

Document file

SOW_NG-Croce_T6.2

Document's coding

Tender document for technical specifications

Tender's goal in brief

Supply of components for setting up the real-time compute system for the "Canadian Hydrogen Observatory and Radio-transient Detector (CHORD)" radio telescope.

Award mode

Open tender call pursuant to art. 70, 71 and 108, legislative decree number 36 of 31 March 2023.

CUP

C53C22000880006

CIG

Lotto 1: A038627B9B Lotto 2: A0386707DA

Kick off Act

Determinazione n. 212 del 4 dicembre 2023

Tender value

2.982.570,00 €

Lot 1: 2.332.570,00 €

Lot 2: 650.000,00 €

Funding

Funded by the European Union - NextGenerationEU

Piano Nazionale di Ripresa e Resilienza (PNRR) - Avviso n. 3264 del 28.12.2021 - Missione 4, "Istruzione e Ricerca" - Componente 2, "Dalla ricerca all'impresa" - Linea di investimento 3.1, "Fondo per la realizzazione di un sistema integrato di infrastrutture di ricerca e innovazione", finanziato dall'Unione europea - NextGenerationEU
Proposta progettuale "NG-Croce: NextGeneration Croce del Nord (NG-Croce)", Codice Identificativo IR0000026, Area ESFRI "Physical Sciences and Engineering", ammesso a finanziamento nell'ambito degli Interventi M4C2 - Investimento 3.1.

Soggetto proponente "Istituto Nazionale di Astrofisica", importo complessivo pari a 18.952.289,40 EUR a valere sulle risorse PNRR, con Decreto Direttoriale - Ministero dell'Università e della Ricerca, 27 ottobre 2022, numero 415. Autorizzazione alla sottoscrizione dell'atto d'obbligo connesso all'accettazione del finanziamento concesso al Progetto "NG-Croce" con Delibera n. 115/2022 del 15 dicembre 2022, del Consiglio di Amministrazione dello "Istituto Nazionale di Astrofisica"

*Chief Procurement Officer
(RUP)*

Ignazio Enrico Pietro Porceddu

Table of contents

1.	Definitions, glossary.....	3
1.1.	Glossary.....	3
2.	Tender description.....	5
2.1.	Contracting Authority.....	5
2.2.	Background - Summary of the tender.....	5
2.3.	Tender's Lots.....	5
2.4.	General timeline.....	8
2.5.	Delivery.....	8
3.	Lot 1 – X-Engine.....	9
3.1.	Summary and background.....	9
3.2.	Lot 1 - Value.....	10
3.3.	X-engine node requirements.....	10
3.4.	Simplified system diagram (each 2U X-Engine node).....	14
3.5.	Submission checklist.....	15
4.	Lot 2 – “FRB-Search”.....	16
4.1.	Summary and background.....	16
4.2.	Lot 2 - Value.....	16
4.3.	FRB-Search node requirements.....	17
4.4.	Simplified system diagram (each 2U FRB-Search node).....	20
4.5.	Switch requirements.....	21
4.6.	High Speed Network cables.....	22
4.7.	Storage System Requirements.....	23
4.8.	Management Computer Requirements.....	24
4.9.	Submission checklist.....	25
5.	Warranty conditions and related support.....	25

1. Definitions, glossary

Technical specifications provide to the bidder technical details of the material, equipment and related delivery specifications, which the bidder is to supply to INAF if he becomes a successful bidder; technical specifications will become an annex of the contract once the actual bid is accepted. This document presents as well as describes

- The functional requirements, which indicate the purpose, objective and function of the supply;
- The technical requirements, which define the characteristics and technical specifications of the supply;
- The performance requirements. Which define what performance and level of service the supply must have;

The requested supply has to obey to the set of minimum functional, technical and performance requirements below listed. Furthermore, reward requirements identify the characteristics of a technical and/or functional and/or performance that improve these minimum requirements set by the contracting authority (INAF), subject to discretionary and "tabular" evaluation by the judging commission.

1.1. Glossary

2DPC	2 DIMMs per Channel
AVX	Advanced Vector Extensions
BIOS	basic input/output system
BMC	Baseboard Management Control
COTS	Commercial Off The Shelf
CPU	Central Processing Unit
DDR4	Double Data Rate 4 (memory standard)
DDR5	Double Data Rate 5 (memory standard)
DEC	Chief of the Execution Phase
DIMM	Dual In-line Memory Module (memory module)
DPDK	Data Plane Development Kit
DWPD	Drive Writes Per Day
ECC	Error Correction Code
FAT	Factory Acceptance Test
FE	Frontend Node
FPGA	Field Programmable Gate Array
FRB	Fast Radio Burst (an astronomical signal)
GPU	Graphics Processing Unit
HDD	Hard Disk Drive
HPC	High Performance Computing
HW	Hardware
ICT	Information and communications technology
IOMMU	Input-Output Memory Management Unit

IOPS	Input/Output operations Per Second
IPMI	Intelligent Platform Management Interface
KVM	Keyboard Video Mouse
LAN	Local Area Network
MM	Multi mode (multi mode)
NRZ	Non-return-to-zero
MTBF	Mean Time Between Failure
NIC	Network Interface Card
NVMe	Nonvolatile Memory Express
OACA	Osservatorio Astronomico di Cagliari
OCP 3.0	Open Compute Project 3.0 (a network card interface)
OEA	Operatore Economico Aggiudicatario (proposed winner)
OF	Optical fiber
OS	Operating System
PAM4	Pulse Amplitude Modulation with Four Levels
PCIe	Peripheral Component Interconnect Express
QSFP	Quad Small Form-factor Pluggable
RDMA	Remote Direct Memory Access
RDIMM	Registered Dual In-line Memory Module
RoCEv2	RDMA over Converged Ethernet version 2
RUP	Responsabile Unico del Progetto (Chief Procurement Officer)
SA	Stazione Appaltante (Contracting Authority)
SATA	Serial Advanced Technology Attachment
SFP	Small Form-factor Pluggable
SIMD	Single instruction, multiple data
SS	Storage Scratch
SSD	Solid-state Drive
SSE4.2	Streaming SIMD Extensions 4.2
SSH	Secure Shell
SW	Software
TBW	Terabytes Written
TDP	Thermal Design Power
TFLOPS	Tera(10^{12}) Float Point Operations per Second
TOPS	Tera(10^{12}) Operations per Second (used for integer operations)

2. Tender description

2.1. Contracting Authority

The contracting authority is the "National Institute of Astrophysics - Astronomical Observatory of Cagliari (INAF-OACA)", with headquarters in via della Scienza 5 - 09047 Selargius (CA). Tax code is 97220210583, VAT number 06895721006, ISTAT code 092011. Website: <http://www.oa-cagliari.inaf.it/>, client profile http://www.oa-cagliari.inaf.it/page.php?id_page=101&level=3.

