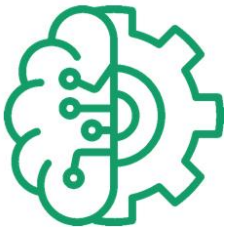
BRAINE



BRAINE - Big data pRocessing and Artificial Intelligence at the Network Edge

Project Title:	**BRAINE - Big data pRocessing and Artificial Intelligence at the Network Edge**
Contract No:	876967 – BRAINE
Instrument:	ECSEL Research and Innovation Action
Call:	H2020-ECSEL-2019-2-RIA
Start of project:	1 May 2020
Duration:	36 months

Deliverable D2.2

Second project report on the status of WP2

Due date of Internal Report:	28 February 2022
Actual submission date:	10 April 2022
Version:	2

ECSEL Joint Undertaking

Project funded by the European Community under the H2020 Programme for Research and Innovation.

Project ref. number	876967
Project title	BRAINE - Big data pRocessing and Artificial Intelligence at the Network Edge

Deliverable title	**WP2 Components for BRAINE Release 1.0**
Deliverable number	D2.2
Deliverable version	Version 3
Previous version(s)	Version 2
Contractual date of delivery	30 July 2021
Actual date of delivery	10 April 2022
Deliverable filename	*BRAINE_MS7_v4.docx*
Nature of deliverable	Report
Dissemination level	PU
Number of pages	95
Workpackage	WP2
Task(s)	T2.1, T2.2 and T2.3
Partner responsible	MLNX
Author(s)	J.J. Vegas Olmos (NV-MLNX), Gianluca Rizzi (WI3), Adam Flizikowski (ISW), Munjure Mowla (ISW), Patrick Moder (IFX), Antonino Albanese (ITL), Mats Hellman (HID), Philippe Nguyen (SIC), Janos Lazanyi (PCBD), Gautier Rouaze (JJC), L. Valcarenghi (SSSA), F. Cugini, F. Paolucci (CNIT), Peter Szanto (BME), B. Cimoli (TUE)
Editor	J.J. Vegas Olmos (NV-MLNX)

Abstract	WP2 HW and embedded SW components for BRAINE 1.0 solution - schematics on EMDC comprising switches, PCBs and peripherals. - technical requirements layer by layer to support AI. - components for processes at edge computing supporting AI - performance metric indicating which components are selected for acceleration - 5G software stack to be supported by the EMDC in BRAINE - Definition of the cryptography technology to be implemented in

	BRAINE and description of the roadmap for HW acceleration
Keywords	AI, Security, EMDC, Networking, P4, SOC, GPU, FPGA

Copyright

Deliverable history

Version	Date	Reason	Revised by
0.1	28/01/2022	First version	J.J. Vegas Olmos (NV-MLNX)
0.2	15/02/2022	Partner contributions	J.J. Vegas Olmos (NV-MLNX)
1	25/02/2022	Updated version	J.J. Vegas Olmos (NV-MLNX) F. Cugini (CNIT)
2	10/04/2022	Final revised version	J.J. Vegas Olmos (NV-MLNX) F. Cugini (CNIT)

List of abbreviations and Acronyms

Abbreviation	Meaning
AI	Artificial Intelligence
ANSD	Active Nucleation Site Density
ASIC	Application-Specific Integrated Circuits
BMC	board management node/controller
COM	computer-on-module
CPS	Cyber Physical Systems
CPU	central processing unit
DPU	data processing unit
EMDC	Edge Micro Data Center
FPGA	field-programmable gate array
GPU	graphics processing unit
KPI	Key Performance Indicator
LTS	Loop Thermosyphon
MWCNT	Multiwall carbon nanotubes
NIC	Network Interface Card
PCIe	peripheral component interconnect express
QoS	Quality of Service
SBC	single-board computer
SMT	simultaneous multithreading
SOC	System on Chip
TCO	Total cost of ownership
TDP	Thermal Design Power
TIM	Thermal Interface Material
VTU	Virtual Transcoding Unit
WPs	Work Packages

Table of Contents

1. Executive summary

This deliverable contains a detailed description of the hardware subsystems, system and embedded software components of BRAINE release 1.0, which built an edge micro data center (EMDC) able to support intensive artificial intelligence (AI).

The components have been designed and implementation towards the final release BRAINE 2.0 is in progress. This document describes each single component as well as its current status development.

2. Introduction

This document details the components for BRAINE, comprising a list of HW and SW components. They include:

- Components in the Hardware or firmware domain
  - Electrical system
  - CPU board
  - GPU board
  - NVMe board
  - PCIe switch
  - Ethernet switch design
  - Power supply
  - Backplane system
  - Cooling system
  - Enclosure and ejection tool
  - FPGA and board management firmware
- Components in the software domain
  - P4 programs and programmable switch operating system
  - Switch connectivity
  - 5G accelerators
  - Quantum communication system

This document is self-contained yet it is advisable to be considered together with the reports up to Month 20 of BRAINE in order to have a deep dive into the technologies developed in BRAINE.

3. List of WP2 components

3.1 List of WP2 HW components

#	Component name	Description	Status	Licence	Partner
C2.1	Monobloc evaporator	Evaporator of the loop thermosyphon cooling system. Ensure the mechanical alignment of the electronics boards.	First 4 slots version is manufactured and being tested. Second 11 slots version is being developed.	-	JJC
C2.2	MPT louvered fins condenser	Condenser of the loop thermosyphon cooling system.	Fist 600W version is being tested in BRAINE 1.0 testbench. Second 1.5kW enclosed version is under development.	-	JJC
C2.3	Thermal Equivalent PCB for COMe Modules	A set of 4 slot backplane, COMe Carrier and COMe Module PCBs are developed to mimic the heat load of the real modules	First prototypes being manufactured, and sent to JJC for integration	-	PCB
C2.4	BRAINE Connector test system	Mechanical dummy to test tolerance of the TE Sliver connector.	Sent to Helder ID for integration	-	PCB
C2.5	BRAINE EXAMAX SWITCH Test connector	Mechanical dummy to test tolerance of the EXAMAX connector.	Sent to Helder ID for integration	-	PCB
C2.6	BRAINE System CARDS	CPU, GPU, NVME, BMC, PCIe Switch,	Design ready, under manufacturing		PCB
C2.7	BRAINE System CARDS	Ethernet switch, Backplane	Schematics ready, PCB routing		PCB

C2.8	BRAINE System	ARM Node	Under development. The Original target was Ampere EMAG processor, but it become obsolete.		PCB
C2.9	Remote Radio Unit	O-RAN compliant 5G-NR Radio Unit	Initial use case operational in local lab. Functional split gap on fronthaul I/F being addressed	Proprietary	COM
C2.10	Nano-coolants	Working fluids containing nano-particles (nano-refrigerants)	Preliminary, stable samples of nano-coolants with H2O and EG as base fluids.	-	SYN
C2.11	Mechanical test model narrow slot	Mechanical testing of pressure applied to the cooling bridge by wedging, while maintaining a connection of the TE Sliver connector.	Testing finished.		HID
C2.12	Mechanical test model wide slot	Mechanical testing of pressure applied to the cooling bridge by wedging, while maintaining a connection of the EXAMAX connector.	Ongoing. Small changes to model needed to enable measurement of connector connection		HID
C2.13	Ejector tool P1	Mechanical tool for securely removing pcb assemblies from their slot in the monobloc, overcoming the resistance of the connector. First mechanical test model.	Testing finished. Lessons learned and implemented in P2.		HID
C2.14	Ejector Tool P2	Optimized version of P1.	Testing finished.		HID

C2.15	Heatspreader CPU	Custom Aluminum parts for the heat transition between the CPU and the cooling bridge of the monoblock.	Mechanical testing finished.		HID
C2.16	Quantum safe fiber link	AES encrypted fiber link with security based on quantum key distribution (QKD) for high speed communications between edge nodes.	Quantum and service channels completed. Waiting for real time encryptor for the data channel.		TUE

3.2 List of WP2 embedded SW components

#	Component name	Description	Status	Licence	Partners
C2.17	P4 program	Software component for programming in P4 language the programmable switch	Key functionalities supported in 1.0. Advanced functionalities in progress (See Sect. 3.5.1)	Open source	CNIT, MLNX
C2.18	Switch operating system	Software component for operating the programmable switch	HW dependent. Completed on Spectrum1 ASIC standalone. In progress for Spectrum1 EMDC and Spectrum2 standalone (See Sect. 3.5.2)	Open source with proprietary component	MLNX, CNIT
C2.19	SW Agent for transceiver configuration and monitoring	Software component for the switch connectivity and monitoring through pluggable modules	Key functionalities supported in 1.0. Advanced functionalities in progress (see Sect. 3.5.3)	Open source	CNIT

C2.20	5G NR model	5G NR model including PHY layer high and PHY layer low.	gNB transmitter side (downlink) 90% ready. gNB receiver (5%). Integration with USRP RF – ongoing.	Proprietary	ISW
C2.21	vRAN adjustments prototypes	vRAN modifications of selected layers to support acceleration (e.g. PDCP) as well as simulation for AI/ML acceleration	PDCP layer encryption for FPGA has been evaluated to support the real-time stack. Now another round is planned to address split 2 and SmartNIC. AI/ML acceleration options for models life-cycle are being prepared	Proprietary /	ISW
C2.22	MBMC software	Software component for the nodes' management microcontroller	Partially node dependent. Common functions will be supported in BRAINE 1.0, node specific functions will be implemented later. (see Sect. 3.5.1)		BME
C2.23	BMC software, first release	First release of the AST2500 CPU based, openBMC derivate acting as system supervisor	First proof of concept on Congatec evaluation board. iKVM (Keyboard Video Mouse) demonstrated		PCB
C2.24	FPGA node embedded firmware	FPGA architecture and software stack for the FPGA node.	Development in progress using development boards, as BRAINE 1.0 does not have FPGA node. (see Sect. 3.5.3)		BME

C2.25	OpenCL program for Low-PHY	FPGA-based Hardware Acceleration for Terasic DE10-pro FPGA	Implemented up to 2048 IFFT-points	Proprietary	SSSA
C2.26	VTU – virtual transcoding unit	Video transcoding SW component	Two versions are available: one for managing a single stream and another for multi-stream environments. Demonstrated in the context of the smart city use case UC2) during the first-year review.	Open source	ITL
C2.27	N2Net	Software toolchain to implement binary neural networks in network devices' data plane	Development in progress, first lab results available	Proprietary	NEC
C2.28	Quantum-safe readiness	Studies of required modification of component to reach post-quantum status and standard	Standardization competition follow-up and reporting to partner. PQ conversion guideline (ETSI & NIST) used and reported to partner.		SIC
C2.29	Low-bit-rate blockchain protocols	Studies of required security protocols to enable IoT devices to participate in a blockchain without trusted intermediaries	Developed PLS and SLVP protocols, completed security analysis	Proprietary	IMC
C2.30	Transport Layer Security accelerator	Security SW component	Implemented autonomous offloads for AES based TLS in-line encryption	Open source	MLNX
C2.31	Smart sensors integration	DPS and PALS integration for wearable devices with edge functionality	Cloud (i.e. Arrowhead framework) integration progressing, edge	Partly proprietary, partly public	IFX

			integration in preparation		
C2.32	Distributed sound sensors	Distributed sound sensor for smart city use case	In progress, need to move to final HW patform	Propr.	MAI
C2.33	Quantum-Safe Cryptographic Solutions Evaluation	An evaluation of the possible quantum-safe communication encryption algorithms to determine the one best suited for BRAINE	In progress; several algorithms investigated and KPI's measured. Some KPI tests remaining	N/A	DELL

4. Hardware components BRAINE 1.0

This section focuses on the HW components on BRAINE 1.0, starting from the architecture, which inherently defines the hardware specifications. It should be noted that BRAINE holistically builds an EMDC, including the enclosure (which comprises the cooling system, ejection system, power system, among other). Hence, this section is very heterogeneous in its nature and in relation to the described technologies, ranging from mechanical and thermal engineering to electronic engineering.

4.1 BRAINE Edge Micro Data Center (EMDC) final architecture

This section describes the design choices for the hardware components of BRAINE 1.0. The architecture choice decided among the partners is shown in the next figure. This design was conducted over the first six months of the project and balanced three dimensions: technical feasibility, performance and executability during the project lifetime.

EMDC Architecture was defined together with the key partners of WP2. Several options were discussed, and key challenges were identified. After common conceptual, system level design the design the detailed designs were performed by each individual partners, with a closed loop synchronization among the partners.

Before final architecture decision several competitors were analyzed, including M2DC, IBM Dome project, and BAMBOO Systems and HP Moonshot. Major findings were the following:

- Many designs use SOM based approach → we also considered this option
- Performance is limited by cooling resources
  Thermal design is critical → We introduced novel cooling technology
- Heterogenous architecture is required -> X64, GPU, ARM, FPGA nodes were defined
- AI Acceleration / data /image processing is a key usage scenario
- Performance / cm3 is critical in Micro Data Center applications
- Network bandwidth increase is required -> Added 3.2 TB Mellanox switch

After competitor analyses, and market research, each partner contributed to the overall concept, and defined the state-of-the-art innovative architecture of the EMDC system:

Architecture

- Configurable, customizable heterogeneous architecture
- AI hardware acceleration (GPU, FPGA)
- Optimal architecture for multiple small workloads, with high bandwidth requirements.
- SOM Based approach (faster time to market)
- SOC based approach (Each node holds CPU too)
- Minimalistic, compact design
- No redundancy implemented currently, but architecture supports.

Mechanical & thermal concept:

- 3U standard rack size, depth varies based on card number.
- Individual replaceable / reconfigurable cards, compact design.
- Monoblock design, which acts as both cooling bridge and mechanical frame

- Max 150 W / card thermal characteristics
- Two phase colling / passive cooling possible on smaller designs
- Two-phase cooling enhancement using nanotechnology

Electrical design

- Each processing card is same sized, with identical pinout
- SOM based design where possible (COMe Type VII, Xavier AGX, etc)
- Dual layered switching: Both (10/25/ 100G) Ethernet and PCIe Gen 4 (fabric)
- High bandwidth / processing ratio
- 48 V power input support, dual 12/ 48V rails, highly efficient 48 / 12 V conversion.
- No hot swap currently, but architecture supports.

Low level software components

- Centralized Open source BMC (OpenBMC)
- Each card with similar BMMC (Board Management Micro Controller)
- Communication between BMC- BMMC over CDC- Ethernet tunnels (USB)
- TPM is used on all boards.

In the EMDC architecture, it can be observed that an Ethernet switch and a PCIe switch are interconnected to 8 nodes; this number of nodes allows a good mix between processing and storage blocks with a feasible execution plan within the BRAINE budget.

Initially, 24 nodes were also considered in order to add redundancy, but it was decided at early stage that such redundancy didn't help to demonstrate any particular EMDC capability yet it would increase the cost of the prototypes and complexity of the control system. The nodes are controlled through a management board (BMC), which is designed and implemented considering the full vertical stack (explained in this document under the software section). The Ethernet switch is connected through high-speed lanes to QSFP transceivers, enabling large capacity connectivity to other EMDC platforms or cloud computing platforms.

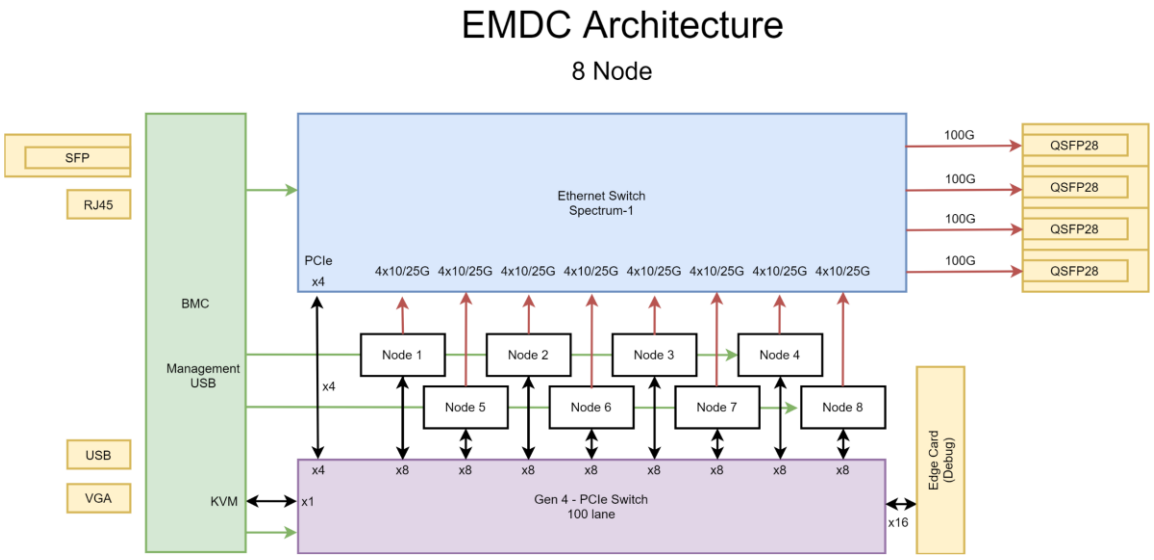


Figure 4.1: 8-node BRAINE EMDC Architecture

4.2 BRAINE EMDC Detailed technical requirements

Once the architecture was decided, an effort to transfer it to a set of specifications was conducted. Such requirements range from power supply needs to size of the enclosure. This section describes these requirements.

