ngVLA Electronics Memo #22

ngVLA SBP Hardware Downselection: Technical Trade Study Update

O. A. Yeste Ojeda (NRAO)

September 11, 2026

1 Executive Summary

This memo evaluates alternative hardware technologies for the ngVLA Central Signal Processor (CSP) Subband Processor (SBP). The implementation of the frequency channelization and beamforming (B&C) stage is the open question. The study compares two approaches that implement this function separately from the cross-correlation engine, using either a custom application-specific integrated circuit (ASIC) or commodity graphics processing units (GPUs) in PCIe servers, with two integrated approaches that combine both functions in a single system, using either a GPU rack-scale platform or field-programmable gate arrays (FPGAs). These are referred to in the study as the ASIC B&C, PCIe GPU B&C, NVL72 SBP, and FPGA SBP, respectively.

The PCIe GPU B&C has the lowest estimated acquisition cost at $27.8M and the lowest estimated 20-year electricity cost at $5.6M. The GPU implementation also provides greater algorithm flexibility and lower lifecycle engineering effort because the signal processing is software-defined. Prototype benchmarking has demonstrated that the B&C workload can be implemented in real time on a GPU. The principal remaining uncertainty is whether the assumed PCIe 6.0 GPU can sustain the B&C workload assigned to each processing node, which determines the final node count, operating power, and system size.

The ASIC B&C has comparable estimated operating power but substantially higher NRE cost and less algorithm flexibility. The FPGA SBP and NVL72 SBP have substantially higher hardware acquisition cost and operating power, together with additional lifecycle or platform dependencies. The estimated full-system costs, including acquisition and 20 years of electricity, are $33.4M for the PCIe GPU B&C, $48.9M for the ASIC B&C, $46.6M for the NVL72 SBP, and $42.7M for the FPGA SBP. These figures exclude technology refresh, maintenance, and spares, which are not quantified by this study.

The PCIe GPU B&C is the preferred candidate under the current trade-study assumptions. Final selection is subject to completion of the targeted B&C workload benchmarking, which will determine whether the assumed GPU implementation can sustain the required workload within the planned GPU power envelope.

Contents

2 Introduction & Scope

This memo provides an updated technical assessment to support the hardware downselection of the SBP within the Central Signal Processor during the Preliminary Design phase. The candidate technologies have been assessed in prior studies as capable of meeting the system requirements and are evaluated here under a common set of assumptions to enable direct comparison. This update also introduces the PCIe GPU B&C mapping, extending to the B&C stage the commodity GPU technology already selected for the X-Engine. The objective of the Preliminary Design downselection is to identify the processing technology to carry forward into the development phase and implement in the qualification model for Final Design Review (FDR) testing. The final technology downselection is confirmed at FDR. The selection may be revised after the Preliminary Design Review (PDR) only in response to a significant technological development.

The downselection target is the SBP hardware, and the B&C-Engine implementation is the open question. The ASIC B&C and PCIe GPU B&C candidates implement the B&C-Engine separately from the X-Engine and are referred to as the split architectures. The NVL72 SBP and FPGA SBP candidates implement both functions in a single architecture and are referred to as the integrated architectures. The standalone X-Engine is included in the analysis so that the split architectures can be compared with the integrated architectures on a complete SBP basis.

3 Trade Study Methodology

This trade study compares the candidate architectures using a common set of system assumptions and assessment criteria. The detailed design of each candidate is described in Section 4, and the corresponding technical analysis is presented in Section 5.

The comparison is made at the full SBP level. The split architectures are therefore evaluated together with the standalone X-Engine required to provide the complete SBP function. Unless otherwise stated, the comparison uses the full ngVLA array and the maximum processed bandwidth considered in this study. When estimating the number of nodes required, beamforming and cross-correlation are treated as separate workloads rather than simultaneous workloads. Any candidate-specific assumptions that affect the comparison are stated explicitly in the relevant analysis.

The candidates are evaluated using five criteria: power efficiency, engineering and programmatic risk, lifecycle maintainability, cost, and algorithm flexibility. Power efficiency is an important criterion because lower power consumption reduces the energy required to operate the system and its associated environmental impact over the operating life. Cost includes the total cost of ownership, including non-recurring engineering, capital expenditure, operating expenditure, and energy cost. Technology refresh is not included in the cost model. Routine replacement hardware is assumed to be provided through spares, while a full technology refresh would require a separate proposal and cost assessment and is outside the scope of this study.

The final recommendation considers all five criteria. Requirement compliance is a prerequisite, and the recommended architecture is the one that provides the strongest overall combination of power efficiency, engineering and programmatic risk, lifecycle maintainability, cost, and algorithm flexibility.

4 Hardware Architecture Candidates

This section describes the four hardware candidates. All candidates ingest the coarsely channelized subband stream produced by the ngVLA Digital Back-End (DBE) and perform fine frequency channelization, beamforming, and cross-correlation. The ASIC B&C is built on custom silicon, the PCIe GPU B&C uses commodity GPUs in standard servers, the NVL72 SBP is based on NVIDIA GB300 NVL72 racks, and the FPGA SBP uses an FPGA-based architecture. The standalone X-Engine required by the split architectures is analyzed in Section 5.

4.1 ASIC B&C Design

The ASIC B&C candidate implements the B&C stage using a custom ASIC. Each chip performs geometric delay and phase tracking, fine frequency channelization, and participates in the distributed beamforming process.

Beam-specific geometric delay compensation is implemented directly in the ASIC. Each B&C node contains a delay-phase tracker for each generated beam, which applies the required delay and phase correction to the antenna signal before its contribution is accumulated in the distributed beamforming process. The architecture supports 10 simultaneous beams across the full bandwidth using true-time-delay compensation, with beamforming implemented through the on-chip ring topology.

The hardware is organized into B&C nodes, B&C Units, and Line Replaceable Units (LRUs). Each chip is a B&C node. Eight nodes form a B&C Unit that processes 8 antennas and 4 subbands. A COTS FPGA gateway aggregates the eight nodes and provides the unit's network interface. Each B&C Unit is implemented on a single motherboard. Two B&C Units form an LRU, the field-replaceable building block of the deployment.

The architecture produces two output streams. Channelized per-antenna voltages are sent to a separate GPU-based X-Engine for cross-correlation. The corner turn is performed by the network, redistributing the per-antenna data streams from the B&C nodes into per-frequency-channel streams for the X-Engine nodes. Beamformed outputs are delivered on an independent path to the downstream backends. The design is entirely air-cooled. Although its signal processing algorithms are fixed in silicon, the architecture supports all required observing modes by reconfiguring operating parameters rather than the algorithms themselves.

4.2 PCIe GPU B&C Design

The PCIe GPU B&C candidate assumes NVIDIA PCIe 6.0 GPUs that have not yet been released, as successors to the current L40S [1] and RTX PRO 6000 Blackwell [2] products. The GPUs are hosted in standard servers. Each GPU is paired with an 800GbE-class network interface card and operates as an independent processing node. Channelization is distributed across processing nodes by antenna, with each node receiving the subband streams of one or a few antennas from the DBE and channelizing them locally.

Channelized data from the B&C nodes is redistributed by a corner turn so that all antennas for a given frequency channel are processed together. Each B&C node tracks the bulk geometric delay of its antenna toward the pointing center, and this correction serves all beams. The remaining beam-specific differential delay is applied as a per-channel phase shift in the fine frequency channelization. The delay-compensation requirements are described in Appendix B.

The number of beams affects compute requirements on the processing nodes but does not increase network traffic, because the corner-turn volume is set by the antenna and bandwidth product and does not depend on the beam count. The system is sized for the full array operating in one observing mode at a time. Smaller subarrays can operate simultaneously in both modes when sufficient SBP resources are available. Beamformed outputs are delivered from the processing nodes to the downstream backends.

The candidate is entirely air-cooled and follows standard server thermal practice. Its signal processing runs in software on general-purpose GPUs, so the processing algorithms can be revised after deployment without hardware changes.

