ASTRON
Netherlands Institute for Radio Astronomy

Implementation scenarios for ARTS

	Organisatie / Organization	Datum / Date
Auteur(s) / Author(s): Eric Kooistra	ASTRON	2015
Controle / Checked: Joeri van Leeuwen	ASTRON	
Goedkeuring / Approval: Joeri van Leeuwen	ASTRON	
Autorisatie / Authorisation: **Handtekening / Signature** Joeri van Leeuwen	ASTRON	

Distribution list:

Group:	Others:
Jeanette Bats Joeri van Leeuwen Roy Smits Alessio Sclocco Gert Kruithof Andre Gunst Stefan Wijnholds	Wim van Capellen Hajee Pepping Daniel van der Schuur Alwin Zanting

Document history:

Revision	Date	Author	Modification / Change
0.1	2015-08-5	E. Kooistra	Creation
0.2	2015-08-13	E. Kooistra	• Processed review comments from Stefan and Allesio. • Added Table 6 with preliminary design choices that were made for Arts. • Added Table 12 with an overview of the scenarios and section 2.8 that provides a conclusion • Added scenario 2b with 8 UniBoard^2-A10 • Added section 6.3.3 on external memory operation. • Removed cost for HEM, because HEM is only nice to have in the scenarios with UniBoard^2. • Updated cost estimate for UniBoard^2 with Stratix10 FPGAs. • UniBoard^2 can fit 32 GByte DDR4 modules (section 4.4.2, 6.3) • Clarified that SC2 only needs the Apertif X functionality for the central CB (section 6.1). • Added cost estimates for the Arts PL (section 2.6).

Table of contents:

References:

[1] “Arts Requirements Specification”, ASTRON-RS-020, J. van Leeuwen
[2] “Analysis of tied-array beamforming for Arts”, ASTRON-MEM-191, S.J. Wijnholds
[3] “Impact analysis of change request to use 1 MHz beamlets instead of 0.78125 MHz beamlets in Apertif and Arts”, ASTRON-CR-032, E. Kooistra
[4] “Apertif Beamformer Output Interface Specification”, ASTRON-SP-061, E. Kooistra
[5] “Comparison of FPGA, GPU and network switch”, ASTRON-MEM-193, E. Kooistra

Terminology:

ADC	Analogue to Digital Conversion
ADU	Analogue to Digital Unit (board with 8 ADC)
Apertif	APERture Tile In Focus
Arts	Apertif Radio Transient System
beam	Group of beamlets that point in the same direction
beamlet	Beam formed subband, a small beam spanning one subband
BF	BeamFormer
BG	Block Generator
BN	Back Node FPGA on UniBoard
bps	Bits per second
BSN	Block Sequence Number (time stamp)
BW	BandWidth
CB	Compound Beam, formed at dish level over the FPA
channel	Unit frequency band within a beamlet
CoBI	Correlator Backplane Interface board (connects 8 Uniboards with 8 OEB)
DB	Data Buffer
DM	Dispersion Measure
DT	Delay Tracking
FF	Flip Flop
FIR	Finite Impulse Response (digital filter)
FN	Front Node FPGA on UniBoard
FoV	Field of View
FPGA	Field Programmable Gate Array
FS	Fringe Stopping (is DT + PT)
GbE	Gigabit Ethernet
Gbs	Gigabit per second
GPU	Graphics Processing Unit
HEM	HMC Extension Module (provides UniBoard^2 with one HMC and extra optical IO per PN)
HMC	Hybrid Memory Cube
IAB	Incoherent array beam, formed by incoherently combining dishes
IO	Input Output
LE	Logic Element
LSbit	Least Significant bit
MAC	Multiply and Accumulate, Medium Access, Monitoring and Control
MM	Memory Mapped (control and status register interface)
MSbit	Most Significant bit
MTps	Mega transfers per second
NC	Not Connected
NIC	Network Interface Card for 10GbE
node	Processing node (PN), typically one FPGA chip
Nof	Number of
OEB	Optical-Electrical Board (provides UniBoard BN with same optical IO as the FN)
PAC	Power and Control board
power beam	Full Stokes power values: I, Q, U, V
PL	Pipeline processing
PN	Processing Node (FN or BN), PN0:3 = FN0:3, PN4:7=BN0:3
PT	Phase Tracking
QSFP	Quad Small Form-factor Pluggable transceiver (to connect a 4*10G link)
RF	Radio Frequency
SC	Science Case
SFP	Small Formfactor Pluggable transceiver (to connect a 10G link)
SP	Signal Path, 1 CB consists of N_{pol} = 2 SP, 1 SP per Apertif BF subrack
SR	Science Requirement
sps	Samples per second

subband	Frequency band, unit output of the filterbank
TAB	Tied array beam, formed by coherently combining dishes
T_{ant}	Transpose to group data from all S = 64 (≥ N_{ant}) antenna elements in the FPA
T_{dish}	Transpose to group data from all N_{dish} = 12 dishes
T_{pol}	Transpose to group data from both N_{pol} = 2 polarizations
T_{sp}	Transpose to group data from all N_{sp} = N_{pol} * N_{dish} signal paths, so combines T_{dish} and T_{pol}
T_{band}	Transpose to group data from all N_{band} = 16 bands
$T_{integration}$	Transpose to group data from an integration interval of N_{int_x} values in time
T_{FoV}	Transpose to group data from all N_{CB} = 37 beams for the full FoV
VLBI	Very Long Baseline Interferometry
voltage beam	Dual polarization sample values with phase information: X_{re}, X_{im}, Y_{re}, Y_{im}
WSRT	Westerbork Synthesis Radio Telescope
X	Correlator

Definitions:

$N_{complex}$	2	Two part of a complex number, the real and imaginary part
N_{pol}	2	Number of polarizations, X and Y
N_{Stokes}	4	Number of power values in the Stokes vector [I, Q, U, V]
N_{dish}	12	Number of WSRT dishes in Apertif
N_{sp}	24	Number of signal paths = N_{dish} * N_{pol} at the output of the Apertif BF
CB_{BW}	300 MHz	Full bandwidth of the CB and also of the TAB and IAB (SR-0.2)
B_{sub}	781250 Hz	Subband bandwidth in Apertif BF, = beamlet bandwidth
N_{band}	16	= nof_fn_bf, Number of bands in the Apertif BF to process the full CB_{BW}
N_{CB}	37	Required number of compound beams
N_{gr}	12	Required number of TAB grating lobe patterns to cover the full CB (SR-0.41)
N_{VLBI}	12	Required number of TABs in the central CB for VLBI, choose = N_{gr} (SR-0.23)
K_{TAB}	12	Implemented number of TABs per beamlet (≥ N_{gr})
N_{TAB}	444	= N_{CB}*K_{TAB}, number of TABs
N_{IAB}	37	= N_{CB}, number of IABs
N_{link}	384	= N_{PN}, number of physical 10G output links of the Apertif BF, so 1 link per PN
N_{PN}	384	= N_{sp} * N_{band}, total number of parallel processing nodes in the Apertif BF
M_{PN}	128	= N_{band} * nof_un, total number of parallel processing nodes in the Arts
M_{uni}	16	= N_{band} = nof_fn_bf, total number UniBoards in the Arts BF and in Apertif X
M_{uni2}	4	Total number UniBoard^2 in the Arts BF
N_{chan}	4	Number of channels per beamlet, for SC3 and SC4
B_{chan}		= B_{sub}/N_{chan}, channel bandwidth within a beamlet, for SC3 and SC4
N_{int}	≈ 10	Number of Stokes channel power values that are integrated in Arts
T_{Stokes}	≈ 50 µs	Minimum required sample period for the Stokes power values
f_{Stokes}	≈ 20 kHz	= $1/T_{Stokes}$, minimum required sample frequency for the Stokes power values
nof_uni	4	Number of UniBoards per polarization and dish in the Apertif BF
nof_bn	4	Number of back node FPGAs (BN) per UniBoard
nof_fn	4	Number of front node FPGAs (FN) per UniBoard
nof_un	8	= nof_fn + nof_bn, number of processing node FPGAs per UniBoard
nof_10g	3	Number of 10G links per FPGA node on UniBoard
nof_pn		Number of processing nodes (BN or FN on UniBoard or PN on UniBoard^2)
nof_bn_fb	16	= nof_uni*nof_bn, number of subband filterbank BN per SP in the Apertif BF
nof_fn_bf	16	= nof_uni*nof_fn, number of beamformer FN per SP in the Apertif BF
$N_{workstation}$	24	Number of workstations in the GPU cluster of the Arts PL
$W_{beamlet}$	6	Word width in number of bits of a beamlet voltage sample
W_{tab}	6	Word width in number of bits of a TAB voltage sample
W_{power}	8	Word width in number of bits of a IAB or TAB power sample

1 Introduction

1.1 Purpose

In chapter 2 this document describes several implementation scenarios for Arts. The scenarios should provide sufficient information to plan the development of Arts. Table 12 provides an overview of the scenarios and section 2.8 contains a conclusion that summarizes it all. Chapter 3 provides a system overview of the four science cases (SC) of Arts [1]. The other chapters provide background information and more details regarding the critical requirements and hardware constraints that influence the implementation options for Arts.

1.2 Scope

Main focus is on the implementation scenarios for the Arts FPGA beamformer, however the Arts pipeline GPU cluster is also discussed, because it influences the overall implementation choices, especially regarding the transient data buffer and the Arts BF output format and network. The UniBoard is already in use for the Apertif BF and for the prototype Apertif X. The CoBI, OEB, UniBoard^2 and HEM boards are new developments that are currently being tested or designed (section 7).

2 Scenarios

2.1 Overview of the Arts data path

Figure 1 shows the Arts data path and its main interfaces. The subsequent sections describe the functions in the data path and their implementation scenarios.

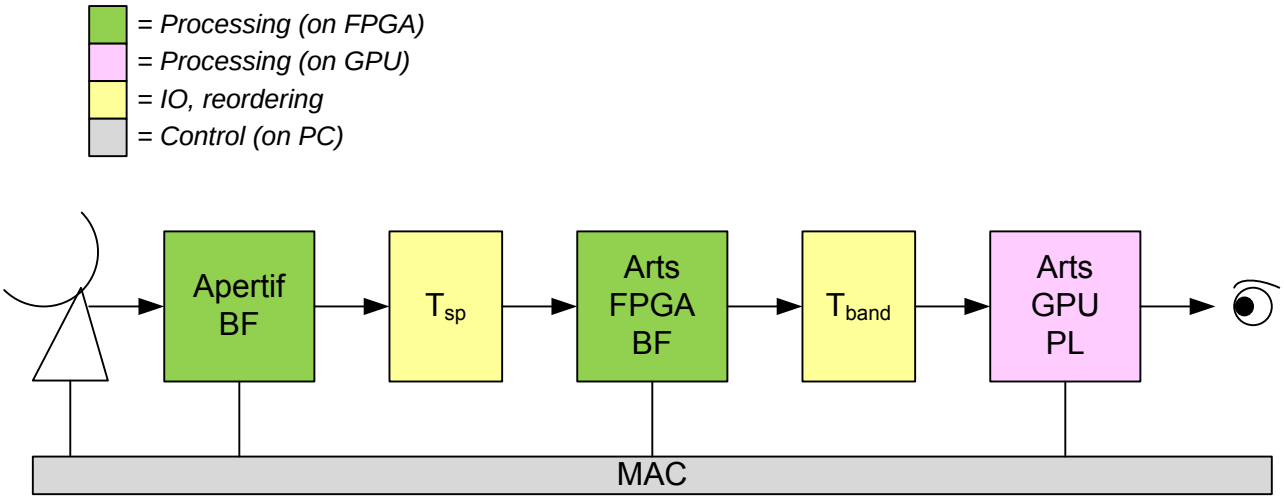

Figure 1: Arts top level data path

2.2 Apertif BF

Table 1 lists the functions that relate to the Apertif BF and are needed for Arts.

Function	Description	Remark
1 MHz subbands	Subband filterbank with 1 MHz subbands instead of 781250 Hz to fit SC1 and SC2. After the change the 1 MHz subbands will be used by Apertif X and by all Arts SC.	Change Request [3]
Smooth delay tracking (DT)	Recalculate subbands for duration of the filter impulse response after each DT step.	Section 6.2. Also applies to the Apertif X.
Phase tracking (PT)	Provide fringe stopping (FS) per CB by adding phase tracking (PT) in combination with the true sample delay tracking (DT).	Section 6.2. Also needed for the Apertif X.
Bypass T_{int}	Add programmable bypass of the integrating interval transpose T_{int} that is used for the Apertif X.	Section 6.4
Parallel CB-444 output	Provide parallel CB-444 offload via second 10G port to dedicated Arts FPGA processing boards. To avoid dependencies with the Apertif X it seems better to use dedicated Apertif BF output fibers for Arts than to let Apertif X pass on the CB-444 data.	Section 6.9, 6.10

Table 1: Arts functions that relate to the Apertif BF

2.3 Transpose T_{sp}

Table 2 lists the functions that relate to the transpose T_{sp} and are needed for Arts.

Function	Description	Remark
Input redistribution via UniBoard mesh		Section 6.5. Reuse from Apertif X.
Input align on UniBoard[2]		Section 6.5. Reuse existing DP component.

Table 2: Arts functions that relate to the CB-444 input transpose T_{sp}

2.4 Arts FPGA beam former (BF)

2.4.1 Data processing functions

Table 3 lists the functions that relate to the data processing functions that are needed for Arts.

Function	SC	Remark
TAB-1	1	Reuse BF unit from Apertif BF
TAB-12	2	Reuse BF unit from Apertif BF
TAB-444	4	Reuse BF unit from Apertif BF
CB-12 pass on	2	
IAB-37 component	3	New component
Channel filterbank	3,4	Reuse from Apertif X
IQUV Stokes powers component	3,4	New component
Integrate powers component	3,4	New component
Test logic with generators and monitors	All	Reuse BG, DB and BSN monitor

Table 3: Arts functions that relate to data path processing per SC

2.4.2 Transient data storage

Table 4 lists the functions that relate to the transient data storage functions that are needed for Arts (section 6.3).

Function	SC	Description	Remark
Transient data buffer	3,4	Streaming write access to DDR3 memory. When a trigger occurs the write stops. A selection of the DDR3 memory contents (ie. one CB, a few TABs) can then be read via the MM interface or streamed out via direct offload to 1GbE.	New component
Try 16 GByte DDR3	3,4	Verify that UniBoard can use two large DDR3 modules per FPGA. This is needed for the 4 TByte voltage buffer scenario 1 and 2.	Reuse UniBoard test design for DDR3
Try 16 GByte DDR4	3,4	Verify that UniBoard^2 can use two large DDR4 modules per FPGA. This is needed for the 0.5 TByte power buffer scenario 4.	Reuse UniBoard^2 test design for DDR4

Table 4: Arts functions that relate to transient data storage

Initially assume that SC3 is not needed because SC4 can run commensal on dedicated hardware in case of scenario 3 or 4 (section 2.7). If SC3 is not needed then the 4 TByte voltage buffer is not needed (section 6.3.1). For SC4 still a 0.8 TByte power data buffer is needed (section 6.3.2, 6.6). Scenario 3 can offer up to 4 TByte. Assume the 0.5 TByte that scenario 4 can offer (section 4.4.2) are also sufficient. If the full Stokes power data for SC4 are transported to the GPU cluster then the power data buffer can be implemented on the GPU cluster (section 6.3.4). Initially assume that only the Stokes I power data are transported (section 6.7) and that the power data buffer needs to be implemented on the DDR3 (scenario 1, 3) or DDR4 (scenario 4) of the FPGA beamformer. After a trigger the transient data buffer can be read out via the 1GbE interface, because only a few TAB need to be read out (section 6.18.5).