4.2.1 Requirements

The high-level requirements are listed in the following list (functional and non-functional requirements):

The overall requirements for the system:

- Maximum 150W (continuous) power per card
  - Primary power supply 48VDC
  - Secondary 12V DC for each card
- Can be integrated into a 19”rack (3U)
- Compute nodes must have passive cooling
- Two-phase cooling was selected
- All processing cards must have a BMMC (Board Manager) microcontroller and a TPM (Security) device.
- Hot swap is not a requirement
- Redundant switching / power not required at current level (project TRL 6-7)
- Target node size: 116 x 139 x 26 mm

The 8 slot backplane requirements are:

- 4x 100G QSFP28 external Ethernet interface
- 8 x Compute node slot
  - COMe CPU support
  - GPU card support
  - FPGA card support
  - ARM card support
  - NVMe card support
- 1 x PCIe switch card
- 1 x Ethernet switch card
- 1x Power supply
- 1x BMC
  - 1x RJ45 + 1x SFP connection
  - VGA interface
  - USB 2.0

COMe CPU card requirements:

- Must support ComExpress Type VII CPU modules
- Must support AMD (EPYC) and Intel (Xeon D) modules
- Must support at least 64 GByte memory , minimum 2 memory channels
- Must have on-board NVMe storage
- Must have at least 32 GB of system memory
- Must have minimum 2x10 G Ethernet interface
- Must have minimum x8 Gen3 PCIe connection

GPU card requirements:

- Must support Xavier AGX (32GB) AI SoC module

    - 512-Core Volta GPU Tensor Core-al
    - 8-Core ARM v8.2 64-Bit CPU
- Must have NVMe storage (M.2 2280)
- Must have at least 32 GB of system memory
- Must have at least an 8xGen3 PCIe host interface
- Must have a 2x 10 G (25 G) Ethernet interface

NVMe card

- Must have NVMe storage (M.2 2280)
- Minimum capacity: 8 TB / card
- Minimum 8x Gen 3 host interface

PCIe Switch card

- Must have a 96 lane Gen3 (4) PCIe interface
- Multi host support
- Non-Tranparent Bridge support
- SR-IOV support.
- Dynamic reconfiguration option support

Ethernet Switch card

- Internal connection: With 4 x 10G / 25G Interface
- External connection: minimum 4x QSFP28
- Software defined networking support
- P4 language support

Power supply card

- 48 V DC input voltage
- 12V DC output voltage
- Redundancy is not a requirement.
- Minimum 1800 W power @ 12 V output power. (continuous)

BMC card

- USB 2.0 interface for all cards in the system (8 pcs)
- I2C / IPMI interface for card identification
- KVM support (VGA, USB Device Emulation)
- remote access (API)
- remote monitoring (WEB)

4.2.2 Applicable standards

The final EMDC hardware must meet the following standards.

Directives

- 2014/35 / EU Directive - Safety of electrical equipment ("Low-VoltageDirective" - LVD)
- 2014/30 / EU Directive - Electromagnetic Compatibility (EMC)
- 2011/65 / EU Directive - Restriction on the use of Hazardous Substances in the EEA ("RoHS")

Standards

- Electrical Safety (to LVD-Directive) EN 62368-1: 2014
- Electromagnetic Compatibility EN 55022: 2010

- EN 55024: 2010
- EN 61000-3-2
- EN 61000-3-3
- ROHS EN 50581: 2012

4.2.3 Physical size / Operation range

- Operating Temperature range up to 10..30 degrees Celsius
- Cooling air temperature Max 50 C inlet, Max 60 C degree exit (estimated)
- Storage temperature 0 .. 60 degrees filled with coolant, -30… 60 without coolant
- Humidity 10… 90% Relative humidity / non-condensing
- Noise Does not contain moving parts
- Weight under 20 kg / 8 slot system
- Weight under 50 kg / 24 slot systems
- Input voltage range 40-58 V, DC
- Current consumption: Less than 50 A.
- Maximum power for a 2kW / 8 slot system with full deployment

4.2.4 Detailed Introduction of electrical system components

EMDC is a modular, heterogeneous, freely configurable, liquid-cooled server. The main interchangeable computing elements are:

- CPU unit: Com Express AMD (Epyc) or Intel (Xeon -D) CPU
- GPU Unit: NVIDIA Xavier AGX Soc
- NVMe unit: 4x M.2 SSD
- ARM CPU unit: still under definition
- FPGA unit: FPGA containing Xilinx Versal AI. Still under definition.

The computing elements are connected by the following components:

- PCIe switch card: 128 lane Gen3 PCIe switch
- Ethernet switch card: Ethernet switch with 3.2 Tb / s aggregate bandwidth.
- 8 slot backplane: 8 motherboards connecting heterogeneous computing elements + 1x PCIe switch and 1 x Ethernet switch +

Other additional hardware elements:

- BMC (Board Management Controller) that oversees the system.
- Power supply circuit that performs 48V -12V conversion

The following document describes the architecture and preliminary 3D layout of the main PCBs

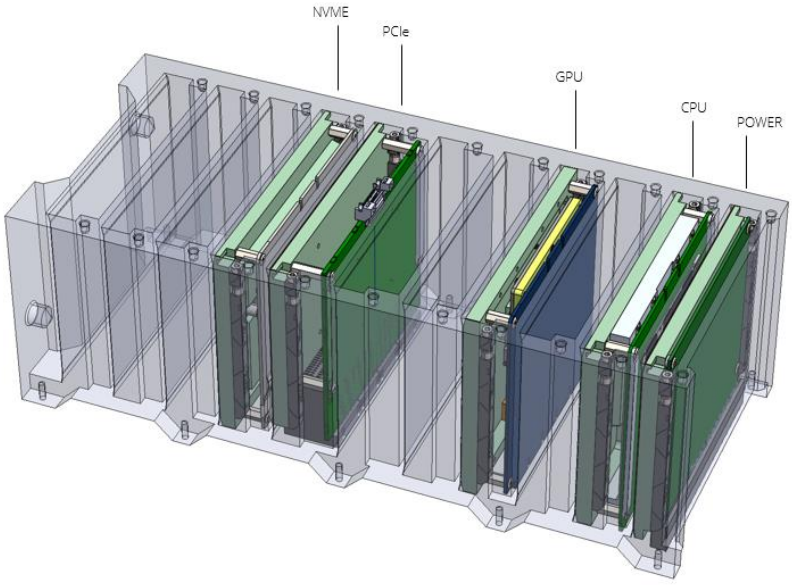


Figure 4.2: BRAINE EMDC 3D Layout

4.2.5 CPU Board design

Preliminary market research (e.g. M2DC, or Bamboo) has shown that most companies which are developing similar edge computing platforms use standard-sized CPU modules instead of developing their own module. This approach is very sensitive as there is a large pool of CPU processors available and hence designing an ad-hoc solution is not sensitive. This approach has many advantages, such as:

- Can be built with ready-made modules (COTS)
- Standardized connectors
- Intel, AMD, ARM processors and FPGA cards are also available
- Faster time to market, less development time
- Modules with extended operating temperature are also available (-40 ° C..85 ° C)

4.2.5.1 COMe TYPE VII

The figure below shows an example of a Congatec CPU module. The standard defines various sizes. Basic (125x95mm) form factor was selected in this project. It can be clearly seen that the PCB surface is used heavily mainly by: the processor, DDR4 memory and power supplies. This approach ensures that an optimized solution is found for both consumption and space utilization.

Figure 4.3: Congatec - COM Express VII CPU Module
AMD EPYC 3451 CPU- 16 Core, 96 GB RAM
Size: 125x95 mm

COM Express (COMe) based cards are widespread and standardized in the market (by PCI Industrial Computer Manufacturers Group - PICMG).

The COMe family has different types. The project uses a Type VII connection module:

The main interfaces of the TYPE VII COMe module.

- Up to 32x PCIe interface (Gen3)
- Up to 4x 10 GBase KR interface
- 4x USB 3.0, 4x USB 2.0
- 2x SATA, Gigabit Ethernet
- Other interface: LPC / eSPI

Type VII form factor has been optimized specifically for high performance embedded applications, focusing on high-speed data interfaces (10G Ethernet, Up to 32 PCIe lanes) but lacking e.g. display interface.

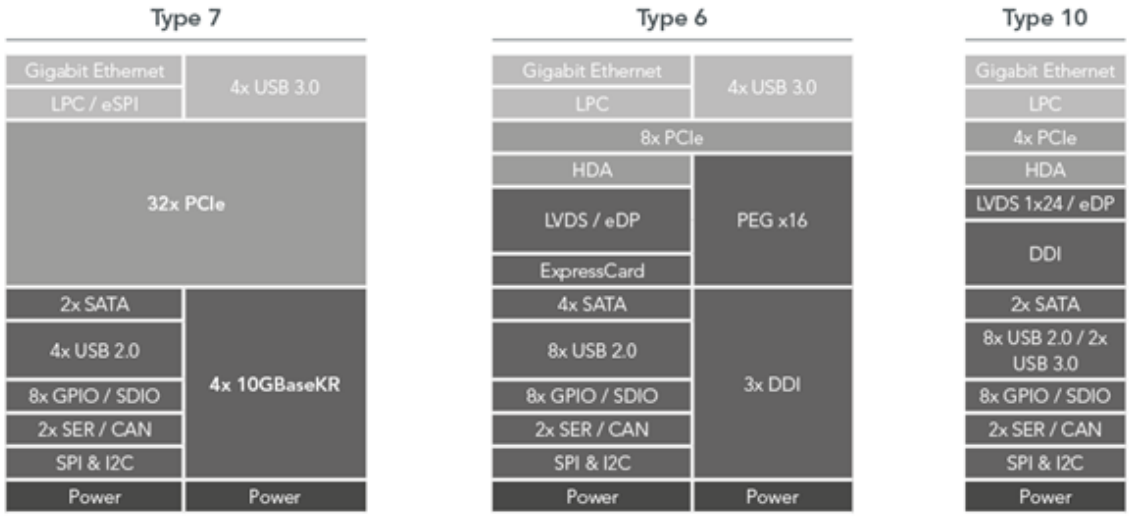


Figure 4.4: COMe module connector types

Based on market research, considering the smallest size and maximum computing capacity (24 PCIe ports, 3 DDR memories, 2x10G KR interface, we found the following manufacturers:

Manufacturer	Card Name	CPU	KR Ports	PCIe	Storage	RAM
ADLINK	Express-BD7	Intel Xeon / Pentium	2x	24x Gen3 (x16+x8)	None	2x SO-DIMM
ADLINK	Express-DN7	Intel Atom	4x	16x Gen3 (2x x8)	Option	3x SO-DIMM
Adlink	Express-BD74	Intel Xeon / Pentium	2x	24x Gen3 (x16+x8)	None	4x SO-DIMM
Advantech	SOM-5962	Intel Atom	4x	12x Gen3 (any conf)	Solder	4x SO-DIMM
Advantech	SOM-5992	Intel Xeon / Pentium	2x	24x Gen3 (any conf)	None	4x SO-DIMM
Advantech	SOM-9590	Intel Xeon	2x	24x Gen3 (x16+x8)	Solder	Soldered-on
Axiomtek	CEM700	Intel Xeon / Pentium	2x	24x Gen3 (x16+x8)	None	3x SO-DIMM
Congatec	conga-B7E3	AMD EPYC	4x	32x Gen3	Option	3x SO-DIMM
Congatec	conga-B7AC	Intel Atom	4x (2x)	12x Gen3	Option	3x SO-DIMM
Congatec	conga-B7XD	Intel Xeon / Pentium	2x	24x Gen3	None	3x SO-DIMM
Eurotech	CPU-162-23	Intel Xeon / Pentium	2x	24x Gen3 (x16+x8)	None	3x SO-DIMM
iEi	ICE-BDE-T7	Intel Xeon	2x	24x Gen3 (x16+x8)	None	2x SO-DIMM
iEi	DV970	Intel Atom	2x	16x Gen3 (weird confs)	None	2x SO-DIMM
Kontron	COMe-bBD7	Intel Xeon / Pentium	2x	24x Gen3 (any conf)	Option	2x SO-DIMM
Kontron	COMe-bDV7	Intel Atom	4x	14x Gen3	Option	2x SO-DIMM
MSC	MSC C7B-DVL	Intel Atom	4x	14x Gen3	Option	3x SO-DIMM
Seco	COMe-C42-BT7	AMD EPYC	4x	24x Gen3	None	4x SO-DIMM
X-ES	XPedite7650	Intel Xeon	2x	24x Gen3 (x16+x8)	None	Soldered-on

For the development, we chose the product of the Congatec company, based on the best price / value. (Both AMD and Intel based cards are available with the same connector and PCIe lane layout)

4.2.5.2 COMe CARRIER PCB

In order to interface the COMe Type VII CPU module, a carrier PCB was developed to interface COMe standard signals to “BRAINE backplane connector”

The design of the main system interface (BRAINE connector) was highly matched to Come Type VII module as shown in the figure below.

The following table summarizes the CPU module interfaces and their terminals.

COMe Type VII Interface	Use on COMe Carrier PCB
2 (4) x 10 G bit Etherent (KR)	Terminal to BRAINE connector
Max 32x PCIe Lane (Gen x3)	x8 lane BRAINE connector x8 lane BRAINE connector x1 lane on board i210 1 G ETH PHY x4 lane on board BGA SSD (OS) x4 lane on board M.2 SSD (Storage)
1x Base-T Ethernet	Debug connector
SATA	Debug connector
USB3	1 x Debug connector
USB3	1 x BRAINE connector
LPC / eSPI bus	1 x BRAINE connector
UART	STM32 BMMC
SPI (BIOS flash)	On board flash with BMMC access

4.2.5.3 High level block diagram

Following the above considerations, a high-level block diagram of the card was developed.

The figure below shows the block diagram and the preview 3D image of the card.

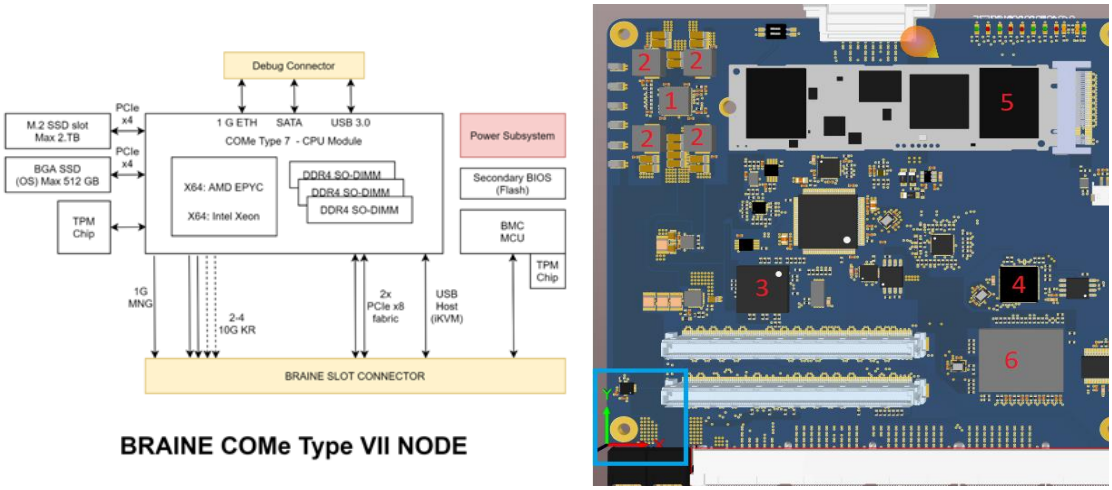


Figure 4.5: Block diagram and preview of the COM3e card

4.2.5.4 PCIe Lane allocation

Conga B7xD internal structure and PCIe LANE allocation.

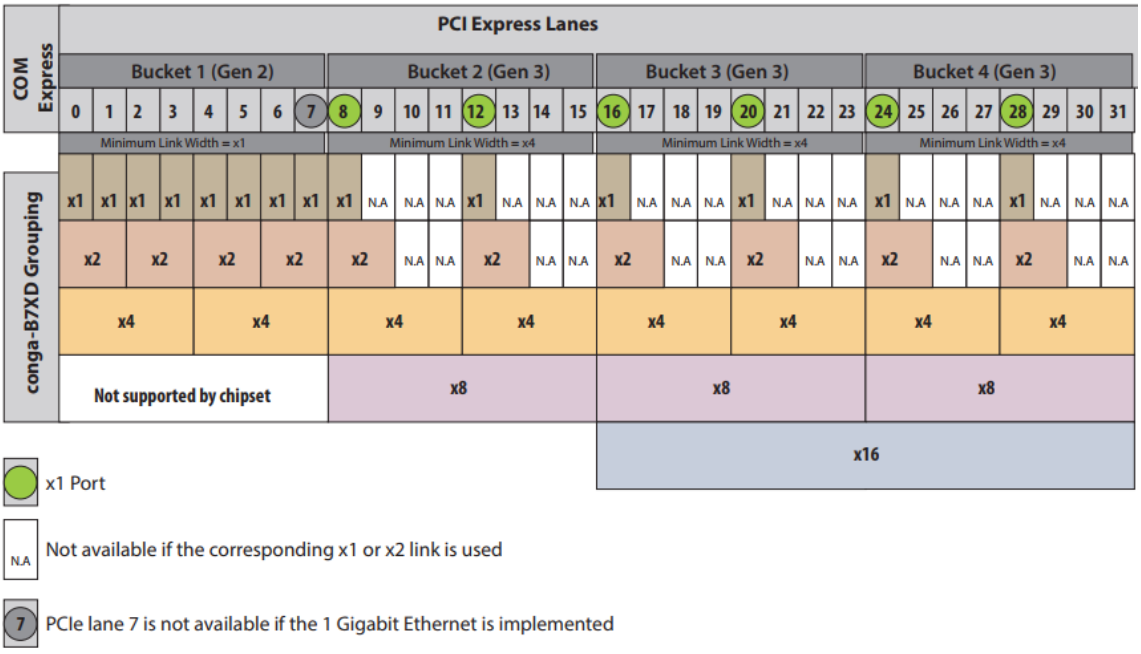


Figure 4.6: Conga B7xD internal structure and PCIe LANE allocation

From the possible room future PCIe lane configurations, we selected the ones that give the most flexibility, so the final configuration was formed.

4.2.5.5 Ethernet Interface

Their COMe modules typically have a 2-4 10 G BASE KR interface. To make it compatible with external SFP + modules, a 10 G retimer unit is placed on the card, which converts

KR signals into an SFI interface. It is designed only on the first two lanes, the other two interfaces are designed in 10 G BASE KR, while many CPU modules support only 2x10 G option.

The SFP interface often requires optical module identification via an I2C bus, and a “fake” EEPROM is optionally placed on the card to solve this.

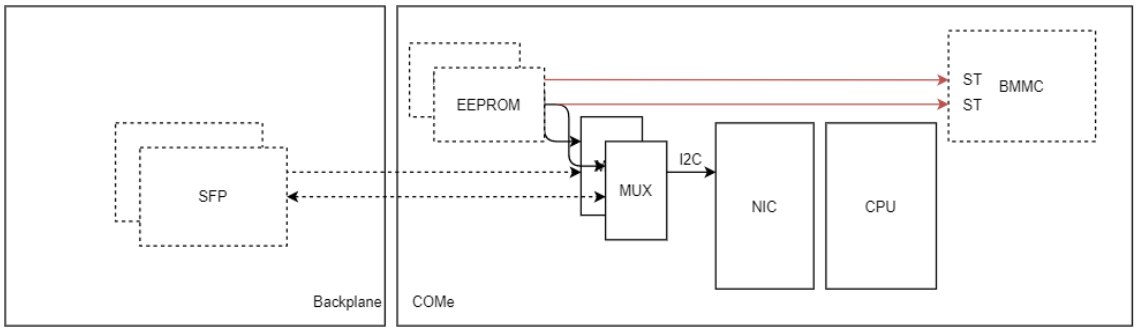


Figure 4.7: Ethernet component

4.2.5.6 Thermal Simulations

Helder ID and JJ Cooling consortium partners were optimizing thermal structure and performed thermal simulations to design proper cooling. Based on these, an “inner cooler” unit was designed between the COMe module and the substrate to dissipate the heat generated on the card in the direction of the monobloc. It is clear from the simulation that the system performs excellently even at a flow temperature of 50 degrees.