4.3 NVL72 SBP Design

The NVL72 SBP candidate implements the complete SBP using NVIDIA GB300 NVL72 racks. Each rack integrates 36 Grace CPUs and 72 Blackwell Ultra GPUs connected through a rack-scale NVLink domain. Each GPU operates as an independent processing node and provides 800GbE external network connectivity.

Channelization is distributed across processing nodes by antenna, whereas the beamforming and X-Engine workloads are distributed by frequency channel. The NVLink fabric performs the corner turn within the rack, allowing the same processing nodes to execute both the B&C and X-Engine kernels without traversing the external network. Beamforming follows the same approach as the PCIe GPU B&C candidate described in Section 4.2.

The architecture supports dynamic allocation of processing resources to subarrays and all required observing modes through software. The GB300 NVL72 platform is liquid-cooled and fully software-defined, providing algorithm flexibility.

4.4 FPGA SBP Design

The FPGA SBP candidate implements the complete SBP using an FPGA-based FX correlator architecture derived from the Advanced Technology ALMA Correlator (ATAC) developed for the ALMA2030 Wideband Sensitivity Upgrade (WSU) [3, 4, 5]. The ATAC design provides the reference architecture and sizing basis for this candidate, which is scaled to the ngVLA requirements.

The reference architecture consists of a Very Coarse Channelizer (VCC) followed by multiple Frequency Slice Processors (FSPs), each processing a single frequency subband independently. In the ngVLA architecture, the DBE performs the VCC function. Consequently, only the FSPs are considered in this study, replacing both the B&C and X-Engine functions of the SBP. The reference design processes 200 MHz of non-redundant bandwidth per FSP, which differs from the subband definition used by the current ngVLA DBE. This study therefore assumes an updated DBE capable of producing subbands compatible with the reference architecture.

The reference hardware platform is the BittWare TeraBox 1501B FPGA server [6], with two FPGAs per server. At ngVLA scale, each FSP spans four servers. The corner turn therefore crosses servers but remains confined within each FSP, where it is carried over a passive optical mesh interconnecting the FPGAs of the FSP.

The FSPs perform the frequency channelization, correlation, and beamforming processing. The available processing resources therefore constrain the frequency channelization used for beamforming (see Appendix B).

The architecture supports all required observing modes through assignment of FSPs to the corresponding mode, with each FSP programmed for a single observing mode and sized for the full array. Different FSPs may therefore operate in different observing modes simultaneously. The design is liquid-cooled and provides algorithm flexibility through FPGA reprogrammability.

5 Technical Analysis

The following subsections present the technical analysis of the four hardware candidates for the full-scale ngVLA configuration of 263 antennas and 20 GHz of processed bandwidth, based on the current design descriptions of Section 4. The same aspects are considered for each candidate: rack count, operating power, cooling approach, and primary unresolved assumptions.

5.1 ASIC B&C Analysis

The B&C Unit is the basic hardware building block of the ASIC B&C architecture. Each B&C Unit processes 8 antennas and 4 subbands. Supporting the full array therefore requires 33 antenna groups ($\lceil 263/8 \rceil$)[1], while supporting the full 20 GHz processed bandwidth (92 subbands) requires 23 subband groups ($\lceil 92/4 \rceil$). The resulting implementation therefore requires 759 B&C Units.

[1] $\lceil x \rceil$ denotes the ceiling function, the smallest integer greater than or equal to x.

Each LRU contains two B&C Units, resulting in a total of 380 LRUs for the full implementation. Each B&C Unit consumes an estimated 150 W, yielding a total B&C-Engine power consumption of 113.85 kW. The implementation occupies 13 racks, based on 30 of the 1U LRUs per 42U rack. This corresponds to 9 kW per rack, which is consistent with conventional air-cooled rack power levels.

The 150 W figure is a design-budget value rather than a measured production value. A bottom-up review of the device-level power budget indicates that it is plausible but leaves limited margin, with gateway FPGA power representing the dominant uncertainty. The value is therefore retained as the documented planning baseline rather than replaced by an independently derived estimate.

5.2 PCIe GPU B&C Analysis

The basic hardware building block of the PCIe GPU B&C architecture is a processing node consisting of one PCIe 6.0 GPU paired with an 800GbE-class network interface card. Each node receives and channelizes the complete 92-subband stream from a single antenna. At 8 Gb/s per subband, this corresponds to an ingest rate of 736 Gb/s, or approximately 92% of the 800GbE interface. Supporting the full array therefore requires 263 processing nodes.

The candidate has been benchmarked using an NVIDIA L40 GPU [7] provided by NVIDIA as a prototype platform. The prototype meets the required 250 MS/s input rate at 19 dual-polarization subbands per GPU, with the compute stage using approximately 83% of the real-time processing budget. At 23 subbands, the compute stage uses approximately 99% of the budget, but the host-staged input path does not sustain the required input rate. The limit at that point is the PCIe 4.0 link of the L40 rather than compute capacity. The 23-subband result therefore provides a reference for GPU compute capacity, while the 19-subband result is the demonstrated end-to-end real-time operating point. The difference comes from the prototype using host-memory staging instead of the GPUDirect [8] path planned for production.

Each processing node budgets 300 W for the GPU, 75 W for an 800GbE-class network interface card, and 100 W of amortized server overhead covering the CPU, memory, fans, and power-supply losses, assuming four processing nodes per server. This gives an estimated power consumption of 475 W per node, or approximately 125 kW for the complete B&C-Engine. At 8 of the 4U servers per 42U rack, the B&C-Engine occupies 9 racks and draws 15 kW per rack, within typical air-cooled rack power levels. At the 23-subband operating point, the L40 recorded approximately 299 W of GPU board power and operated at its 300 W limit throughout the measured steady-state interval. This reading covers GPU board power only, not the network-interface or server-overhead terms. The measured GPU board power therefore supports budgeting the full 300 W board-power envelope for the production GPU. As a sensitivity case, a 600 W GPU envelope would raise the B&C-Engine estimate to approximately 204 kW.

The 263-node configuration is the planning baseline, contingent on a production PCIe 6.0 GPU sustaining

the complete 92-subband workload for one antenna. This requires approximately four times the compute capacity measured at the 23-subband operating point, within the same 300 W board power envelope. Host I/O is not expected to be limiting with the intended production data path. Two uncertainties remain. First, PCIe 6.0 GPUs are not yet on the market, so the candidate assumes they will be available on the ngVLA procurement timeline. Second, it has not yet been demonstrated that the production GPU can sustain the required real-time throughput and memory bandwidth. If it cannot, additional processing nodes can be added at the expense of higher power consumption.

5.3 NVL72 SBP Analysis

The NVL72 SBP candidate is sized first by the aggregate ingest bandwidth of the array. Each antenna-subband produces 8 Gb/s, giving a total input rate of approximately 194 Tb/s for 263 antennas and 92 subbands. Each GB300 GPU is paired with an 800GbE network interface, giving a 72-GPU NVL72 rack an aggregate ingest capacity of 57.6 Tb/s. The array therefore requires at least four NVL72 racks based on aggregate ingest bandwidth alone.

A representative mapping that fits within this four-rack lower bound assigns 8 subbands and 11 antenna streams to each GPU. Each GPU therefore receives 704 Gb/s of input, or approximately 88% of the available 800GbE interface. Covering all 263 antennas for one 8-subband frequency slice requires $\lceil 263/11 \rceil = 24$ GPUs. Covering all 92 subbands requires $\lceil 92/8 \rceil = 12$ such frequency slices, for a total of 288 GPUs, exactly filling four NVL72 racks. Other antenna/subband partitions satisfying the same network constraint are possible. This mapping is used as a representative configuration for the subsequent resource analysis.

The NVL72 is a liquid-cooled rack-scale system, with the cooling infrastructure integrated into the rack. The four-rack implementation corresponds to an estimated operating power of 528–560 kW, based on the published 132–140 kW power rating for a GB300 NVL72 rack [9]. Because actual device utilization for the combined B&C and X-Engine workload has not been established, the published rack power rating is treated as a conservative upper-bound estimate. The upper bound of 560 kW is carried into the operating cost estimate in Section 6.2.4.