The certified electronic e-mail address (PEC) is inafoacagliari@pcert.postecert.it

2.2. Background - Summary of the tender

This tender is for computer systems to process astronomical data in real-time from a new radio telescope located near Penticton, British Columbia, Canada. Unlike a more generic computer cluster which might perform multiple tasks for a variety of users and use cases over time, this hardware will run the same code continuously. In addition, the system is considered a soft-real-time cluster, meaning it must keep up with the incoming data rate at all times. The tender consists of two lots. The first lot covers the so-called "X-engine" which takes data from a custom FPGA array and processes it for various science cases before sending it to various secondary processing clusters. The second lot covers one of those secondary processing clusters (called the "FRB-Search"), which also must operate as a soft-real-time system. An overview of the entire system (not all of which is in this tender), is shown in **Table 1**.

Please note that the overview/diagram only shows the main "core" array. Additional nodes are needed for other telescope sites, and as spares.

5

2.3. Tender's Lots

The IT components to be supplied were split into two lots, which in principle can be supplied from two different Companies. Below a short summary of both lots.

2.3.1. Summary of Lot 1

We are looking to acquire high-performance, GPU accelerated computer systems to do soft-real-time processing of radio data from a new radio telescope being constructed in Canada. The system is unique from a general-purpose computer cluster as it must process a very high rate of data (>10 Tb/s) with high computational requirements on a soft-real-time basis, 24 hours a day, year round. This lot will consist of just the GPU accelerated computers with site infrastructure being provided by the telescope site operators. We are looking for between 70 and 82 of these high performances GPU accelerated systems, each of which will have 2 CPUs, 2 GPUs, a large RAM buffer on each CPU, and 4 high-speed NICs.

2.3.2. Summary of Lot 2

We are looking to acquire high-performance, GPU accelerated computer systems to do a soft-real-time search for astrophysical events in sky images generated by a new radio telescope being constructed in Canada. The system is unique from a general purpose computer cluster in that it must process a very high rate of data and search it with complex algorithms on a soft-real-time basis, 24 hours a day, year round.

The bulk of the equipment will be 18 to 23 of these GPU accelerated computers, each with 1 CPU, 2 GPUs, a large RAM buffer, and 2 high-speed NICs.

In addition to the soft-real-time computers, we will require a large, full-mesh, high-speed switch supporting at minimum 256 by 25 Gb/s links via 100G breakout cables. As well as 2 support computers, 3 storage computers, and cabling for the high-speed network. The telescope site operators will provide the remaining site infrastructure.

Lot	Lot short IDs	CIG	Value
1	"X-engine" correlation	A038627B9B	2.332.570,00 EUR
2	cluster FRB	A0386707DA	650.00,00 EUR

Condensed System Diagram

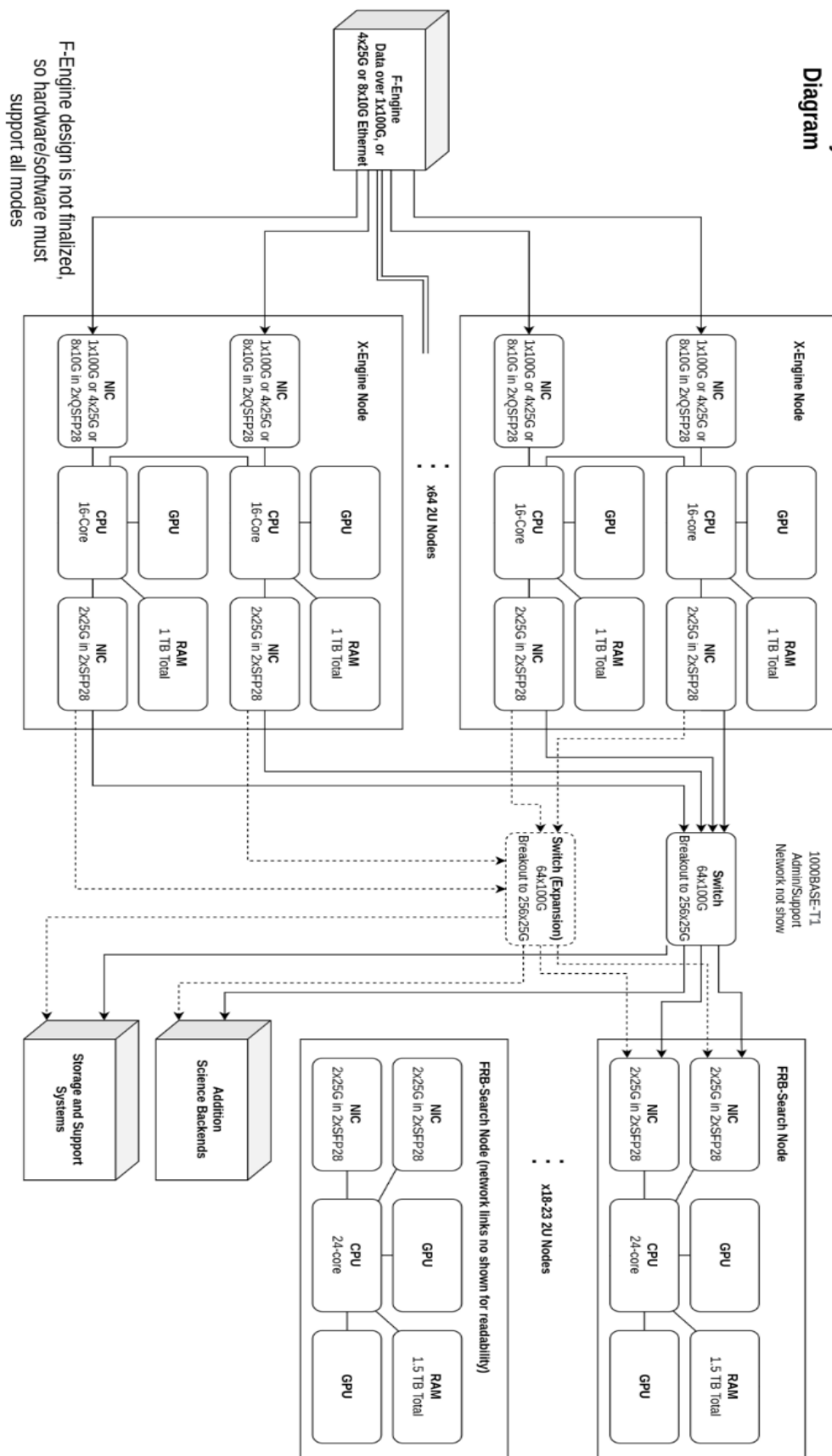


Table 1 – Overview of the entire compute system

2.4. General timeline

Times are relative to the start of the tender process:

- 1) Weeks 0-5 (35 days): Tender open
- 2) Weeks 5-10 (35 days): Evaluation of proposals
- 3) Weeks 10-14 (30 days): Announcement of tentative winner and public posing period
- 4) Week 14: Final selection of vendor(s).

2.5. Delivery

2.5.1. Delivery timeline

Delivery to take place within 4 months of contract award, subject to any unforeseen external hardware shortages causing longer lead times.