2.5 Transpose T_{band}

Table 5 lists the functions that relate to the transpose T_{band} and are needed for Arts.

Function	SC	Remark
Output reorder and offload via 1GbE	1, 3	Section 6.15, 6.16
Output reorder and offload via 10GbE	2, 4	Section 6.15, 6.16
Output redistribution via UniBoard mesh	2, 4	Section 6.15.2
Try Tx only 10GbE links to NIC and via switch	2, 4	Section 6.17
Pass on meta data including flagging	All	

Table 5: Arts functions that relate to the Arts BF output transpose T_{band}

Assume that SC1 will use output via 1GbE (section 6.18.16.18.4). If SC3 is needed then SC3 can also use the 1GbE network for output (section 6.18.4). For SC2 and SC4 output via 10GbE is needed. Initially assume that for SC4 only the Stokes I data are transported (section 6.7, 6.8). This then implies that only one 10GbE output per Uniboard (scenario 1, 3) or per PN (scenario 4) is needed (section 6.18.2, 6.18.3).

The transpose T_{band} can be done by a 10GbE switch, by one UniBoard^2 with Arria10 or by the infiniband switch in the GPU cluster, see section 6.18.2. Assume that a UniBoard^2 will be used, see Table 6. The cost of this are then 15 Euro.

2.6 Arts pipeline GPU cluster

The main architecture choices for the GPU cluster (section 5) concern:

- Arts BF output data format specification concerning need for data reordering (section 6.16), need for packing into bytes (section 6.8.1)
- The number of workstations in the Arts PL GPU cluster and the number of 10GbE ports
- Streaming transport full Stokes data or only Stokes I.

- When only Stokes I data is transported then the transient data buffer for SC4 needs to be on the Arts BF
- The CB-444 voltage data transient data buffer for SC3 can only be on the Arts BF
- How much DDR memory does the Arts PL need for processing and what is the access rate per CPU, GPU.
- Having an Infiniband switch in the GPU cluster may be valuable anyway even if it is not used for the transpose T_{band}.

The cost for the GPU cluster are estimated at 20k for SC1, 20k for SC2 and 240k for SC3 and SC4, so about 10k per workstation assuming $N_{workstation}$ = 24.

2.7 Hardware scenarios

2.7.1 Preliminary design choices

Table 6 lists the preliminary design choices that have been made for Arts.

Design choice	SC	Description
Use dedicated CB-444 output from Apertif BF for Arts BF.	All	Advantage is that dedicated Arts BF hardware can operate independently from the Apertif X hardware, see section 6.9
Use 16*1GbE→10GbE switch between Arts BF and Arts PL for SC1	1	Section 6.18.1
Implement transient data buffer in Arts BF.	3,4	Advantage is that only the stokes I data need to be streaming transported from the Arts BF ot the Arts PL, see section 6.3, 6.7, 6.18.5.
Use UniBoard^2-Arria10 as a 10GbE switch between Arts BF and Arts PL.	2,4	Advantage is that this UniBoard^2 can map the unidirectional inputs to full duplex 10GbE outputs so no need for HEM or to investigate Tx only capabilities of switches, it can do additional reording within data packets or between data packets to map 16 inputs ports to 24 output ports, and it can implement the transpose T_{band}., see section 6.16, 6.17, 6.18.2. The Infiniband switch in the GPU cluster is then not needed (at least not for the transpose T_{band}).

Table 6: Preliminary design choices

2.7.2 Scenarios for the Arts FPGA beamformer

The following tables list several scenarios for the Arts FPGA beamformer.

Scenario	Reuse 16 UniBoards
Description	Reuse the 16 UniBoards of Apertif X for all four Arts SC (section 6.11): • SC1 and SC4 fit on the 16 UniBoards, because they are not running simultaneously with Apertif X. • The SC2 with local interferometer data for only the central CB can fit on the 16 UniBoards, see section 6.1 • Commensal SC3 will have to run together with the Apertif X on the same FPGAs.
Issues	• SC3 cannot fit together with Apertif X on the UniBoard FPGAs, see section 4.4.5 • Combining Apertif X with Arts SC3 in one FPGA will make the firmware design and maintenance more complicated.

Table 7: Arts BF scenario 1

Scenario	Use dedicated 4 UniBoard^{2}s with Arria10 FPGAs (section 4.2)
Description	Reuse 16 UniBoards for SC1 and SC4 and use dedicated 4 UniBoard^{2}s with Arria10 for SC2 and SC3: • SC1 can remain on the 16 UniBoards, but it could also be ported to the 4 UniBoard^2. • SC4 needs to remain on the 16 UniBoards, because regarding processing SC4 does not fit on the 4 UniBoard^2 with Arria10 (see section 4.5).
Issues	• UniBoard^2 only has maximum 1 TByte for transient data storage, whereas SC3 requires 4 TByte for 10 s of CB-444 voltage data.

Table 8: Arts BF scenario 2

Scenario	Use dedicated 8 UniBoard^{2}s with Arria10 FPGAs (section 4.2)
Description	Compared to scenario 2 the SC4 can now run on the UniBoard^2. • With SC4 on a dedicated system the SC3 is no longer needed and can be omitted.
Issues	• The 4 extra UniBoard^{2}s increase the cost, complexity and the power consumption.

Table 9: Arts BF scenario 2b

Scenario	Use dedicated 16 UniBoards (section 4.1, 6.12)
Description	All Arts SC can run on dedicated 16 UniBoards. • With SC4 on a dedicated system the SC3 is no longer needed and can be omitted.
Issues	• Extra cost of about 312 kEuro for UniBoard + OEB hardware, see Table 22.

Table 10: Arts BF scenario 3

Scenario	Use dedicated 4 UniBoard^{2}s with Stratix10 FPGAs (section 4.2)
Description	All Arts SC can run on dedicated 4 UniBoard^{2}s with Stratix10. • SC1 may remain on 16 UniBoard that are reused when the Apertif X is not used, but SC1 could also be ported to the 4 UniBoard^2. • Commensal SC2 and SC4 can fit on 4 dedicated UniBoard^2 with Stratix10. • With SC4 on a dedicated system the SC3 is no longer needed and can be omitted.
Issues	• Extra cost of about ??? kEuro for UniBoard^2 (Stratix10) + HEM hardware, see Table 23. • Stratix10 production samples are expected in Q3 2016, section 7.2. However firmware development can already start earlier because Stratix10 can be simulated.

Table 11: Arts BF scenario 4

2.7.3 Evaluation of the scenarios

Table 12 summarizes the science capabilities of the scenarios including the hardware cost and the limitations.

Scenario	Hardware Arts BF	Cost for Arts BF	Science	Limitation
1	Use the 16 UniBoards of Apertif X	**10k** (= 1 TByte for SC4, section 4.4.2)	• All Arts SC can run on 16 UniBoards.	• Maximum 4 TByte buffer • No commensal modes together with Apertif X due to lack of FPGA resources (section 4.4.5).
2	Dedicated 4 Uniboard^2-Arria10	**132k** (= 116k from Table 23 + 16k for 0.5 TByte for SC4, section 4.4.2)	• Commensal SC4, but with limited FoV due to less than 12 TABs.	• Maximum 1 TByte buffer • No commensal SC3 due to only 1 TByte (instead of 4 TByte). • For SC4 maximum 7 TABs per CB instead of 12 (section 4.4.6).
2b	Dedicated 8 Uniboard^2-Arria10	**228k** (= 212k from Table 24 + 16k for 0.5 TByte for SC4, section 4.4.2)	• Commensal SC4 for maximum 14 TABs • No need for SC3.	• Maximum 2 TByte buffer
3	Dedicated 16 UniBoards	**322k** (= 312k from Table 22 + 10k for 1 TByte, section 4.4.2)	• Commensal SC4 for maximum 20 TABs • No need for SC3, but can fit SC3.	• Maximum 4 TByte buffer
4	Dedicated 4 Uniboard^2-Stratix10	**252k** (= 236k from Table 23 + 16k for 0.5 TByte for SC4, section 4.4.2)	• Commensal SC4 for maximum 58 TABs • No need for SC3.	• Maximum 1 TByte buffer

Table 12: Overview of the Arts BF scenarios

For all scenarios applies that using less boards means less bandwidth. For example if instead of 16 Uniboard only 1 UniBoard is used than $CB_{BW}/16$ = 19 MHz and similar if instead of 4 UniBoard^2 only 1 UniBoard^2 is used then $CB_{BW}/4$ = 75 MHz. The firmware validation on hardware can be done with 1 board, because the boards operate independently and do the same. Firmware can already be developed before hardware is available, because it can be verified in simulation.

From a technical point of view scenario 4 is preferred because it easily fits SC4 (section 4.5), but the UniBoard^2 with Stratix10 FPGAs is not available yet (section 7.2). A prototype UniBoard^2 with Arria10 FPGAs is available now. UniBoard production boards are available now. Therefore first for SC1 and then for SC4 the Arts development can start with scenario 1. Porting SC4 from StratixIV to Arria10 or Stratix10 is relatively small, because the application components are designed in a technology independent way. However SC4 on Uniboard requires some extra development effort to output the data via 10GbE (section 6.5, 6.15.2) that is not needed for UniBoard^2.

For SC2 scenario 2 or 4 is preferred, because then the Apertif X can be used for the local interferometer data of the central CB. Whereas using scenario 1 requires adding a central CB correlator to the Arts SC firmware.

Commensal SC3 will not fit together with Apertif X on the 16 UniBoards (section 4.4.5). Therefore dedicated hardware will be needed for commensal Arts. Scenario 2 using 4 UniBoard^2 with Arria10 cannot fit the 4 TByte memory required for the transient voltage data buffer for SC3 (section 4.4.2) and it does not have sufficient processing resources to fit SC4 (section 4.4.6). Using 8 UniBoard^2 can fit SC4 but comes at a cost regarding hardware, power consumption and development.

The advantage scenario 3 and 4 is that they can fit SC4 and then SC3 and the 4 TByte transient voltage data buffer are no longer needed. The advantage of scenario 4 is that the UniBoard^2 hardware is cheaper than the UniBoard hardware for scenario 3 and that scenario 4 can provide many more TABs (section 4.4.6). The energy consumption of scenario 4 using UniBoard^2 is also less and may save > 28 kEuro over 5 years (section 4.4.8). With UniBoard^2 the firmware architecture does become simpler (section 6.5.3, 6.15.2, 6.17.2), because UniBoard^2 has more IO than Uniboard.

2.7.4 Arts SC1 hardware infrastructure in scenario 1

Figure 2 shows Arts SC1 implemented according to scenario 1 reusing the 16 UniBoards of the Apertif X. Between the Arts BF and Arts PL there needs to be an Ethernet switch to lead the data to a single 10GbE port. This Ethernet switch effectively performs the transpose T_{band} to bring together the entire CB_{BW} = 300 MHz at a single GPU node (section 6.18.1). For SC1 one GPU workstation suffices to do the pipeline processing. Later on the GPU workstation may remain standalone or it can become part of the GPU cluster.

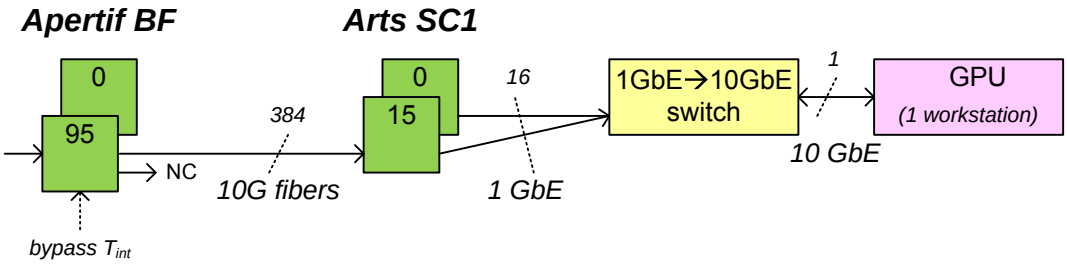

Figure 2: Arts SC1 reusing 16 UniBoard from Apertif X and 1GbE switch for Arts BF output to PL

2.7.5 Arts SC4 hardware infrastructure in scenario 1

Figure 3 shows Arts SC4 implemented according to scenario 1 reusing the 16 UniBoards of the Apertif X. This scheme assumes that only the TAB-444 Stokes I power data is streamed from the Arts BF to the Arts PL. This implies that only one 10GbE port per UniBoard is needed (section 6.7, 6.18.2) and that the UniBoards will need to buffer the TAB-444 Stokes I power data in DDR3 memory (section 6.3, 6.7). Using e.g. 4 GByte DDR3 modules 16 Uniboards can store 1 TByte. The number of 10GbE output links from the Arts BF to the Arts PL is N_{band} = 16. The GPU cluster has e.g. 24 workstations and a UniBoard2 with Arria10 is used as switch to connect the Arts BF to the Arts PL.

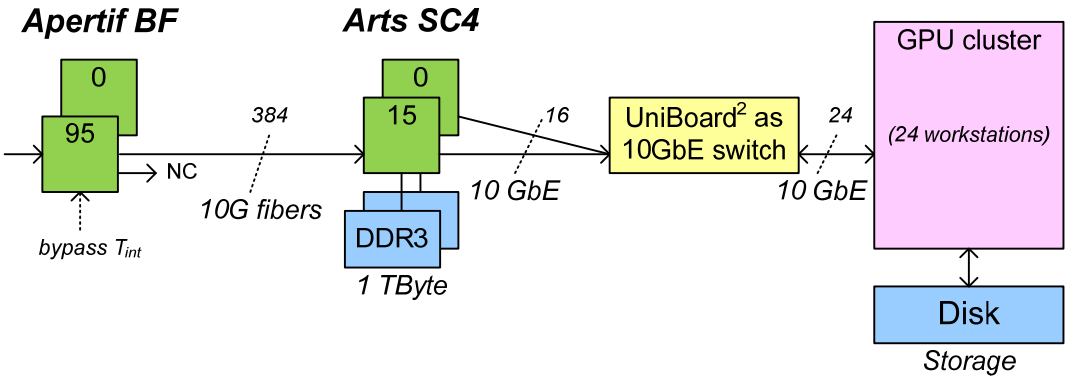

Figure 3: Arts SC4 reusing 16 UniBoards from Apertif X and Stokes I output to PL

2.7.6 Arts SC2 and SC4 hardware infrastructure in scenario 4

Figure 4 shows Arts SC2 and SC4 implemented according to scenario 4 using 4 dedicated UniBoard^{2}s. By using dedicated hardware for Arts and by using the second output port of the Apertif BF the Arts hardware is independent from the Apertif X hardware. In Figure 4 SC1 has been ported to UniBoard^2, however it could also remain on the Apertif X UniBoards to avoid this effort. The single GPU workstation for SC1 may also be part of the GPU cluster. The 10GbE link for SC1 may be connected via the UniBoard2 with Arria10 that is used as a switch to connect the Arts BF to the Arts PL.