The combined B&C and X-Engine workload is estimated at approximately 80 TOPS per GPU, based on operation counts rather than measured application throughput [10]. The PCIe GPU B&C prototype described in Section 5.2 provides some evidence that system-level constraints can affect sustained application performance, although the specific transport and power constraints observed in that prototype do not apply to the GB300. No sustained-throughput measurement of the combined workload exists for the GB300. Memory bandwidth is not expected to be the primary constraint, given the 8 TB/s of HBM3e bandwidth available per GB300 GPU [11]. The selected mapping also confines each corner-turn domain to a single NVLink fabric, with the resulting corner-turn traffic corresponding to approximately 8% of the available per-GPU NVLink bandwidth [12]. Compute throughput and the actual memory-bandwidth requirement therefore remain the

principal resources requiring validation on the target platform.

The four-rack configuration is consistent with the available I/O, memory bandwidth, and NVLink resources of the GB300 NVL72 platform. It is used as the planning baseline, subject to implementation-level benchmarking to establish whether the combined B&C and X-Engine workload can be sustained in real time.

5.4 FPGA SBP Analysis

The FPGA SBP candidate is sized by scaling the reference ATAC FSP architecture to the ngVLA requirements. Each FSP processes 200 MHz of bandwidth, so covering the required 20 GHz processed bandwidth requires 100 FSPs. In the reference design, each FPGA processes up to 33 antennas. Supporting the full 263-antenna array therefore requires $\lceil 263/33 \rceil = 8$ FPGAs per FSP, resulting in a total requirement of 800 FPGAs. Assuming the selected hardware platform with two FPGAs per server, as used in the ATAC design [13], the complete implementation requires 400 servers.

Assuming a representative power consumption of approximately 1 kW per liquid-cooled dual-FPGA TeraBox server, including the FPGAs, CPUs, memory, network interfaces, cooling, and power-supply losses, the 400-server implementation corresponds to an estimated operating power of approximately 400 kW. The assumed packing of 5 FSPs per rack results in 20 racks at 20 kW per rack. The packing assumption is a planning value for this study and should be revisited when the allowable rack power for the assumed cooling configuration is established.

Compute throughput is expected to increase relative to the ATAC reference implementation. The per-FPGA X-Engine workload is estimated to increase by approximately a factor of four relative to the reference design. Demonstrating that the selected FPGA platform can sustain this increased computational density is therefore one of the principal technical uncertainties of the candidate, together with adaptation of the updated ngVLA DBE interface, implementation of the inter-server corner turn within each FSP, and the beamforming channelization requirements.

5.5 Standalone X-Engine Analysis

The standalone X-Engine (XE) is common to both split architectures and is therefore not evaluated as an independent candidate in this trade study. Its architecture and technology selection are adopted from the dedicated XE trade study [10]. This section summarizes the resulting implementation used in the system-level comparison.

The dedicated XE trade study concludes that the processing node is limited by data ingest rather than by GPU compute throughput [10]. The B&C-Engine outputs critically sampled channelized data, reducing the aggregate input rate from approximately 194 Tb/s at the DBE interface to approximately 168 Tb/s at the XE input. Recent GPUDirect results show that the network interface card delivers received data into GPU

memory at line rate [14], so the network interface itself sets the node ingest limit. Using the same 800GbE GPU processing node assumed for the PCIe GPU B&C candidate in Section 5.2, with 736 Gb/s of payload capacity per node, the standalone XE requires approximately 230 processing nodes.

The standalone XE is estimated at approximately 109 kW using the 475 W per-node power assumption from the PCIe GPU B&C candidate. With 8 servers per 42U rack, it occupies 8 racks at 15 kW per rack.

6 Cost, Networking, and Lifecycle Assessment

This section compares the four candidates on network infrastructure, cost, and lifecycle. It begins with the external network each candidate requires, including its power and rack space. It then estimates non-recurring engineering, hardware acquisition, and 20 years of operating cost. It closes with the lifecycle and obsolescence considerations that bear on the comparison.

6.1 Network Infrastructure

The external networking requirements vary significantly among the four candidates because of their architectural organization. The split architectures require an external Ethernet fabric to connect the DBEs, B&C stage, and X-Engine. The integrated architectures keep the corner turn within the processing system and therefore require fewer external connections.

The ASIC B&C architecture requires 873 external network connections, comprising 263 DBE ingest links, 380 B&C gateway links, and 230 X-Engine links. The 380 gateway links correspond to the 759 B&C Units described in Section 4.1, with two gateways sharing each 800 GbE switch port. The PCIe GPU B&C architecture requires 756 connections, comprising 263 DBE ingest links, 263 B&C node links, and 230 X-Engine links. In both architectures, the external network carries both DBE ingest and the corner-turn exchange between the B&C and X-Engine stages.

The NVL72 SBP architecture requires 551 external network connections, consisting of 263 DBE ingest links and 288 SBP processing-node links. The corner turn is performed within the NVLink fabric, so the X-Engine does not introduce additional external network connections. The FPGA SBP architecture uses the external network only for DBE ingest, with 263 DBE ingest links and 400 FPGA server links, for a total of 663 connections. Each FSP contains four servers, with each server using one 800 GbE external connection.

For a consistent comparison, all four candidates are modeled using the same 800 GbE switch technology and a two-tier non-blocking folded-Clos topology. The reference switch is the NVIDIA SN5600, which provides $N_{\text{port}} = 64$ ports of 800 GbE [15]. A first-order estimate of three switch ports per external connection is used, corresponding to one endpoint-facing port and two fabric-facing ports. Thus, for N_{conn} external

Table 1. Preliminary external-networking requirements for the candidate CSP architectures. Power covers switches and switch-side transceivers. Rack space covers switch equipment only.

Architecture	N_{conn}	Switches	Optics	Power (kW)	Rack (U)	Switch (M$)	Optics (M$)	Total (M$)
ASIC B&C	873	41	2619	83.1	82	3.08	3.93	7.00
PCIe GPU B&C	756	36	2268	72.4	72	2.70	3.40	6.10
NVL72 SBP	551	26	1653	52.5	52	1.95	2.48	4.43
FPGA SBP	663	32	1989	63.9	64	2.40	2.98	5.38

connections, the estimated number of switch ports is $3N_{\text{conn}}$, and the required number of switches N_{sw} is

$$N_{\text{sw}} = \left\lceil \frac{3N_{\text{conn}}}{N_{\text{port}}} \right\rceil .$$

This is a preliminary infrastructure estimate rather than a detailed network implementation. The analysis assumes one pluggable optical transceiver on each of the $3N_{\text{conn}}$ switch ports. Endpoint transceivers are excluded because the corresponding network interface hardware is already included with the processing nodes. Direct-attach copper connections are not modeled at this stage.

NVIDIA specifies a typical power of 940 W for the SN5600 with passive cables using the ATIS methodology [15]. The analysis uses this value for switch power and accounts separately for the optical transceivers. The selected transceiver is the NVIDIA MMA4Z00-NS-T, with a published maximum power of 17 W [16]. Additional fan and power-supply overhead associated with the optical load is not included because NVIDIA does not provide a corresponding value. The resulting power estimates are therefore first-order estimates.

The resulting network requirements are summarized in Table 1. The NVL72 SBP has the lowest external network burden, followed by the FPGA SBP, PCIe GPU B&C, and ASIC B&C. The same ordering is obtained for both network power and network hardware cost. This ordering is the primary result of the network analysis.

The network hardware cost uses rounded planning values of $75k per switch and $1.5k per transceiver. Both are taken from current reseller listings [17, 18]. These are single-unit reseller prices rather than vendor quotations and are therefore planning-level estimates. Volume pricing is not included.

Passive cabling and network spares are excluded from the network estimate. The cabling requirements have not been defined at this stage, and there is no established policy for network spares.

6.2 Cost Assessment

6.2.1 Cost Basis

The acquisition-cost comparison includes the non-recurring engineering (NRE) and hardware required to implement each candidate architecture. Hardware costs include the processing hardware and external

Table 2. Planning estimate of NRE by candidate architecture.