More information in section 3.2.

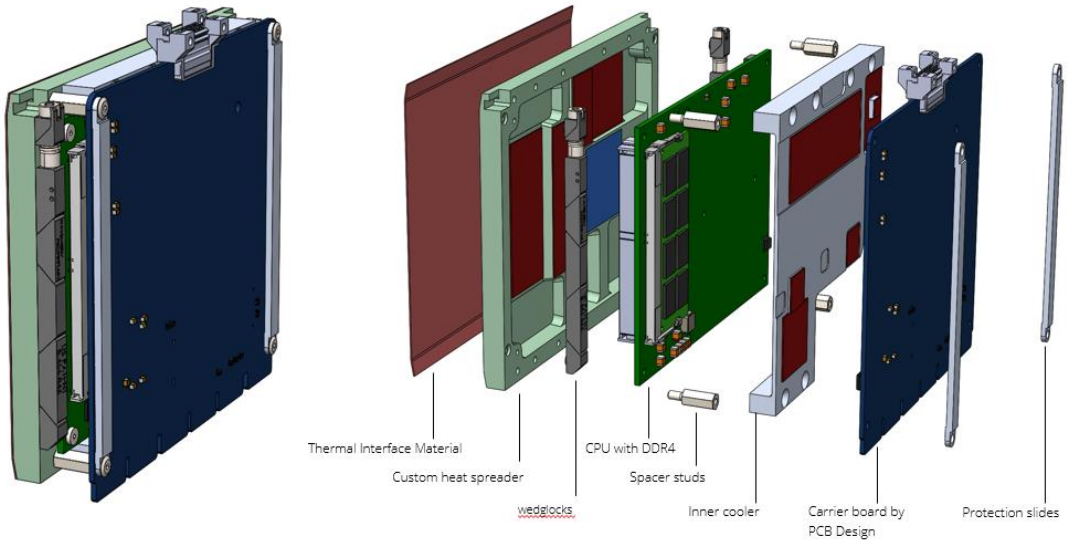


Figure 4.8: “Innecooler” unit

4.2.5.7 CPU card prototype

The CPU card for BRAINE has been designed and fabrication, and it is shown in the following figure:

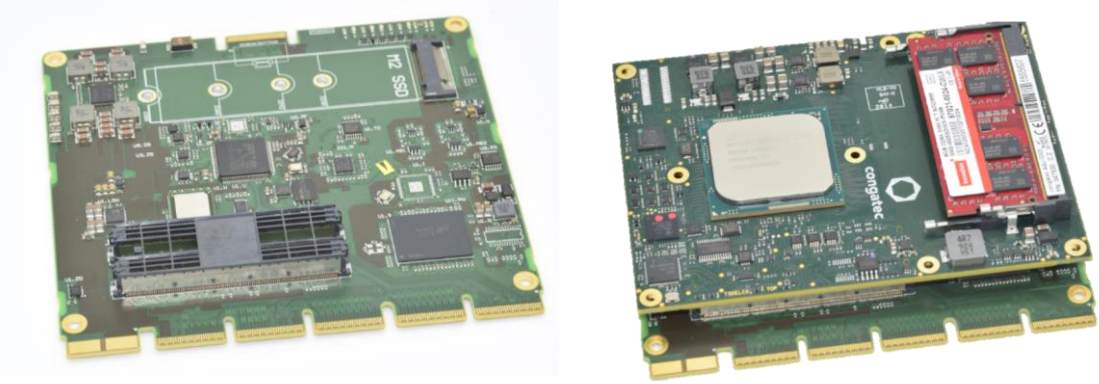


Figure 4.9: Actual photo: Prototype CPU card, with and without Congatec module (heatsinks removed)

4.2.6 GPU board design

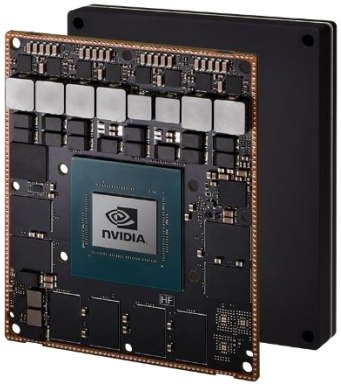
The core of the GPU node is the NVIDIA Xavier AGX module.

Like the COM Express module, this card has a modular design. Xavier AGX card was being used and a carrier card has been designed for it.

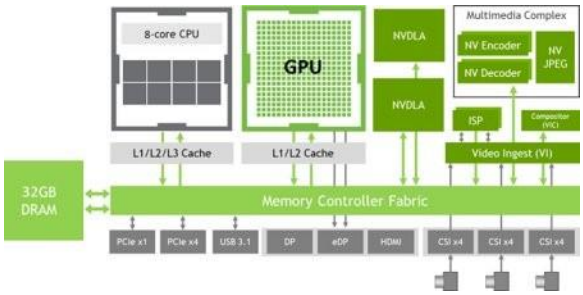
The main parameters of the GPU module:

- GPU 512-core Volta GPU with Tensor Cores
- CPU 8-core ARM v8.2 64-bit CPU, 8MB L2 + 4MB L3
- Memory 32GB 256-Bit LPDDR4x | 137GB / s
- Storage 32GB eMMC 5.1
- DL Accelerator (2x) NVDLA Engines
- Vision Accelerator 7-way VLIW Vision Processor
- Encoder / Decoder (2x) 4Kp60 | HEVC / (2x) 4Kp60 | 12-Bit Support
- Size 105 mm x 105 mm x 65 mm
- Deployment Module (Jetson AGX Xavier)

The main advantage of the Xavier AGX is the large number of Tensor Core, which is specifically designed to help speed up AI applications.



GPU Node

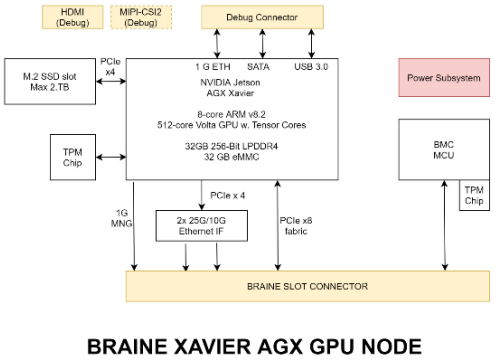


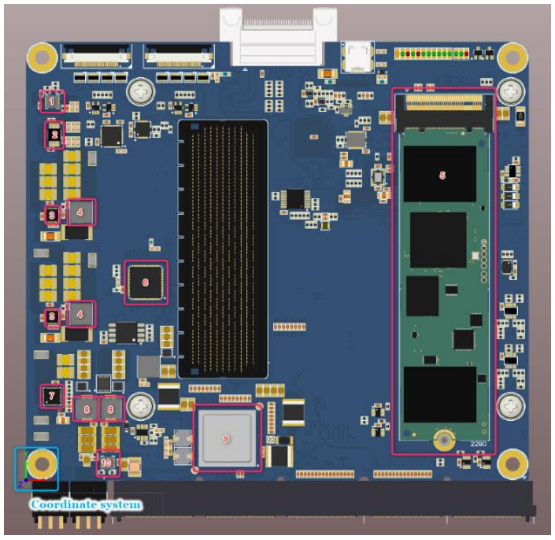
NVIDIA Xavier AGX

Figure 4.10: GPU node

4.2.6.1 GPU Carrier PCB

Since the Xavier AGX module has only one 1G BASE-T Ethernet controller, it is connected to the debug connector. To allow massive data transfer and flowless integration into the EMDC system, it became necessary to build a high-speed Ethernet connection, so the card also had a 2 x 10/25 G Ethernet NIC controller.





GPU Node NVIDIA Xavier AGX

Figure 4.11: GPU Carrier PCB

When allocating PCI e Lanes, an x 8 Gen4 compatible interface was added to the EMDC connector.

An M.2 SSD was also placed on the card.

The mechanical concept design of the GPU is very similar to the COMe module.

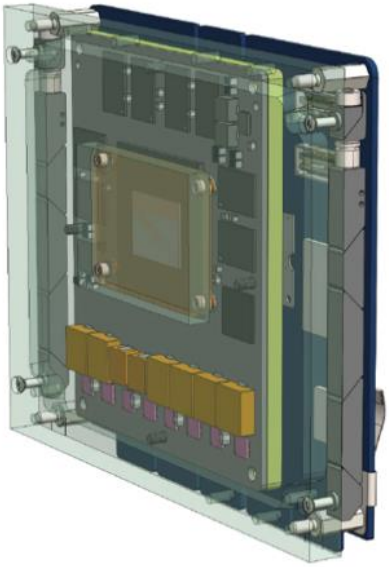


Figure 4.12: Layout

4.2.7 NVME board design

Because COMe cards have only a limited amount of integrated storage, dedicated storage card was defined.

A 4x M2 2280 storage SSD module is located on the module, each of which is available on the rear dual x8 PCIe bus. In this case, the PCIe configuration is 4 x4 lane.

As long as the CPU / GPU / FPGA / ARM module acts as a root complex, the NVME module acts as a PCIe device. It is important to identify the individual cards at system startup and reconfigure the central PCIe switch accordingly.

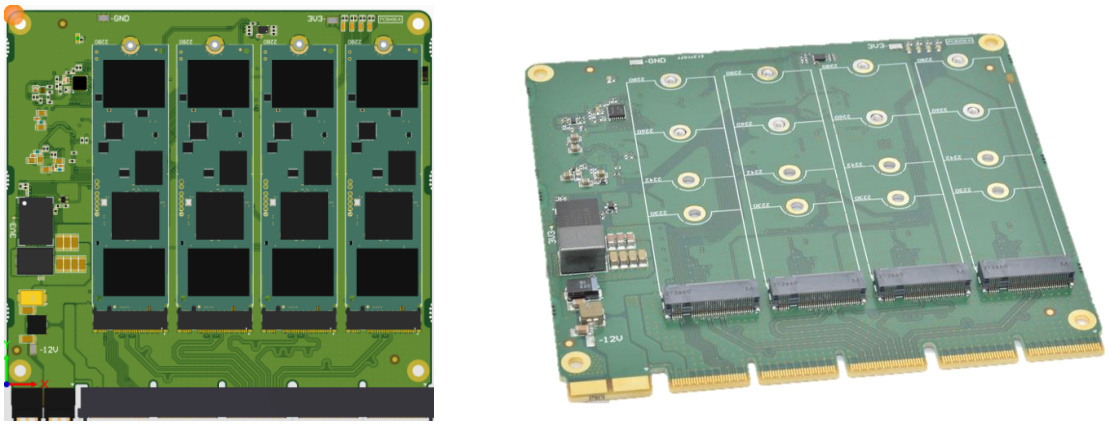


Figure 4.13: Rendered image / actual photo of the NVMe Module

4.2.8 PCIe Switch

The function of the PCIe switch is to connect the central nodes.

In a traditional desktop computer, the function of a PCIe switch is to connect a single CPU module to variety of peripherals.

In the EMDC system, the PCIe has a **fabic design**, allowing the following functions:

- Capable of supporting multiple PCIe initiator units (eg CPUs)
- Ability to communicate between multiple PCIe initiators (Non transparent bridge)
- Ability to dynamically allocate target devices (eg NVMe drive) between multiple PCIe initiators
- Supports IO virtualization SR_IOV.

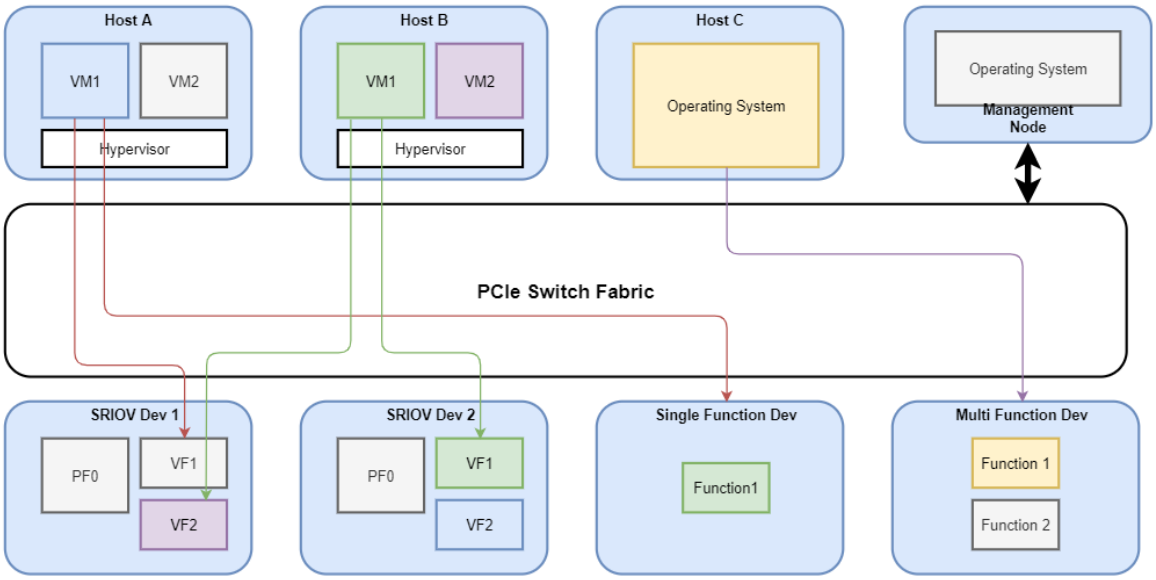


Figure 4.14: switch

The selected PCIe switch is made by Microchip / Microsemi. (PM40100)

Main specification:

- 100 PCIe Gen 4 lane
- Max 52 ports,
- Max 48 NTBs,
- 26 Virtual Partitions
- Flexible port binding: x2, x4, x8, and x16

4.2.8.1 Mechanical /Thermal design

The design of the module is slightly different from the standard CPU / GPU module.

Since the 100 PCIe lane was routed directly, it became necessary to use a high-speed connector here. The selected backplane connector is the SAMTEC Examax family, which is considered almost an industry standard and can be used up to 56/112 Gbit / s (PAM4). For PCIe Gen 4, the line speed is 16 Gbit / s (NRZ).

The physical floor area of the card is the same as that of the CPU cards, but thanks to the 4 Examax connectors, its thickness had to be increased by a few millimeters.

The BMMC microcontroller is designed in the same way on the card.

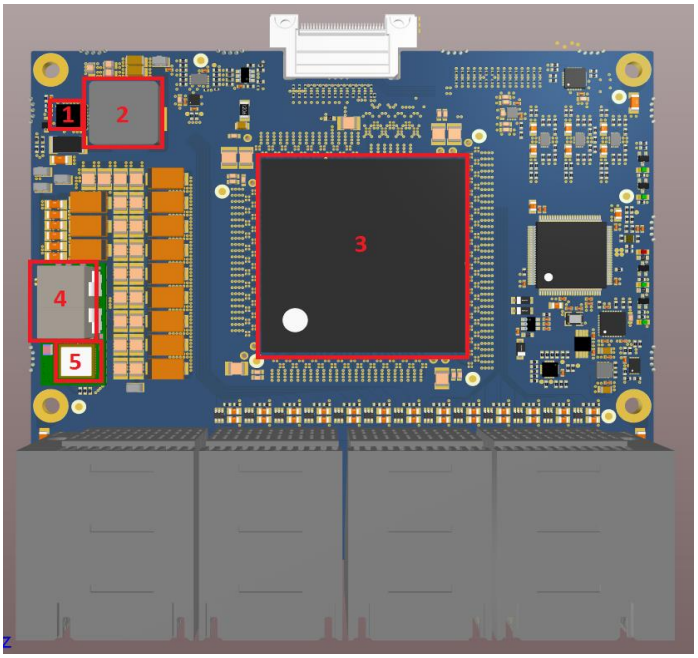


Figure 4.15: layout

Since all lanes have been routed out, the use of the card is universal, e.g. in the 24-slot system, it provides communication between 8 groups of cards, but a 4 card can also be inserted into the system with the hierarchical fabric swich function.