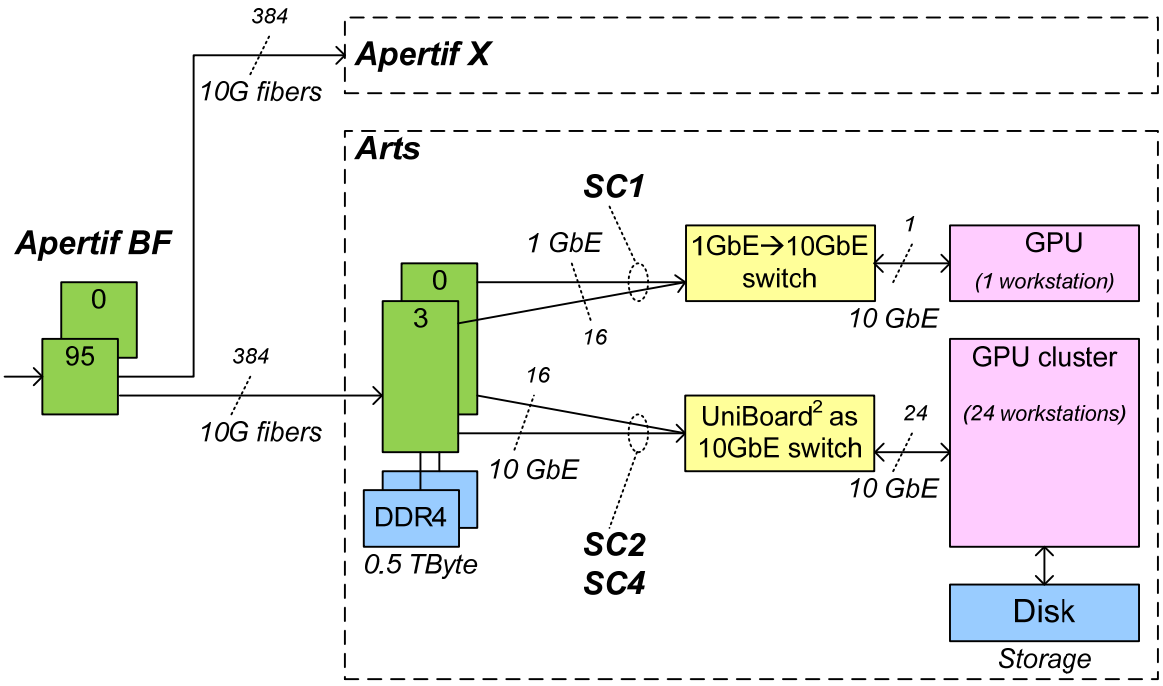

Figure 4: Arts SC2 and SC4 using dedicated UniBoard^2 with Stratix10 and independent from Apertif X

2.8 Conclusion

Scenario 1 is the cheapest regarding Arts BF hardware cost, but it implies that Arts cannot run commensal with Apertif X. Running Arts commensal requires adding dedicated hardware for the Arts BF as in the other scenarios.

Scenario 2 can fit SC2. Scenario 2 cannot provide sufficient TABs for SC4, but it can provide a development path towards scenario 4 using the one UniBoard^2 with Arria10 that we will have in 2016 for initial tests (providing $CB_{BW}/4$ = 75 MHz and maximum 7 TABs / CB).

Scenario 2b fits SC4 but it is not much cheaper than scenario 4 and provides far less TABs than scenario 4.

Scenario 3 fits SC4 (and SC3) but it is more expensive than scenario 4 and provides far less TABs than scenario 4.

Scenario 4 easily fits SC4 because it can provide about a factor 4 more TABs than required. Being able to run SC4 commensally makes that SC3 is not needed.

The estimated cost for the Arts BF using scenario 4 is 252k (Table 12).

The estimated cost for the Arts PL is 280k (section 2.6).

The network between Arts BF and Arts PL will cost about 20k (1/10GbE switch for SC1 and a Uniboard^2 with Arria10 for SC2 and SC4, see Figure 4).

3 System overview

3.1 Apertif

Arts [1] implements the tied array and VLBI functionality of Apertif. Figure 5 shows Arts within Apertif. Both the Apertif correlator (X) and Arts use the beam data from the Apertif beamformer (BF).

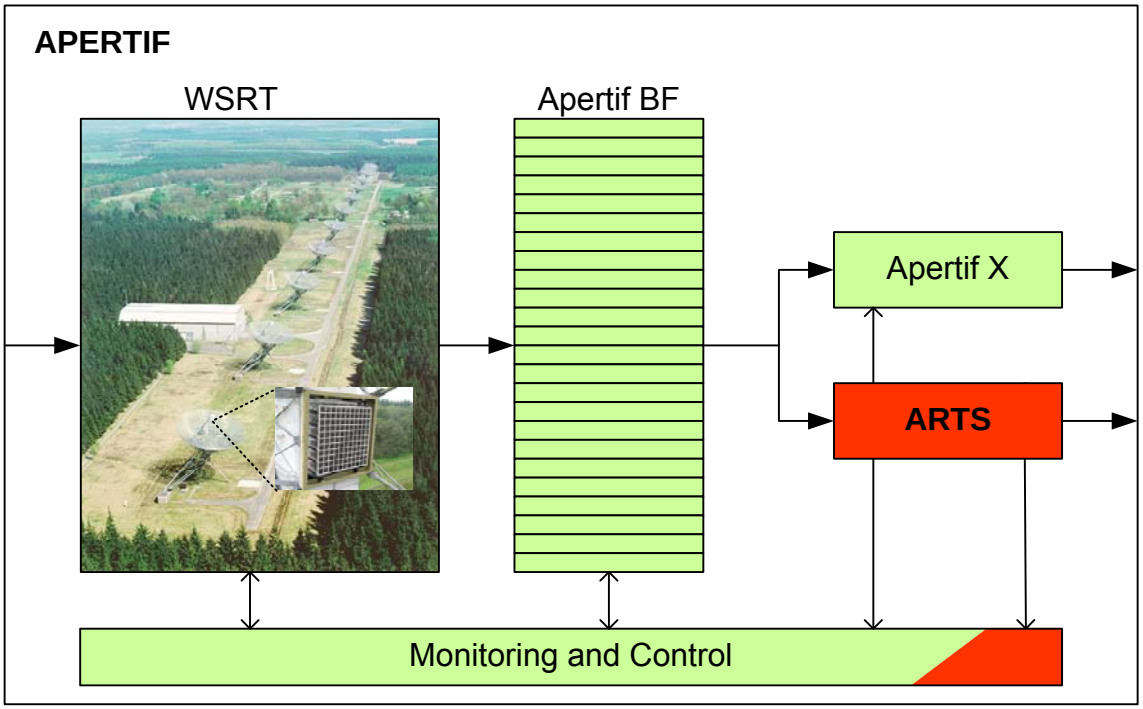

Figure 5: Top level overview of Apertif with Arts included

3.2 Arts

Within Arts the processing consists of an FPGA beamformer and a GPU pipeline, as shown in Figure 6. Arts has four science cases (SC) and for all four SC the FPGA beamformer will be implemented on Uniboards and the pipeline processing will be implemented on GPUs. The mapping on FPGAs or GPUs is discussed generally in [5].

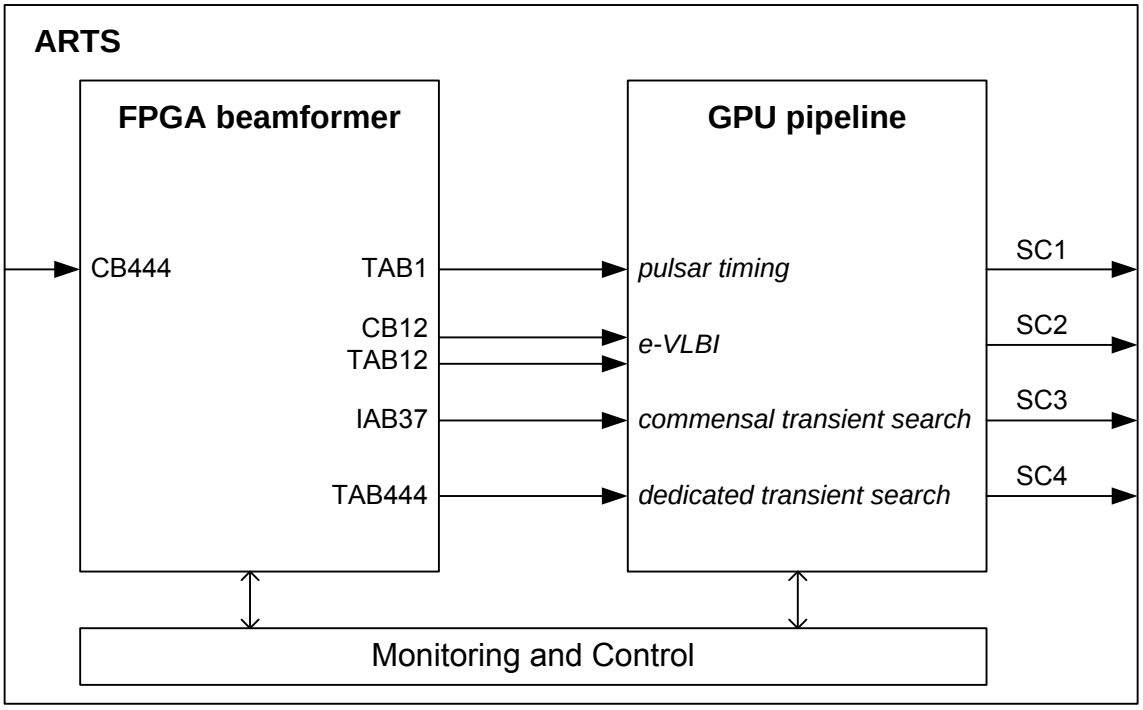

Figure 6: Arts FPGA beamformer and GPU pipeline

3.3 Interfaces

3.3.1 Apertif BF output

The Apertif BF processes each dish and each polarization independently. The beamformed output for one polarization of one dish is called a signal path (SP). Figure 7 shows the Apertif BF subrack that can beamform the N_{CB}=37 compound beams (CB) and 1 CB consists of N_{pol}=2 SP. The subrack consists of 4 Uniboards and N_{band}=nof_fn_bf=16 outputs. Note that each FN on UniBoard has nof_10G=3 ports available, but for the Apertif BF output only one 10G port is used as shown in Figure 7.

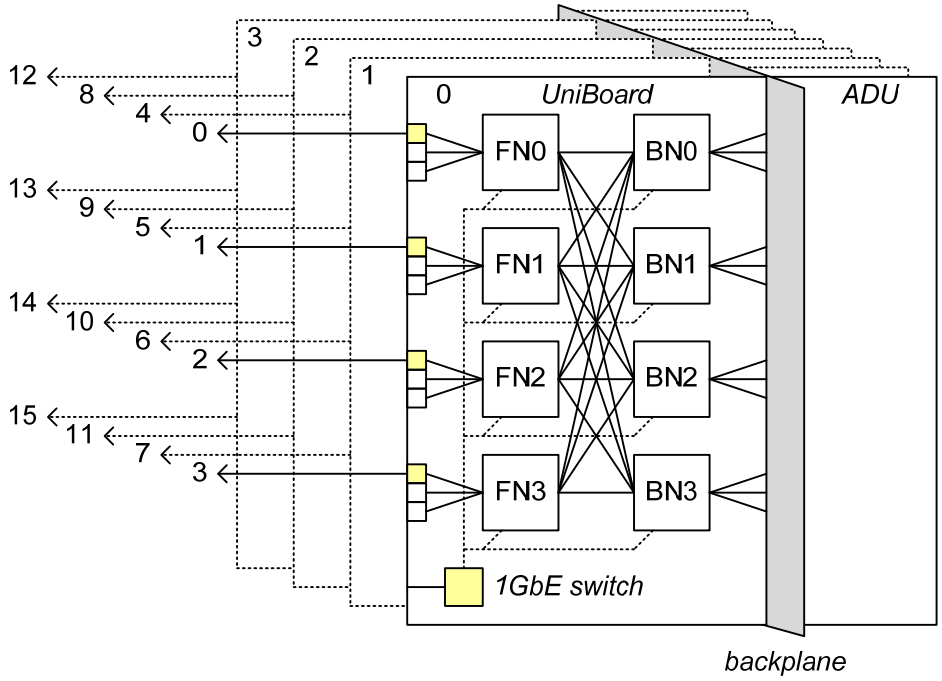

Figure 7: One Apertif BF subrack per signal path with nof_uni=4 UniBoards and N_{band}=16 FN

In total the Apertif BF consists of $N_{sp}=N_{pol}*N_{dish}$ = 2*12 =24 subracks, as shown in Figure 5 and Figure 8. Figure 8 shows that the Apertif BF output is carried via $N_{link} = N_{sp}*N_{band}$ = 24 * 16 = 384 10G fibers to the N_{band}=16 UniBoards that implement the Apertif X.

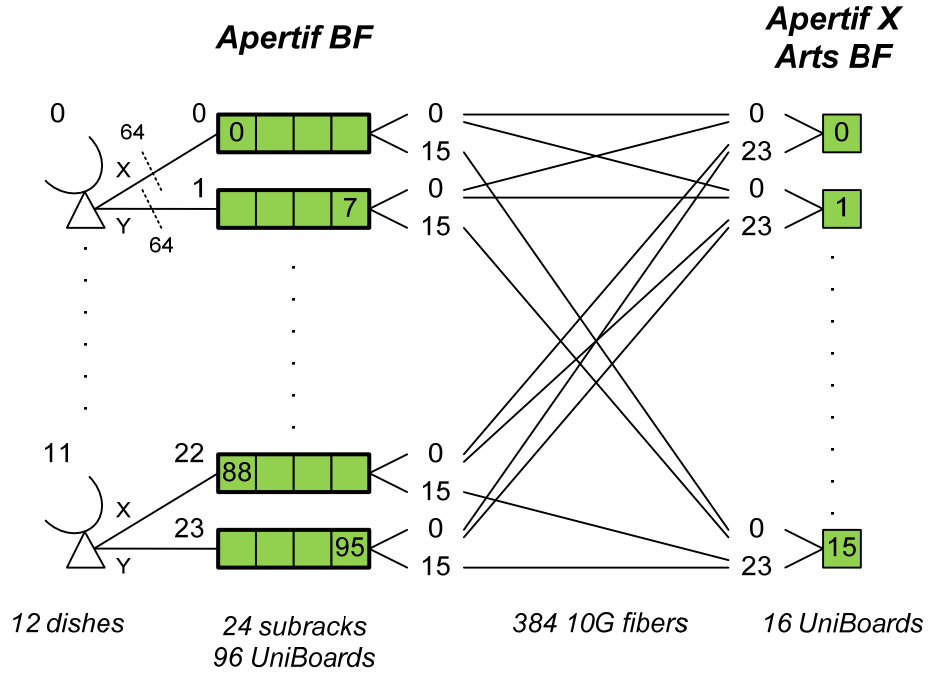

Figure 8: Apertif BF transpose T_{sp} interconnect to Apertif X and Arts

Functionally the Apertif BF output load L_{BF_CB444} consists of $N_{CB}*N_{dish}$ = 37*12 = 444 compound beams that each have a bandwidth of CB_{BW}=300 MHz. Hence each UniBoard in the Apertif X receives $1/N_{band}$ = 1/16

part of the CB_{BW}. Table 13 lists the load definitions for the Apertif BF output with $W_{beamlet}$ = 6 bit. The total load L_{BF_CB444} = 3.2 Tbps and the number of 10G links N_{link}=384 are so large that the Apertif X and Arts BF need to be implemented on FPGAs rather than on GPUs.