Architecture	FTE-years	Labor NRE (M$)	Non-labor NRE (M$)	Total NRE (M$)
ASIC B&C + standalone X-Engine	63	13.9	9.7	23.6
PCIe GPU B&C + standalone X-Engine	48	10.6	0.0	10.6
NVL72 SBP	32	7.0	0.0	7.0
FPGA SBP	34	7.5	0.0	7.5

network infrastructure identified in the preceding analysis. Pricing inputs are based on vendor or project sources where available. Where no such source is available, an explicitly identified planning estimate is used.

6.2.2 Non-Recurring Engineering

Non-recurring engineering is estimated from engineering effort expressed in full-time-equivalent (FTE) years using a planning rate of $220k per FTE-year. The effort estimates are engineering planning assumptions rather than project-derived staffing levels or vendor quotations. The activities are based on the documented development scope and the principal technical work identified for each candidate.

The NRE estimates are summarized in Table 2. The resulting labor estimates are 63 FTE-years ($13.9M) for the ASIC B&C path and 48 FTE-years ($10.6M) for the PCIe GPU B&C path. The integrated NVL72 and FPGA SBP architectures require 32 FTE-years ($7.0M) and 34 FTE-years ($7.5M), respectively.

The PCIe GPU B&C estimate does not include a credit for potential shared development between the B&C and X-Engine. The FPGA SBP NRE does not include any redesign of the DBE, which is outside the SBP acquisition-cost scope.

The $9.7M ASIC non-labor NRE is a planning allowance for the costs of developing and bringing the ASIC into production, none of which can yet be established from project sources or vendor quotations. It covers silicon fabrication, mask costs, physical IP, tooling, external design services, package and test development, qualification, and a planned design respin. The estimate assumes a 28 nm process, a 50 mm^2 die, and production of 6,072 ASICs. These assumptions are provided to support a sanity check against industry costs rather than to imply an itemized cost model. The external design-services allowance covers only the premium over the engineering labor already included in the labor estimate.

6.2.3 Hardware Acquisition Cost

Hardware acquisition cost represents the recurring hardware required for each architecture. Network hardware is treated separately in Section 6.1 and is not included in the hardware costs below. No production spare factor is applied to any estimate.

Table 3. Estimated hardware acquisition cost for the ASIC B&C architecture.

Hardware item	Quantity	Unit Cost	Total Cost (M$)
ASIC processing nodes	6,072	$50	0.30
Gateway FPGA devices	759	$5,500	4.17
B&C Unit production motherboards	759	$3,300	2.50
B&C LRU chassis and mechanical assemblies	380	$1,200	0.46
B&C rack infrastructure	13	$6,500	0.08
ASIC B&C hardware subtotal			**7.52**

6.2.3.1 ASIC B&C

The estimated hardware acquisition cost is summarized in Table 3, based on the documented B&C Unit architecture. The ASIC processing-node cost is an engineering planning estimate for packaged and tested devices. The motherboard estimate assumes direct mounting of the packaged ASICs on the B&C Unit motherboard.

The gateway estimate uses the AMD Versal Prime VM2152 as a representative COTS planning device. Current distributor pricing for the VM2152 is approximately $5.5k per device [19]. The LRU estimate covers the custom 1U mechanical assembly, cooling, and local power hardware required to house two B&C Units. The rack estimate covers the documented COTS 42U racks and rear-mounted AC-AC PDUs. Facility power distribution and cooling infrastructure are outside the acquisition scope. The ASIC B&C hardware subtotal is approximately $7.5M, with gateway devices accounting for approximately 56%.

6.2.3.2 PCIe GPU B&C

The PCIe GPU B&C hardware acquisition cost is summarized in Table 4, using the processing-node and rack quantities from Section 5.2. The server quantity follows from four processing nodes per server, giving 66 servers for 263 nodes. Each processing node is paired with one 800GbE SuperNIC.

The GPU processing-device cost is based on the current manufacturer list price for the NVIDIA RTX PRO 6000 Blackwell, which is used as a reference device for planning purposes [20]. The $16,000 price is a current planning value rather than a procurement quotation and does not imply that the RTX PRO 6000

Table 4. Estimated hardware acquisition cost for the PCIe GPU B&C architecture.

Hardware item	Quantity	Unit Cost	Total Cost (M$)
GPU processing devices	263	$16,000	4.21
800GbE SuperNICs	263	$2,500	0.66
Host servers	66	$15,000	0.99
Rack infrastructure	9	$6,500	0.06
PCIe GPU B&C hardware subtotal			**5.91**

Table 5. Estimated hardware acquisition cost for the standalone X-Engine.

Hardware item	Quantity	Unit Cost	Total Cost (M$)
GPU processing devices	230	$16,000	3.68
800GbE SuperNICs	230	$2,500	0.57
Host servers	58	$15,000	0.87
Rack infrastructure	8	$6,500	0.05
Standalone X-Engine hardware subtotal			**5.18**

Blackwell is the assumed production device. The list price has risen from $8,565 at launch in March 2025 to $13,250 in June 2026 and $16,000 in September 2026, an increase of 87% in eighteen months driven by graphics memory supply. The eventual production device is assumed to be a future PCIe 6.0 GPU, which is not yet available. The 800GbE SuperNIC estimate uses the identified distributor price of $2.5k for the NVIDIA ConnectX-8 SuperNIC [21]. The host-server estimate of $15k is an engineering planning assumption covering the server platform, CPU, memory, and boot storage, with the GPUs and SuperNICs costed separately. The rack estimate uses the same fitted COTS rack infrastructure established for the ASIC B&C candidate.

The PCIe GPU B&C hardware subtotal is approximately $5.9M. GPUs account for approximately 71% of the cost, with the host servers and SuperNICs providing most of the remainder. The standalone X-Engine is not included in this subtotal. Its hardware is costed separately because the same X-Engine is shared by the PCIe GPU B&C and ASIC B&C split architectures.

6.2.3.3 Standalone X-Engine

The standalone X-Engine hardware acquisition cost is summarized in Table 5. It uses the same PCIe GPU processing platform as the PCIe GPU B&C implementation. The processing-node and rack quantities are taken from Section 5.5. The server quantity follows from four processing nodes per server, giving 58 servers for 230 nodes.

The processing hardware, SuperNIC, host-server, and rack price bases are identical to those used for the PCIe GPU B&C estimate. The standalone X-Engine hardware subtotal is approximately $5.2M. This hardware is common to both split architectures and is therefore counted once when the candidate-level acquisition costs are reconciled.

6.2.3.4 NVL72 SBP

The NVL72 hardware estimate is based on the four-rack configuration from Section 5.3 and is summarized in Table 6. Each rack is treated as a complete system, including the GPUs, host processors, memory, NVLink infrastructure, rack and power hardware, and cooling.

NVIDIA does not publish a list price for the NVL72. A planning value of $6.0M per rack is used, based

Table 6. Estimated hardware acquisition cost for the NVL72 SBP architecture.

Hardware item	Quantity	Unit Cost	Total Cost (M$)
NVL72 rack-scale systems	4	$6,000,000	24.00
NVL72 SBP hardware subtotal			**24.00**

on a March 2026 third-party estimate of $6.0M to $6.5M for an AI-inference-optimized GB300 NVL72 system [22]. This is a market reference rather than a vendor quotation. At the $6.0M planning value, the four-rack configuration has a hardware subtotal of approximately $24.0M. Using the upper-end estimate would increase this to approximately $26.0M.

6.2.3.5 FPGA SBP

Table 7 costs the accelerator-card, server, and rack quantities established in Section 5.4. The FPGA SBP implementation is based on the COTS FPGA platform evaluated in ATAC Memo 009 [5].

The FPGA accelerator card is priced at $18.5k based on the current Mouser listing for the BittWare IA-860m [23]. BittWare does not publish a price for the TeraBox 1501B. A value of $14k per server is used as a planning assumption, based on published prices for similar 1U EPYC-based servers [24]. These servers cannot hold the same four dual-slot accelerator cards as the TeraBox 1501B, so $14k is only a rough estimate, not a market price or vendor quotation. The estimate should be revisited if a vendor quotation becomes available.

The TeraBox includes the host processor, memory, storage, power supplies, and chassis. These components are therefore included in the server cost and are not costed separately. The FPGA accelerator cards are costed separately because they are distinct products.