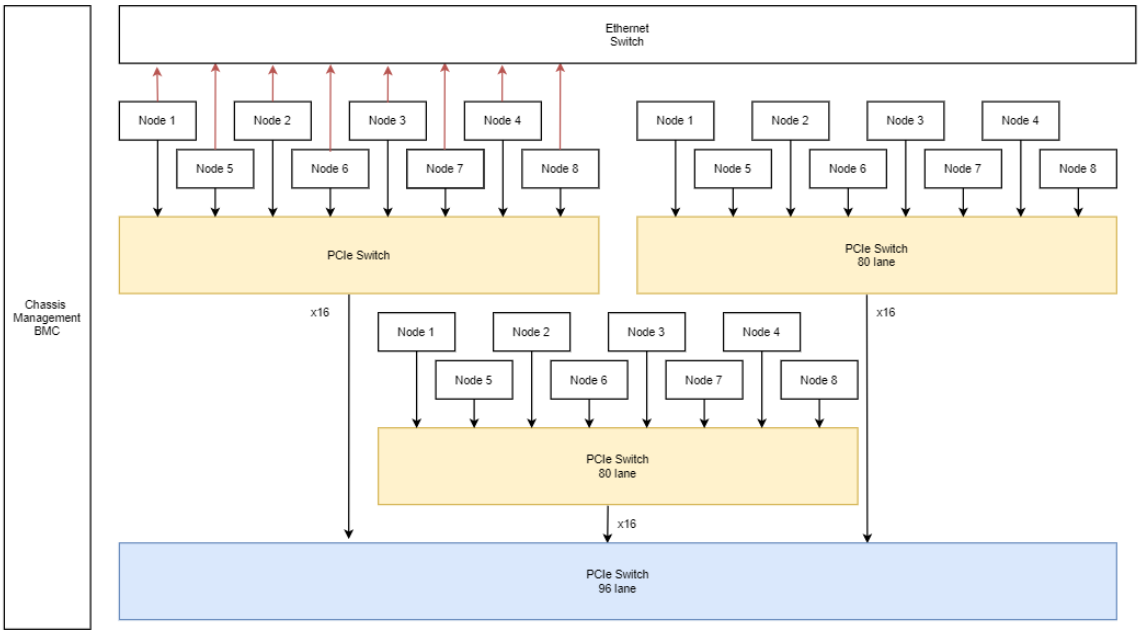


Figure 4.16: 24-slot system

More detailed block diagram of the card:

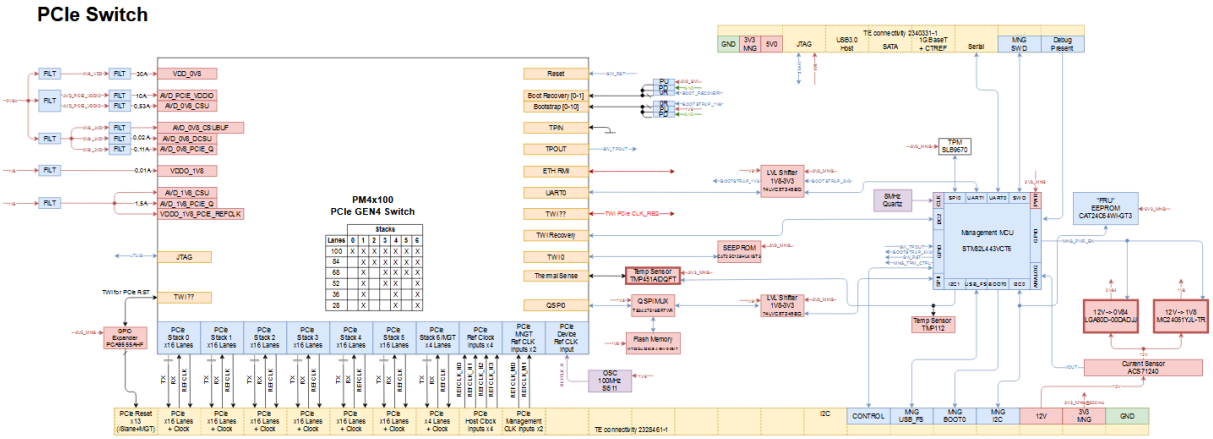


Figure 4.17: detailed block diagram

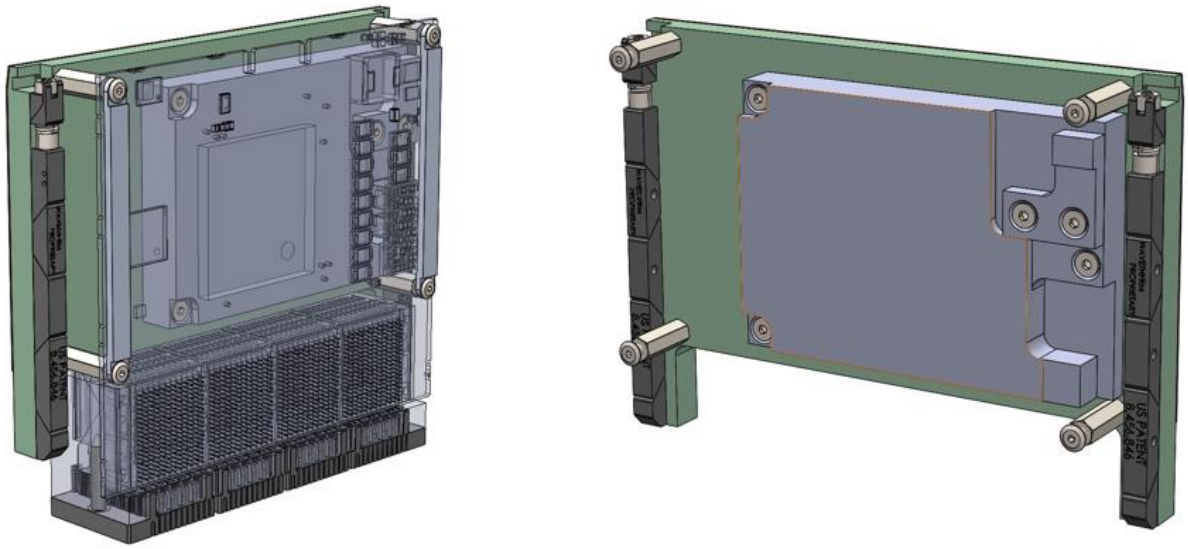


Figure 4.18: Mechanical construction

4.2.9 Ethernet switch

The function of the Etherent switch is to connect the cards to the outside world, thus providing data to the server. The Mellanox / NVIDIA Spectrum family has been selected as the central switch as the Mellanox consortium partner.

Main features of the Spectrum switch.

- Max capacity: 3.2 Tbit / s

- 128 x 1/10/25/28 G PHY
- Possibility of 32 x 40/56 / 100GbE connection
- Software Defined Networking function

The design of the switch card is also universal, which means that all 128 lanes are routed to 4 Examax connectors.

Intended use of lanes in the 24-slot system:

- Each 24 node has 4 lane (Max 100 Gbit / s) - 96 lane
- 4 x 100 G QSFP28 modules to the outside world (4 lane / QSFP28) - 32 lane

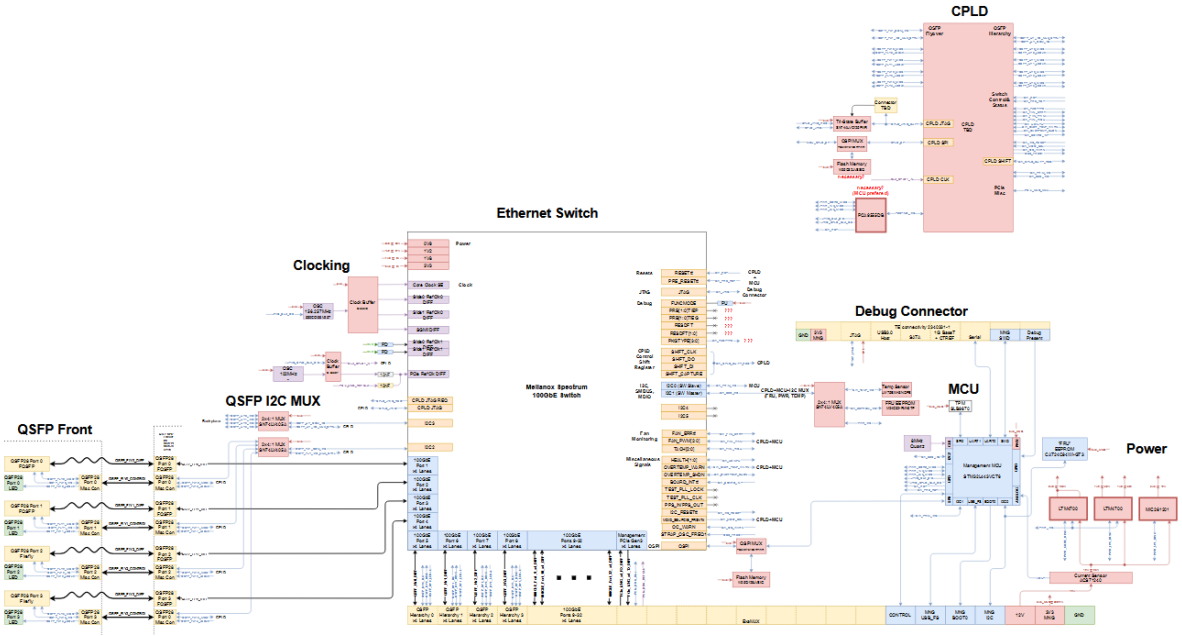


Figure 4.19: Block diagram of the Ethernet switch

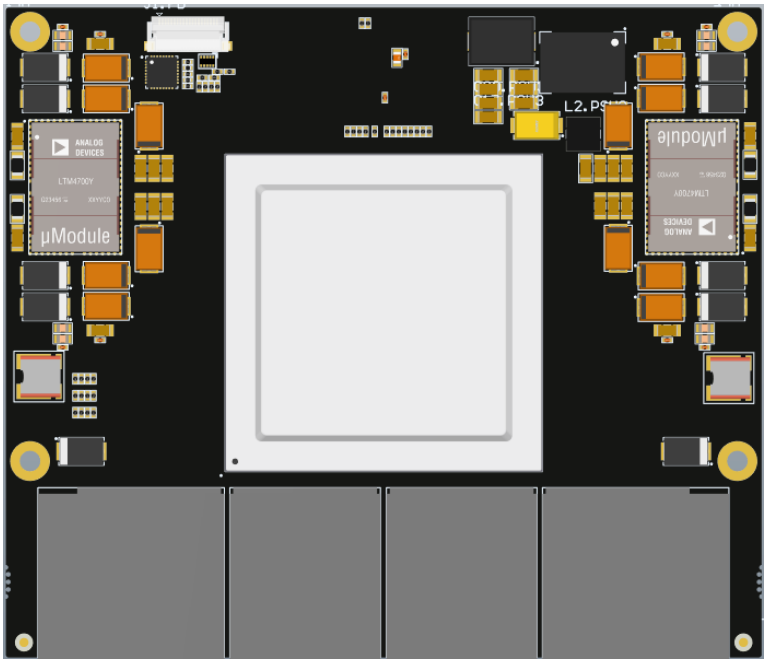


Figure 4.20 Rendered image of the Ethernet switch

4.2.10 8 SLOT backplane

After a long consultation with the members of the consortium, it was agreed that the 3 sample systems would be completed in an 8-slot configuration. It is also necessary to design a backplane card to connect the main modules.

The high-level diagram of the card is shown in the figure below:

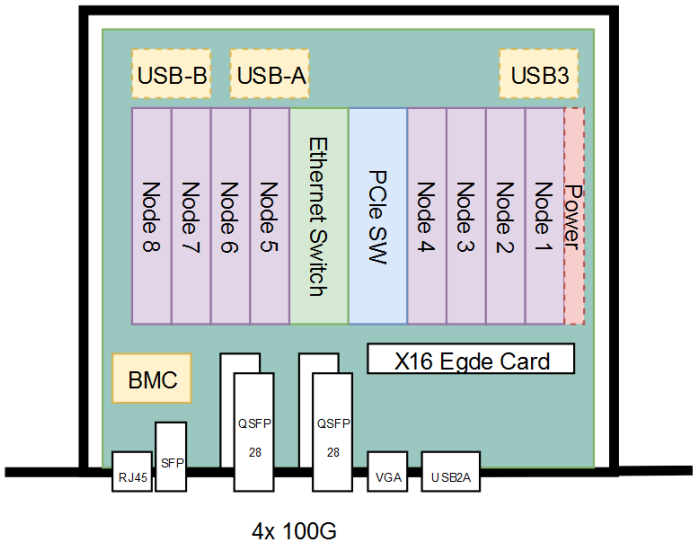


Figure 4.21 high-level diagram

Key features:

- Accomadate 8 CPU / ARM / GPU / FPGA / NVMe cards
- 1 PCIe switch
  - All cards with x8 Gen 4 connection
- 1 Ethernet switch
  - All cards with x4 10/25 connections + 1x 1 G connection (Management)
  - 4x front panel QSFP28 - 100 Gbit / s Ethernet interface
  - Debug PCIe x16 edge connector for development purposes.
- 1 BMC card
  - 1 x RJ45 and 1 x SFP management Ethernet connection
  - 1x VGA and USB 2.0 to implement the KVM (Keybard Video Mouse) function
- 1 power card (48V-12V) for conversion
- Debug connections to the back panel
- Additional standard x16 Pcie slot for further extension.

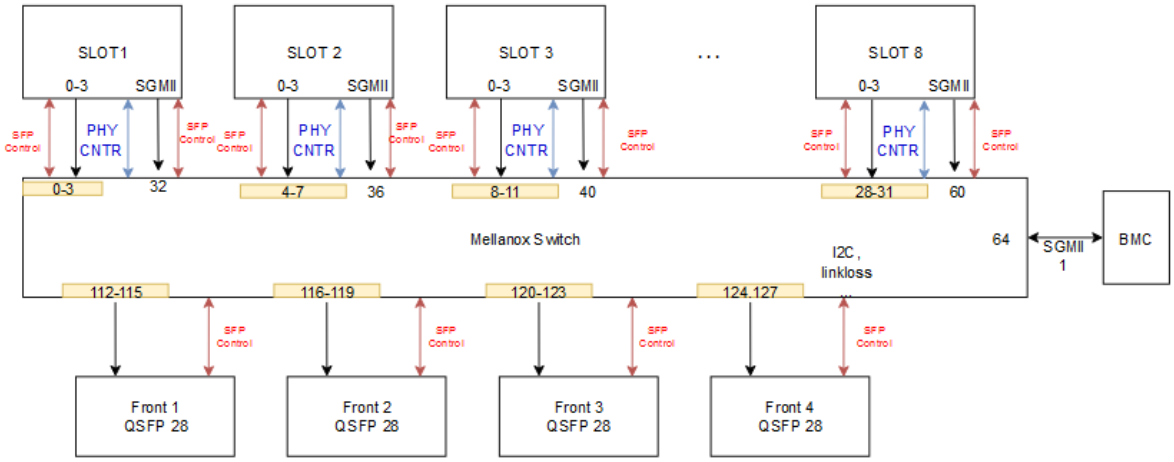


Figure 4.22: Wiring of Ethernet lanes

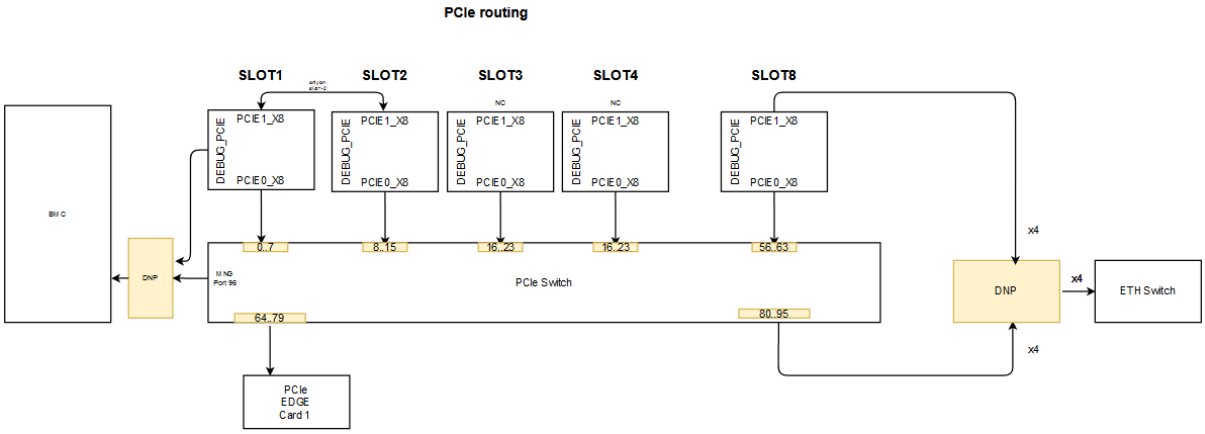
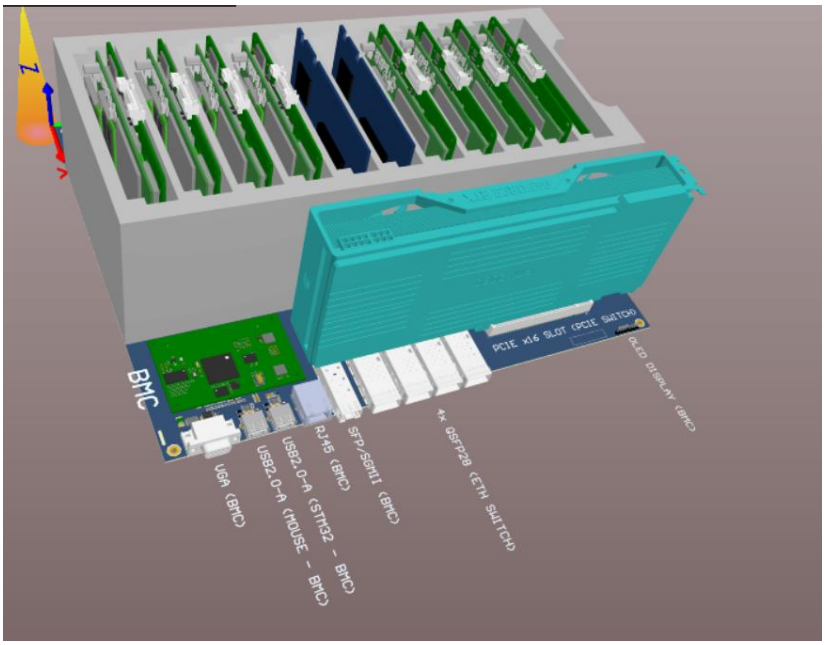


Figure 4.23: Wiring of PCIe Lanes



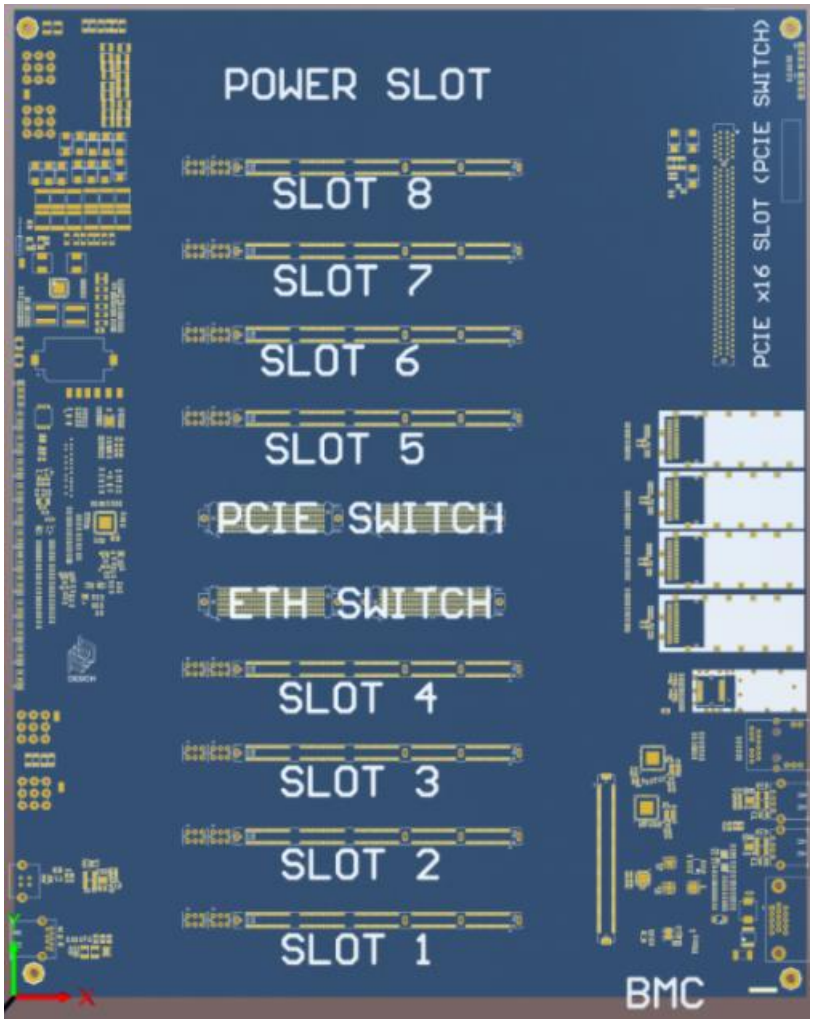


Figure 4.24: Location of the cards in Altium

4.2.11 BMC module

Each card has a so-called BMMC (Board Management Micro Controller) is a microcontroller-based unit that connects to a central Linux BMC card via USB.
The BMMC is responsible for identifying the cards, enabling power supplies, measuring consumption data, setting firmware upgrades / parameters, and supporting iKVM (Keyboard Video Mouse).