Load	Equation	Value	Description
L_{BF_SP1}	= CB_{BW} * $N_{complex}$ * $W_{beamlet}$	3.6 Gbps	Load for 1 SP
$L_{BF\_SP1\_band}$	= L_{BF_SP1} / N_{band}	225 Mbps	Load for 1 SP per band (= per BF node)
$L_{BF\_SP37\_band}$	= N_{CB} * $L_{BF\_SP1\_band}$	8.325 Gbps	Load for N_{CB} = 37 SP per band (= per BF node)
L_{BF_CB1}	= N_{pol} * L_{BF_SP1}	7.2 Gbps	Load for 1 CB (= 2 SP, N_{pol} = 2)
L_{BF_CB12}	= N_{dish} * L_{BF_CB1} = N_{PN} * $L_{BF\_SP1\_band}$	86.4 Gbps	Total load from N_{dish} = 12 dishes, for 1 CB
L_{BF_CB444}	= N_{CB} * L_{BF_CB12} = N_{PN} * $L_{BF\_SP37\_band}$	3.2 Tbps	Total load from N_{dish} = 12 dishes, for N_{CB}= 37 CB
L_{BF_link1}	= $L_{BF\_SP1\_band}$	225 Mbps	Link load for 1 SP per band (= per BF node)
L_{BF_link37}	= $L_{BF\_SP37\_band}$	8.325 Gbps	Link load for N_{CB}= 37 SP per band (= per BF node)

Table 13: Load definitions for Apertif BF output interface with $W_{beamlet}$ = 6 bit

3.3.2 Arts FPGA BF output

The input CB data width from the Apertif BF is $W_{beamlet}$=6 bits. The Arts BF output can be:

- CB voltage data for SC2
- TAB voltage data for SC1 or SC2
- TAB or IAB power data for SC3 and SC4

For CB voltage data the Arts BFoutput data width remains $W_{beamlet}$=6 bit. For TAB voltage beam data the Arts BF output data width is W_{tab}=4 bit and at least 5 bit to include strong sources like Cas A [2]. Choose to output W_{tab}=6 bit to have some more dynamic range. For TAB or IAB power beam data the Arts BF output data is W_{power}=8 bit in a semi-floating point format [2]. Table 14 derives the Arts BF output loads given these data widths. The power beam data can be full Stokes I,Q,U,V or only I. The power beam data is integrated over N_{int}= 10 channel samples.

Load	SC	Equation	Value	Load for
L_{TAB1}	1	$= CB_{BW} * N_{po} * N_{complex} * W_{tab}$	7.2 Gbps	Voltage TAB-1
L_{TAB1_band}	1	$= L_{TAB1} / N_{band}$	450 Mbps	Voltage TAB-1 per PN0
L_{TAB12}	2	$= N_{VLBI} * L_{TAB1}$	86.4 Gbps	Voltage TAB-12
L_{TAB12_band}	2	$= L_{TAB12} / N_{band}$	5.4 Gbps	Voltage TAB-12 per PN0
L_{CB12}	2	$= L_{BF\_CB12}$	86.4 Gbps	Voltage CB-12
L_{CB12_band}	2	$= L_{CB12} / N_{band}$	5.4 Gbps	Voltage CB-12 per PN0
L_{IAB1_stokes}	3	$= CB_{BW} * N_{Stokes} * W_{power}$	9.6 Gbps	IAB-1 without integration
$L_{IAB1\_stokes\_int}$	3	$= L_{IAB1\_stokes} / N_{int}$	0.96 Gbps	IAB-1 with N_{int}=10
L_{IAB37_stokes}	3	$= N_{CB} * L_{IAB1\_stokes}$	355.2 Gbps	IAB-37 without integration
$L_{IAB37\_stokes\_int}$	3	$= N_{CB} * L_{IAB1\_stokes\_int}$	35.52 Gbps	IAB-37 with N_{int}=10
$L_{IAB37\_stokes\_I\_int}$	3	$= L_{IAB37\_stokes\_int} / N_{Stokes}$	8.88 Gbps	IAB-37-I with N_{int}=10
$L_{IAB37\_stokes\_int\_band}$	3	$= L_{IAB37\_stokes\_int} / N_{band}$	2.22 Gbps	IAB-37 with N_{int}=10 per UNB
$L_{IAB37\_stokes\_I\_int\_band}$	3	$= L_{IAB37\_stokes\_I\_int} / N_{band}$	0.555 Gbps	IAB-37-I with N_{int}=10 per UNB
$L_{TAB1\_stokes\_int}$	4	$= L_{IAB1\_stokes\_int}$	0.96 Gbps	Power TAB-1 with N_{int}=10
L_{TAB444}	4	$= N_{CB} * N_{gr} * L_{TAB1}$	3.1968 Tbps	Voltage TAB-444
L_{TAB444_stokes}	4	$= L_{TAB444} * (W_{tab}/W_{power})$	4.2624 Tbps	Power TAB-444 without integration
$L_{TAB444\_stokes\_int}$	4	$= N_{CB} * N_{gr} * L_{TAB1\_stokes\_int}$	426.24 Gbps	TAB-444 with N_{int}=10
$L_{TAB444\_stokes\_I\_int}$	4	$= L_{TAB444\_stokes\_int} / N_{Stokes}$	106.56 Gbps	TAB-444-I with N_{int}=10
$L_{TAB444\_stokes\_int\_band}$	4	$= L_{TAB444\_stokes\_int} / N_{band}$	26.64 Gbps	TAB-444 with N_{int}=10 per UNB
$L_{TAB444\_stokes\_I\_int\_band}$	4	$= L_{TAB444\_stokes\_I\_int} / N_{band}$	6.66 Gbps	TAB-444-I with N_{int}=10 per UNB

Table 14: Load definitions for Arts BF output interface (with $W_{beamlet}$ = 6 bit, W_{tab} = 6 bit, W_{power} = 8 bit)

4 Uniboard hardware

4.1 UniBoard and OEB

The front nodes on UniBoard have optical 10G interfaces and the UniBoard uses an OEB to also provide the back nodes with optical 10G interfaces. In total the UniBoard + OEB then has nof_un * nof_10G = 8*3 = 24 optical 10G interfaces, so just enough to connect to N_{sp}=24 links from the Apertif BF as shown in Figure 8. One UniBoard can process $1/N_{band}$=1/16 part of the CB_{BW}=300 MHz band. To process the whole CB_{BW} there are N_{band} = 16 UniBoards as shown in Figure 9.

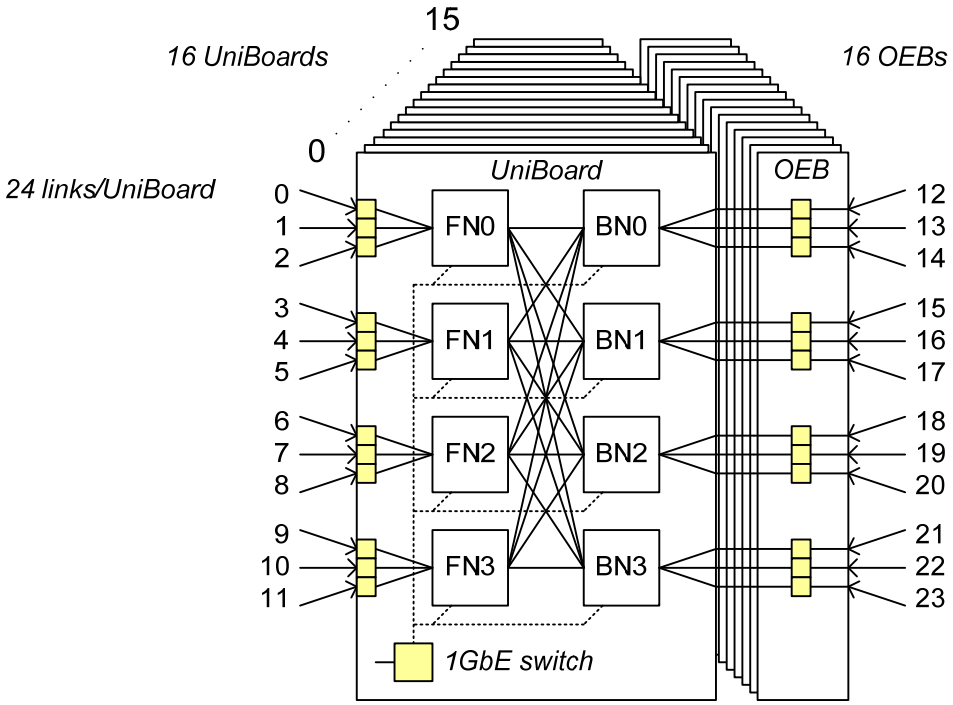

Figure 9: One UniBoard to process $1/N_{band}$ =1/16 part of the CB_{BW} =300 MHz for N_{sp}=24 signal paths

Figure 9 also shows the full duplex mesh interconnect between the FN and BN on UniBoard. In fact UniBoard and OEB are connected in groups of 8 boards via a backplane called CoBI. Figure 9 does not show the CoBI board.

4.2 UniBoard^2 and HEM

The UniBoard^2 has nof_pn=4 FPGA and each PN has 24 optical 10G links. Therefore from an IO point of view UniBoard^2 is 4 times more powerful than a UniBoard and hence M_{uni2} = 4 UniBoard^2 are sufficient to receive the output from the Apertif BF as shown in Figure 10.

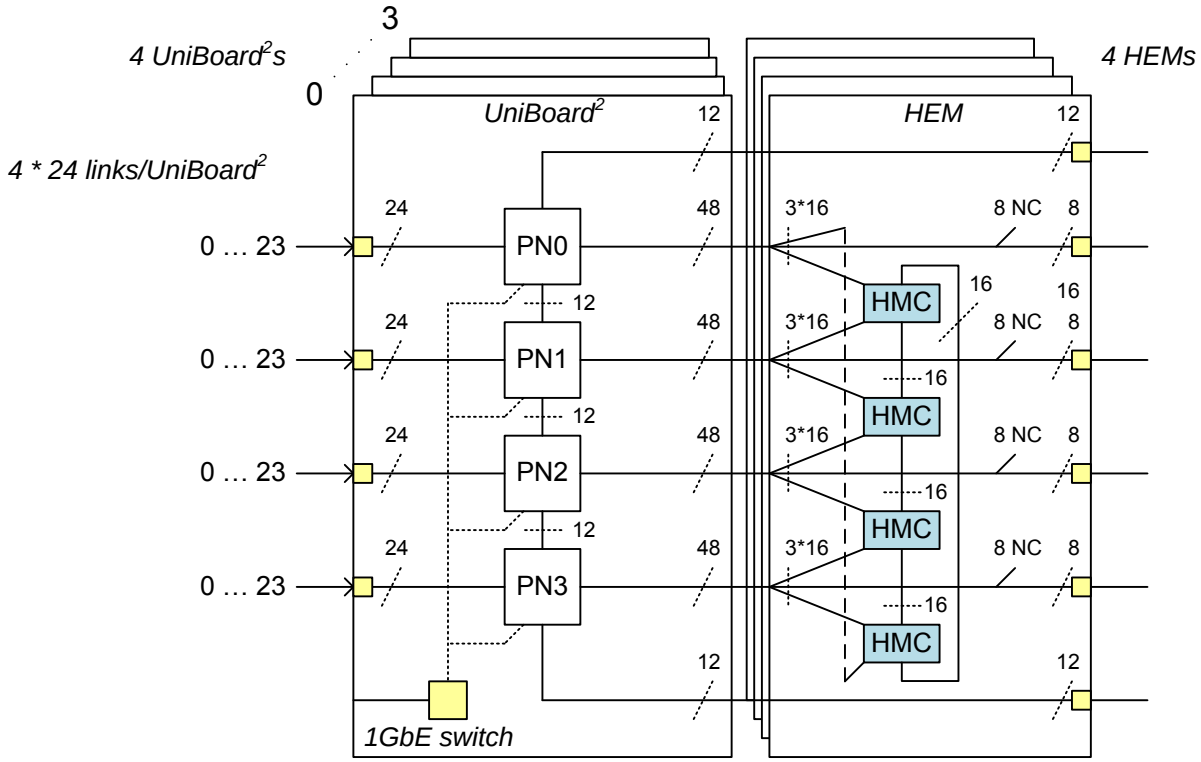

Figure 10: One UniBoard^2 to process 4/N_{band} =1/4 part of the CB_{BW} =300 MHz for N_{sp}=24 signal paths

The Hybrid Memory Cube (HMC) Extension Board (HEM) for UniBoard^2 in Figure 10 is similar to OEB for UniBoard, because it provides 8 extra optical 10G interfaces to the board. In addition HEM also provides extra external memory. This HMC memory is accessed via transceiver links. UniBoard^2 itself has two DDR4 memory modules per PN. With HEM the amount of external memory per PN about doubles.

Without HEM the UniBoard^2 needs to use unidirectional Tx only links to output 10G data to the GPU PL. With HEM the UniBoard2 has 8 extra full duplex 10G (or 2 40G) links available per PN for output to the GPU PL.

The UniBoard^2 also has 12 full duplex links so 120 Gbps, between the PN. The HEM board also closes this full duplex 120 Gbps ring between PN3 and PN0. These inter PN links can be used to exchange (reorder) input data or output data between PN.

4.3 Control interface and data offload via 1GbE

UniBoard in Figure 9 and UniBoard^2 in Figure 10 have an onboard 1GbE switch for access via 1GbE. The 1GbE network is primarily used for control but it can also be used to offload streaming data from the PN (or to the PN). Table 15 summarizes the IO capabilities of the 1GbE interface of UniBoard and UniBoard^2.