The current FPGA SBP architecture assumes liquid cooling, following the ATAC design [13]. The IA-860m configuration used for the card price is specified with a passive heatsink, so the compatibility of the priced card configuration with the assumed cooling architecture remains an open design and costing issue. The rack liquid-cooling allowance is retained as a planning estimate and should be revisited when the final card and server cooling configuration is established.

Table 7. Estimated hardware acquisition cost for the FPGA SBP architecture.

Hardware item	Quantity	Unit Cost	Total Cost (M$)
BittWare IA-860m FPGA accelerator cards	800	$18,500	14.80
BittWare TeraBox 1501B servers	400	$14,000	5.60
Rack liquid-cooling infrastructure	20	$40,000	0.80
Rack infrastructure	20	$6,500	0.13
FPGA SBP hardware subtotal			**21.33**

Table 8. Estimated operating load and electricity cost by candidate architecture, at 8,760 operating hours per year.

Architecture	Load (kW)	Energy (MWh/yr)	Annual cost (k$/yr)	20-year cost (M$)
ASIC B&C + standalone X-Engine	306	2,681	278	5.56
PCIe GPU B&C + standalone X-Engine	306	2,681	278	5.56
NVL72 SBP	613	5,370	557	11.14
FPGA SBP	464	4,065	422	8.44

The FPGA SBP hardware subtotal is approximately $21.3M. Accelerator cards account for approximately 69% of this cost, with servers accounting for most of the remainder. The server estimate carries the lowest confidence among the priced FPGA components.

6.2.4 Operating Expenditure

The ngVLA lifecycle is assumed to span 30 years, comprising 10 years of construction followed by 20 years of operation. Operating expenditure is estimated over the 20-year operating period. Electricity is the only recurring operating cost quantified in this study. Maintenance, spares, personnel, and other operating costs are not included. The processing and network power estimates from Section 5 and Section 6.1 are combined to estimate the operating load and electricity cost for each candidate, as summarized in Table 8.

The processing system is assumed to be located at the VLA site on the Plains of San Agustin in Socorro County, New Mexico. Electricity cost is taken as $909/kW-year based on the applicable local utility rate [25]. This value is applied consistently to all candidate architectures. The estimates assume continuous operation at full load, with no availability or observing duty-cycle factor. Facility cooling, power conversion losses, and other building loads are excluded.

The resulting electricity costs are therefore planning estimates rather than forecasts of the actual electricity cost. The primary purpose of the estimate is to provide a common basis for comparing the candidates over the 20-year operating period.

6.2.5 Overall Cost

Table 9 summarizes the estimated acquisition and 20-year electricity costs for the four candidate architectures. The network infrastructure is included in the acquisition cost.

The PCIe GPU B&C has the lowest estimated acquisition cost at $27.8M. The FPGA SBP and NVL72 SBP are close at $34.2M and $35.4M, respectively, while the ASIC B&C has the highest acquisition cost at $43.3M. The small difference between the FPGA SBP and NVL72 SBP should be considered in the context of the uncertainty in the underlying planning assumptions.

Table 9. Estimated acquisition and 20-year electricity costs by architecture, in M$. Totals exclude technology refresh, maintenance, and spares.

Architecture	NRE +	Hardware +	Network =	**Acquisition** +	Electricity =	**Total**
ASIC B&C + standalone X-Engine	23.6 +	12.70 +	7.00 =	**43.30** +	5.56 =	**48.86**
PCIe GPU B&C + standalone X-Engine	10.6 +	11.09 +	6.10 =	**27.79** +	5.56 =	**33.35**
NVL72 SBP	7.0 +	24.00 +	4.43 =	**35.43** +	11.14 =	**46.57**
FPGA SBP	7.5 +	21.33 +	5.38 =	**34.21** +	8.44 =	**42.65**

The cost structure also differs across the candidates. The ASIC B&C has substantially higher NRE than the other architectures, while the integrated architectures have higher hardware acquisition costs. Including 20-year electricity cost does not change the overall cost ranking, although it increases the separation between the FPGA SBP and NVL72 SBP because of the higher operating load of the NVL72 SBP.

6.3 Lifecycle Considerations

The 20-year operating period extends beyond the commercial life of individual processing platforms. Lifecycle risk therefore depends primarily on the ability to adopt successor hardware, device and supplier dependence, and the engineering effort required when components become obsolete.

The PCIe GPU B&C and NVL72 SBP architectures offer greater refresh flexibility because their signal processing is software-defined. Their processing algorithms can be migrated to successor GPU hardware without redesigning the underlying algorithms, although recompilation, requalification, and performance revalidation would still be required. The PCIe GPU B&C uses individually replaceable accelerator cards, whereas the NVL72 is a single-vendor rack-scale platform with no equivalent second source, making hardware refresh a larger and less incremental action.

The FPGA SBP has greater exposure to FPGA device lifecycle risk. The selected Agilex 7 M-Series devices include integrated HBM2E, which is explicitly excluded from Altera's 2045 lifecycle extension because of the shorter lifecycle of HBM components [26]. A transition to a different FPGA device family would require firmware migration and requalification, adding engineering effort to a hardware refresh.

The ASIC B&C avoids dependence on the commercial evolution of processing devices because its processing is fixed in silicon. However, changes to the processing requirements cannot be accommodated through software or firmware updates and would require a new ASIC development. The COTS FPGA gateway introduces the same general device lifecycle dependency as the FPGA SBP, with replacement of the gateway device potentially requiring migration and requalification of the gateway design.

Serviceability also differs between the architectures. The air-cooled architectures use conventional rack infrastructure, while the liquid-cooled architectures require additional cooling infrastructure and facility interfaces that must be maintained over the operating period. Unit replacement in liquid-cooled systems

also requires handling the coolant connections.

7 Summary and Recommendation

This trade study compares four SBP architectures using a common set of technical, cost, lifecycle, and engineering assumptions. The PCIe GPU B&C currently provides the strongest overall balance. Prototype benchmarking has demonstrated real-time end-to-end operation at 19 of the 92 subbands per antenna on an NVIDIA L40 GPU (PCIe 4.0, no GPUDirect, PCIe limited). The production implementation assumes a PCIe 6.0 GPU sustaining the full per-antenna workload, and that assumption remains subject to benchmarking. The current estimate gives the PCIe GPU architecture the lowest acquisition cost and lowest 20-year electricity cost, with operating power comparable to the ASIC architecture. Its software-defined implementation also provides greater flexibility for future algorithm changes and reduces the engineering effort associated with hardware refresh.

The ASIC B&C remains a valid alternative because it builds on the existing B&C implementation approach and provides comparable estimated operating power to the PCIe GPU architecture. Its principal disadvantages are substantially higher NRE cost and the fixed-function nature of the implementation.

The FPGA and NVL72 architectures integrate the B&C and X-Engine in a single architecture, with both having higher estimated operating power than the split architectures. Both also have higher estimated acquisition cost than the PCIe GPU B&C. The FPGA architecture builds on a mature correlator design, which reduces architectural risk, but its HBM2E-equipped devices have a documented lifecycle concern. The NVL72 provides a highly integrated rack-scale implementation, but introduces additional dependence on a single vendor. On balance, these tradeoffs make both architectures less attractive than the PCIe GPU B&C.

The cost and power results should be interpreted as engineering planning estimates rather than procurement-level values. Several important quantities depend on representative devices, assumed future hardware, or implementation-level benchmarking. Power efficiency remains an important consideration in the downselection as a responsible and sustainable system-design objective, but the present estimates should not be treated as sufficiently precise to establish small differences in power as decisive.

Based on the current evidence, the PCIe GPU B&C is recommended as the preferred architecture. If the PCIe GPU B&C implementation is found not to be viable at the required workload, power envelope, or system scale, the broader trade study should be revisited rather than treating the ASIC B&C as an automatic fallback. The FPGA and NVL72 architectures should remain part of that reassessment, since the present study does not establish that the ASIC B&C is clearly preferable to them under changed assumptions.

Final selection remains subject to the targeted PCIe GPU B&C workload benchmarking and the resulting validation of processing capacity, power, and system sizing.