The task of the BMC is:

- Card identification and management over I2C.
  - FRU EEPROM
  - Enabling feeds
  - Reset
  - Firmware upgrade on BMMC units.
- Communication with BMMCs via USB after booting
  - The BMC is the USB host and each card is a USB Device
  - Implementation with USB HUBs
- KVM support (Keybard Video Mouse)

  - PCIe target within the BMC, emly can be assigned to each node via the PCIe network. Thus, the root complex there sees BMC as a video tool.
  - USB device emulation within the BMC, i.e. virtual mouse and keyboard.
    - The BMC is the USB device, the cards are USB Hosts
    - Implementation using high-speed analog MUX.
- ESPI / LPC support (for testing)

The main functions of the BMC card is summarized in the diagram below

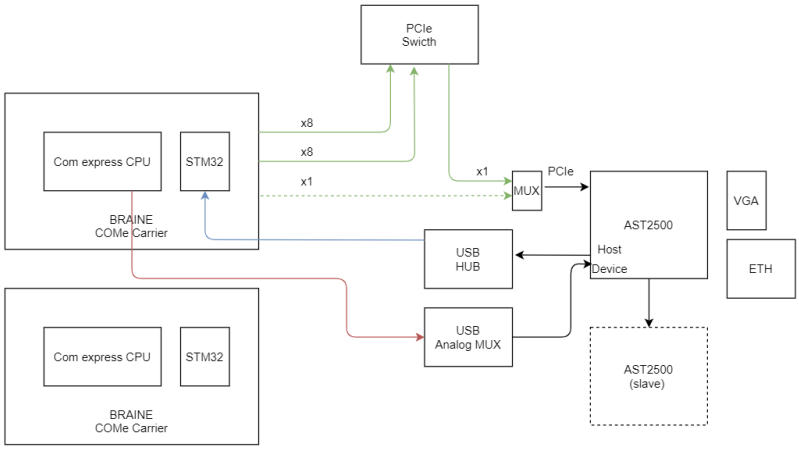


Figure 4.25: BMC card block diagram

8 Slot backplane BMC functions are listed in the follwing drawing.

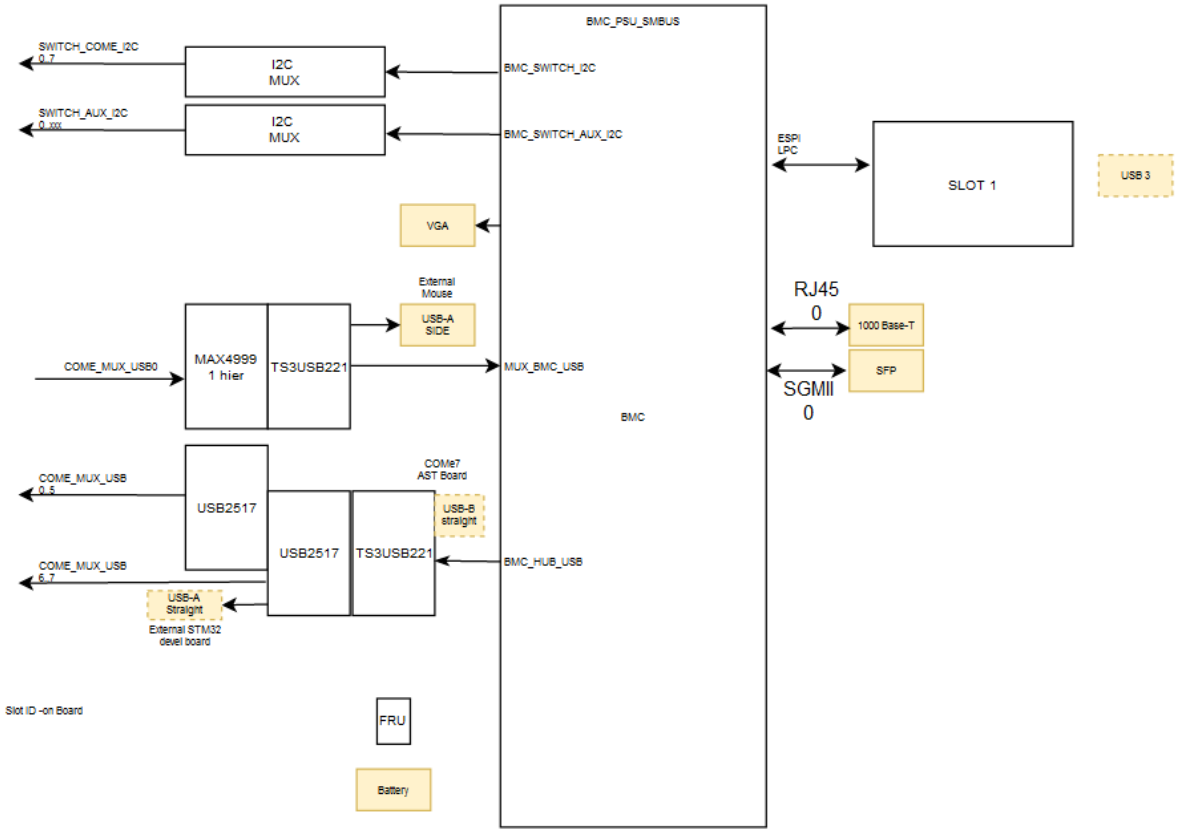


Figure 4.26: 8 slot backplane BMC connection

As the BMC microcontroller, we chose the AST2500 (industry standard) microcontroller, which is an ideal choice for both hardware and software (OpenBMC)

Main features of the AST2500 MCU

Embedded CPU	800MHz ARM11
SDRAM Memory	800Mbps DDR3/1600Mbps DDR4 SDRAM
Flash Memory	SPI flash memory
Video-Over-IP	Video Redirection up to 1920x1200
USB-Over-IP	USB 2.0 virtual hub controller with 5 devices supported USB 1.1 HID device controller
VGA	PCIe VGA/2D Controller / 1920x1200@60Hz 32bpp
LAN	Dual 10/100/1000M bps MAC

In addition to the AT2500 microcontroller, the card contains the following peripherals:
• DDR3 memory
• 2 x QSFPI flash (Linux + VGA BIOS)
• Power supply
• TPM chip
• 2x 1 G Base-T / SFP PHY.

Figure 4.27: BMC Module picture

4.2.12 Power Supply PCB

The following high level requirement must be considered when powering the entire EMDC system. Each card has a maximum allowable (operational) heat output of 150 watts.

The total current consumption of the 24-slot system is around 4.5-5 kW.

Conventional 12 V systems at this load would mean a current of about 400 A. Neither traditional power supply infrastructures (eg CRPS power supply) nor other infrastructures are prepared for this. Looking at market trends, we can conclude that 48V rated power is often used for higher performance servers. Thus, the current and copper loss can be reduced by a quarter. The 48V system is also well-suited to HVDC (High Voltgae DC) power supply systems, the main advantage of which is that the UPS solution does not require 230V AC regeneration, so losses can be further reduced.

Traditional system:

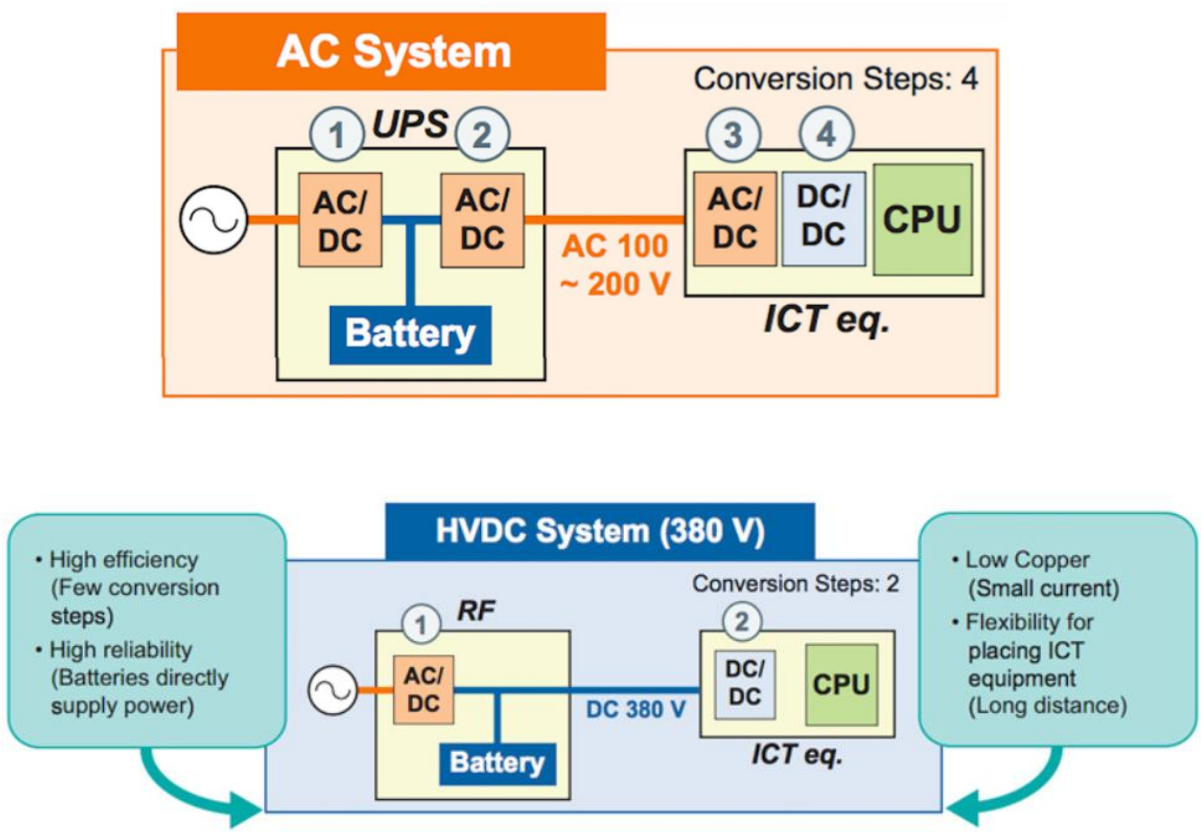


Figure 4.28: HVDC system

As a conclusion the system is designed for a 48V supply.

In many cases (eg COMe CPU card) it is advisable to continue to use 12V on the cards, so it is necessary to design a high current / high efficiency 48V-12V DC / DC power supply card. For the performance of the card to be designed, we calculated the following data:

- 8 CPU cards max 150 W / card
- 1 PCIe / 2 Ethernet switch 150 W / card
- BMC etc: 300 W

A total of 1800 W / 8 slot system.

The following Vicor unit was selected as the DC / DC power supply: DCM3717S60E14G5TN0

Main parameters:

- 40 - 60VDC input voltage range
- 10.0 - 13.5VDC output voltage
- 97.0% peak efficiency
- Max 750W / 62.5A continuous operation
- Max 900W / 75A transient
- > 1MHz switch frequency
- PMBus telemetry
- Up to 4 units can be paralleled

It can be clearly seen that for the production of 1.8 kW energy min. 3 modules need to be parallelized. During the design, we parallelized 4 units to make the units operate at a more optimal working point.

The card was designed to use a backplane connector, pinout was defined to accommodate future redundant power supply.

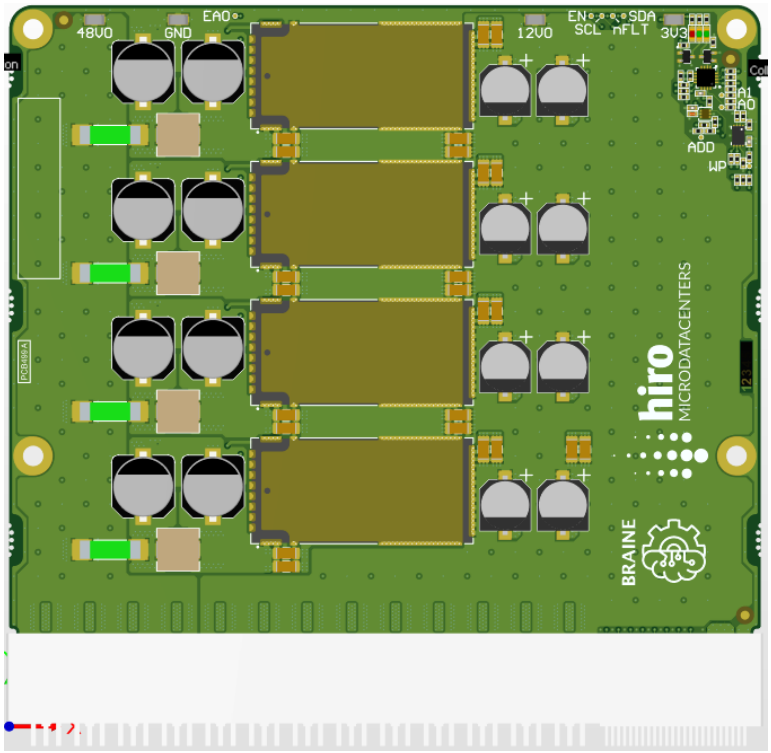


Figure 4.29: Rendered image of the Power supply Board.

The following figure shows a picture of the fabricated power supply board to be used in BRAINE EMDC.

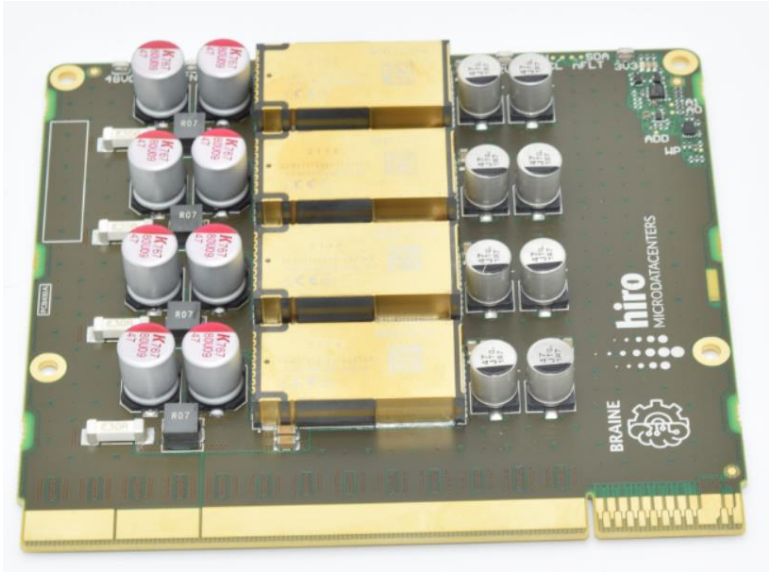


Figure 4.30: Actual photo of the Power Supply Board

4.3 Cooling system design

The cooling system in BRAINE is key to sustain the high level of processing density aimed to support intense AI processes with high connectivity. This section describes the burden on the cooling system and the innovations related to this area.

4.3.1 Requirements

The hardware team worked closely to define and report the thermal requirements. As detailed in the previous electrical section, the whole system is modular with 8 compute node slots, one ethernet switch slot, one PCIe switch slot and a power slot. The maximum heat dissipation per slot is set to 150W with a total for the 8-slot system of 1.5kW. All the CPU compute modules are based on the standard COMe Type 7 boards mounted on custom carrier boards. The cooling system designed for BRAINE must be adapted around the defined node dimension of 116 x 139 x 26mm and fit inside a 3U 19” rack enclosure.

Important requirements for the cooling part are recalled here:

- Heat dissipation of 1.5kW for the 8-slot system
- Maximum of 150W per slot
- Two-phase cooling with non-uniform heat in between the module
- Passive cooling at the node level (no moving part in the 3U enclosure)
- Modular (the 8 compute modules can be interchangeable in between the slots)
- Node size target of 116 x 139 x 26mm
- Hot swap is not a requirement
- Maximum inlet air temperature of 50°C (max. outlet air temperature 60°C)
- Relative humidity ranges from 10 to 90%
- Weight under 20kg for the 11-slot system

The modular aspect of the compute node led to the design of a common cooling interface, independent of the type of node installed in the slot (CPU, GPU, FPGA or NVME). The overall cooling system has to handle the 150W per node over this standard cooling interface but also handle the different hotspots on each module. This kind of modular system needs a board-specific heat spreader (and inner cooler for the cases of the COMe CPU modules) to bring the heat from the high-heat dissipation components’ junction to the cooling interface. Simulations and tests have to be performed to ensure that the junction temperature of each of these high heat flux components is not reaching its maximum for the worst ambient temperature of 50°C.

For this, the critical components have been reported for both the custom and off-the-shelf boards. Several versions of the chosen Congatec AMD and Intel boards are available. From the cooling point of view, the one with the highest TDP has been chosen as the most critical board for the requirements.

The overall power consumption per module has been updated and summarized in the Table 3.1. Those values are used in the overall cooling system design to ensure that it can handle the non-uniform heat in between the nodes.

Table 3.1: Summary of the maximum dissipation per node

Node	Max. power dissipated
Intel CPU Module (w/ Congatec B7E3)	112 W

AMD CPU Module (w/ Congatec B7XD)	152 W
GPU Module (Xilling AGX)	68 W
NVME 16 TB Module	35 W
PCIe Switch	49 W
Ethernet Switch	80 W
Power Supply PCB	76 W

To ensure that all the high-heat dissipation components are well cooled on each module, their position and local heat dissipation have been reported. Defining those requirements is essential for the design of an appropriate heat spreader and inner cooler, to bring the heat to the cooling interface. The following example shows the most critical module to be installed in the system, the dual-die AMD CPU module and its carrier board. This analysis has been made on each module defined in the previous electrical section to ensure that all the critical components are well reported for the design of the cooling system.

