPU bandwidth than previously, unfortunately with a high price tag. We discussed other accelerator types, a few of which we may want to pursue later on in the


This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101093934

project. Suitable CPUs and CPU systems are readily available, but the 400 GbE network solution with an 800 GbE switch that we desire, is difficult to realize. The project goals can still be met using previous-generation network equipment, at the expense of using converter cables and using more cables to achieve the same bandwidth. The cluster will allow us to develop the envisioned "radio blocks" and applications, and to demonstrate how new technology redefines the state of the art in GPU correlators.

# References


[1] *4th Gen Intel Xeon Scalable Processors*. https://www.intel.com/content/www/us/en/products/docs/processors/xeon-accelerated/4th-gen-xeon-scalable-processors.html.

[2] *4th Generation AMD EPYC Processors*. https://www.amd.com/en/products/processors/server/epyc/4th-generation-9004-and-8004-series.html.

[3] *5th Gen Intel Xeon Scalable Processors*. https://www.intel.com/content/www/us/en/products/docs/processors/xeon/5th-gen-xeon-scalable-processors.html.

[4] N. Denman et al. "A GPU-based Correlator X-engine Implemented on the CHIME Pathfinder". In: *IEEE International Conference on Application-specific Systems, Architectures and Processors*. Toronto, ON, Canada, July 2015, pp. 35–40.

[5] *AMD ACAP*. https://www.xilinx.com/products/silicon-devices/acap/versal-ai-core.html.

[6] *AMD CDNA Architecture*. https://www.amd.com/en/technologies/cdna.html.

[7] *AMD Instinct MI300*. https://www.amd.com/en/products/accelerators/instinct/mi300.html.

[8] *AMD Radeon PRO W7000 series*. https://www.amd.com/en/graphics/workstations#W7000-Series.

[9] *AMD rocFFT*. https://github.com/ROCm/rocFFT/.

[10] *Arista 7060DX5-64E*. https://www.arista.com/en/products/7060x5-series.

[11] *ASRockRack 4U4G-TR5/2T*. https://www.asrockrack.com/general/productdetail.asp?Model=4U4G-TR5/2TAQUA.

[12] C.G. Bassa et al. "Fourier-Domain Dedispersion". In: *Astronomy and Astrophysics* 657(A46) (Jan. 2022), pp. 1–7.

[13] *Bittware IA-860m*. https://www.bittware.com/products/ia-860m/.

[14] P. Chris Broekema et al. "Cobalt: A GPU-based correlator and beamformer for LOFAR". In: *Astronomy and Computing* 23 (Apr. 2018).

[15] *Chelsio Terminator 7*. https://www.chelsio.com/terminator-7-asic/.

[16] M.A. Clark, P.C. La Plante, and L.J. Greenhill. "Accelerating Radio Astronomy Cross-Correlation with Graphics Processing Units". In: *International Journal of High Performance Computing Applications* 27.2 (May 2013), pp. 178–192. DOI: 10.1177/1094342012444794.

[17] *ConnectX-7 Firmware*. https://forums.developer.nvidia.com/t/connectx-7-firmware-28-37-1014-download-for-mcx715105as-weat/268651/3.

[18] *CPU Benchmarks*. https://www.cpubenchmark.net/high_end_cpus.html.




This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101093934

[19] *Data Plane Development Kit*. https://www.dpdk.org/.

[20] Nolan Denman et al. "A GPU Spatial Processing System for CHIME". In: *Journal of Astronomical Instrumentation* 9 (3 Sept. 2020).

[21] *Edge-Core AIS800-64D*. https://www.edge-core.com/product/ais800-64d/.

[22] *European Processor Initiative Accelerator Processor*. https://www.european-processor-initiative.eu/accelerator/.

[23] *European Processor Initiative General Purpose Processor*. https://www.european-processor-initiative.eu/general-purpose-processor/.