Acknowledgments

The authors gratefully acknowledge NVIDIA for providing the NVIDIA L40 GPU loaner used for B&C prototype development and characterization. We particularly thank A. Thompson and E. Eshelman for their generous and continued technical support throughout this work.

We also thank the digital design team at the Central Development Laboratory for their assistance in drafting and reviewing this memorandum. In particular, we thank N. Denman for the development of the X-Engine trade study and the GPU prototype, which enabled the characterization and benchmarking of the B&C implementation.

The authors also acknowledge the use of generative AI tools in the preparation of this memorandum. Anthropic Claude was used to assist with wording and language refinement and with development of the B&C prototype software used for the benchmarking described in Section 5.2. OpenAI ChatGPT was used to assist with wording, language usage, and editorial refinement throughout the memorandum. The authors reviewed and remain responsible for all content, analysis, technical conclusions, and results presented in this work.

The National Radio Astronomy Observatory and Green Bank Observatory are facilities of the U.S. National Science Foundation operated under cooperative agreement by Associated Universities, Inc. This work was supported by awards AST-2034328 (MSIP Prototype Antenna) and AST-2334267 (ngVLA Design Activities); NRAO related activities are funded under award AST-1647378 (NRAO Operations/Development).

8 References

[1] NVIDIA, “L40S GPU for AI and Graphics Performance.” [Online]. Available: https://www.nvidia.com/en-us/data-center/l40s/

[2] PNY, “NVIDIA RTX PRO 6000 Blackwell Server Edition.” [Online]. Available: https://www.pny.com/nvidia-rtx-pro-6000-blackwell

[3] ALMA Observatory, “ALMA2030 WSU Technology.” [Online]. Available: https://www.almaobservatory.org/en/scientists/alma-2030-wsu/wsu-technology/

[4] NRAO, “ALMA Wideband Sensitivity Upgrade (WSU).” [Online]. Available: https://science.nrao.edu/facilities/alma/science_sustainability/wideband-sensitivity-upgrade

[5] S. Harrison, A. Banwait, M. Grainger, M. Pleasance, and F. Zhang, “Evaluate Suitability of Next Generation FPGA Technology and 400 Gigabit Ethernet Switches for Use in a 2nd Generation ALMA Correlator Design,” National Research Council of Canada, Herzberg Astronomy and Astrophysics Research Centre, NA ALMA Study Report and ATAC Memo 009, Jul. 15, 2024.

[6] BittWare (Molex), “TeraBox 1501B 1U FPGA Server,” datasheet Rev. 2024.1.16. [Online]. Available: https://www.bittware.com/products/servers-systems/terabox1500b/

[7] NVIDIA, “L40 GPU for Data Center.” [Online]. Available: https://www.nvidia.com/en-us/data-center/l40/

[8] NVIDIA, “GPUDirect RDMA,” CUDA Toolkit Documentation. [Online]. Available: https://docs.nvidia.com/cuda/gpudirect-rdma/

[9] Supermicro, “Supermicro NVIDIA GB300 NVL72 SuperCluster Datasheet.” [Online]. Available: https://www.supermicro.com/datasheet/datasheet_SuperCluster_GB300_NVL72.pdf

[10] N. Denman, “XE and PSE Hardware Class Down-Selection,” ngVLA Electronics Memo #20, May 22, 2025.

[11] NVIDIA, “GB300 NVL72.” [Online]. Available: https://www.nvidia.com/en-us/data-center/gb300-nvl72/

[12] NVIDIA, “NVLink and NVLink Switch.” [Online]. Available: https://www.nvidia.com/en-us/data-center/nvlink/

[13] C. Brogan, “The Advanced Technology ALMA Correlator,” presented at The Promises and Challenges of the ALMA Wideband Sensitivity Upgrade, ESO Garching, Germany, Jun. 24–28, 2024.

[14] NVIDIA, “DAQIRI: Data Acquisition for Integrated Real-time Instruments,” accessed September 2026. [Online]. Available: https://nvidia.github.io/daqiri/

[15] NVIDIA, “Spectrum-4 SN5000 2U Switch Systems Hardware User Manual.” [Online]. Available: https://docs.nvidia.com/nvidia-spectrum-4-sn5000-2u-switch-systems-hardware-user-manual.pdf

[16] NVIDIA, “MMA4Z00-NS-T 800Gb/s Twin-port OSFP, 2x400Gb/s Multimode, 50m.” [Online]. Available: https://networking-docs.nvidia.com/800gmma4z00nst

[17] Server Supply, “NVIDIA Spectrum SN5600 64-Port 800GbE Ethernet Switch,” SKU 920-9N42F-00RI-7C0, accessed August 2026. [Online]. Available: https://www.serversupply.com/NETWORKING/SWITCH/64%20PORT/NVIDIA/920-9N42F-00RI-7C0_383148.htm

[18] Server Supply, “NVIDIA MMA4Z00-NS-T 800Gb/s Twin-port OSFP 2x400Gb/s Multimode 50m Transceiver,” accessed August 2026. [Online]. Available: https://www.serversupply.com/NETWORKING/TRANSCEIVER/800%20GIGABIT/NVIDIA/MMA4Z00-NS-T_398413.htm

[19] DigiKey, “AMD XCVM2152-1MSENFVM1369, Versal Prime adaptive SoC, 757K logic cells, 1369-FCBGA,” accessed August 2026. [Online]. Available: https://www.digikey.com/en/products/detail/amd/XCVM2152-1MSENFVM1369/27803902

[20] Tom’s Hardware, “Nvidia doubles RTX PRO 6000 Blackwell’s MSRP to a staggering $16,000,” Sep. 2026. [Online]. Available: https://www.tomshardware.com/pc-components/gpus/nvidia-doubles-rtx-pro-6000-blackwells-msrp-to-a-staggering-usd16-000-96gb-card-started-pre-orders-below-usd8-000-last-year

[21] Cloud Ninjas, “NVIDIA ConnectX-8 800GbE Single-Port VPI OSFP SuperNIC,” part number 900-9X81E-00EX-ST0, accessed August 2026. [Online]. Available: https://cloudninjas.com/products/nvidia-connectx-8-800gbe-vpi-single-port-osfp-nic

[22] Tom’s Hardware, “NVIDIA rack-scale system pricing,” Mar. 2026. [Online]. Available: https://www.tomshardware.com/tech-industry/artificial-intelligence/price-of-nvidias-vera-rubin-nvl72-racks-skyrockets-to-as-much-as-usd8-8-million-apiece-but-server-makers-margins-will-be-tight-nvidia-is-moving-closer-to-shipping-entire-full-scale-systems

[23] Mouser Electronics, “BittWare IA-860m-0037 Accelerator Card, Intel Agilex M-Series PCIe Card, QSFP-DD Config 3, Passive Heat Sink, 2 Slot,” Mouser part 538-IA-860M-0037, accessed August 2026. [Online]. Available: https://www.mouser.com/en/new/bittware/bittware-ia-860m-fpga-card/

[24] “Configured 1U AMD EPYC 9004 server pricing survey,” August 2026, [Online]. Available: https://cloudninjas.com/collections/cloud-ninjas-servers, https://www.thinkmate.com/systems/servers/rax/amd.

[25] Socorro Electric Cooperative, Inc., “Seventh Revised Rate No. 3, Large Power Service,” Advice Notice No. 72, effective Jun. 3, 2022. [Online]. Available: https://www.socorroelectric.com/sites/default/files/downloads/Rates/AN%2072%20%20-%20Rate%203%20-%20Large%20Power%20Service%20%206.3.2022.pdf, https://www.socorroelectric.com/rates.

[26] Altera, “Altera Extends Lifecycle Support for Several FPGA Families Through 2045,” press release, Apr. 9, 2026. [Online]. Available: https://www.altera.com/newsroom/news/press-release/altera-extends-fpga-product-lifecycle-support-2045

[27] “ngVLA Central Signal Processor Requirements Specification,” 020.40.00.00.00-0001_REQ.

[28] A. R. Thompson, J. M. Moran, and G. W. Swenson Jr., *Interferometry and Synthesis in Radio Astronomy*, 3rd ed. Springer, 2017.