2.5.2. Delivery place

The winner / entrusted Company must be aware that the Contracting Authority requires the whole supply must have as "Delivery Place" of the DDP Incoterm the following address:

CHORD Project c/o Leo Belostotski
University of Calgary, Electrical Engineering
3838 24 Ave NW
Calgary, AB - T2N 4V5 CANADA

8

2.5.3. Delivery mode and terms

The winner / entrusted Company must be aware that the Contracting Authority requires the **Incoterms Delivered Duty Paid (DDP)** to be used. As known, DDP is a delivery agreement whereby the **Seller** assumes all of the responsibility, risk, and costs associated with transporting goods until the Buyer receives them at the above describe "**Destination place**".

This agreement includes paying for shipping costs, export and import duties, insurance, and any other expenses incurred during shipping to the University of Calgary, the established "Destination place" according to Incoterms DDP.

3. Lot 1 – X-Engine

3.1. Summary and background

We are building a Graphics Processing Unit (GPU) accelerated real-time compute cluster to process raw voltage data from a radio telescope. The system will receive data over Ethernet links at a fixed rate from a large number of FPGAs, then process it in real-time on the GPUs, and then transmit the resulting data products to additional computer clusters either onsite, or to remote locations.

Unlike a traditional scientific computer cluster, which might have a wide range of users, this one will be purpose built for exactly this task. In addition, because the data is arriving at the system continuously the system must be able to process the incoming data in real-time. In a normal cluster if the system happens to be too slow by say 10%, then jobs simply take 10% longer, however for this cluster being 10% (or even 1%) slower than required results in a complete failure of the system to achieve the science goals of the project. Also unlike a traditional cluster, adding more computer nodes will not trivially add additional performance, since the network connections to the FPGAs are likely to be one-to-one, and the data cannot be easily broken down to cover more nodes without considerable costs in additional switching infrastructure and software or FPGA firmware design changes. As a result the specifications are more strict than for a general purpose compute cluster.

The simplified description of our application is:

9

- 1) We receive data from a custom FPGA array, with a total system data rate around 10 Tb/s. The design of this FPGA array is not yet finalised, but it will transfer this data using:
 - 100G Ethernet (QSFP28/56 transceivers)
 - 25G Ethernet (4x25G links in QSFP28 transceivers)
 - or 10G Ethernet (4x10G links in QSFP+ transceivers)

Therefore the Input NICs need to be able to support all of the above modes.

Each CPU will handle about 78 Gb/s of input data, a copy of this data is buffered for about 100 second in system memory (requiring 1 TB of RAM for each CPU), and another copy is transferred to the GPU and processed, then the output products from the GPU are sent to other systems via a second network card at a rate of up to 45 Gb/s using 2xSFP28 ports. Periodically the system will also copy out parts of the 100 second buffer over the network, and save them for offline processing. We require hardware of sufficient performance to run this application, along with enough overhead to allow for future system expansions.

Due to space constraints on site, we require a compact design with 64 nodes which each fit into standard 2U rack spacing. (64 nodes are at the core site, additional nodes will be used to support smaller outtrigger sites and as spares). Thus, each node will receive around 156 Gb/s, and output around 90 Gb/s. Each of these nodes will have 2 CPUs, 2 Input NICs, 2 GPUs, two output NICs, and 2 TB of RAM. Additional nodes will be used to support additional remote sites, and provide hot spares.

3.2. Lot 1 - Value

Lot maximum cost: **2,332,570 euro**. Including node hardware, assembly, shipping and handling, and customs and taxes associated with delivering the items to the University of Calgary in the province of Alberta, Canada. Installation by the vendor is not required at either the University of Calgary or the telescope site.

3.3. X-engine node requirements

The required X-engine hardware is as follows:

Item	Minimum Quantity	Maximum Quantity
GPU Compute Server "X-Engine node"	70	82 (nodes for additional sites, and spares)

We have spent considerable time optimising our application, and testing its performance on various systems. The specifications below will meet our system requirements provided the equipment performs according to its specification.

Each "X-engine" is a 2U sized GPU optimised server with the following components and requirements:

- **2x CPUs each with the following properties:**
 - Minimum of 16 cores
 - Base clock of at least 2.0 GHz
 - Support for at least 8 memory channels of DDR5 memory running at a minimum of 4400 MT/s in 1 DIMMs per Channel (1DPC) configuration
 - Support for running at least 16 DIMMs in 2 DIMMs per Channel (2DPC) configuration with a clock speed of at least 4400 MT/s (if using 64GB RDIMMs, see below for memory requirements)
 - Minimum 30 MB of cache
 - Support for running in dual-socket (2P) configurations
 - Support for at least 48 PCI Express 5.0 lanes per CPU
 - Natively support the x86-64 instruction set, with support for the following vector instruction extensions:
 - ✧ All standard AVX2 and SSE4.2 instructions
 - ✧ AVX-512 Foundation (AVX512F)
 - ✧ AVX-512 Byte and Word Instructions (AVX512BW)
 - ✧ AVX-512 Vector Length Extensions (AVX512VL)
 - ✧ AVX-512 Vector Neural Network Instructions (AVX512_VNNI)
 - ✧ AVX-512 Half-Precision Floating-Point Instructions (AVX512_FP16)
 - Uses a monolithic core design to minimise core to core cache latency for packet processing pipelines
 - Parts with a Thermal Design Power (TDP) of less than 200W are preferred
 - Direct to cache support for I/O devices like network cards
 - Must not be a vendor locked part to a specific motherboard vendor