Board	Number of 1GbE links per PN	Number of external 1GbE ports board	Maximum IO rate per board
Uniboard	1	4	4 Gbps
Uniboard^2	2	6	6 Gbps

Table 15: 1GbE IO interface for UniBoard and UniBoard^2

4.4 Comparison

4.4.1 Optical links

Table 16 summarizes section 4.1 and 4.2 regarding the 10G optical IO capabilities of UniBoard = EOB and UniBoard^2 + HEM.

	UniBoard + OEB	**UniBoard^2 + HEM**
Number of optical 10G links	4*3 + 4*3 = 24	4*(24 + 8) = 96 + 32

Table 16: Number of optical 10G links per board

4.4.2 External memory

Table 17 shows that 4 UniBoard^2 have a factor 8 less external memory than 16 Uniboards (assuming maximum 16 GByte DDR4 modules). The total access rate is only about a factor 3 less, because DDR4 is faster than DDR3.

External memory type:	**DDR3 on UniBoard**	**DDR4 on UniBoard^2**
Size and access rate:		
Memory module size	16 GByte (max. Micron)	16 GByte (max. Micron)
Memory module access rate	800 MTps (max. 1066 MTps)	2400 MTps (max. 2666 MTps)
Total number of memory slots 16 UniBoard	16 * 8 * 2 = 256	-
Total number of memory slots 4 UniBoard^2	-	4 * 4 * 2 = 32
Total memory size	256 * 16G = 4 TByte	32 * 16G = 0.5 TByte 32 * 32G = 1 TByte
Total access rate	256 * 800M * 64b = 13.1 Tbps	32 * 2400M * 64b = 4.9 Tbps

Table 17: External memory size and access rate

External memory costs about 37 Euro per 4 GByte DDR3 module and 281 Euro per 16 GByte DDR3 module (July 2015). Assume the cost for DDR3 memory on a GPU workstation are a little less, because we do not have to use the largest modules then and for UniBoard we typically need to use modules from Micron. Assume that 16 GByte DDR3 and 16 GByte DDR4 will both cost about 250 Euro per module, then:

- GPU cluster: 4 TByte = 50 kEuro
- UniBoard : 4 TByte = 256 * 250 Euro = 64 kEuro
- UniBoard: 1 TByte = 256 * 40 Euro = 10 kEuro
- UniBoard^2: 0.5 TByte = 32 * 250 Euro = 16 kEuro

4.4.3 Processing resources

Table 18 shows the logic, multiplier and RAM resources that are available for the different FPGAs that are used on UniBoard and Uniboard^2. UniBoard uses 8 Stratix IV FPGAs. UniBoard^2 uses 4 Arria10 FPGAs. End 2016 UniBoard^2 can be fitted with 4 Stratix10 FPGAs that are pin compatible (1932 pins, 45 x 45 mm).

FPGA type:	Stratix IV - EP4SGX230	Arria10 - 10AX115		Stratix10 – 10SG280	
Resource:			x		x
ALM [1])	91200	427200	4.6	933120	10.2
Registers (FF)	182400	1708800	9.3	3732480	20.4
Logic Elements (LE)	228k	1150k	5.0	2753k	12.0
M9K	1235	-		-	
M144K	22	-		-	
M20K [2])	-	2713	2.2	11721	9.5
Mbit	14.283	53	3.7	229	16.0
18x18 multipliers	1288	-		-	
18x19 multipliers [3])	-	3036	2.3	11520	8.9
Clock max [4])	200 MHz	200 MHz	1	400 MHz	2

Table 18: FPGA processing resources

Notes for Table 18:

1) One ALM (Adaptive Logic Module) on Stratix IV contains 2 flipflops (FF) while on Arria10 and Stratix10 it contraints 4 FF. The ALM can be represented by a certain number of logic elements (LE). The LE figure can be used to compared the general purpose logic resources between Altera FPGAs.
2) Compare the number of M20K block RAM with the number of M9K block RAM. The unit block RAM size on Stratix IV is M9K = 9 kBit while on Arria10 and Stratix20 it is M20K = 20 kBit. Often at least one block RAM is needed therefore a worst case comparison compares the number of block RAMs and a best case comparison compares the amount of Mbits.
3) Compare the number of 18x19 multipliers with the number of 18x18 multipliers.
4) The maximum core clock rate Fmax for Stratix IV and Arria 10 is about 550 MHz but in practise this means that for applications only 200 MHz is easily achievable and more only with extra design effort. For Stratix10 Fmax is about 1 GHz, hence assume a practical clock rate is 400 MHz, so a factor 2 more.

From Table 18 it follows that the Arria10 FPGA is about a factor 2 - 4 more powerful than the Stratix IV FPGA whereas the Stratix10 FPGA is about a factor 16 - 32 more powerful than the Stratix IV.

4.4.4 Processing load estimates for Apertif X

Table 19 shows an estimate of the number of FPGA resources that will be needed for the Apertif X. The estimate is based on a synthesis result of Apertif X with channel filterbank (using 9b FIR coefficients) and the correlator. For the input reorder section via the mesh (T_{sp}) the synthesis results from another design were used. The sum of these results is an estimate for the final Apertif X resource usage. The notes for Table 18 also apply to Table 19. The fact that the number of ALMs > 100 % implies that the final synthesis will have to pack the logic more efficiently into the ALMs, which is feasible, because the number of FF is still < 100 %. Nevertheless the Apertif X FPGA is almost full.

FPGA resources	Stratix IV - EP4SGX230	Apertif X	Usage
Logic (ALM)	91200	108k	120 %
Registers (FF)	182400	175k	96 %
Block RAM (M9K)	1235	960	78 %
Multipliers (DSP 18x18)	1288	1084	84%

Table 19: Estimated FPGA resources for Apertif X

4.4.5 Processing load estimates for Arts SC3

Table 20 shows an estimate of the extra number of FPGA resources that will be needed for the Arts SC3. The estimate is based on the synthesis results of:

- a BF unit without weights and without BST (to represent the IAB-37)
- the node_unb1_ddr3_reorder function (to represent the transient buffer access control)
- one extra io_ddr component to represent that both DDR3 memory slots are needed
- a mesh reorder function (to represent the output of the Stokes I data via the mesh)
- a tr_10GbE component (to represent the output of the Stokes I data via 10GbE)

FPGA resources	Stratix IV - EP4SGX230	Arts SC3	Usage
Logic (ALM)	91200	22k	24 %
Registers (FF)	182400	28k	15 %
Block RAM (M9K)	1235	106	8 %
Multipliers (DSP 18x18)	1288	96	7 %

Table 20: Estimated FPGA resources that are needed extra for commensal Arts SC3

Adding the numbers in Table 19 and Table 20 shows that the Stratix IV FPGA cannot fit Apertif X and SC3 together on the same FPGA. The bottleneck are the logic resources (the number of ALM and FF).

4.4.6 Processing load estimate per TAB

The channel filterbank requires 288 18bx18b multipliers, assuming 8 FIR filter taps and 9 bit FIR filter coefficients, and N_{chan}=4 channels. The number of multipliers that is needed to beamformer 1 TAB per PN is $N_{sp} * P_{cmult} * f_{data}/f_{clk}$ = 24 * 4 * 0.5 = 48 18bx18b multipliers. Here N_{sp} = 24, P_{cmult} = 4 for one complex multiply and f_{data}/f_{clk}=100MHz / 200MHz=0.5, so 1 TAB requires 48 multipliers. The maximum available number of TABs K_{TAB} given in Table 21.

Platform	Number of multipliers per PN	Relative number of PN	Relative f_{clk}	K_{TAB}	K_{TAB} / N_{gr}
16 UniBoard with StratixIV	1288	1	1	(1288-288)*1*1/48 = 20	1.66
4 UniBoard2 with Arria10	3036	1/8	1	(3036-288)* 0.125*1/48 = 7	0.58
8 UniBoard2 with Arria10	3036	2/8	1	(3036-288)* 0.25*1/48 = 14	1.16
4 UniBoard2 with Stratix10	11520	1/8	2	(11520-288)*0.125*2/48 = 58	4.83

Table 21: Maximum number of TABs

The required number of TABs for SC4 is N_{gr}=12. Clearly 4 UniBoard^2 with Stratix10 can implement SC4. The 16 UniBoard can also implement SC4. Four UniBoard^2 with Arria10 running at f_{clk} = 200 MHz can only achieve 7 TABs instead of 12.

4.4.7 Hardware cost

Table 22 lists the costs of using dedicated hardware using 16 UniBoards in a similar configuration as for the Apertif X.

Component	Cost per component [Euro]	Number of components	Total cost [Euro]
UniBoard	12.5k	16	200k
OEB	2.5k	16	40k
PAC	4k	2	8k
Mini-PAC	1k	2	2k
CoBI	3k	2	6k
subrack	2k	2	4k
48V power supply	6k	2	12k
SFP fiber connector	50	2*384 [2])	40k
Total:			**312k**

Table 22: Cost for Arts BF using dedicated hardware using 16 UniBoards

Table 23 lists the cost when 4 UniBoard^2 with Arria10 FPGAs are used.

Component	Cost per component [Euro]	Number of components	Total cost [Euro]
UniBoard^2	15k (using Arria10)	4	60k
UniBoard^2	45k (using Stratix10)	4	180k
HEM	5k	4	20k
box	1k	4	4k
48V power supply	6k	2	12k
SFP fiber connector	50	384 [1])	20k
QSFP fiber connector	200 [2])	96 [1])	20k
Total:			**116k with Arria10 [3])**
			236k with Stratix10 [3])

Table 23: Cost for Arts BF using dedicated hardware using 4 UniBoard^{2}s with Arria10 or Stratix10

Notes for Table 23:
1) The number of SFP connectors is 2 * 384 to allow the scheme of Figure 13.
2) Assume that 1 QSPF connector for Uniboard^2 costs as much as 4 SFP connectors.
3) Cost without HEM.

Table 24 lists the cost when 8 UniBoard^2 with Arria10 FPGAs are used. The extra 4 UniBoard^{2}s provide extra processing and are connected in a daisy chain to the first 4 UniBoard^{2}s (similar as in Figure 14).

Component	Cost per component [Euro]	Number of components	Total cost [Euro]
Complete system of 4 UniBoard^2 with Arria10	116k	1	116k (from Table 23)
UniBoard^2	15k (using Arria10)	4	60k
box	1k	4	4k
48V power supply	6k	2	12k
QSFP fiber connector	200	96	20k
Total:			**212k**

Table 24: Cost for Arts BF using dedicated hardware using 8 UniBoard^{2}s with Arria10 (and no HEM)

4.4.8 Power consumption and cost

Exact power consumption estimates are difficult to give because they depend on the processing and the number of IO links. However assume that UniBoard takes 250 W and that UniBoard^2 with Arria10 or

Stratix10 will take 300 W, then 16 Uniboards take 4 kW and 4 UniBoard^2 take 1200 W. The difference is 2800 W. Assume Arts is powered continuously for 5 years. A year has ~10k hours and assume that 1 kWh costs 0.20 Euro, then the energy cost savings by using UniBoard^2 are 2800 * 10k * 0.00020 * 5 = 28 kEuro. The savings will even be more because the energy for cooling will also be less.

The difference in between using 4 or 8 UniBoard^2 with Arria10 is 1200 W. Hence the extra power consumption cost of using 4 more UniBoard^{2}s are 1200 * 10k * 0.00020 * 5 = 12 kEuro.

4.5 Conclusion

A system with 16 UniBoards has 16 * 8 =128 FPGAs. A system with 4 UniBoard^2 has 4 * 4 = 16 FPGAs. Hence for the same processing a UniBoard^2 FPGA needs to be 8 times as powerful regarding IO (Table 16), external memory (Table 17) and logic, DSP multipliers, on chip RAM (Table 18). SC4 requires nearly all FPGA resources of the 16 UniBoards. SC1, SC2 and SC3 need less FPGA resources, because they make less beams (TAB-1, TAB-12, IAB-37) and because on UniBoard^2 they do not need the logic to redistribute the CB data for the T_{sp} transpose that UniBoard does need. Conclusion:

- 4 UniBoard^2 with Stratix10 FPGAs can easily replace 16 UniBoards also for SC4 (Table 21), with 1 TByte they also have sufficient external memory.
- 4 UniBoard^2 with Arria10 FPGAs can replace 16 UniBoard from an IO point of view and for SC2 and SC3 this is also enough from a processing point of view. However from a processing point of view 8 UniBoard^2 with Arria10 FPGAs are needed for SC4 (assuming f_{clk}=200 MHz).

5 GPU cluster workstations

5.1 Input data

From Table 14 is follows that the integrated full Stokes data rate for SC4 that comes from the Arts BF is $L_{TAB444\_stokes\_int}$ = 444 (TABs) * 300 MHz (TAB bandwidth) * 4 (Stokes parameters) * 8 bit (power data) / 10 (number of samples per integration period to have T=50 us per channel) = 426.24 Gbps = 53.3 GByte/s.

If only the Stokes I data is transported then $L_{TAB444\_stokes\_I\_int}$ = 106.56 Gbps.

5.2 Transient data buffer

If the full Stokes data is transported then the transient data buffer can be implemented in the Arts BF or in the Arts PL. If only the Stokes I data is transported, then the transient data buffer needs to be implemented on the Arts BF.

5.3 GPU cluster

The Arts pipeline (PL) processing will be done on a cluster of workstations with GPUs. Each workstation (machine) in the GPU cluster has a PCI express bus that connects:

- CPUs with DDR3 memory
- GPUs with DDR3 memory
- 1*1GbE interface for control
- 2*10GbE interface for data
- 40G Infiniband interface for IO between the workstations

Assume a cluster with $N_{workstation}$ = 24 GPU workstations and that the workstations are interconnected via a 24 port 40G Infiniband switch.

5.3.1 Input data rate per work station

With 24 workstations and a total data rate of 426.24 Gbps TAB-444 full Stokes data each workstation gets 17.76 Gbps = 2.22 GByte/s, so two 10GbE interfaces should be enough. If only the TAB-444 Stokes I data needs to be transported, then the data rate is a factor N_{stokes}=4 less, so 4.44 Gbps = 0.555 GByte/s and then one 10 GbE interface per workstation suffices.

5.3.2 Transient data buffer storage

The TAB-444 full Stokes data needs to be buffered (SR-0.35). If this is not done on the UniBoards in the Arts BF then it needs to be done in the Arts PL. For 15 s this requires in total 15 s * 53.3 GByte/s = 0.8 TByte, so about 40 Gbyte/ workstation and about 53.3 G / 24 = 2.22 GByte/s (write) access rate per workstation.

5.3.3 T_{band} data transpose to bring together the 300 MHz band

The data arrives from 16 UniBoards and each UniBoard provides 1/16 th of the 300 MHz band width. To bring these N_{band}=16 bands together requires either an 10GbE Ethernet switch in between the Arts BF and the Arts PL or a 24 port Infiniband switch in the Arts PL. For SC1 and SC3 it is also feasible to use a 1GbE/10GbE switch.