[29] J. D. Kraus, *Radio Astronomy*, 2nd ed. Cygnus-Quasar Books, 1986.

[30] J. W. M. Baars, “Characteristics of the Paraboloidal Reflector Antenna,” NRAO Electronics Division, Green Bank, WV, Internal Report No. 57, Aug. 1966.

[31] R. N. Bracewell, *The Fourier Transform and Its Applications*, 3rd ed. McGraw-Hill, 2000.

A Acronyms, Units, and Symbols

A.1 Acronyms

- ALMA: Atacama Large Millimeter/submillimeter Array
- ASIC: Application-Specific Integrated Circuit
- ATAC: Advanced Technology ALMA Correlator
- ATIS: Alliance for Telecommunications Industry Solutions
- B&C: Beamformer and Channelizer
- COTS: Commercial Off-the-Shelf
- CPU: Central Processing Unit
- CSP: Central Signal Processor
- DBE: Digital Back-End
- FDR: Final Design Review
- FPGA: Field-Programmable Gate Array
- FSP: Frequency Slice Processor
- FTE: Full-Time Equivalent (unit of engineering effort)
- FX: F-Engine / X-Engine correlator architecture
- GbE: Gigabit Ethernet
- GPU: Graphics Processing Unit
- HBM: High-Bandwidth Memory; HBM2E and HBM3e denote successive generations
- HPBW: Half-Power Beamwidth
- I/O: Input/Output
- IP: Intellectual Property
- LRU: Line Replaceable Unit
- ngVLA: Next Generation Very Large Array
- NRAO: National Radio Astronomy Observatory
- NRE: Non-Recurring Engineering
- PCIe: Peripheral Component Interconnect Express
- PDR: Preliminary Design Review
- PDU: Power Distribution Unit
- SBP: Subband Processor
- SoC: System on Chip
- SuperNIC: NVIDIA designation for its network interface cards
- ULA: Uniform Linear Array
- VCC: Very Coarse Channelizer
- VLA: Very Large Array
- WSU: Wideband Sensitivity Upgrade

- XE: X-Engine, the cross-correlation stage of the SBP

A.2 Units

- Gb/s, Tb/s: gigabits and terabits per second.
- TB/s: terabytes per second.
- MS/s: megasamples per second.
- kHz, MHz, GHz: kilohertz, megahertz, and gigahertz.
- W, kW: watts and kilowatts.
- MWh: megawatt-hour.
- μs: microseconds.
- m, nm, mm, km: meters, nanometers, millimeters, and kilometers.
- U: rack unit, 44.45 mm of vertical rack space.
- TOPS: tera-operations per second.

A.3 Symbols

- $A(g)$: array factor, the normalized sum of the per-antenna residual phases at frequency offset g.
- **B**: baseline vector between two antennas.
- c: speed of light.
- D: antenna aperture diameter.
- f, f_0: frequency, channel-center frequency.
- g: frequency offset from the channel center, $g = f - f_0$.
- N_conn: number of external network connections.
- N_sw: number of network switches.
- $\hat{s}$: unit vector toward the source direction.
- x: generic real-valued argument, e.g., in sinc(x) or the ceiling function $\lceil x \rceil$.
- $\Delta\nu$: frequency channel width.
- $\Delta\nu_{\text{max,band}}$, $\Delta\nu_{\text{max,edge}}$: widest channel width admissible under the minimum guaranteed efficiency, band averaged and at the channel edge.
- $\Delta\nu_{\text{ULA,band}}$, $\Delta\nu_{\text{ULA,edge}}$: the same quantities for the uniform linear array reference case.
- η: beamforming efficiency, the ratio of achieved beam signal power to its ideal value.
- η_{band}: band-averaged efficiency, for a broadband signal filling the channel.
- η_{edge}: efficiency at a frequency component located at the channel edge.
- $\eta_{\text{min,band}}$, $\eta_{\text{min,edge}}$: minimum guaranteed efficiency over all antenna configurations of a given extent, band averaged and at the channel edge.
- $\eta_{\text{ULA,band}}$, $\eta_{\text{ULA,edge}}$: efficiency of a uniform linear array, band averaged and at the channel edge.
- θ_{HPBW}: half-power beamwidth.

- λ: wavelength.
- ν: frequency.
- π: the constant pi.
- τ, τ_g: geometric delay.
- τ_{diff}: differential delay between the pointing center and a beam.

B Beamforming Delay Compensation: Phase-Shift versus True-Time-Delay

This appendix determines the beamforming channel width required for a per-channel phase shift to compensate the beam-specific geometric delay to a stated beamforming efficiency, as a function of array extent, observing frequency, and signal type. The driving requirements are at least 10 simultaneous beams placed anywhere within the antenna half-power beamwidth, subarray extents up to 700 km, and a beamforming efficiency of at least 95% [27].

The result is a requirement on the beamforming channelization, not a conclusion about any candidate. The channel width at which each candidate forms beams is set by that candidate's architecture and by requirements outside the scope of this appendix. Whether a given candidate meets the width derived here by phase shifting alone, or must instead provide an explicit true-time-delay stage, is therefore not determined here.

B.1 Geometry and the Worst-Case Residual Delay

The geometric delay of a plane wave from direction $\hat{s}$ across a baseline $\mathbf{B}$ is [28]

$$\tau_g = \frac{\mathbf{B} \cdot \hat{s}}{c}. \tag{B.1}$$

The bulk geometric delay toward the pointing center is tracked per antenna before fine channelization and is common to all beams. The beamformer therefore needs to compensate only the differential delay between the pointing center and the beam direction.

For a uniformly illuminated circular aperture of diameter D, the half-power beamwidth is [29]

$$\theta_{\mathrm{HPBW}} \simeq 1.02 \, \frac{\lambda}{D}. \tag{B.2}$$

A beam within the primary beam can be offset from the pointing center by at most $\theta_{\mathrm{HPBW}}/2$. For an array extent B, the corresponding worst-case differential delay is therefore

$$\tau_{\mathrm{max}} \simeq \frac{B}{c} \frac{\theta_{\mathrm{HPBW}}}{2} \simeq 0.51 \, \frac{B\lambda}{cD}. \tag{B.3}$$

The actual primary beam depends on aperture illumination. The half-power beamwidth increases with edge taper [30]. A -10 dB edge taper increases the beamwidth coefficient from 1.02 to approximately 1.14, while a -18 dB edge taper increases it to approximately 1.20. Because the ngVLA illumination taper is not specified in the material used for this study, the uniform-illumination value is retained. The resulting τ_{max}

should therefore be regarded as a lower-bound estimate.

B.2 Residual Delay from Phase-Shift Compensation

Within a channel of width $\Delta\nu$ centered at f_0, let $g = f - f_0$ denote the frequency offset from the channel center. A delay τ produces a linear phase shift according to the Fourier shift theorem [31]

$$x(t-\tau) \;\longleftrightarrow\; X(f)\, e^{-j2\pi f\tau}. \tag{B.4}$$

Phase-shift beamforming applies the correction $e^{+j2\pi f_0\tau_i}$ for antenna i, which exactly compensates the delay at the channel center. The remaining frequency-dependent phase is

$$e^{-j2\pi f\tau_i}\, e^{+j2\pi f_0\tau_i} = e^{-j2\pi g\tau_i}. \tag{B.5}$$

At the channel center, $g = 0$ and the residual phase is zero. Away from the channel center, the residual phase depends on both the frequency offset and the differential delay. A true-time-delay correction removes this residual across the channel. The remainder of this appendix therefore quantifies the beamforming efficiency loss associated with phase-shift compensation.