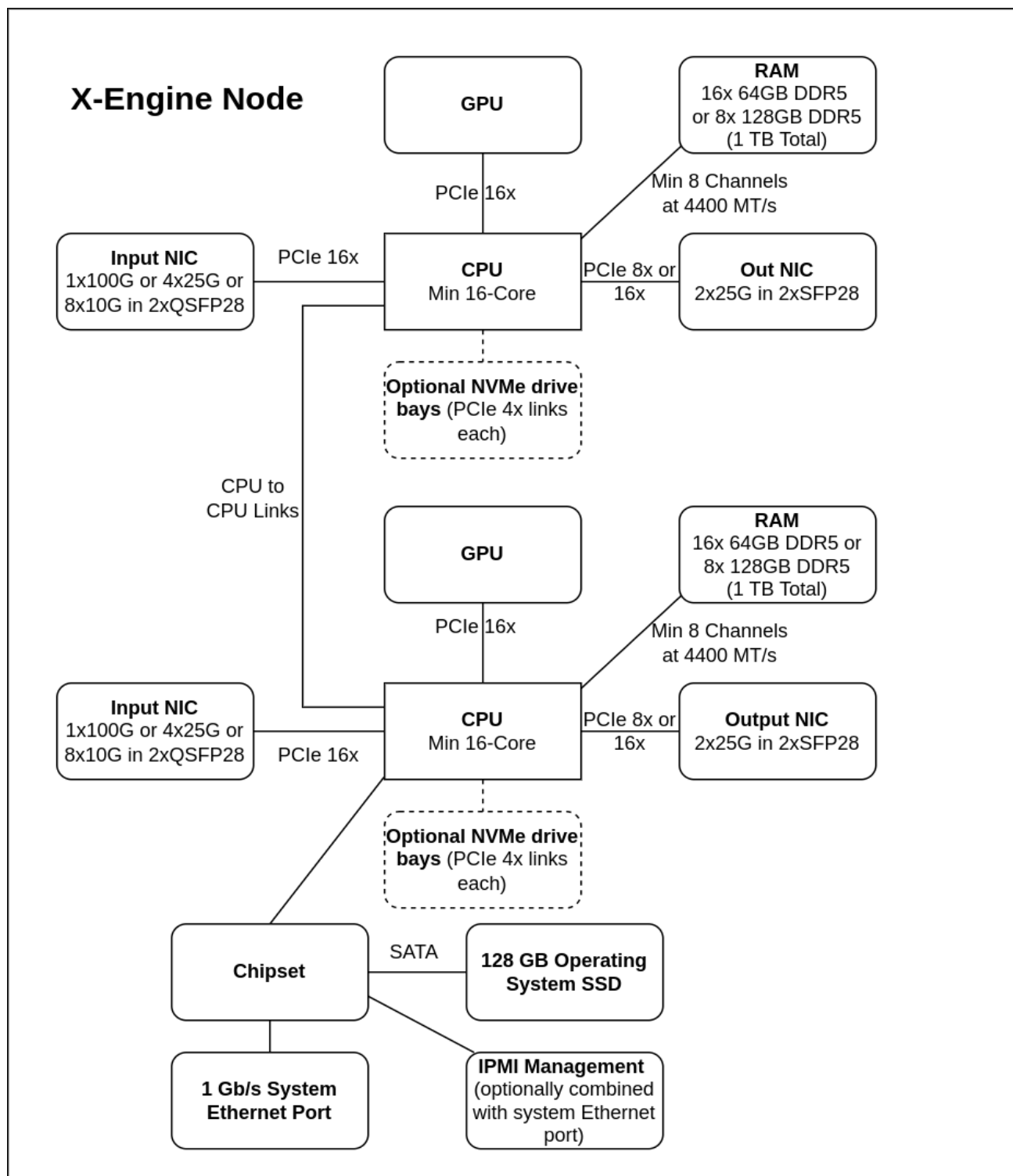
- **2x GPUs each with the following properties:**
 - Minimum PCIe 4.0 bus speed with 16x lanes
 - Peak global memory bandwidth of at least 690 GB/s
 - Peak FP32 performance of at least 37 TFLOPS
 - Peak INT4 performance in programmable matrix cores of at least 590 TOPS
 - Peak INT8 performance in programmable matrix cores of at least 290 TOPS
 - Peak FP16 performance in programmable matrix cores of at least 140 TFLOPS with an FP16 accumulator
 - At least 48 GB of GPU global memory with EEC support
 - Thermal Design Power (TDP) of at most 300W
 - Driver with EULA allowing data centre use
 - Designed to operate 24/7 with a passive cooler design (no active fan)
- **2x NICs (Input NICs) each with the following properties:**
 - Must have 2x QSFP28 Ethernet ports
 - Minimum PCIe 4.0 bus speed with 16x lanes
 - Can use PCIe or OCP 3.0 form factor
 - Maximum throughput of at least 100 Gb/s in one port mode (note it is not necessary that both QSFP28 ports run at full 100 Gb/s rate, see modes of operation below)
 - Each card must support the following 4 modes of operations:
 - ✧ 1x100Gb/s: with 4xNRZ encoding in one QSFP28 port
 - ✧ 1x100Gb/s: with 2xPAM4 encoding in one QSFP28(56) port
 - ✧ 4x25Gb/s: with 4xNRZ encoding in one QSFP28 port to support breakout QSFP28 to 4x SFP28 cables. NIC must provide 4 separate MAC addresses
 - ✧ 8x10Gb/s with 4xNRZ 10G in two QSFP28 ports (operating as QSFP+) to support breakout cables QSFP+ to 4xSFP+. NIC must provide 8 separate MAC addresses
 - Card must support the open source Data Plane Development Kit (DPDK) kernel by-pass library
 - Support for jumbo frames up to 9K.
 - Card must be capable of line rate 100GbE performance in a dpdk-l3fwd zero frame loss test with frame sizes of 1024 KB or less on at least one port. The vendor can reference public benchmarks at <https://core.dpdk.org/perf-reports/> for DPDK versions 22.11 or higher, or similar publicly posted benchmarks, for proof this requirement is met by the proposed NIC.
 - No vendor locks on transceiver support. The vendor does not need to guarantee support for third party transceivers, but must not block their operation with a software or firmware vendor check.
 - As an example of cards that meet these requirements, we have lab tested both of these units and found them to meet our requirements:
 - ✧ Silicom P4CG2I81-QX4
 - ✧ Intel E810-CQDA2
- **2x NICs (Output NICs) each with the following properties:**
 - Must have 2x SFP28 25Gb/s Ethernet ports
 - Minimum PCIe 4.0 bus speed with at least 8x lanes

- Can use PCIe or OCP 3.0 form factor
 - Must be able to use both ports at line rate simultaneously
 - Support for the following technologies:
 - ✧ Remote Direct Memory Access (RDMA) over converged Ethernet version 2 (RoCEv2)
 - ✧ Support for RoCEv2 memory access directly into the GPU global memory space without a copy through the host system memory
 - ✧ TCP/UDP offload
 - ✧ Support the open source Data Plane Development Kit (DPDK) kernel by-pass library
 - No vendor locks on transceiver support. The vendor does not need to guarantee support for third party transceivers, but must not block their operation with a vendor check.
- **32x 64GB PC5-38400 4800MHz DDR5 ECC RDIMMs or 16x 128 GB PC5-38400 4800MHz DDR5 ECC 3DS RDIMMs**
- Memory model must work reliably with the selected motherboard and fully populate all CPU channels according to the motherboard requirements.
 - Memory per CPU must be at least 1 TB total, with a system total of at least 2 TB per node.
- **1x 128 GB (or larger) enterprise class SSD**
- Can be on a SATA or NVMe interface
 - Used just for the operating system
- **1x Chassis and motherboard**
- Must be compatible with, fit, and power, all of the above equipment. Note: the equipment above does not need to be listed in the QVL table of the chosen motherboard, provided the parts are functionally compatible, and sufficiently powered and cooled at full load.
 - Must be able to cool all of the above equipment sufficiently running at maximum load in an operating range between at least 10C to 35C and a non-condensing humidity of between at least 10 to 80%.
 - Must not exceed 2 standard rack units (2U), or 87mm, in system height, or 850 mm in system depth.
 - Include rails for a standard post-mount server rack.
 - Redundant power supplies which are each capable of providing sufficient power to all hardware at maximum load with reasonable overhead when running on 208V AC power (North American 208V power). The power supplies also require a minimum of an “80 PLUS Platinum” efficiency rating.
 - Have a 2 CPU (2P) layout with links between the two CPUs
 - The GPUs and NICs above must be installed in the system such that:
 - ✧ Each of the two CPUs is directly attached via PCIe (can use standard PCIe slots or OCP 3.0 slots for the NICs) to exactly:
 - 1x GPU with 16x PCIe 5.0 lanes
 - 1x Input NIC with 16x PCIe 5.0 lanes
 - 1x Output NIC with 8x or 16x PCIe 5.0 lanes
 - Note the links must support PCIe 4.0 backwards compatibility to support the GPUs and NICs, *which can be PCIe 4.0 parts.*