5.3.4 Pipeline (PL) processing

The streaming pipeline (PL) processing operates on the TAB-444 Stokes I data. Per workstation the Stokes I input data rate per workstation is 4.44 Gbps = 0.555 GByte/s, this needs to be processed by the GPUs. The GPU processing may also need access to the DDR3 and it will need IO to output results via 1GbE.

6 Critical system requirements and hardware constraints

6.1 Commensal modes

Apertif X, Arts SC1 and SC4 do not have to run at the same time. Hence Apertif X, Arts SC1 and SC4 can reuse the same hardware and then run their dedicated application.

Arts SC2 requires local interferometer data from Apertif X (SR-028 [1]), but only for the central CB. Hence for Arts SC2 it is sufficient to implement only the central CB part of the Apertif X. The purpose of Arts SC3 is to run in parallel with Apertif X (SR-038 [1]) hence Arts SC3 requires running in parallel to the full Apertif X.

6.2 Fringe stopping

Fringe stopping (FS) applies to both Apertix X and Arts BF and is a task that can be done in the Apertif BF. Typically fringe stopping is a two-step process that consists of a true sample delay before a filterbank and phase tracking per frequency channel after the filterbank. For the Apertif the true sample delay tracking (DT) occurs on the ADC samples at the input of the Apertif BF followed by phase tracking (PT) of the beamlet data to stop the residual fringe. The phase tracking needs to be done per beamlet, because it depends on the subband frequency and on the CB direction.

The delay tracking involves duplicating an input sample or skipping an input sample. The phase tracking involves phase rotation by a complex multiplication. The phase tracking must be synchronous to the delay tracking, because a delay step causes a phase step. For the Nyquist frequency a delay step of one sample causes a phase step of 180 degrees. The delay step also disturbs the subband filtering if it is not accounted for. To have a smooth delay tracking the ADC data with the extra sample or the skipped sample needs to be reprocessed by the subband filterbank for the duration of the filterbank impulse response. The subband filterbank in the Apertif BF has 16 taps and an FFT size of 1024, so with 8 bit ADC samples this requires storing at least 16 kByte per ADC input.

Currently the delay tracking is implemented in RAM in the Apertif BF, but only accounts for buffering the number of samples that are expected for the maximum geometrical delay difference between dishes. The phase tracking is not implemented yet in the Apertif BF. The smooth delay tracking to recalculate the subbands after each delay step is also not implemented yet in the Apertif BF. If the internal RAM is limited, then the delay tracking may require using external memory DDR3. After every delay step the data subband filterbank processing needs to catch up with the input data rate of 200 MHz. This implies that for smooth delay tracking the subband filterbank will need to run at slightly more than 200 MHz. An alternative to smooth delay tracking is to flag the 16 beamlet time samples as being disturbed by the delay step. However for each dish the delay step is typically applied at a different instant, so this then causes quite some flagging. It is sufficient to flag the start of a disturbance, it is not necessary to exactly flag all data that will get affected. Without smooth delay tracking it is necessary to have some rudimentary flagging of the delay step instances, to avoid inform the scientists that there may be small artefacts in the data. The smooth delay tracking can then be added in a later stage of the development.

The fringe stopping may best be implemented in the Apertif BF, because that is the central location. When the fringe is stopped then all user applications of the Apertif BF output data can then rely on it.
.

6.3 Transient data buffer

6.3.1 CB-444 voltage data

Arts SC3 requires a transient data buffer that can store 10 s of CB-444 data (SR-040 [1]). The CB-444 voltage data is only available within the Arts BF so it cannot be stored in the Arts PL. From Table 13 it

follows that the Apertif BF output load L_{BF_CB444} = 3.2 Tbps with $W_{beamlet}$ =6 bit. To store 1 s of this CB voltage data requires 0.4 TByte. Hence to store 10 s requires 4 TByte.

6.3.2 TAB-444 integrated power data

Arts SC4 requires a transient data buffer that can store 15 s of integrated full Stokes TAB-444 data (SR-035 [1]). This data is available in the FPGA beamformer and may be available in the GPU PL provided that the full Stokes data is output to the PL. From Table 14 it follows that the integrated full Stokes data rate is $L_{TAB444\_stokes\_int}$ = 426.24 Gbps = 53.3 GByte/s with W_{power} = 8 bit. To store 1 s of this TAB power data requires 53.3 GByte, so to store 15 s requires 0.8 TByte.

The DM search range in the Arts PL is about 5 s. This DM search range is slided in steps of 1 s. Therefore with 1 s margin before and after the detection the minimum required storage time is 5 + 2*1 = 7 s. This would require about 0.4 TByte.

6.3.3 External memory operation

During an observation the external memory is used as a circular buffer that is continuously written with the streaming CB or TAB data. When a trigger occurs then the writing stops to freeze the buffer. If the external memory is in the Arts BF then the Arts PL will provide a trigger to the Arts BF probably via the Arts MAC. All CB or TAB data gets written, but only the one CB or a few TAB that had the trigger need to be read. The read access control is typically done via the Arts MAC. Therefore the read data rate is much lower than the write data rate. The read data may be output to the Arts PL via 1GbE or 10GbE (see section 6.18.5).

6.3.4 External memory in Arts

Table 25 summarizes the external memory that is available within the Arts BF when either UniBoard or UniBoard^2 is used and in the Arts PL when CUDA 4230 GPU Workstations are used.

Platform:	UniBoard	UniBoard^2	CUDA 4230 GPU Workstation
Number of FPGA per board	8	4	-
Number of DDR3 memory modules per FPGA	2	-	-
Number of DDR4 memory modules per FPGA	-	2	-
Number of DDR3 memory modules per workstation			12
Number of boards in Arts BF	16	4	-
Number of work stations in Arts BF	-	-	24
Total number of DDR memory modules in Arts BF	256	32	288
Total memory in Arts BF using 16 GByte DDR memory modules	4 TByte	0.5 TByte	4.6 TByte
Total memory in Arts BF using 32 GByte DDR memory modules	-	1 TByte	-
Transfer rate per memory module	800 MTps	2400 MTps	1866 MTps
Data rate per memory module (64b data, write only, ~80% effective)	40 Gbps	120 Gbps	95 Gbps / 6
Total access rate	10 Tbps = 1.25 TByte/s	3.8 Tbps = 0.47 TByte/s	0.4 Tbps= 0.05 TByte/s

Table 25: External memory in Arts

Conclusion:

- 16 UniBoard can store 4 Tbyte which is sufficient for the CB voltage data (4 TByte for SC3) or TAB power data (0.8 TByte for SC4).
- 4 UniBoard^2 can store 1 TByte which is sufficient for the TAB power data (0.8 TByte for SC4) and when SC4 can run commensal then SC3 is no longer needed. The 32 GByte modules are not available yet from Micron. With 16 Gbyte modules that are available from Micron the 4 UniBoard2 can store 0.5 TByte, which implies storing 0.5/0.8 * 15 s = 9.3 s. This 9.3 s is also sufficient, because the minimum is about 7 s (see section 6.3.2).
- 24 GPU workstations can have sufficient memory to fit the CB voltage data or TAB power data, but the DDR3 access rate store the TAB power data may be an issue. In Table 25 it is assumed that the 12 DDR3 modules per workstation can be accessed via two ports in parallel (12/ 2 = 6).

6.4 Apertif X integation interval transpose T_{int} in the Apertif BF

As shown in Figure 5 and Figure 8 both Apertif X and Arts BF use the same Apertif BF output. The Apertif BF performs a transpose over T_{int} = 1 s to reorder the CB data appropriately for the Apertif X. For Arts SC 3 this T_{int} introduces a delay of 1 s which can be an issue if Arts needs to trigger other systems (SR-034 [1]).

The transpose over T_{int} also influences the order in which the TAB weights need to be applied. However TAB weights are not needed in SC3. For the other SC1, 2, and 4 that do use TAB weights the transpose over T_{int} can be bypassed in the Apertif BF, but that requires a new bypass function in the Apertif BF and a second Apertif BF CB-444 output data format. The Apertif BF output format for Apertif X with transpose over T_{int} is specified in [4]. The Apertif BF output format with transpose over T_{int} includes an additional interleave factor $N_{interleave}$ = 2 to group the nof_un=8 beamlets per subblock. For Apertif BF output without with transpose over T_{int} this additional interleave factor $N_{interleave}$ = 2 is not needed, because then the data blocks already have the beamlets at the fastest array index.

To avoid dependencies with the Apertif X it is better to bypass the T_{int} and to use a dedicated Apertif BF output for Arts. In case the commensal SC2 and SC3 run on the same UniBoards as Apertif X then T_{int} cannot be bypassed, but this scenario is unlikely to fit.

6.5 Transpose T_{sp}

6.5.1 On UniBoard

The CB-444 transpose T_{sp} that brings together all N_{sp}=24 SP on one node on UniBoard is achieved in two steps. The first step is done by the wiring of the N_{link}=384 links in Figure 8. The second step requires an input redistribution function in the FPGAs that uses the mesh interconnect on the UniBoard to keep 1/nof_un= 1/8 part at each node and transport the rest to the other nof_un-1=7 processing nodes (PN). After the complete T_{sp} transpose each PN can process about N_{CB}/nof_uni = 37/8 ≈ 5 CB for one complete band and all SP as shown in Figure 11. Hence the T_{sp} in Figure 11 preserves the full band of CB_{BW}/N_{band} = 300/16 MHz or N_{FN}=24 subbands. An alternative scheme would be to preserve the FoV by letting each PN process all 37 CB, but only 1/nof_un = 1/8 part of the band. For Arts it is necessary to preserve the full band.

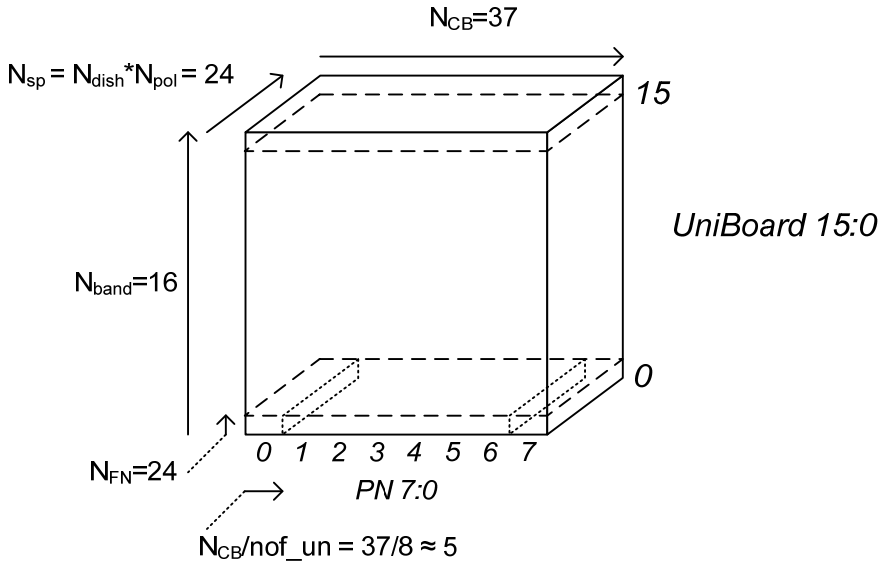

Figure 11: T_{sp} transpose of the CB-444 data to UniBoard via fiber and to each PN via the mesh

6.5.2 Load on the UniBoard mesh

The CB-444 input load per PN on UniBoard is L_{BF_CB444} / M_{uni} / nof_un = 3.2 Tbps / 16 / 8 = 25 Gbps. This CB data needs to be redistributed via the mesh: 1/8 remains at this PN and 7/8 is send across. Then 4/8 remains across and 3/8 needs to be send across again, because a BN can reach an FN directly but another BN only via an FN (and similar for FN to BN or FN). Hence the total data rate to the mesh per PN is (7+3)/8 * 25 Gbps = 31.25 Gbps. This load is transported to nof_fn = nof_bn = 4 nodes across so 31.25 Gbps / 4 = 7.8125 Gbps per BN-FN link. The capacity of the mesh is 12 Gbps per BN-FN link.

6.5.3 Using UniBoard^2

The advantage of UniBoard^2 is that only the first step done by the wiring of the N_{link}=384 links in Figure 8 is already enough to get the N_{sp} links to one node. No further redistribution functionality is needed which makes the firmware design smaller and simpler. The PN on UniBoard^2 only needs to align the N_{sp}=24 inputs. On UniBoard^2 one PN does what one UniBoard does, in total there are N_{band}=16 PN and each PN processes all 37 CB for one band and all SP as shown in Figure 12.

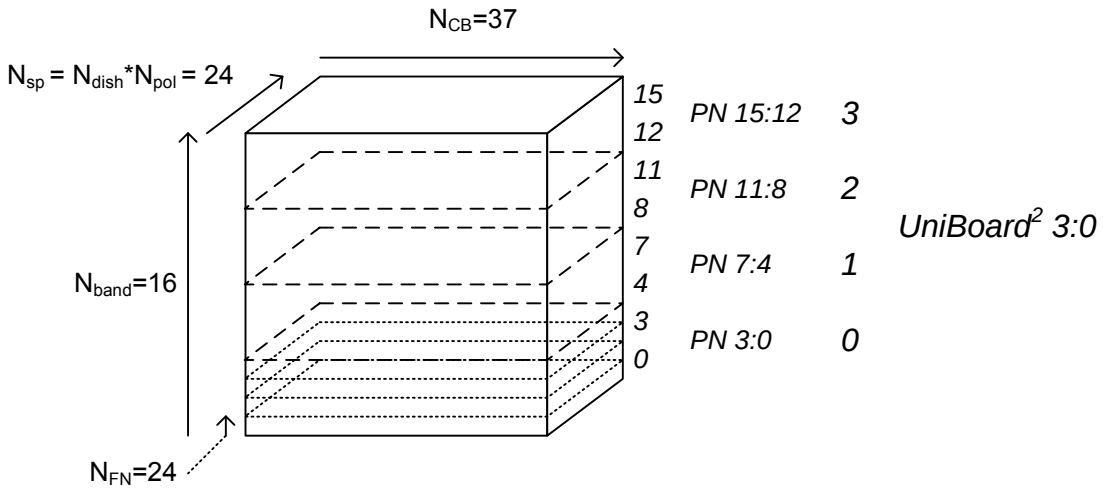

Figure 12: T_{sp} transpose of the CB-444 data on UniBoard^2 via fiber directly to each PN

6.6 Channel band width and time resolution of the Stokes beam data

For SC3 and SC4 the CB beamlets need to be separated into N_{chan}=4 channels (SR-0.30 [1]). The beamlet bandwidth set by the subband bandwidth B_{sub} = 781250 Hz or 1 MHz, so the channel bandwidth is B_{chan} = 195.3125 or 250 kHz. The smaller channel band requirement causes:

- The channel filter is implemented as a polyphase filterbank and consumes substantial resources in the FPGA.
- The resolution of the Stokes power beam data must T_{Stokes} ≤ 50 µs (SR-0.31 [1]). This at most about $N_{int} = T_{Stokes}*B_{chan} \leq$ 9.8 or 12.5 channel samples can be integrated. Therefore set N_{int}=10 as used in Table 14. Hence given a required time resolution T_{Stokes} ≤ 50 µs the data output rate is increases with N_{chan}.