[24] *Gigabyte G383-R80*. https://www.gigabyte.com/Enterprise/GPU-Server/G383-R80-rev-AAM1.

[25] *Intel Agilex*. https://www.intel.com/content/www/us/en/products/details/fpga/agilex.html.

[26] *Intel Data Center GPUs MAX Series*. https://www.intel.com/content/www/us/en/products/details/discrete-gpus/data-center-gpu/max-series.html.

[27] *Intel Habana Gaudi2*. https://habana.ai/products/gaudi2/.

[28] R.V. van Nieuwpoort and J.W. Romein. "Correlating Radio Astronomy Signals with Many-Core Hardware". In: *International Journal of Parallel Programming* 39.1 (Feb. 2011), pp. 88–114. DOI: 10.1007/s10766-010-0144-3.

[29] *Nokia 7750 service router*. https://www.nokia.com/networks/ip-networks/7750-service-router/.

[30] *NVIDIA BlueField-3*. https://www.nvidia.com/en-us/networking/products/data-processing-unit/.

[31] *NVIDIA Converged Accelerator*. https://www.nvidia.com/en-us/data-center/products/converged-accelerator/.

[32] *NVIDIA CUDA toolkit*. https://developer.nvidia.com/cuda-toolkit.

[33] *NVIDIA cuFFTDx*. https://docs.nvidia.com/cuda/cufftdx/index.html.

[34] *NVIDIA DOCA*. https://developer.nvidia.com/networking/doca.

[35] *NVIDIA Grace CPU Superchip*. https://www.nvidia.com/en-gb/data-center/grace-cpu-superchip/.

[36] *NVIDIA Grace Hopper Superchip*. https://www.nvidia.com/en-gb/data-center/grace-hopper-superchip/.

[37] *NVIDIA Jetson*. https://www.nvidia.com/en-eu/autonomous-machines/embedded-systems/.

[38] *NVIDIA MCX715105AS-WEAT*. https://docs.nvidia.com/networking/display/connectx7vpi.

[39] *NVIDIA Tensor Cores*. https://www.nvidia.com/en-us/data-center/tensor-cores/.

[40] Peeyush Prasad et al. "The AARTFAAC All-Sky Monitor: System Design and Implementation". In: *Journal of Astronomical Instrumentation* 5.4 (Dec. 2016), pp. 1641008-1–1641008-17.

[41] *QCT S74G-2U*. https://www.qct.io/product/index/Server/rackmount-server/GPGPU-Xeon-Phi/QuantaGrid-S74G-2U.



This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101093934

[42] J.W. Romein. “An Efficient Work-Distribution Strategy for Gridding Radio-Telescope Data on GPUs”. In: *ACM International Conference on Supercomputing (ICS’12)*. Venice, Italy, June 2012, pp. 321–330.

[43] John W. Romein. “A Comparison of Accelerator Architectures for Radio-Astronomical Signal-Processing Algorithms”. In: *Int. Conf. on Parallel Processing (ICPP’16)*. Philadelphia, PA, Aug. 2016, pp. 484–489.

[44] John W. Romein. “The Tensor-Core Correlator”. In: *Astronomy and Astrophysics* 656(A52) (Dec. 2021), pp. 1–4.

[45] Bram Veenboer, Matthias Petschow, and John W. Romein. “Image-Domain Gridding on GPUs”. In: *IEEE International Parallel and Distributed Processing Symposium (IPDPS’17)*. Orlando, FL, May 2017, pp. 545–554.

[46] Bram Veenboer and John W. Romein. “Radio-Astronomical Imaging on Graphics Processors”. In: *Astronomy and Computing* 32 (July 2020).

[47] Bram Veenboer and John W. Romein. “Radio-Astronomical Imaging: FPGAs vs GPUs”. In: *Euro-Par’19 (Best-Paper Award)*. Göttingen, Germany, Aug. 2019.

RADIO BLOCKS

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101093934