- Refer to the diagram below for a visual representation of the required layout.
- ✧ Each device must be in its own PCIe root complex
- ✧ Each device must have its own dedicated (not shared) lanes to the CPU
- ✧ ***Vendor is to provide a detailed block diagram of the motherboard along with proposed locations for the GPUs and NICs, clearly showing the PCIe link arrangement to the CPUs.***
- If using 64 GB RDIMMs:
 - ✧ Have at least 16x DIMM slots per CPU socket (32 DIMM slots total) connected to at least 8 memory channels per CPU
 - ✧ Support a memory clock of at least 4400 MT/s in this configuration
- If using 128 GB 3DS RDIMMs:
 - ✧ Have at least 8x DIMM slots per CPU socket (16 DIMM slots total) connected to at least 8 memory channels per CPU
 - ✧ Support a memory clock of at least 4400 MT/s in this configuration
- No enforced locking of the CPU, GPU, NIC, RAM, or SSD compatibility to a single vendor/supplier, either by use of firmware checks or physically non-industry-standard slots, which would artificially limit the reasonable expandability, or reparability of the node. Note, the vendor is *not* required to guarantee compatibility with third party parts, nor provide warranty support if third party equipment is used.
- Have at least 1x 1 Gb/s Ethernet built-in NIC (which can be combined or separate from the out-of-band management port).
- Must support out-of-band management (e.g. IPMI) with remote Keyboard Video Mouse (KVM) support via a web browser with an HTML5 based interface without additional licensing costs (or included perpetual licenses).
- Motherboard must support the following BIOS options: virtualisation support and Input-Output Memory Management Unit (IOMMU) support (for DPDK Poll-mode drivers).
- Additional points are available for the following features:
 - ✧ Support for up to 4x NVMe U.3 SSD drives (each with a 4x PCIe 4.0/5.0 link). No SSDs for these drive bays are required in the bid, just the drive bays for future expansion.
- Examples of chassis which could meet requirements (subject to exact part selection above) are:
 - ✧ Gigabyte R283-S93 (rev. AAF1)
 - ✧ Asus RS720-E11-RS12U (with the GPU and OCP 3.0 options)

In addition, the nodes must support the open source Linux operating system, specifically Ubuntu 22.04 LTS. All hardware must have drivers or native support for that operating system.

3.4. Simplified system diagram (each 2U X-Engine node)



3.5. Submission checklist

Vendors **must provide** the following documentation to support their design (in addition to all other tender documentation requirements):

- Detailed list of all proposed components for each X-engine node
- Detailed block diagram of the motherboard showing how all the components will be connected to the CPUs
- Reference a benchmark showing the DPDK performance of the input NICs on a standard dpdk-l3fwd zero frame loss test.
- Detailed spec sheets for each proposed component
- Description of the Quality Assurance process
- (Optional but preferred) Copy of the proposed motherboard manual

4. Lot 2 – “FRB-Search”

4.1. Summary and background

We are building a Graphics Processing Unit (GPU) accelerated real-time compute cluster to search for phenomena known as Fast Radio Bursts (FRBs). This cluster (the search cluster) will receive data from the other cluster in this bid package (the “X-engine”) over an Ethernet network. It will search the incoming data for the unique signature of FRBs, and if it finds one, save a copy of the incoming data from its RAM buffer to a high-speed SSD. The data will then be pushed into slow archival storage later. The majority of the search process will take place in the GPUs, however since the data set is very large, it will require a lot of bandwidth between the GPU and the host system memory, along with a large amount of host memory.

Unlike a traditional scientific computer cluster, which might have a wide range of users, this one will be purpose built for exactly this task. In addition, because the data is arriving at the system continuously, the system must be able to process the incoming data in real-time. In a normal cluster if the system happens to be too slow by say 10%, then jobs simply take 10% longer, however for this cluster being 10% (or even 1%) slower than required results in a complete failure of the system to achieve the science goals of the project.

The bulk of this lot is the FRB search system. Which are a number of GPU accelerated 2U compute nodes with 1 or 2 CPUs each, and 2 GPUs, 2 Network Cards, 2 high-speed NVMe drives, and 1.5 TB of RAM. The number of CPUs and choice of DIMM size (64GB or 128GB) is left to the vendor to optimise for cost.

16

The contents of this lot can be summarised as follows:

Item	Minimum Quantity	Maximum Quantity
GPU Compute Server “FRB Search node”	18	23
Core 64x100G Switch	1	1
Breakout Cables 100G to 4x25G (15m)	48	48
Storage Systems	3	4
Management Computers	2	2

4.2. Lot 2 - Value

Lot maximum cost: **650.000,00 euro**. Including node hardware, assembly, shipping and handling, and customs and taxes associated with delivering the items to the University of Calgary in the province of Alberta, Canada. Installation is not required at either the University of Calgary or the telescope site.

4.3. FRB-Search node requirements

Each node must have the following components:

- **One of the following CPU configurations:**
 - **1x CPU with the following properties:**
 - Minimum of 24 cores
 - Base clock of at least 2.5 GHz
 - Support for 12 memory channels of DDR5 memory running at a minimum of 4800 MT/s in 1 DIMMs per Channel (1DPC) configuration and 3600 MT/s in a 2 DIMM per channel (2DPC) configuration
 - Minimum 64 MB of L3 cache
 - Minimum of 128 PCIe 5.0 lanes
 - Natively support the x86-64 instruction set, with support for the following vector instruction extensions:
 - ✧ All standard AVX2 and SSE4.2 instructions
 - ✧ AVX-512 Foundation (AVX512F)
 - ✧ AVX-512 Byte and Word Instructions (AVX512BW)
 - ✧ AVX-512 Vector Length Extensions (AVX512VL)
 - **Or 2x CPUs each with the following properties:**
 - Minimum of 16 cores
 - Base clock of at least 2.5 GHz
 - Support for 12 memory channels of DDR5 memory running at a minimum of 4800 MT/s in a (1DPC) configuration
 - Minimum 64 MB of L3 cache
 - Minimum of 64 PCIe 5.0 lanes
 - Natively support the x86-64 instruction set, with support for the following vector instruction extensions:
 - ✧ All standard AVX2 and SSE4.2 instructions
 - ✧ AVX-512 Foundation (AVX512F)
 - ✧ AVX-512 Byte and Word Instructions (AVX512BW)
 - ✧ AVX-512 Vector Length Extensions (AVX512VL)
- **2x GPUs each with the following properties:**
 - Minimum PCIe 4.0 bus speed with 16x lanes
 - Global memory bandwidth of at least 690 GB/s
 - Peak FP32 performance at least 37 TFLOPS
 - Peak INT4 performance in programmable matrix cores of at least 590 TOPS
 - Peak INT8 performance in programmable matrix cores of at least 290 TOPS
 - Peak FP16 performance in programmable matrix cores of at least 140 TFLOPS with an FP16 accumulator
 - Thermal Design Power (TDP) of at most 300W