6.7 Streaming output full Stokes or output only Stokes I data

If only Stokes I data is output then $L_{TAB444\_stokes\_I\_int\_band}$= 6.66 Gbps for SC4, so factor N_{Stokes}=4 less. However the disadvantage is then that the transient data buffer with full Stokes data has to be on the Uniboards (see section 6.3).

6.8 Number of bits per sample

The number of bit per sample $W_{beamlet}$= 6 bit, W_{tab}= 6 bit and W_{power} = 8 bit have a direct impact on the Arts BF output load concerning:

- The total output data rate
- Whether the data can be transported via an 1GbE links or has to use 10GbE links
- The integer number of links that are needed per UniBoard

The Arts BF data output rates for SC1, 2, 3 and 4 are listed in Table 14. The L_{TAB1_band} = 450 Mbit for SC1 and $L_{IAB37\_stokes\_int\_band}$ = 2.22 Gbps for SC3 are small enough to be transported via the 1GbE switch on UniBoard. The L_{CB12_band} = 5.4 Gbps and L_{TAB12_band} = 5.4 Gbps for SC2 is too much for the 1GbE on UniBoard, so it requires using at least one 10GbE link. The $L_{TAB444\_stokes\_int\_band}$= 26.64 Gbps for SC4 requires using at least 3 10GbE link per Uniboard. If only the Stokes I data needs to be transported then $L_{TAB444\_stokes\_I\_int\_band}$= 6.66 Gbps and then 1 10GbE link per UniBoard is enough for SC4.

6.8.1 Data packing and unpacking

To avoid too much unpacking overhead in the Arts PL is may be necessary to transport the CB voltage data for $W_{beamlet}$ = 6 bit and TAB voltage data for W_{tab} = 6 bit with 6 bit per bytes instead of as packed 6 bit. In that case L_{TAB1_band} = 600 Mbps for SC1 (still fits on a 1GbE link), and L_{CB12_band} = 7.2 Gbps and L_{TAB12_band} = 7.2 Gbps for SC2 (still fits on a 10GbE link). The 6 bit should not be simply sign-extended to 8 bit, because the extra 2 bits are then better used to transport 2 more MSbits of the internal TAB data to reduce the number of clipped samples.

6.9 Duplicate Apertif BF output

Dedicated UniBoards or UniBoard^2 for Arts can be connected in parallel to the UniBoards for Apertif X using a second 10G output on the FN of the Apertif BF as shown for 16 UniBoards in Figure 13.

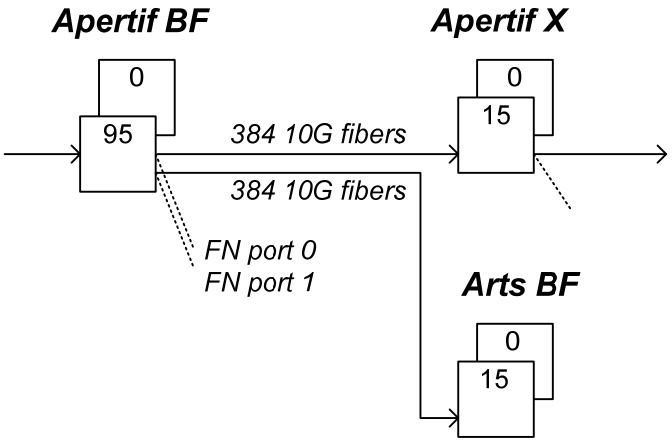

Figure 13: Duplicate Apertif BF output to 16 Uniboards for Arts BF

The pro and con of duplicating the Apertif BF output are:

Pro:

- Dedicated beamlet output for Arts can be tapped of before the T_{int} integration period transpose that is used for Apertif X (see section 6.4)
- Arts and Apertif X are independent apart from that they share the same Apertif BF

Con:

- Development to provide the extra 10G output in Apertif BF
- Extra set of 384 long distance fiber optics needed. The fibers are already available.

6.10 Pass on Apertif BF output in daisy chain

Dedicated UniBoards or UniBoard^2 for Arts series by letting the Apertif X pass on the data as shown for 16 UniBoards in Figure 14. Note that the pass on scheme can be repeated in a daisy chain to add even more boards in series in case more processing is needed or other applications need access to the CB-444 data.

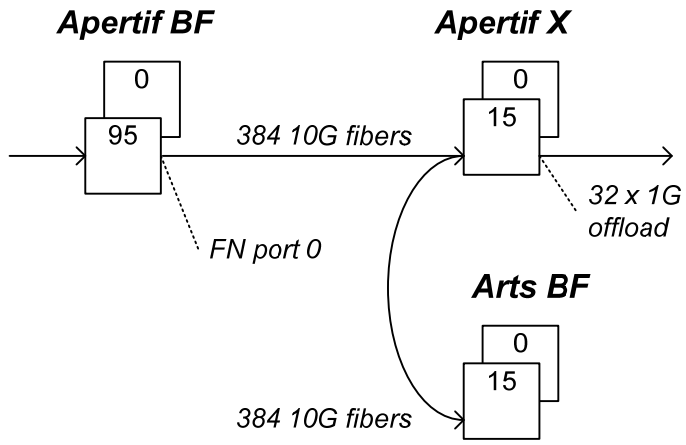

Figure 14: Pass on Apertif BF output via Apertif X to 16 Uniboards for Arts BF

The pro and con of passing on the Apertif BF output are:

Pro:

- Apertif X may pass on the input CB data or Arts can benefit from transpose T_{sp} that is done on Apertif X.
- Vice versa Arts may pass on the data to Apertif X, which can be used to save resources in either Arts BF or Apertif X.

Con:

- Set of 384 short fiber cables needed
- Development of pass on function in firmware of Apertif X
- Arts can only operate if the Apertif X is also active and relies on Apertif X to always pass on the data.

6.11 Using the same 16 UniBoards of Apertif X also for Arts

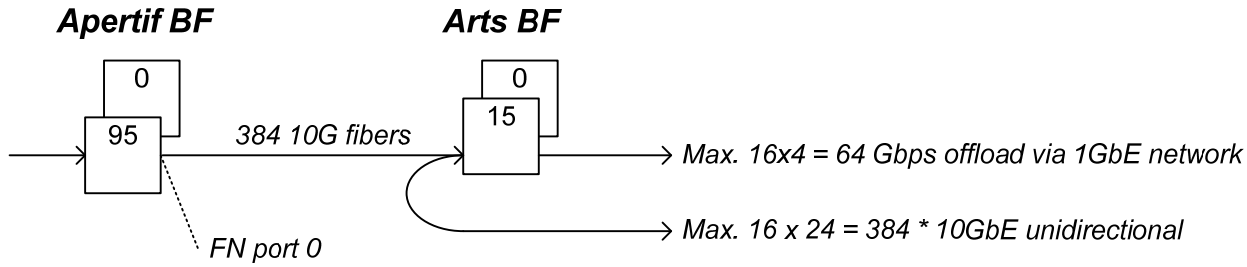

Figure 15: Arts BF on 16 UniBoards

Pro:

- Saves the cost of 16 extra UniBoards for Arts
- Fits SC1 and 4 because these are not active at same time as Apertif X.

Con:

- SC2 and SC3 commensal mode and Apertif X probably will not fit together with Apertif X regarding IO, processing and memory.
- Combining SC2 and SC3 with Apertif X in one FPGA firmware design complicates the development.
- Switching between dedicated Apertif X FPGA image and Arts FPGA image requires that they both fit in the flash, because rewriting the flash too often wears the flash (>10000x erase/write is allowed according to datasheet).

6.12 Using 16 dedicated UniBoards for Arts

Pro:

- Arts SC1,2,3,4 can always run in parallel with Apertif X
- If SC4 can run commensal then SC3 is not needed, which saves the development effort for SC3.

Con:

- Cost of 16 extra UniBoards
- Provide the extra 10G output in Apertif BF (section 6.9) or pass on CB data in Apertif X (section 6.10).

6.13 Using more than 16 dedicated UniBoards for Arts processing

Pro:

- With 32 Uniboards and N_{band}=16 inputs used still 8 10G ports are available for Arts BF output
- With 32 Uniboards there is a factor 2 more DDR3 memory available then with 16 UniBoards

Con:

- Extra cost
- Arts can only operate if the Apertif X is also active.
- Only feasible for multiple of 16 extra UniBoards, so eg. 32

For the original CB-444 data the number of UniBoards needs to be a multiple of 16, whereby the data is then passed on in a daisy chain from Apertif-X to the first set of 16 UniBoards and then to the next set of 16 UniBoard. For the T_{sp} transposed CB-444 data the number of UniBoards also needs to be a multiple of 16, it is not possible to use eg. only 24 UniBoards for Arts then because each PN physically has only nof_10G = 3 ports and functionally it can only parallelize the data further over 5 CB (see section 6.5.1). The band (N_{FN}) and the SP (N_{sp}) need to be kept together, so parallelization can only occur for the 5 CB. However the main restriction is that without a switch it is not possible to distribute the data to other than nof_10G = 3 destinations.

6.14 Using more than 4 dedicated UniBoard^{2}s for Arts processing

To increase the processing and memory capabilities for Arts it is possible to use more than 4 UniBoard^2 for Arts. However similar as in section 6.13 the number of Uniboard^2 needs to be a multiple of 4 so that they can be connected in a daisy chain to pass on all N_{band}=16 bands that form the CB_{BW} = 300 MHz.

Adding only one extra UniBoard^2 can only add extra processing for ¼ part of the CB_{BW} = 300 MHz, so 75 MHz. One extra Uniboard^2 cannot add 25 % more TABs for the entire band, because that would require a transpose T_{band} for which one UniBoard^2 does not have sufficient IO ports.

6.15 Output redistribution inside the Arts BF

6.15.1 Via 1GbE

On UniBoard and UniBoard^2 all PN are connected directly to a 1GbE switch with 4 external ports. The 1GbE network is also used for MAC, but the spare capacity is still close to 100% and can be used for data output offload by the PN.

6.15.2 Via the UniBoard mesh for 10GbE output

If there need to be less than nof_un=8 10GbE output per UniBoard, then the PN have to use the mesh to pass on their output to another PN. To redistribute not only the T_{sp} input data but also the T_{band} output data via the UniBoard mesh complicates the implementation but is possible. The capacity of the mesh is 12 Gbps per BN-FN link. From section 6.5.2 it follows that there is still about 4 Gbps spare capacity per BN-FN link on the mesh. The output redistribution scheme depends on the number of 10GbE outputs that are needed per UniBoard. Table 26 shows the maximum load on per BN-FN link on the mesh for 1, 2, 3 or 4 10GbE output ports per UniBoard. For the lowest data rate per BN-FN link it is necessary have all outputs on one side (eg. at the FN) to use only one output port per node and to distribute the data to all 4 BN for the first hop and to all output FN. Separating the data over multiple streams complicates the implementation. If the output data is not separated then the load per BN-FN link equals the load per PN.

Number of 10GbE outputs per UniBoard	Maximum load per PN	Output node	Transport scheme for the mesh	Maximum load per BN-FN link
1	1.25 G	FN0	BN3:0 send directly to FN0 FN3:1 send to FN0 via BN3:1	2 * 1.25G = 2.5 G
1	1.25 G	FN0	BN3:0 send directly to FN0 FN3:1 send to FN0 via BN3:0	(8-1) / 4 / 1 * 1.25G = 2.1875 G
2	2.5 G	FN0, BN0	BN2:0 send directly to FN0 FN2:0 send directly to BN0	2.5 G
2	2.5 G	FN1:0	BN-i sends to FN1:0 FN3:2 send to BN3:0	(8-2) / 4 / 2 * 2.5G = 1.875 G
3	3.75 G	FN2:0	BN-i sends to FN2:0 FN3 sends to BN3:0	(8-3) / 4 / 3 * 3.75G = 1.5625 G
4	5 G	FN3:0	BN-i sends to FN-i	5 G
4	5 G	FN3:0	BN-i sends to FN3:0	5 G/ 4 = 1.25 G

Table 26: Load per BN-FN link for output redistribution via the mesh in Gbps

6.16 Output data reorder

6.16.1 In time

The data in an Ethernet packet may need to be reordered in the Arts FPGA beam former to ease the processing in the Arts PL GPU cluster. Reordering data takes RAM resources of the FPGA, because the data needs to be stored before it can be read in the required output order.

6.16.2 Per final destination

The Arts BF will have to reorder the output data such that the packet payloads are already prepared for their final destination. For the option with the Infiniband switch the workstation could unpack the incoming packets and distribute their contents to the final destinations, but it probably does not have time to do this.

6.17 Unidirectional 10GbE links

For UniBoard + OEB and for UniBoard^2 all optical 10G ports are used for receiving the Aperif BF data. These ports can still be used for transmitting data to the GPU PL, but only in a unidirectional way. Typically an Ethernet link is full duplex to support e.g. ARP and ping, therefore it needs to be verified that a unidirectional Tx only 10GbE link between UniBoard and a GPU server is possible (both directly or via 1 switch).

6.17.1 Using UniBoard^2 to convert unidirectional 10GbE to full duplex 10GbE

If unidirectional links cannot easily be connected to 10GbE network interfaces then a UniBoard^2 can be used to convert an unidirectional link into a full duplex 10GbE link as shown Figure 16. Uniboard^2 with HEM has in total 24 + 8 = 32 optical links per PN. For SC4 the full Stokes output has $L_{TAB444\_stokes\_int}$ = 426.24 Gbps, so then 3 ports per UniBoard are needed as shown in Figure 16.

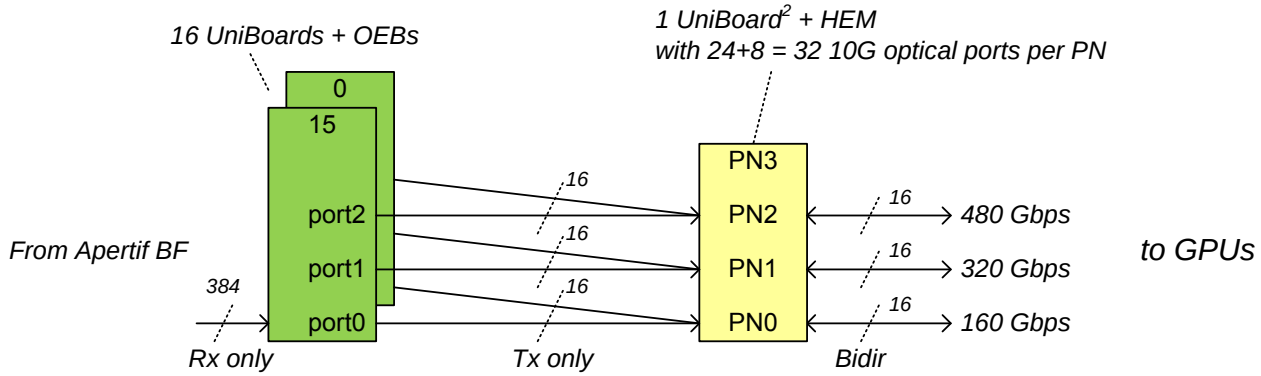

Figure 16: Using UniBoard^2 + HEM to convert Tx only to bidir 10GbE (3 ports for full Stokes)

If only the Stokes I is output then $L_{TAB444\_stokes\_I\_int}$ = 106.56 Gbps and then using 1 port per UniBoard and only PN0 on UniBoard^2 suffices as in Figure 16 or alternatively using two PN on UniBoard^2 without HEM suffices as shown in Figure 17. The data exchange between PN0 and PN1 in Figure 17 uses the ring IO that is available between the PN on UniBoard^2 (see Figure 10).