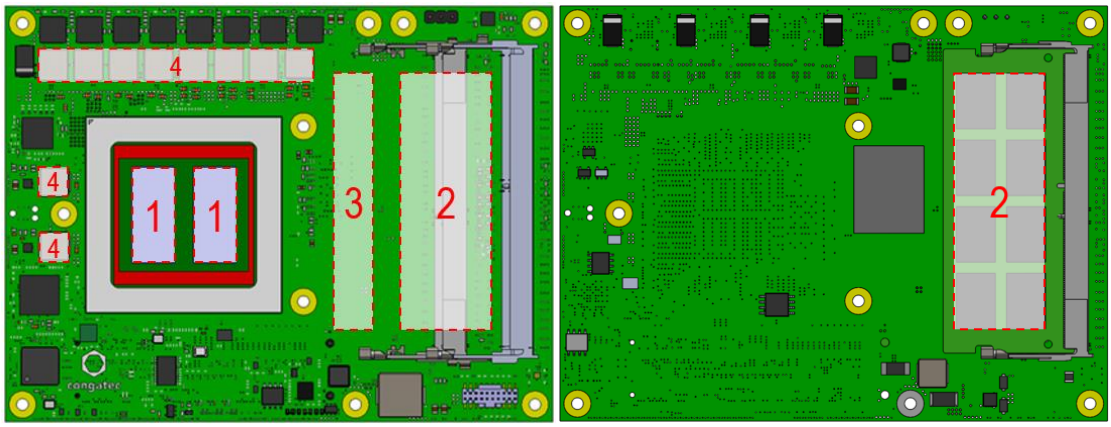


Figure 4.31: Highlighted high-heat dissipation components on the Congatec - COM Express VII CPU Board with AMD EPYC 3451 CPU

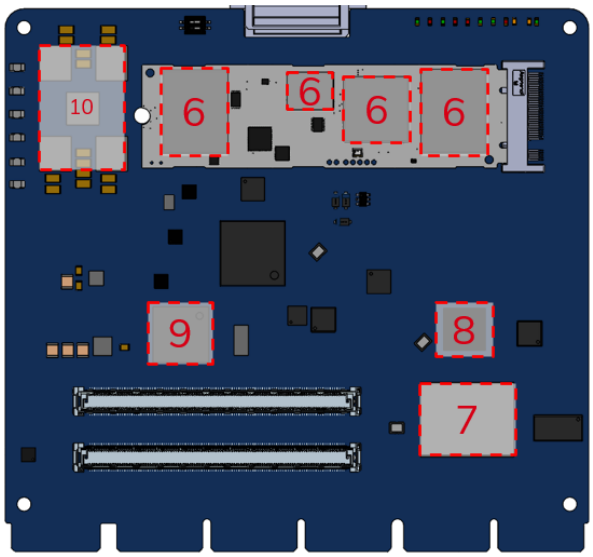


Figure 4.32: Highlighted high-heat dissipation components on the CPU carrier board

Table 3.2: Detailed report of the local heat dissipation on the AMD CPU module

ID	Node	Max. power dissipated	Max. Temp.
1	CPU AMD EPYC 3451	100 W	85 °C
2	DDR4 Upper-level front	7 W	85 °C
3	DDR4 Lower-level front	7 W	85 °C
4	Power supply (COMe)	9 W	85 °C
5	DDR4 back	7 W	85 °C
6	M2. NVMe SSD	7 W	85 °C
7	BGA NVMe SSD	7 W	85 °C
8	Ethernet IC	3 W	85 °C
9	PHY	1.5 W	85 °C
10	Power supplies (carrier)	4 W	85 °C
Total		**152.5 W**	

4.3.2 Design choices in the cooling solution

The major design choices leading to the current cooling solution are reported and justified. The choices are split in three parts:

1. Cooling system architecture,
2. Working fluid selection,
3. Technological choices.

Although design choices are presented as independent in the following paragraphs, an iterative procedure was carried out to find the optimal trade-off between performance, system weight and manufacturability aspects.

4.3.2.1 Cooling system architecture

The cooling system architecture has been selected at an early stage in the project. As stated in the initial proposal (number 876967-2), the target high-heat dissipation rate per unit volume, integration and congestion constraints associated with the EMDC, as well as the willingness to develop a low energy-consumption cooling system drove the selection of its architecture. It is a two-phase loop thermosyphon (LTS) with integrated micro-channel cold plates as evaporator.

This type of passive two-phase, gravity-driven cooling system has already demonstrated its efficiency in multiple applications. The high target in terms of cooling capacity prevented the LTS to be entirely passive, and an air-cooled condenser with integrated fans was selected. Still, the passive nature of the two-phase flow of refrigerant within the LTS makes it the most attractive solution, for both performance and compactness reasons.

While standard heat pipes rely on surface tension within capillary wick structures to sustain working fluid flow, LTS are gravity-driven cooling systems. Figure 4.33 shows the working principle of a loop thermosyphon in its simplest representation. Its main components are:

1. An evaporator located at the lowest elevation in the system,
2. A riser where the liquid-vapor mixture exiting the evaporator will flow upwards,
3. A condenser located at the highest elevation in the system,
4. A downcomer where liquid exiting the condenser will flow downwards,

5. For some applications, an accumulator is added to ensure optimal performance and operability within the range of operating conditions

Under the application of heat at the evaporator (typically generated by electronic components), the liquid will undergo partial evaporation and exit as a low-density liquid-vapor mixture. Buoyancy forces will make this mixture rise up through the riser, before reaching the condenser. Taking the heat away will change the fluid back to single-phase liquid exiting at a high density. The difference in elevation and fluid density between evaporator and condenser outlets results in a natural driving potential, sustaining fluid flow with a higher capacity than standard heat pipes.

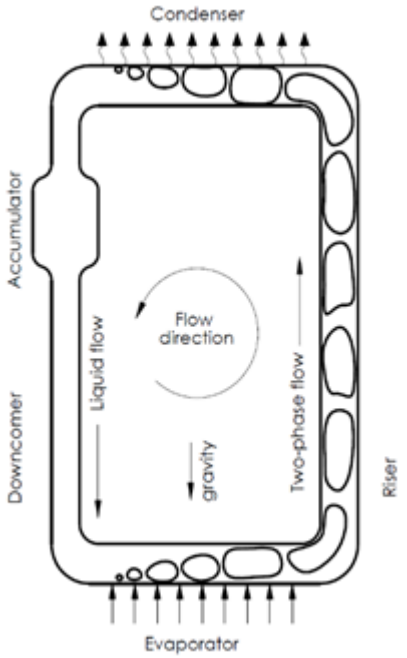


Figure 4.33. Loop Thermosyphon working principle

LTS systems are widely seen as the future of electronics cooling for the following (non-exhaustive) reasons:

1. Passive working fluid flow under the application of heat. The benefits of this are two-fold:
    a. No mechanical driver such as a pump or compressor to cool the system. Cooling related electrical consumption, noise and vibrations are significantly lowered,
    b. Scalability when increasing thermal dissipation needs, since LTS are self-regulating cooling systems, with an actual increase in performance under crescent heat load within a relatively broad range
2. Higher thermal performances compared to state-of-the-art air-cooled, liquid-cooled, and even two-phase immersion cooling solutions. While liquid-cooling is typically limited to heat fluxes lower than 20W/cm^2, the two-phase flow boiling in LTS systems can handle up to 100 W/cm^2 or higher, giving the opportunity to cool the next CPU generation using the same or a similar design. Additionally, two-phase heat transfer occurs isothermally (fluid undergoing phase change rather than temperature change), resulting in a more spatially uniform temperature field in the cooled components. This reduces thermal gradients along the electronics' junction, yielding increased components' lifetime compared to air and liquid-cooled technologies.

3. Strong dielectric properties of the refrigerants typically used. In the unlikely event of leakage, the risk of damaging the electronics is eliminated,

In the framework of the BRAINE project, the objective is to design a LTS to reach high thermal performance with low energy consumption. The thermo-mechanical design of the whole EMDC is split into four levels:

1. Overall loop thermosyphon design,
2. Air-cooled condenser design,
3. Monobloc evaporator design,
4. Heat spreader design at the node level. Each node has a custom heat spreader to accommodate its specific cooling requirements.

Figure 4.34 shows a schematic representation of the cooling loop. The refrigerant circulates within a closed loop between the evaporator and condenser, that is connected via the riser and downcomer. The electronic boards are mounted on modules that spread the heat to cooling plates integrated on the evaporator.

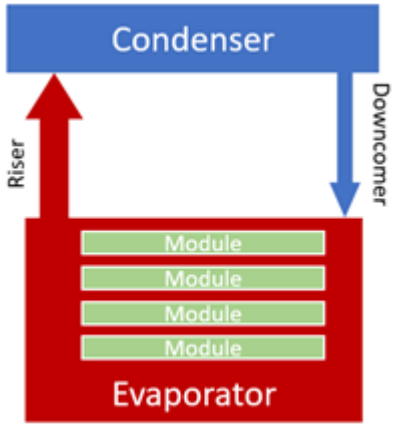


Figure 4.34: Schematic representation of the BRAINE cooling loop

The condenser of a LTS constitutes the cold source in the loop, where heat is rejected to the environment, either with an air-cooled solution or to a cold liquid bus line. If the secondary fluid at the condenser side is not mechanically driven, so that heat removal is done via natural convection, the LTS become a fully passive solution. The area needed for a 1.5 to 4.5kW condenser (as a design criterion, a maximum heat load of 1.5kW per monobloc is taken into account leading to a total of 4.5kW for a 3 monobloc EMDC design), would be prohibitively large for the current stage of the system. For BRAINE 2.0, the EMDC is designed to fit inside a 3U 19” rack enclosure, so that a specific air-cooled condenser has been designed with integrated fans. The condenser design is custom, with multiport tubes and louvered fins chosen to reduce the refrigerant side’s pressure drop, while increasing the air side’s thermal performance and fitting within a 2U 19” rack.

Figure 4.35. BRAINE 2.0 custom designed MPT and louvered fins air-cooled condenser

The evaporator of the LTS is made in one monobloc shown in Figure 4.36. For the future BRAINE 2.0 design, 11 slots with non-uniform heat load will be cooled in parallel. Each slot has a corresponding cooling 'bridge' with integrated micro-channels. The maximum heat dissipation requirement for each node is summarized in Table. 3.1, including the different COMe Type 7 CPU boards which will dissipate the highest level of heat. The EMDC monobloc is designed to be a modular system with 8 custom boards that are interchangeable (CPUs, GPUs, FPGA board etc..), one power supply board, one ethernet switch board, and one PCIe switch board. This modularity is a challenge from the cooling point of view for a passive two-phase cooling system, but the thermosyphon flow simulations show it will be operable for the non-uniform flow distribution to the multiple bridges.

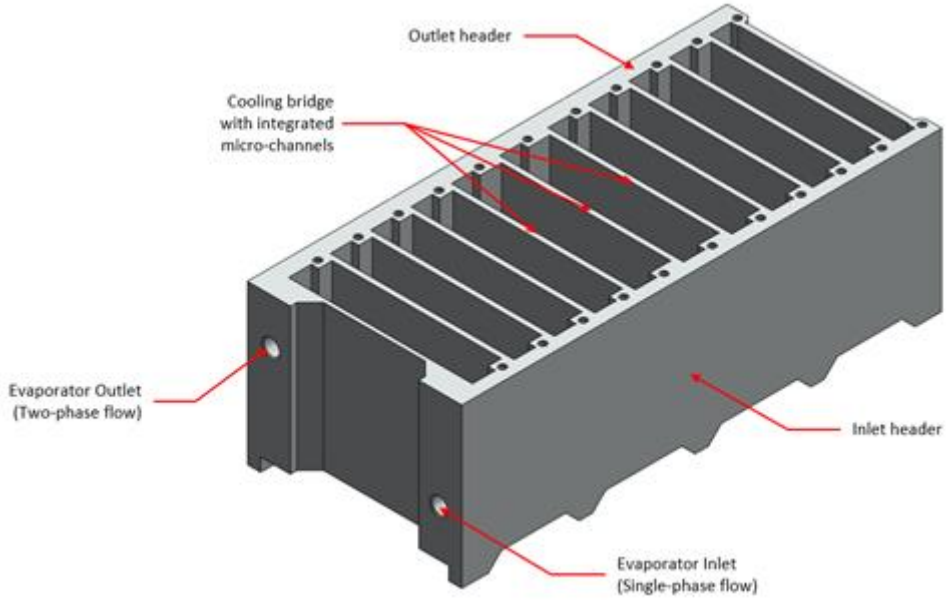


Figure 4.36. Last version of the BRAINE 2.0 Monobloc, evaporator of the LTS

At the node level, each module is cooled via heat conduction to the micro-channel bridge interface. A highly conductive Thermal Interface Material (TIM) is compressed between the bridge and the module using a wedgelock system to ensure good thermal contact. The highest heat flux components on each board are generally placed close to the interface to reduce its thermal path as much as possible. For some more complex modules like the COMe Type 7 CPU board and its carrier board, an inner cooler is necessary to bring the heat from the second PCB level to the bridge interface.

4.3.2.2 *Working fluid selection*

Table 3.3 shows working fluids typically encountered in LTS systems with some relevant thermo-physical properties. Saturation data are taken at 60°C, the maximum expected working saturation temperature for the cooling system.

Table 3.3: Relevant thermo-physical properties of typical refrigerants, evaluated at 60°C saturation temperature

Type	Refrigerant	GWP	Tcrit °C	Psat MPa	ρl kg/m³	ρg kg/m³	μl Pa.s	μg Pa.s
HFO	R1224yd(Z)	<1	156	0.44	1260	27.1	1.97E-04	1.19E-05
HFO	R1233zd(E)	1	166	0.39	1170	20.7	2.04E-04	1.16E-05
HFO	R1234yf	4	95	1.64	941	99.8	9.70E-05	1.37E-05
HFO	R1234ze(E)	6	109	1.28	1003	70.1	1.23E-04	1.45E-05
HFO	R1336mzz(Z)	2	171	0.25	1270	16.0	2.45E-04	1.14E-05
HFC	R134a	1300	101	1.68	1005	87.4	1.24E-04	1.36E-05
HFC	R236fa	8060	125	0.76	1230	51.8	1.84E-04	1.26E-05
HFC	R245fa	858	154	0.46	1240	25.4	2.57E-04	1.17E-05

The selection of the appropriate working fluid in LTS systems is driven by multiple criteria, listed below:

1. Development of a sustainable cooling solution – selection of a low Global Warming Potential (GWP) refrigerant, eliminating the hydrofluorocarbons (HFC) family of refrigerants altogether,
2. Development of a low-weight cooling solution – selection of a low saturation pressure refrigerant. At this stage, R1336mzz(Z), R1233zd(E) and R1224yd(Z) were potential candidates. Minimizing internal pressure results in reducing components' deformation and thus thickness. The choice of a low-pressure solution in this instance also allowed to use of flexible plastic hoses for the piping instead of rigid copper tubes for ease in system installation in a 19" computer rack environment, reduced weight and components' costs. The use of plastic hoses can also enable visual inspection of the system's operation, making maintenance diagnostics and analysis easier,
3. Development of a robust and high-performance cooling solution – selection of a performant refrigerant. As gravity-driven, passive two-phase cooling systems, LTS performance and operability are directly linked to the thermodynamic properties of refrigerants under the entire operating conditions range. Selecting a working fluid in this aspect is a non-trivial task since buoyancy-driven driving potential, frictional pressure losses in the entire system as a function of operating conditions and design, as well as phase change properties of the working fluid, must be accounted for.

Preliminary LTS thermal-hydraulic simulations using JJ Cooling Innovation's simulation code identified Honeywell's R1233zd(E) as the best candidate for this specific application. Additionally, this is a refrigerant with which we have had significant experience in the recent past, successfully designing and testing multiple LTS prototypes specifically designed for electronics cooling.

4.3.2.3 Technological choices

This paragraph focuses on some lower level yet important design choices.

1. Thermal Interface material (TIM),
2. Cooling modules from both sides,
3. Quick disconnect valves.

Thermal Interface Materials (TIMs)

One of the main attention points in the thermal design is to optimize the thermal path from component to the refrigerant. Thermal Interface Materials (TIMs) are used to facilitate the passage of heat where different surfaces are in contact. Their selection must be based on:

1. Performance (often characterized by a thermal resistance versus applied pressure curve) – in this instance, the required performance for the various nodes can differ since their power dissipation requirements and dimensions vary. A high thermal conductivity TIM was selected for the critical components, namely CPUs and bridge interface,
2. Available means to apply pressure (if not sufficiently pressed together, even with a layer of TIM between two surfaces, contact will not be made properly, resulting in poor heat transfer performance) – multiple technological solutions were here considered,
3. Thickness – based on the geometrical layout of the boards, nodes and monobloc faces, the dimensions of the gaps to be filled between the various components can vary significantly. The magnitude of said gaps is about 0.5mm. An added complexity was thus to select TIMs with the required thermal resistance to optimize performance, and sufficiently compressible to maintain efficient surface contact when accounting for the tolerance chain between cold system and during operation.

Multiple commercially available TIMs were investigated (including the associated implications in terms of the mechanical system for pressure application between the various components) and are shown in Table 3.4.

Table 3.4: Properties of the different chosen TIMs

Supplier	Type	Name	Therm. Conduc. W/mK
HITEK	Gap filler pad	HEM STC035	14.0
TGlobal Technology	Gap filler pad	TG-A1250	12.5
Laird Technologies	Gap filler pad	Tgon 800	5.0
Parker Chromeric	Gap filler pad	HSC 579	3.0
LiPOLY	Putty	SH-putty3	8.0
Fujipoly	Putty	Sarcon SPG-50A T	5.0
LiPOLY	Putty	N-putty2	5.0
LiPOLY	Putty	N-putty	3.5

Based on quotations provided by the suppliers and thermo-mechanical simulations using commercial software, a thermo-economic trade-off was performed for TIM candidates' selection. At this stage, since the thermal testing of BRAINE 1.0 is ongoing, the selection is not finalized, although significant advancements have been made in the design and decision process. The planned way forward is to select high-efficiency TIMs for the critical components in terms of thermal performance (CPUs and bridge) and use highly compressible, cost-effective ones for the remaining interfaces.

Cooling modules from both sides

In the early concept stage, the idea of cooling each module from both sides was investigated. The need for an inner cooler between boards (COMe Type 7 and carrier board for example) would have been removed, and the thermal conduction path from the electronics junction would then be reduced. As detailed in the next section, simulations show that the critical temperature for each module will be located at the inner cooler as the heat of both the M2 NVMe and the COMe's backside DDR4 is concentrated in this location. In this preliminary concept, both the interface between the cooling bridge and the module heat spreader are inclined with a 2.6° angle. The wedge support (in light blue on the Figure 4.37 holds a trapezoidal screw that pushes the module inside its slot when turned anti-clockwise. Although promising from the performance point of view, this concept has been abandoned for several reasons. The main issue came from the tolerances needed to connect the module's carrier board to its corresponding PCI connector. The gap between the bridge and the module should be large enough to let the module slide inside the slot with limited force and small enough to align with the connector and avoid a large displacement while tightening the wedge (possible damage on the PCI connector).