- At least 48 GB of GPU global memory with EEC support
- Driver with EULA allowing data centre use
- Designed to operate 24/7 with a passive cooler design (no active fan on the GPU)
- **2x NICs each with the following properties:**
 - 2x 25Gb/s Ethernet ports in the SFP28 form factor
 - Must be able to use both ports at line rate simultaneously
 - Support for the following technologies:
 - ✧ Remote Direct Memory Access (RDMA) over converged Ethernet version 2 (RoCEv2)
 - ✧ Support for RoCEv2 memory access directly into the GPU global memory space without a copy through the host system memory
 - ✧ TCP offload
 - ✧ Support the open source Data Plane Development Kit (DPDK) kernel by-pass library
 - Minimum PCIe 4.0 bus speed with 8x lanes
 - Can use PCIe or OCP 3.0 form factor
- **Total of at least 1.5 TB of system memory in one of the following configurations:**
 - **24x 64GB PC5-38400 4800MHz DDR5 ECC RDIMMs**
 - ✧ This configuration can be used either with the 1 CPU per chassis design in a 2DPC configuration, or with the 2 CPU per chassis design with a 1DPC configuration. In 2DPC configurations, the memory must operate at a speed of at least 3600 MT/s.
 - **12x 128 GB PC5-38400 4800MHz DDR5 ECC (3DS) (L)RDIMMs**
 - ✧ This configuration can only be used in the 1 CPU per chassis design in a 1DPC configuration.
 - Memory model must work reliably with the selected motherboard and fully populate all CPU channels according to the motherboard requirements.
- **2x Enterprise class SSDs each with:**
 - Minimum 3.2 TB.
 - Endurance in total bytes written of at least: 17500 TB
 - Stated peak sequential write speed of at least 5000 MB/s
 - Stated peak sequential read speed of at least 6500 MB/s
 - Stated peak IOPS Read/Write of at least: 1,000,000/350,000 IOPS respectively
 - U.3 form factor
- **1x 128 GB (or larger) enterprise class SSD**
 - Can use a SATA or NVMe interface
 - Used for the operating system.
- **1x Chassis and motherboard:**
 - Must be compatible with, and fit, all of the above equipment. Note: the equipment above does not need to be listed in the QVL table of the chosen motherboard, provided the parts are functionally compatible, and sufficiently powered and cooled at full load.

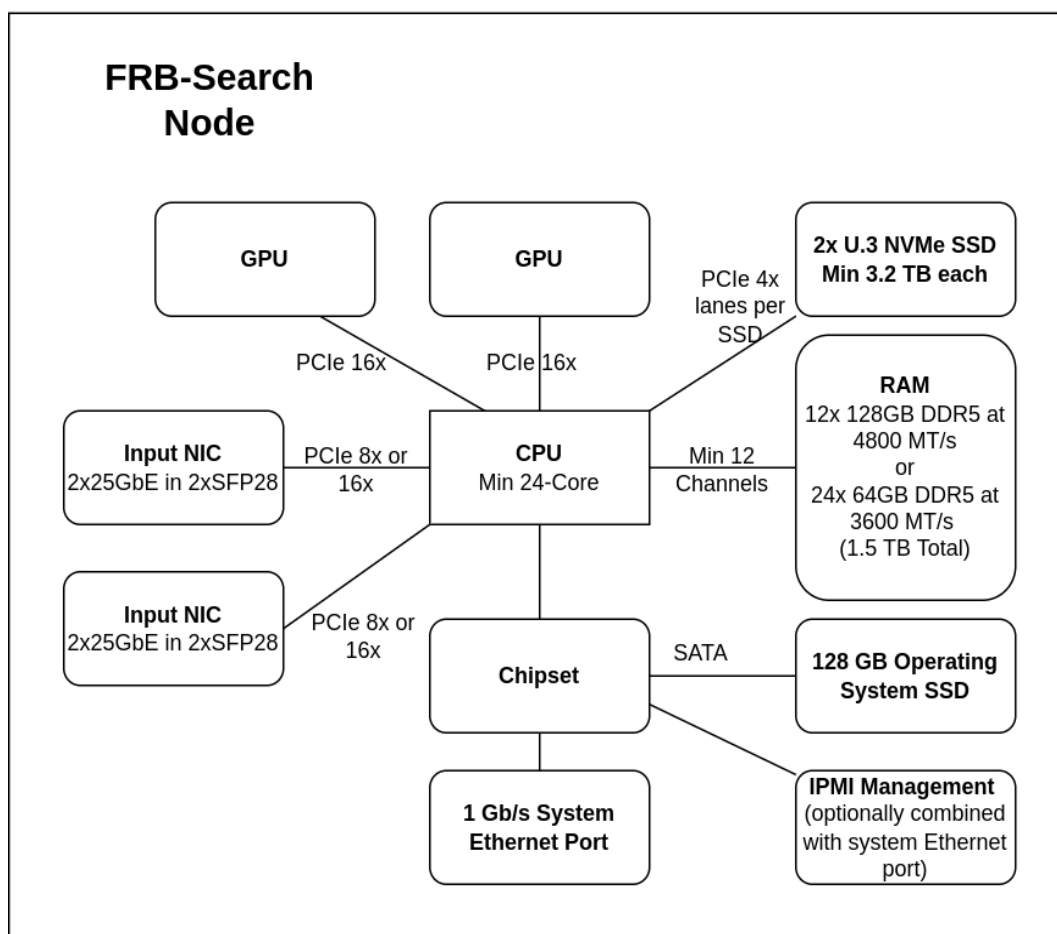
- Must be able to cool all of the above equipment sufficiently running at maximum load in an operating range between at least 10C to 35C and a non-condensing humidity of between at least 10 to 80%.
- Must not exceed 2 standard rack units (2U), or 87mm, in system height, or 850 mm in depth, and include rails for a standard post-mount server rack.
- Have at least 4x front loading U.3 drive bays
- There are two possible layouts for the system:
 - ✧ With 1 CPU per chassis:
 - Each GPU, NIC, and SSD must directly connected to the CPU with the following minimum lane requirements:
 - ✧ 2x GPUs with 16x PCIe 5.0 lanes each
 - ✧ 2x NICs with at least 8x PCIe 5.0 lanes each
 - ✧ 2x SSD with at least 4x PCIe 5.0 lanes each
 - Refer to the first diagram below, “Option 1” for a visual representation of the required layout.
 - ✧ With 2 CPUs per chassis:
 - Each CPU must be directly connected to the following:
 - ✧ 1x GPUs with 16x PCIe 5.0 lanes each
 - ✧ 1x NICs with at least 8x PCIe 5.0 lanes each
 - The 2x SSDs must each have at least 4x PCIe 5.0 lanes and can be connected to either CPU
 - Refer to the second diagram below, “Option 2” for a visual representation of the required layout.
 - ✧ Note the links must support PCIe 4.0 backwards compatibility to support the GPU, NICs and SSDs, *which can be PCIe 4.0 parts.*
 - ✧ NICs can use PCIe or OCP 3.0 form factors
 - ✧ Each device must have its own dedicated lanes to the CPU
 - ✧ ***Vendor is to provide a detailed block diagram of the motherboard along with proposed locations for the GPUs, NICs, and SSDs, clearly showing the PCIe link arrangement.***
- Redundant power supplies which are each capable of providing sufficient power to all hardware at maximum load with reasonable overhead when running on 208V AC power (North American 208V power). The power supplies also require a minimum of an “80 PLUS Platinum” efficiency rating.
- No enforced locking of the CPU, GPU, NIC, RAM, or SSD compatibility to a single vendor/supplier, either by use of firmware checks or physically non-industry-standard slots which would artificially limit the reasonable expandability or repairability of the node. Note, the vendor is *not* required to guarantee compatibility with third party parts, nor provide warranty support if third party equipment is used.