B.3 Beamforming Efficiency

The residual phase in (B.5) determines the coherence of the beamformer across the channel. Define the normalized array factor as

$$A(g) = \frac{1}{N}\sum_{i=1}^{N} e^{-j2\pi g\tau_i}, \qquad A(0) = 1. \tag{B.6}$$

Let η denote the ratio of achieved beam signal power to the ideal beam signal power. Because the antenna noise is uncorrelated and is unaffected by the beamforming phase corrections, this is also the ratio of the achieved beam sensitivity to the ideal sensitivity. For a signal with power spectrum $P(g)$ within the channel,

$$\eta = \frac{\displaystyle\int P(g)\,|A(g)|^2\, dg}{\displaystyle\int P(g)\, dg}. \tag{B.7}$$

For a broadband signal that fills the channel, $P(g)$ is constant and the efficiency is the average of $|A(g)|^2$ across the channel. Expanding (B.6) gives

$$\eta_{\text{band}} = \frac{1}{\Delta\nu}\int_{-\Delta\nu/2}^{\Delta\nu/2} |A(g)|^2\, dg = \frac{1}{N^2}\sum_i\sum_j \frac{1}{\Delta\nu}\int_{-\Delta\nu/2}^{\Delta\nu/2} e^{-j2\pi g(\tau_i-\tau_j)}\, dg. \tag{B.8}$$

The integral is

$$\frac{1}{\Delta\nu}\int_{-\Delta\nu/2}^{\Delta\nu/2} e^{-j2\pi g\Delta\tau}\, dg = \frac{\sin(\pi\Delta\nu\Delta\tau)}{\pi\Delta\nu\Delta\tau} \equiv \text{sinc}(\Delta\nu\Delta\tau), \tag{B.9}$$

where $\text{sinc}(x) \equiv \sin(\pi x)/(\pi x)$ and $\Delta\tau = \tau_i - \tau_j$. Therefore,

$$\eta_{\text{band}} = \frac{1}{N^2}\sum_i\sum_j \text{sinc}(\Delta\nu(\tau_i - \tau_j)). \tag{B.10}$$

The diagonal terms with $i = j$ contribute unity regardless of channel width, since they have zero differential delay. Thus, the band-averaged efficiency depends on the distribution of differential delays across the array rather than only on the largest delay.

For an individual frequency component at offset g_0, the efficiency is instead $|A(g_0)|^2$, evaluated directly at that frequency. The largest residual phase, and therefore the lowest efficiency, occurs at the channel edge, $g_0 = \pm\Delta\nu/2$, which gives

$$\eta_{\text{edge}} = \left|A\left(\pm\tfrac{\Delta\nu}{2}\right)\right|^2. \tag{B.11}$$

Unlike the broadband case, there is no averaging across the channel. The band-averaged and channel-edge cases therefore represent distinct efficiency conditions.

Knowing only that the antenna delays lie within an interval of width τ_{max} gives a configuration-independent lower bound on the beamforming efficiency. If all antenna delays lie within an interval of width τ_{max}, the residual phases at a given frequency offset g span an angular interval of width $2\pi|g|\tau_{\text{max}}$. For $|g|\tau_{\text{max}} \leq 1/2$, the resulting array factor satisfies

$$|A(g)| \geq \cos(\pi g\tau_{\text{max}})\,. \tag{B.12}$$

The bound is attained when the antennas are distributed equally between the two ends of the delay interval. A two-antenna array with $\tau_1 = 0$ and $\tau_2 = \tau_{\text{max}}$ therefore provides the worst case for a given array extent. The corresponding efficiencies are

$$\eta_{\text{min,band}} = \frac{1}{2}\left[1 + \text{sinc}(\Delta\nu\tau_{\text{max}})\right], \qquad \eta_{\text{min,edge}} = \cos^2\left(\frac{\pi\Delta\nu\tau_{\text{max}}}{2}\right). \tag{B.13}$$

These expressions provide lower bounds for any antenna configuration with the same maximum differential delay and do not require knowledge of the individual antenna positions. The condition $\Delta\nu\tau_{\text{max}} \leq 1$ is satisfied over the range considered in this study.

For reference, consider a uniform linear array (ULA) with antennas distributed evenly across the full delay extent. This is not the ngVLA antenna configuration, but provides a useful reference between the conservative two-antenna bound and a fully distributed array.

Equation (B.6) is the Fourier transform of the distribution of antenna delays, evaluated at the frequency offset g. A uniform distribution over an interval of width τ_{max} therefore transforms to a sinc [31], giving

$$A(g) = \text{sinc}(g\tau_{\text{max}}). \tag{B.14}$$

This expression takes the delays centered on the extent, and holds in the limit of many antennas. A finite number of evenly spaced antennas gives a Dirichlet kernel instead. The channel widths of Table B-1 all satisfy $\Delta\nu\tau_{\text{max}} \leq 0.44$, so the array factor need only be evaluated over $|g|\tau_{\text{max}} \leq 0.22$. For as few as 8 antennas, the Dirichlet kernel departs from Equation (B.14) by at most 2.2% of the peak array factor over that range. The departure falls to 0.6% for 27 antennas and to 0.06% for 263 antennas, matching the ngVLA antenna count but not its configuration. This is not significant for a reference case. The corresponding efficiencies are

$$\eta_{\text{ULA,band}} = \frac{1}{\Delta\nu\tau_{\text{max}}}\int_{-\Delta\nu\tau_{\text{max}}/2}^{\Delta\nu\tau_{\text{max}}/2} \text{sinc}^2(u)\,du, \qquad \eta_{\text{ULA,edge}} = \text{sinc}^2\left(\frac{\Delta\nu\tau_{\text{max}}}{2}\right). \tag{B.15}$$

The actual ngVLA efficiency depends on the antenna positions and is not evaluated here. The two-antenna result therefore provides the conservative configuration-independent bound used to establish the channelization requirement, while the uniform linear array provides a reference for the efficiency expected from a more distributed antenna configuration.

B.4 Required Beamforming Channel Width

Inverting (B.13) and (B.15) for $\Delta\nu$ at a target efficiency η gives the maximum admissible channel width,

$$\Delta\nu_{\text{max,band}} = \frac{\text{sinc}^{-1}(2\eta - 1)}{\tau_{\text{max}}}, \qquad \Delta\nu_{\text{max,edge}} = \frac{2\arccos\sqrt{\eta}}{\pi\,\tau_{\text{max}}}, \qquad \Delta\nu_{\text{ULA,edge}} = \frac{2\,\text{sinc}^{-1}\sqrt{\eta}}{\tau_{\text{max}}}. \tag{B.16}$$

Table B-1. Maximum beamforming channel width for per-channel phase-shift compensation at the low edge of Band 1 ($\nu = 1.2$ GHz, $D = 18$ m).

Array extent	τ_{max}	Minimum guaranteed		ULA reference	
		Band averaged	Channel edge	Band averaged	Channel edge
		($\eta_{min,band}$)	($\eta_{min,edge}$)	($\eta_{ULA,band}$)	($\eta_{ULA,edge}$)
Efficiency $\eta \geq 0.99$					
1 km	23.6 ns	4.68 MHz	2.70 MHz	8.12 MHz	4.68 MHz
30 km	708 ns	156 kHz	90.0 kHz	271 kHz	156 kHz
700 km	16.5 μs	6.69 kHz	3.86 kHz	11.6 kHz	6.68 kHz
Efficiency $\eta \geq 0.95$					
1 km	23.6 ns	10.6 MHz	6.08 MHz	18.4 MHz	10.5 MHz
30 km	708 ns	354 kHz	203 kHz	614 kHz	352 kHz
700 km	**16.5 μs**	**15.2 kHz**	**8.69 kHz**	**26.3 kHz**	**15.1 kHz**

The remaining case, $\Delta\nu_{ULA,band}$, has no closed-form inverse and is evaluated numerically. Table B-1 evaluates all four cases at the low edge of Band 1, where the primary beam is widest and the requirement is most demanding. The 700 km row corresponds to the maximum subarray extent specified by the ngVLA beamforming requirements.

The maximum allowable channel width decreases in proportion to array extent and increases in proportion to observing frequency. At the 95% efficiency level, the minimum-guaranteed band-averaged case permits 354 kHz at a 30 km extent and 15.2 kHz at a 700 km extent. The corresponding channel-edge limits are 203 kHz and 8.69 kHz.

These results define the beamforming channelization required when beam-specific geometric delay is compensated by per-channel phase shifts. Each candidate's architecture determines whether its beamforming channelization satisfies these limits or requires an alternative delay compensation approach.