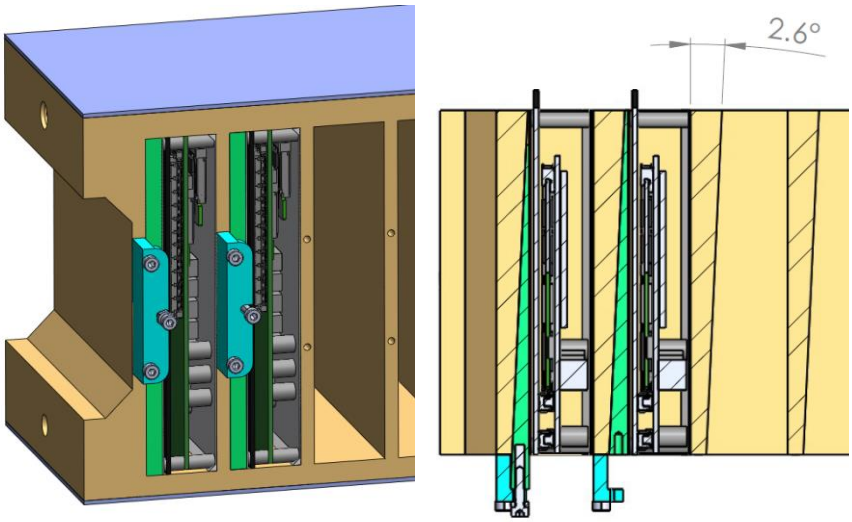


Figure 4.37: Two-sided cooling concept

Quick disconnect valves

The simple idea of having one evaporator per module that can be disconnected from the cooling loop has also been investigated. In this concept, each module has its own evaporator with an inlet and outlet quick-disconnector. Such a solution would have greatly reduced the complexity of the monobloc. However, it greatly increases the overall pressure drop due to the additional connectors. In a passively driven cooling system, reducing pressure losses is of critical importance to both operability and performance.

Additionally, the estimated cost for 22 quick connectors per EMDC made this solution prohibitively expensive.

4.3.3 Thermal-hydraulic simulations

Thermal-hydraulic simulations for the entire LTS have been carried out using JJ Cooling Innovation's inhouse simulation code for loop thermosyphons. Considering the case of one monobloc (up to 1.5kW), they have shown that the node junction temperatures will remain below their operating limit for a critical case of 60°C saturation temperature of the refrigerant and a maximum air ambient temperature of 50°C. The saturation temperature of the normal operation is expected to be lower than 40°C. Improvements and updates of the in-house simulation code for the LTS are ongoing and will be systematically included until the end of the project. More specifically, the non-uniformity in heat dissipated in between the nodes will be studied experimentally and validated. A dedicated 4-slot monobloc with "dummy" heaters to emulate the nodes has been manufactured and is undergoing thermal tests at the time of the redaction of this document. The results of these tests are used to validate the loop thermosyphon solver by comparing, among other quantities, the loop's total pressure drop. Indeed, since it depends on the mass flow rate and vapor quality at evaporator's outlet (two major parameters for heat dissipation), validating its prediction is essential to ensure reliable thermal performance calculation. In order to capture the complex two-phase flow and heat transfer in the monobloc and more particularly in the inlet and outlet headers, numerical coefficients are adjusted and will be used for the simulations of the 11-slot system. The result is shown in Figure 4.38, it can be observed that the solver is able to predict the total pressure drop within ±20%. Adjustments are still ongoing to reduce this relative error.

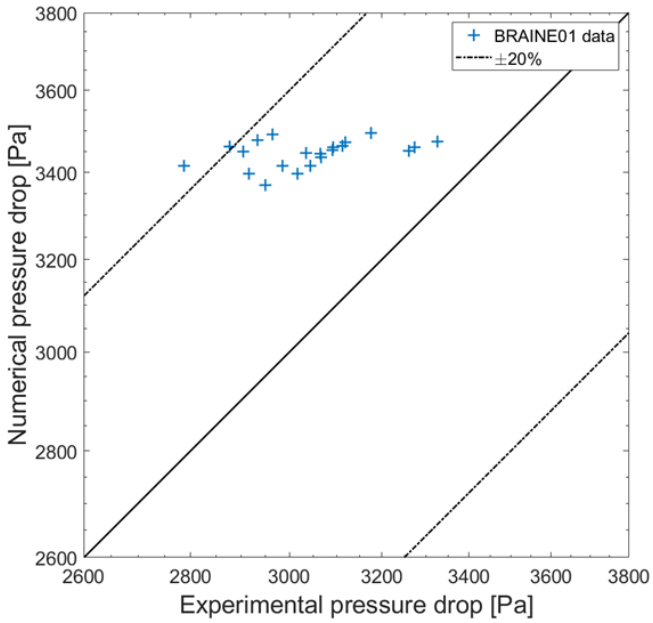


Figure 4.38: Result of numerical coefficients adjustment with a subset of the experimental data for the 4-slot monobloc.

At the node level, 3D conduction simulations have been performed using commercially available software to optimize the location of the high heat flux components on each of the PCBs while considering the effect of the TIM. Those conduction simulations have been done for all the critical modules. A simplified CAD model has been created as the one shown in Figure 4.39, also considering the heat spreader and inner cooler as well as the different TIM planned. The thermal contact between the CPU and the heat spreader is

handled with a high-performance thermal paste (in blue) while for all the lower power components, the thermal contact will be done with gap filler or thermal putty (in red). Such a simulation tool helps to choose a good tradeoff between the TIM performance and cost. However, the simulation takes into account the TIM's seller data which should be validated in-place with the thermal tests.

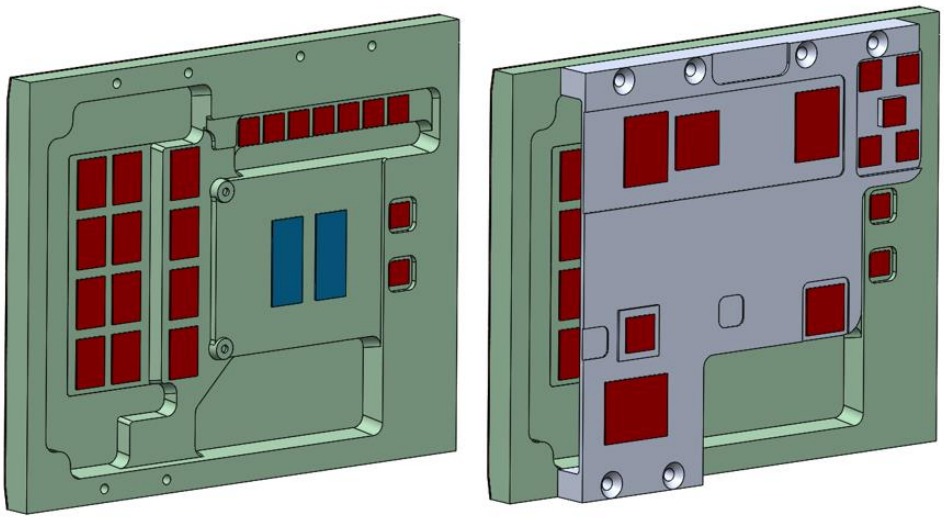


Figure 4.39: AMD CPU module's simplified heat spreader and inner cooler design for 3D conduction simulation

An example of the thermal conduction simulation done for the AMD CPU module is shown in the Figure 4.40. Additional micro-evaporator simulations have been made to ensure that the thermal bridge is able to handle the distributed hotspots of each module.

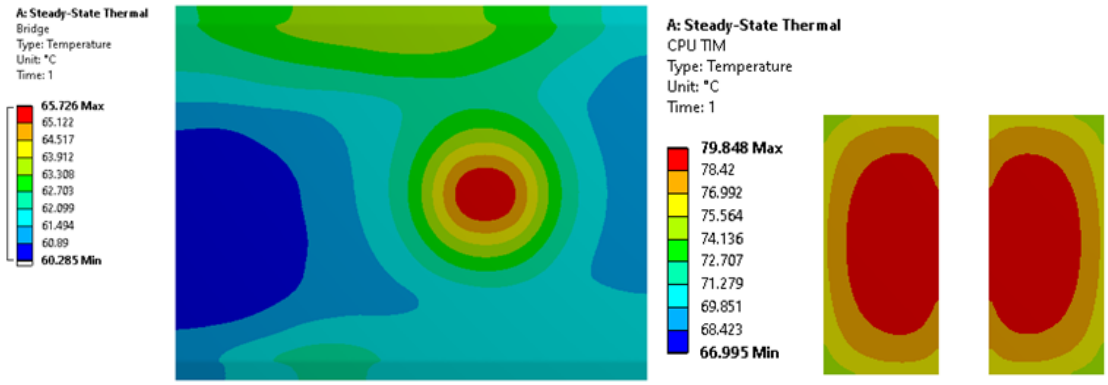


Figure 4.40: Thermal conduction simulation of the AMD COMe Type7 CPU board (left: bridge interface, right: CPU interface)

4.3.4 BRAINE 1.0 supported features

BRAINE 1.0 includes the design of a reduced 600W capacity cooling system to validate the working principle and the design choices for the large 1.5kW prototype. This reduced prototype is a 4 slots monobloc that hosts dummy heating modules. The AMD COMe Type7 module being the most critical one in simulations, the dummy heating module planned for the BRAINE 1.0 is mimicking its local power dissipation. Figure 4.41 shows the design of those 4 identical modules. Similar to the real AMD module, they are composed with the same form factor PCB as well as its carrier board. The aluminium heat

spreader and inner cooler are also similar to the one on the real module to validate the long thermal conduction path from the back carrier board to the bridge interface. Components with high power dissipation are replaced by power resistances properly chosen to get the corresponding proportional power from 0 to 150 W per module with only one power supply. The dummy CPU and the low power components (DDR4, NVMEs..) can be powered separately giving the possibility to overpower the dummy CPU and test the limits of the cooling system. Additionally, the 4 modules can be powered separately as well to test the ability of the loop thermosyphon to operate with non-uniform heat distribution between modules.

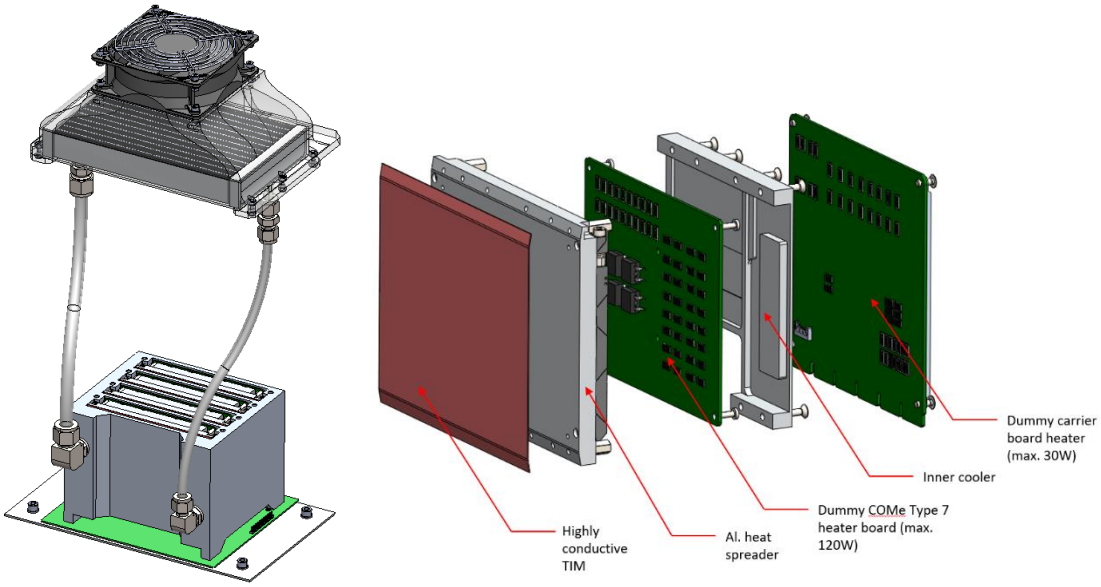


Figure 4.41: BRAINE 1.0 test bench (left) and its dummy heater module (right)

The main objective of such a test bench is to validate the assumptions made during the simulation and design phases. The simulation tool will then be updated if needed to ensure optimal thermal performance on the future larger prototype. The first dummy heater module is instrumented with 12 AWG 36 thermocouples, 4 are placed under the dummy CPU to check any temperature deviation, 2 for the DDR4s and power supply on the front heat spreader and 5 on the inner cooler. The other 3 modules are each instrumented with 5 thermocouples to check the possible non-uniform temperature fields in between the modules. If needed, the best-instrumented module (Module 1) can be easily swapped. On the LTS refrigerant loop, four thermocouples are placed at the inlet and outlet of both the evaporator and condenser. Two thermocouples are monitoring the ambient air temperature. All the thermocouples were calibrated in-house giving an overall uncertainty of 0.1K. Three differential pressures transducers have been installed on the test bench with brazed 2mm capillary stainless-steel tubes. The first one with a maximum range of 70mbar monitors the pressure drop across the evaporator. The second one with a maximum range of 170mbar measures the pressure difference between the evaporator and condenser inlet, thus the pressure drops of both the evaporator and the riser. The third one with a maximum range of 25mbar measures the pressure drop of the condenser. Additionally, an absolute pressure transducer monitors the saturation pressure at the evaporator inlet. With all this instrumentation, the thermal-hydraulic performance of the loop thermosyphon can be completely monitored and compared with the simulated values. The riser and downcomer internal diameters have been properly chosen based on simulation to maximize the refrigerant mass flow rate, 1/2" for the riser and 3/8” for the downcomer. The tubes are in transparent PFA allowing an internal pressure up to 8bar for a flexible pipe, as the system is developed for a maximum saturation temperature of 60°C

or an equivalent saturation pressure of 3.9 bar. The two-phase flow pattern (slug flow, annular flow etc..) can also be monitored during the thermal tests.

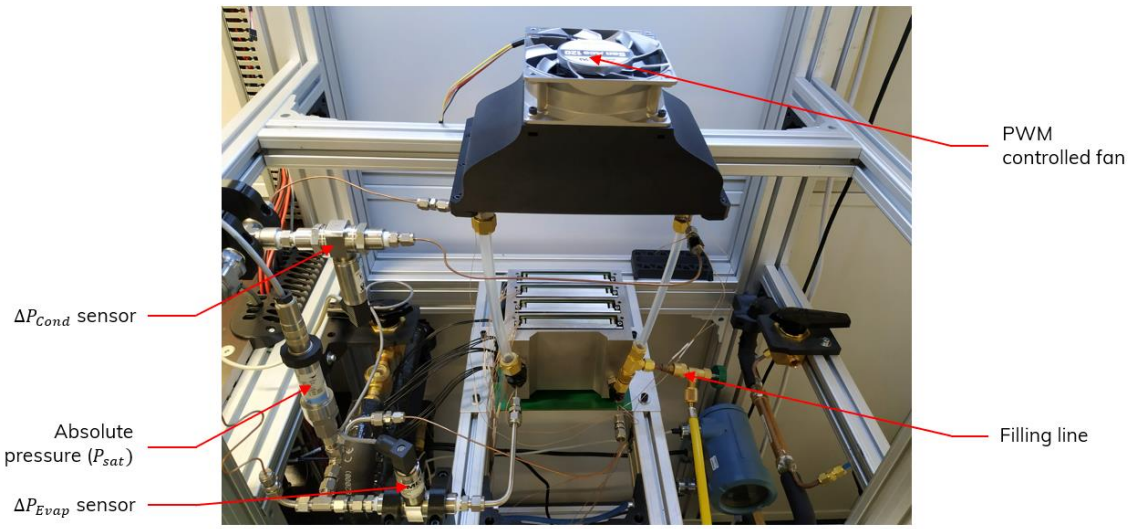


Figure 4.42. BRAINE 1.0 testbench fully populated with 4 dummy heater modules

A large series of tests are planned for the BRAINE 1.0 experimental characterization, both for the LTS thermal-hydraulic performance and for the internal module conduction. First, a complete map of thermal performance versus filling ratio and fan speed will be made to find the optimum point of the system. The effect of the height difference between evaporator and condenser on the system's performance will also be studied. If needed, a liquid accumulator will be added on the downcomer side to stabilize the level of subcooling at the evaporator inlet. On the module level, different thermal interface materials will be tested to assess their thermal performance for the expected compression levels and heat spreader tolerances. For the optimal point, the whole system will then be installed inside a small climatic chamber. The cooling system will then be characterized for a wide range of ambient air temperatures including the maximum ambient temperature of 50°C. At the time of writing, the prototype has been tested up to 250W, thus ~60W per module with a maximum junction temperature of 60°C. Figure 4.43 shows the LTS in circulation.

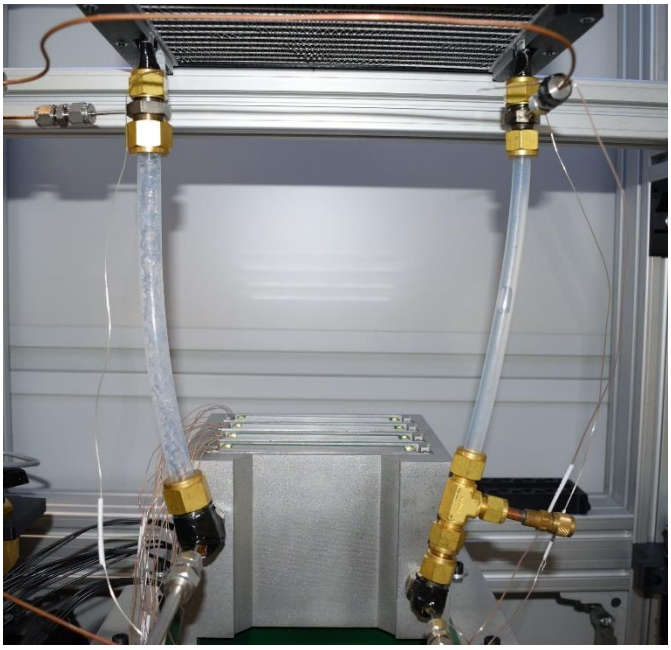


Figure 4.43: BRAINE 1.0 test @250W. Two-phase flow at the riser side (left) and single-phase liquid flow at the downcomer side (right)

4.3.5 Two-phase cooling enhancement using nanotechnology

The R&D for nanoparticles is to intensify the heat transfer of two-phase passive cooling systems (thermosyphons) by utilizing nanoparticles of size below 100 nm. The focus is on enhancing the properties of cooling liquids and takes a look at the evaporator-cooling liquid interface from the perspective of increasing the heat flow, at the same time keeping the architecture of the cooling solution within the frame adopted by the consortium. The KPI related to the proposed solutions includes long-term chemical and physical stability of coolants; their high reliability within the temperature range and operational/environmental conditions of indoor (Use Case 1) and outdoor (Use Case 3) cooling loops.