- Have at least one 1 Gb/s Ethernet built-in NIC (which can be combined or separate from the out-of-band management port).
- Must support out-of-band management (e.g. IPMI) with remote Keyboard Video Mouse (KVM) support via a web browser with an HTML5 based interface without additional licensing costs (or included perpetual licenses).
- Examples of chassis which could meet requirements (subject to exact part selection above) are:
 - ✧ Asus RS520A-E12-RS12U
 - ✧ SuperMicro AS-2025HS-TNR (with some riser and NVMe options)
 - ✧ Gigabyte R283-Z93 (rev. AAF1)

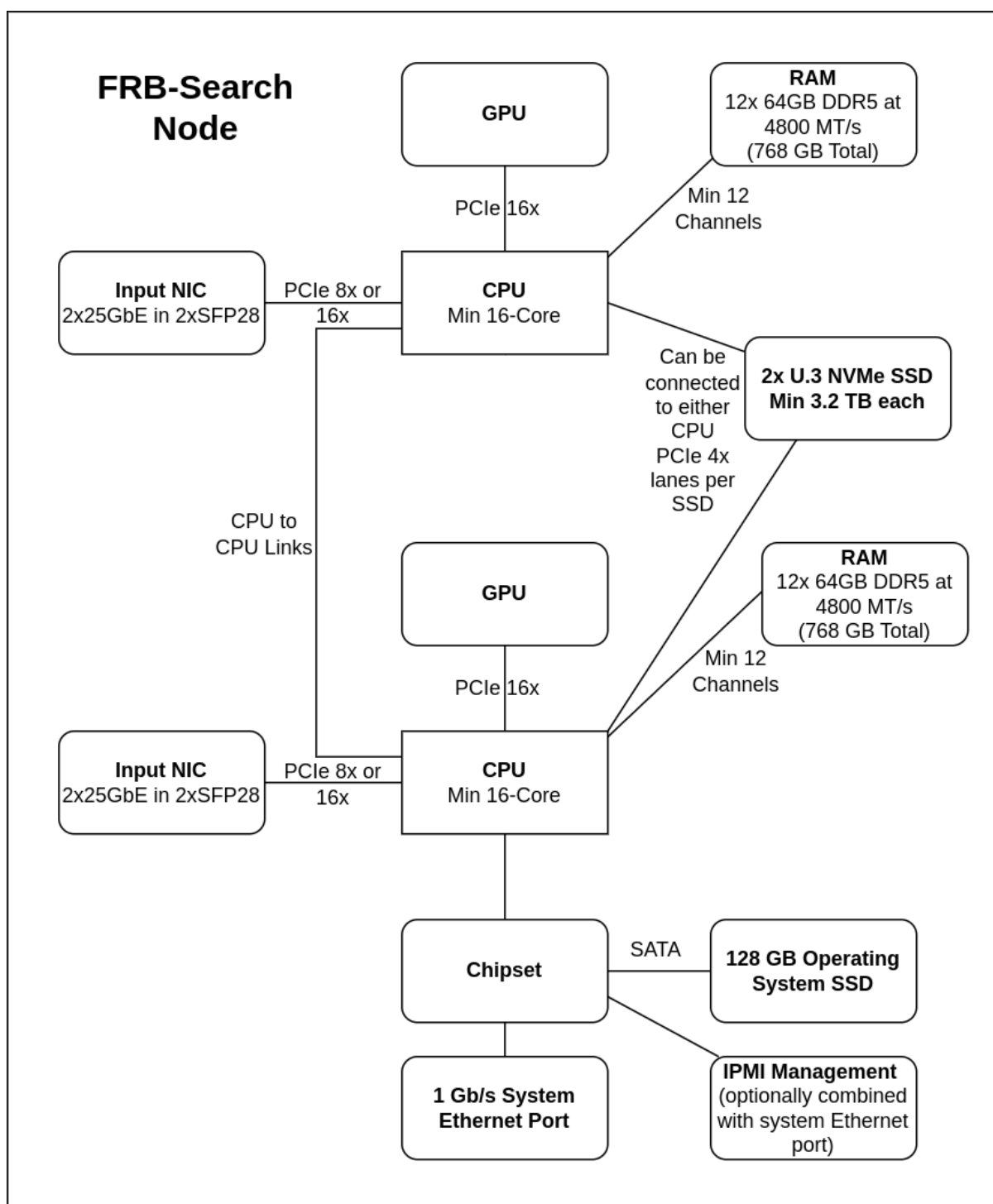
In addition, the nodes must support the open source Linux operating system, specifically Ubuntu 22.04 LTS. All hardware must have drivers or native support for that operating system.

4.4. Simplified system diagram (each 2U FRB-Search node)

4.4.1. Option 1: 1x CPU per chassis



4.4.2. Option 2: 2x CPU per chassis



4.5. Switch requirements

To connect the X-engine to the backends (FRB search, and other systems) we require a switch capable of creating a full mesh network between the two sets of systems. In general each X-engine node sends some proportional fraction of its data to every node in the backends. Therefore the network core must be

capable of full mesh switching near line-rate for all nodes in the system.

The simplest layout involves using one (or more) core switches with 64x 100 Gb/s Ethernet ports supporting QSFP28 breakout to 4x SFP28 on all ports. This effectively creates a 256 port 25 Gb/s Ethernet switch.

The requirements of this switch are listed below:

- 64x QSFP28 ports, each capable of breakout to 4x SFP28 25 Gb/s ports via a supported breakout cable. A total of 256 x 25 Gb/s individually manageable Ethernet ports.
- Support backwards compatibility with QSFP+ 40 Gb/s and breakout to 4x SFP+ 10Gb/s Ethernet modes.
- Capable of full-mesh networking between all ports
- A core with a minimum of 12 Tb/s total throughput (sum of input/output rates on all links)
- Forwarding rate of at least 4 Bpps (billion packets per second)
- Jumbo frame support up to at least 9KB packets
- APR/MAC tables with at least 10,000 entries
- Non-blocking
- Packet buffer of 40 MB or more
- Management system providing:
 - ✧ SSH based command line remote configuration (web and serial configuration optional)
 - ✧ Support for Simple Network Management Protocol (SNMP) v2/v3
- Support for at least 50 VLANs
- Support for Link Aggregation Control Protocol (LACP)
- Support for RoCEv2 congestion control protocols, specifically:
 - ✧ Priority-based flow control (PFC)
 - ✧ Explicit Congestion Notification (ECN)
- Redundant power supplies
- Include all licences and/or service contracts required to operate the switch and update the firm-ware as needed over at least 5 years.
- No vendor lock on transceiver compatibility. The switch does not need to qualify it will work with any vendor of transceiver, but must not block the operation of 3rd party transceivers (or provide an option to enable the use of 3rd party transceivers without additional cost)
- L3 multicast support with a table containing at least 1000 entries
- Support spanning tree protocol (STP)
- Operating temperature range of at least 10C to 35C, and operating humidity range of at least 10% to 85% non-condensing.
- Two examples of switches which satisfy these requirements are:
 - Cisco 9364D-GX2
 - FS N8560-64C

4.6. High Speed Network cables

For the high-speed data network we will require cables to connect the X-engine to the backends. These cables will have a QSFP28 port on one end, and 4x SFP28 ports on the other end. To minimise connection

failure points, all the cables should be fused to their SFP28 and QSFP28 transceivers modules (e.g. all-in-one cables).