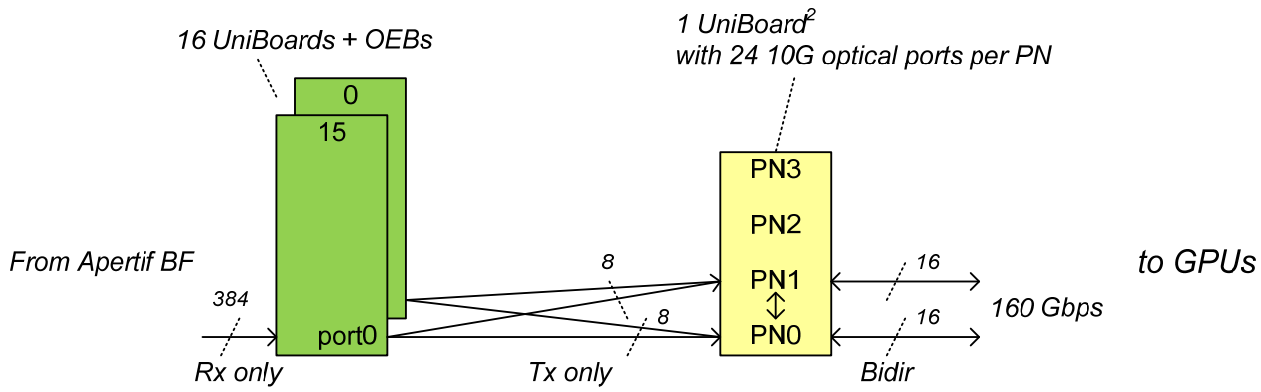

Figure 17: Using two PN on UniBoard^2 without HEM to convert Tx only to bidir 10GbE (1 port for Stokes I)

6.17.2 Using UniBoard^2 for processing

If 4 UniBoard^{2}s with HEM can be used for processing instead of 16 UniBoards then the issue of unidirectional links vanishes as shown in Figure 18, because UniBoard^2 with HEM has still 8 more 10GbE ports available per PN.

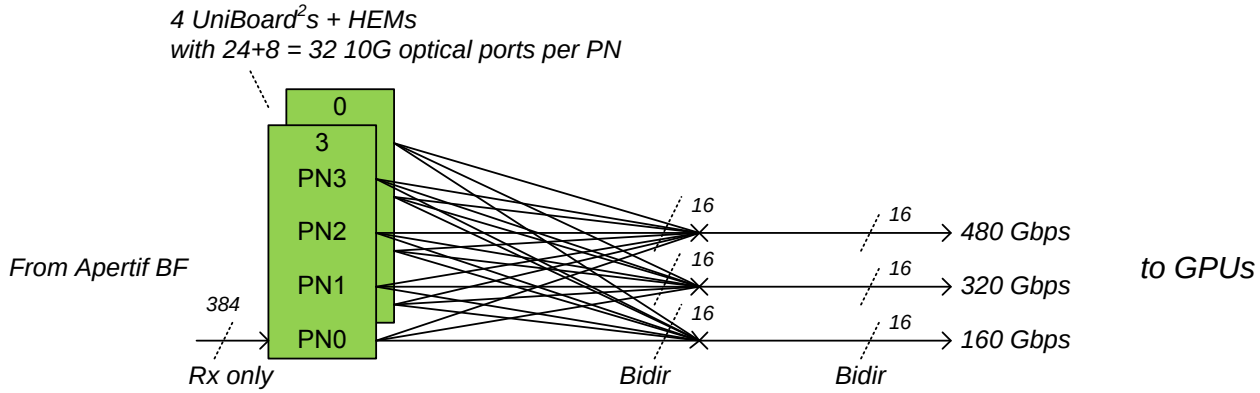

Figure 18: Using UniBoard^2 + HEM for processing to avoid use of Tx only 10GbE ports for output

6.18 Transpose T_{band}

6.18.1 TAB-1 for SC1 using 1GbE and a dedicated switch

For SC1 the load L_{TAB1} = 7.2 Gbps (Table 14) from N_{band} = 16 Uniboards can be collected to a single GPU server by means of a Ethernet switch with at least 16 1GbE ports and 1 10GbE port as shown in Figure 19. The switch also implements the transpose T_{band}, because it brings together the TAB-1 data for the entire CB_{BW} to a single link.

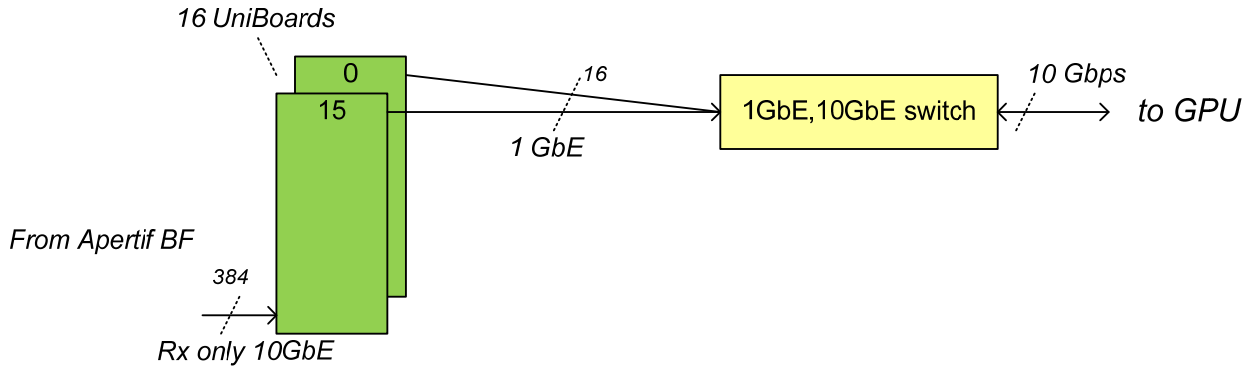

Figure 19: Arts BF output for SC1 via 1GbE network

Remarks:

- The TAB-1 data for SC1 will be calculated on FN0, the other PN on UniBoard are only needed for the transpose T_{sp} to get all SP at FN0. Hence the SC1 output could also be done via the 10GbE output of FN0 on the 16 UniBoards. However then a 10GbE switch is needed with 16 + 1 = 17 ports.
- For SC1 probably only 1 workstation suffices, so then it is beneficial that T_{band} is implemented by the switch as well.

6.18.2 TAB-444 for SC4 via 10GbE

For SC4 the $L_{TAB444\_stokes\_int}$ = 406.24 Gbps (Table 14) from N_{band} = 16 Uniboards or from 16 PN on 4 UniBoard^{2}s. As shown in Figure 16 and Figure 18 this requires a total link capacity of 3 * 16 * 10GbE = 480 bps. If only Stokes I is transported then 1 * 16 * 10 GbE = 160 Gbps is needed. The transpose T_{band} can be done via:

- 10GbE switches each with 32 ports, see Figure 20
- 1 Uniboard^2 + HEM, see Figure 21
- 1 UniBoard^2 using two PN, see Figure 22
- An Infiniband switch that interconnects the workstations in the GPU cluster, see Figure 23.

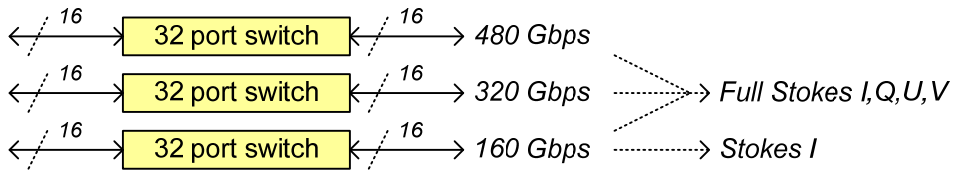

Figure 20: T_{band} for SC4 via 10GbE switches

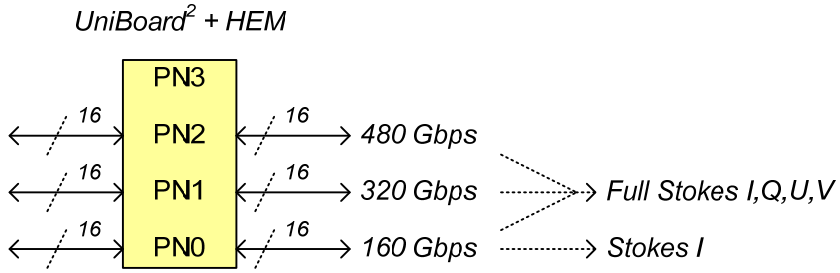

Figure 21: T_{band} for SC4 via one UniBoard^2 + HEM

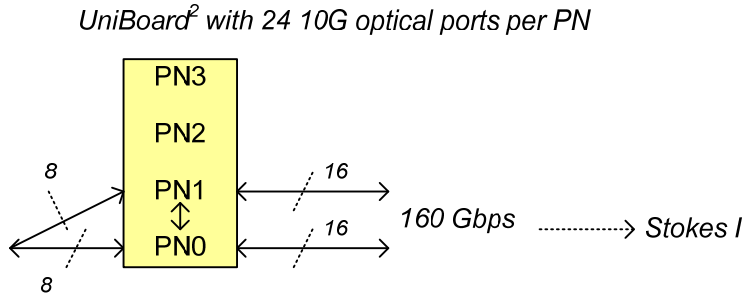

Figure 22: T_{band} for SC4 via one UniBoard^2 without HEM by using two PN

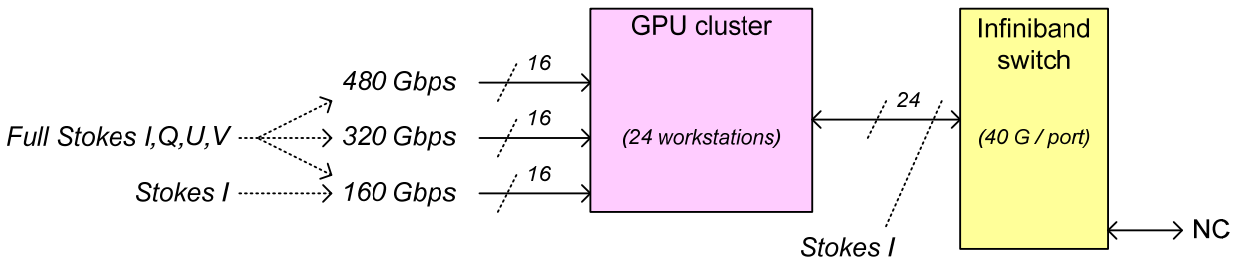

Figure 23: T_{band} for SC4 via an Infiniband switch that interconnects the GPU workstations

If an Infiniband switch performs T_{band} and if the workstations in the GPU cluster have a multiple of N_{band} = 16 10GbE ports, then point-to-point links can be used between Arts BF and Arts PL, see Figure 23.

The advantage of using a UniBoard^2 as in Figure 21 is that the PN on UniBoard^2 can convert unidirectional Tx only input to full duplex output (section 6.17.1) and the PN can also perform output data reordering (see section 6.16). The advantage of using an Ethernet switch or Infiniband switch is that this requires no FPGA firmware development. An Inifiniband switch costs as much as a few workstations, therefore it may be appropriate to have an Infiniband switch available in the GPU cluster anyway, even if it is not used to do the transpose T_{band}.

6.18.3 CB-12 and TAB-12 for SC2

For SC2 the load is L_{CB12} = 86.4 Gbps or L_{TAB12} = 86.4 Gbps. The SC2 only uses the central CB and this data is available on FN0. Similar as for SC1 the other PN on UniBoard are only needed for the transpose T_{sp} to get all SP for the central CB at FN0. With N_{band} = 16 UniBoards the output load per board is 5.4 Gbps so using 1 10GbE port per FN0 suffices. For the interconnect to the GPU cluster SC2 can use the interconnect of SC4 as shown in section 6.18.2.

6.18.4 IAB-37 for SC3

For SC3 the $L_{IAB37\_stokes\_I\_int}$ = 8.88 Gbps (Table 14) from N_{band} = 16 Uniboards. Per UniBoard this requires 1 1GbE link so to collect the data from N_{band} = 16 UniBoards requires 1 Ethernet switch with 16 1GbE ports and 1 10GbE port similar as for SC1. The IAB-37 data for SC3 will be calculate on all PN, so the on board switch aggregates the data from the FN and BN.

6.18.5 Transient buffer data readout for SC3 and SC4

After a trigger the transient data only needs to be read for the CB (in case of SC3) or the few TABs (in case of SC4) in which the transient occurred. The load for 1 CB for N_{dish} = 12 dishes is L_{BF_CB12}=86.4 Gbps, so 10 s is 864 Gbit = 108 Gbyte. This amount is best offloaded via 10GbE interfaces. The load for 1 TAB is $L_{TAB1\_stokes\_int}$=0.96 Gbps, so 15 s is 14.4 Gbit = 1.8 GByte. This data for 1 TAB or a few TABs can be read via the 1GbE control interface or offloaded via the 1GbE interface.

7 Hardware status

7.1 UniBoard

- The 4 GByte DDR3 is working on UniBoard and on all 4 UniBoards in an Apertif BF subracks (used for T_{int} storage of Apertif X).
- The prototype CoBI / OEB is not working yet (necessary to support N_{sp}=24 instead of only 12)
- Order and verify using a 16 GByte DDR3 memory on UniBoard (necessary for 10 sec CB-444 voltage data buffer with SC3)

7.2 UniBoard^2

- The UniBoard^2 prototype is being tested. The 1GbE and 10GbE interfaces are all working, however the DDR4 memory can be written and read but still has some data bit errors. The first production board of UniBoard^2 with Arria10 will arrive December 2015.
- The HEM prototype board is currently in the schematic design layout phase.
- Production samples for the Stratix10 FPGAs are expected in Q3 2016 and engineering samples in Q4 2015. Hence a first production board of UniBoard^2 with Stratix10 can be available end 2016.
- Order and verify using a 16 GByte and a 32 GByte DDR4 memory module on UniBoard^2 (necessary for 15 sec TAB-444 power data buffer with SC4)

7.3 GPU cluster

- The Dragnet GPU cluster for pulsar and fast radio transient search with LOFAR can serve as reference example.