The cables must be compatible with the proposed data switch.

The breakout point "pig-tails" must be at least 1m from the SFP28 end of the cable, and comfortably allow each SFP28 transceiver of a breakout cable to connect to up to 4x adjacent 2U nodes.

4.7. Storage System Requirements

We will require at least three storage servers for temporary storage of data products prior to moving them offsite.

Based on local experience, we will use the ZFS file system running in the Ubuntu 22.04 operating system. This requires the system to expose each disk individually in a so-called Just-a-Bunch-Of-Disks (JBOD) configuration. Hardware RAID systems will not be considered. Each disk should be connected to a Host Bus Adapter (HBA) card with sufficient bandwidth to handle all the attached drives at maximum sustained read/write speeds.

Each of the storage servers should be a self-contained unit of no more than 4U in total, e.g. not a chassis with a separate disk array.

The drives should be accessible from the front and/or back of the unit without needing to remove it from the rack.

The system components for each server should be as follows:

- Minimum of 24x 3.5 inch drive bays accessible from the front and/or back of the unit
- A CPU using the x86-64 instruction set, with at least 24 cores in a 1P configuration, and at least 8-channels of DDR4 or DDR5 memory. Minimum base clock speed of at least 2.0 GHz on all cores.
- Redundant power supplies with sufficient power to handle the inrush current of all drives. In addition they must have an 80 PLUS Platinum efficiency rating, and be able to operate at 208V.
- Maximum height of 4U
- Minimum of 256 GB of ECC DDR4 or DDR5 RDIMM memory with a speed of at least 3200 MT/s, and enough DIMMs to fully populate the CPU's memory controller, e.g. 8x 32 GB DIMMs for an 8-channel system.
- Sufficient Host Bus Adapter (HBA) cards to fully connect the 24 drives and operate them at full SATA speeds.
- 20x 16 TB Enterprise grade SATA HDDs with the following minimum specifications:
 - 2.5 million hour MTBF
 - Unrecoverable Read Error (URE) rate of 1 in 10^{15}
 - Rated for 24x7 continuous operation for 5 years
 - Maximum sustained transfer rate of 240 MB/s
 - Spindle Speed 7200 RPM

- 4x Enterprise grade SATA (or NVMe) SSD with of at least 3.2 TB in size with the following minimum specifications:
 - Endurance of at least 1 drive write per day (1 DWPd) for 5 years, e.g. for a 3.2 TB drive, that would be a total of ~5800 TBW.
 - At least 50,000 steady state random 4k read/write IOPS.
 - 2 million hour MTBF
 - 500 MB/s maximum sustained read/write speeds
 - Note: NVMe based solutions with drives inside the chassis meeting all other specs are also acceptable here.
- 1x 128 GB Enterprise grade SSD on a SATA or NVMe interface for the operating system
- Support out-of-band management (e.g. IPMI) with remote Keyboard Video Mouse (KVM) support via a web browser without additional licensing costs (or included licences).
- A network interface card with 2x25 Gb/s Ethernet network ports in the SFP28 form factor connected to the CPU with at least 8x PCIe 3.0 (or faster) lanes.
- Have at least 1x 1000BASE-T built-in NIC (which can be combined or separate from the out-of-band management port).
- All components are compatible with each other.
- Operating temperature between at least 10C and 35 C.
- Operating relative humidity of up to 80% non-condensing

4.8. Management Computer Requirements

We will require a few systems to host supporting applications (logging services, cluster management software, graphing tools, network monitoring, etc.). It is likely that these systems will have fairly low CPU usage, but lots of small file I/O, therefore systems with NVMe storage are optimal.

The requirements for each system are as follows:

- Minimum of 4x U.3 drive bays with front of chassis access.
- A CPU using the x86-64 instruction set, with at least 16 cores in a 1P configuration, and at least 8-channels of DDR4 or DDR5 memory. Minimum base clock speed of at least 2.0 GHz on all cores.
- Redundant power supplies with sufficient power to handle the inrush current of all drives at startup. In addition they must have an 80 PLUS Platinum efficiency rating, and be able to operate at 208V.
- Maximum height of 1U
- Minimum of 128 GB of ECC DDR4 or DDR5 RDIMM memory with a speed of at least 3200 MT/s, and enough DIMMs to fully populate the CPU's memory controller, e.g. 8x 16 GB DIMMs for an 8-channel system.
- 2x Enterprise class SSDs:
 - Minimum 3.2 TB each.
 - Endurance in total bytes written of at least: 17500 TB
 - Stated peak sequential write speed of at least 5000 MB/s

- Stated peak sequential read speed of at least 6500 MB/s
- Stated peak IOPS Read/Write of at least: 1,000,000/350,000 IOPS respectively
- U.3 form factor
- Support out-of-band management (e.g. IPMI) with remote Keyboard Video Mouse (KVM) support via a web browser without additional licensing costs (or included incenses).
- A network interface card with 2x25 Gb/s Ethernet network ports in the SFP28 form factor connected to the CPU with at least 8x PCIe 3.0 (or faster) lanes.
- Have at least 1x 1000BASE-T built-in NIC (which can be combined or separate from the out-of-band management port).
- All components are compatible with each other.
- Operating temperature between at least 10C and 35 C.
- Operating relative humidity of up to 80% non-condensing

4.9. Submission checklist

Vendors **must provide** the following documentation to support their design (in addition to all other tender documentation requirements):

- Detailed list of all proposed components
- Detailed block diagram of the FRB-Search system motherboard showing how all the components will be connected to the CPUs
- Detailed spec sheets for each proposed component in the FRB-search nodes
- (Optional but preferred) Copy of the proposed FRB-Search system motherboard manual
- Description of the Quality Assurance process

25

5. Warranty conditions and related support

Standard warranty period will be 12 months, when no different obligation is due. The extension of the guarantee period beyond the 12 months period is part of the evaluation process, which will award additional points and a better overall score to the bidder.

The company undertakes to replace at its own expenses those parts of the supply which, for any reason, are found to be unsuitable or defective, as well as to carry out all the consequent services for the entire period of contractual coverage accepted in the phase of award. The warranty and maintenance contract will start from payment of the supply.