During the reported period of time, following tasks were performed:

1. In-depth analysis of thermo-hydrodynamic mechanisms behind the heat transfer process incorporating effects caused by the presence of nanoparticles
2. Formulation of framework for the design and synthesis of nanofluids based on the above analysis
3. Preparation of nanofluids with alumina and graphene nanoparticles
4. Characterization of prepared nanofluids
5. Planning and preparation of in-house experimental system to demonstrate improved boiling heat transfer in presence of nanoparticles for two phase thermosyphons

It is worth mentioning that at low flow quality the nucleate boiling mechanism is dominant, whereas at high flow quality the contribution from the convective boiling prevails in mini/microflow boiling channels for saturated inlet conditions. For nucleate boiling, the presence of nanoparticles at the evaporator-coolant interface may become essential in enhancing the boiling, whereas for the convective boiling mode the particles in bulk of working liquid intensify the heat transfer by Brownian motion, liquid layering, particles aggregation, etc.

Experiments with nanofluids indicate that with increasing the weight concentration of nanofluids, the thermal efficiency improves. Moreover, the rise of the input power leads to

rising the thermal efficiency while reducing thermal resistance. A nanoparticle layering on the evaporator surface seems to cause such changes. As the nanoparticle deposits on the surface of the evaporator, the heat transfer area increases and thus the heat transfer rate. However, quantification is difficult due to the variety and complexity of structures formed at the interface. In addition, the stability of nanofluids becomes critically important for the performance of thermosyphons since the sedimentation of nanoparticles may change the pattern and the properties of the two-phase flow in the evaporator.

We have focused our attention on more predictive points to estimate the effect of nanoparticles at the interface on the nucleate boiling in flow. Namely, the wettability of the evaporator's surface, the nucleation site density, and the porous layer formation.

Experiments and analysis reveal that the best result can be achieved when the contact angle between the surface and the cooling liquid is close to zero.

A nucleate boiling regime is considered to better understand the heat transfer enhancement by deposited nanoparticles. First estimates give an upper limit in 10% enhancement for the heat transfer coefficient for the cooling liquid on copper (5 degrees contact angle) where a decrease in the contact angle is due to the formed layer of nanoparticles with a high affinity to the cooling liquid.

The micro-level irregularities of a heating surface serve as active nucleation sites when heat flux is applied. Apart from the surface microstructures, surface wettability also has a significant effect on nucleation. Finding the golden middle between the surface wettability and the surface structure is the aim of designing the most effective interface with the highest heat flow and thus the heat dissipation from heat-generating electronics. For this, we have estimated the Active Nucleation Site Density (ANSD) as a function of few reduced parameters: the contact angle; the ratio between the roughness of the surface and the size of nanoparticles; the nano-layer temperature, and the temperature of the surface. The ANSD shows the maximum which will be utilized to characterize the best conditions for the heat transfer.

Based on the type of material, shape and affinity towards base fluids, stability and thermal properties, Alumina and Carbon-based nanoparticles were considered. For initial experimentation and proof of concept, water was chosen as the base fluid. For alumina nanofluids (Figure 4.44), the stability over a period of 100 days has been achieved and the enhanced thermal conductivity of nanofluids shows 20%.

Figure 4.44: Stabilized Alumina nanofluid samples with 4% and 5% vol nanoparticles.

Carbon Materials: We consider 4 types of nanoparticles namely, **Graphene Oxide (GO), Graphene (G), Piranha exfoliated Graphite (PexGH), Multiwall carbon nanotubes (MWCNT).**

For stabilization of graphene, we consider strategies based on the covalent and non-covalent functionalization of graphene. These should follow a few basic requirements.

- The total price per unit should be minimal.
- The preparation method should lead to the fast production of graphene nanofluids (exfoliation and dispersibility);

To have a good affinity of nanoparticles to the base fluid (water at this stage), the molecules with proper structure to link nanoparticles with the hydrogen network of water molecules have been used. The basic requirements to linker molecules are such that they exfoliate the graphene-based material, stabilize the particles and enhance the energy transport via the particle-liquid interface. Various functionalization techniques have been adopted to achieve goals and optimal approaches for nanofluid preparation.

Examples of prepared water-based graphene nanofluids are presented in Figure 4.45. The results of observations and measurements of thermophysical properties (stability, thermal conductivity) are shown in Table 3.5.

Figure 4.45: Example of graphene nanoparticles stabilized with various stabilizers. From left to right (1,2) Graphene with PBA (3,4) GO with PBA (5,6) Graphene with 9-ACA.

Table 3.5: Thermo-physical properties and stability chart for prepared carbon based nanofluids

Nanoparticles	volume fraction	Stabilizer/ intercalant	Base fluid	Stability	Thermal cond. enhancement
GO	3.80E-04	PBA-Cs/NaOH	water	stable	No enhancement
Graphene	3.60E-04	9ACA/Ethanol	water	stable	No enhancement
MWCNT	3.06E-04	9ACA/Ethanol	water	stable	No enhancement
MWCNT(-OH)	4.70E-04	PBA-Cs	water	stable	No enhancement
Graphene	0.0083	NA	EG	stable	10%
GO	0.005	NA	EG	stable	10%
Graphite	0.005	NA	EG	not stable	9%
Graphite	0.005	melamine	water	not stable	35%

To get a better understanding of the process of nanoparticle deposition on the evaporator surface during boiling and its subsequent effect on the heat transfer, Synano will develop an in-house test setup. As per the conclusions from the in-depth analysis of thermo-hydrodynamic mechanisms behind the boiling heat transfer process in the presence of nanoparticles, the critical parameters which will be measured are selected, namely,

surface roughness, nanofluid characteristics (size, type, shape and concentration of nanoparticles) and wettability. Initial experimentation will be conducted to select the most optimum nanofluid (having the lowest contact angle or high wettability) for surface roughness generally adopted for commercially used thermosyphon evaporators and in line with the ones manufactured by the consortium partners.

The in-house experimental system (Figure 4.46) will consist of an evaporator tube section which will be provided constant heat flux by passing an electric current through the pipe. The setup will be made in such a way that pipes of different materials, diameters and lengths can be tested. A closed-loop will be constructed with a vacuum pump attached to remove excess air from the loop and create an appropriate pressure difference for starting the flow of working fluid inside the loop. The loop will also include a condenser which will be cooled by a fan. Saturated vapor will become subcooled and exit as a liquid from the condenser. Temperature sensors will be added at appropriate locations to measure evaporator surface temperature, working fluid temperatures in and out of evaporator and condenser. Pressure will also be measured across the evaporator tube. A flow sensor will also be placed before the evaporator to check the mass flow. The entire setup will be properly isolated to minimize heat loss to the environment. The selected base fluid and nanofluid will be tested in the setup to measure surface temperatures, heat flux, mass flows and calculate HTC for respective fluids. Acquired data will be analyzed and validated with literature and a model developed by Synano.

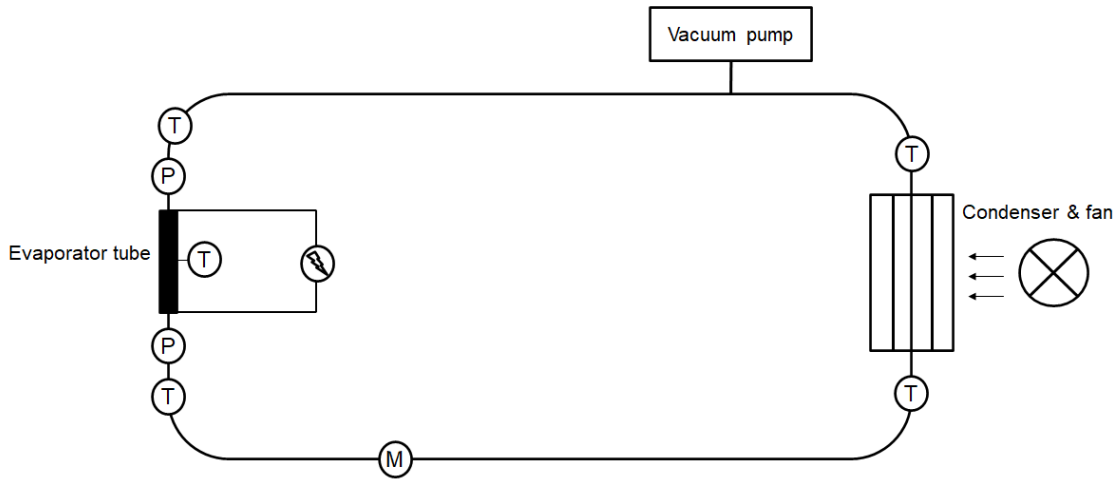


Figure 4.46. Experimental system for the cooling solution.

4.3.6 Roadmap development towards BRAINE 2.0 and KPI validation

The main KPI related to the cooling is an absolute expected power dissipation of 100W per node. While the current SoA solution has a maximum TDP of 75W/node, simulations have shown that the maximum TDP on BRAINE 2.0 can be higher than 150W/node for the same maximum junction temperature. In the BRAINE 1.0 test bench as well as in the large 1.5kW prototype, the dissipated power and the electronic boards' temperature will be accurately monitored to validate the expected efficiency. For a fair comparison with the SoA direct air-cooled cooling solutions, the energy consumption of fans is also measured. This will enable Coefficient of Performance (COP) and Power Usage Effectiveness (PUE) reconstruction on a comparable basis.

Another important requirement included in the KPIs list is a vibration-free cooling system. Vibrations from fans and mechanical drivers on the refrigerant side (compressors or pumps) can cause microcracks in welds and component failure. With the loop thermosyphon solution, the fans used at the condenser side are placed in another level connected with flexible hoses for the downcomer and riser, significantly limiting the vibrations at the EMDC level. The absence of a mechanical driver on the working fluid side eliminates this component of the vibrations altogether. For future potential use cases of EMDC LTS cooling in favorable conditions (specifically the availability of an external space such as the external surface of a building with sufficiently low ambient temperatures), natural convection cooled condenser solution is entirely conceivable, yielding zero vibrations, zero electrical consumption cooling solution.

An important aspect is also the verification of the transient response of the cooling system when electronic components operate (sudden increases in heat load in CPUs for example). Avoiding overshoots in junction temperature above the acceptable levels is crucial. A qualitative analysis of the LTS' dynamic response will be carried out using the BRAINE 1.0 hardware. The analysis will be qualitative in the sense that the installed heater modules' internal characteristics, although allowing for quick heating phases, cannot reach the low time constants typically encountered in CPUs. The first set of de-risk tests will be carried out with this hardware. Before sending the first 1.5kW prototype at TUe and CNIT for the various use case usages in WP5, transient thermal tests will also be carried out to ensure proper cooling system's response with representative components.

During operation, cooling transients occur in each node as their respective electronics' thermal loads vary with time, within a fluid network of 11 units cooled in parallel. Additionally, each slot will have its own level of the transient load depending on what that node has in it (CPU/GPU/memory) and what it is doing, creating significant differences in thermal power in any slot at any one time which the complex evaporating liquid-vapor flow needs to handle. Thus, thermal-hydraulic tests must be undertaken to gather data on the cooling performance of the 11-slot unit and that of the more complex cooling fluid flow network with the three 11-slot units connected and operating in parallel cooling mode.

The complete thermal testing of the planned prototypes (an 11-slot unit in Pisa and a 3 x 11-slot unit TUe) is imperative to the outcome of the project. Without actual full-scale operation cooling tests on these two prototypes, the prototypes cannot be considered to be validated. Therefore, the final prototypes installed at Pisa and Eindhoven must be fully tested thermally with the actual computer nodes inside to demonstrate/validate that the novel cooling system is able to handle the wide variations in cooling load of each node during these nodes'/prototypes' operation during (i) cold start-up and (ii) a wide range of operational work cases.

Current state-of-the-art solutions for cooling high TDP (100W) COMe T7 boards are usually cooled with several heat pipes, a large and heavy aluminum heat sink and a fan per module. The current loop thermosyphon solution allows a significant space reduction and reduces the maintenance needed per module. Only a limited number of fans, installed directly on the condenser, are required to operate the LTS. In addition to the space and maintenance savings, this gravity-driven, passive two-phase cooling solution will thus offer:

1. High capital costs savings due to the reduced number of fans necessary to cool the modules,
2. Unprecedented operating costs savings due to the (quasi) passive nature of the cooling solution,
3. High scalability potential, since the system is designed to operate with nodes requiring various heat dissipation loads,
4. Increased maintenance technician comfort due to low fan noise.

In terms of the European goal of going towards low-carbon footprints of the computer industry, it is mandatory to experimentally measure the total electricity consumed by the electronics, and that by the cooling system's condenser (cooled by a fan), in order to have operational proof with measured values for the low energy consumption of this innovative cooling technology. The ratio between the electrical consumption used for the cooling and the global electrical consumption of the edge datacenter should be compared with the competing solutions on the market, as we hope to demonstrate that we provide higher computing performance per Watt than others (lower operation cost). For defining this computing performance metric or metrics, the partners must discuss and plan for this in the next quarter.

4.4 Mechanical components of BRAINE 1.0

BRAINE sits in a mechanical enclosure and hence decision on the enclosure itself, design and materials have been taken in relation to fabrication aspects, durability, cooling performance and integration in edge environments. Part of the enclosure design consistent of 3D projections, and hence, some variations of BRAINE 1.0 are included in this section (containing more or less modules) - the fabricated prototype remains with 8 modules, yet we included the variations on this document.

4.4.1 Enclosure design

The properties of the enclosure of the EMDC are largely determined by where and how the EMDC will be placed and used. The different Use Cases deliver these input requirements and in conjunction with consortium partners, mainly HIRO, PCB and JJC, HID applies this input to the design of the eventual enclosures of the BRAINE EMDC system.

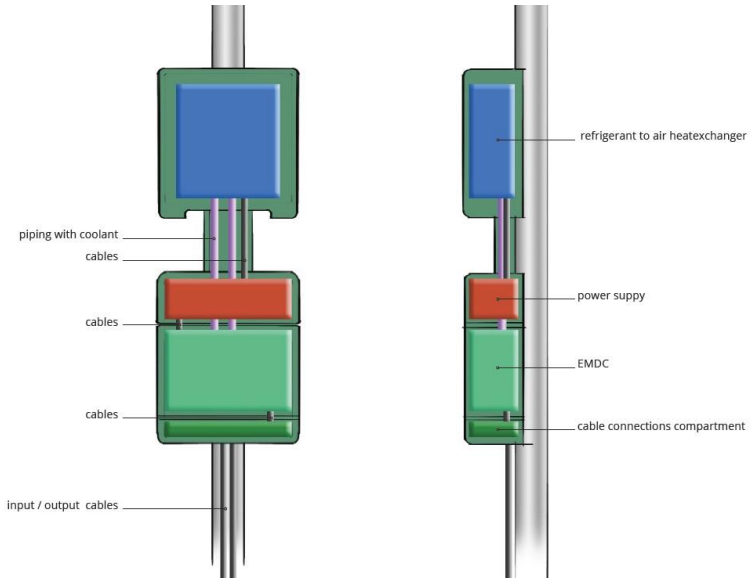


Figure 4.47a general building blocks of the BRAINE enclosure set and their arrangement.

Figure 4.47 shows the high- level product design build-up for the BRAINE enclosures. Such an approach enables all consortium partners involved to contribute and monitor whether the different specific requirements and wishes will be met. The design of the enclosures moves down, parallel with the communication with partners, form a high-level general building block level shown here, to a detailed enclosure design tailored for one or more use cases.

Input has been gathered on how the EMDC will be used, together with HIRO. This has led to the concept design of the EMDC housing. This will enable HIRO and other partners to make further steps on exploitation and implementation.

Figure 4.47b Concept design housing of the 24 node EMDC system as envisioned for the Smart Factory use case. Note the condenser unit is placed outside.

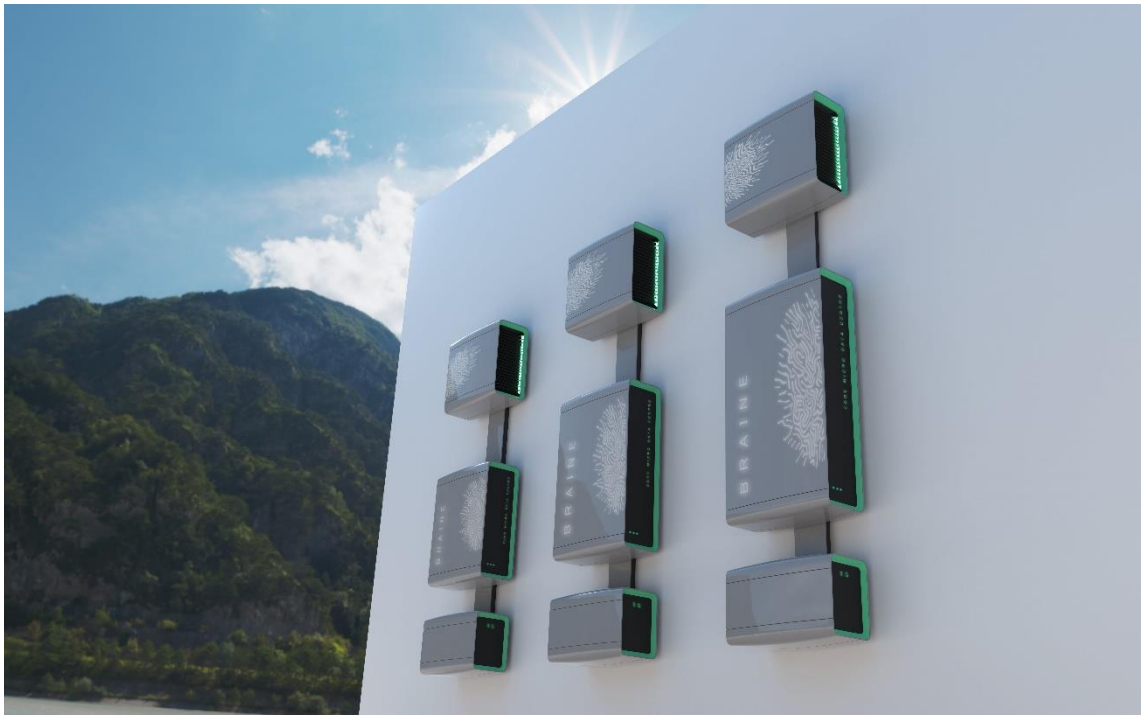


Figure 4.47c Concept design housing of the EMDC system as envisioned when placed outside. Note the EMDC unit is visualized in 3 sizes: 8, 16 and 24 nodes (left to right).

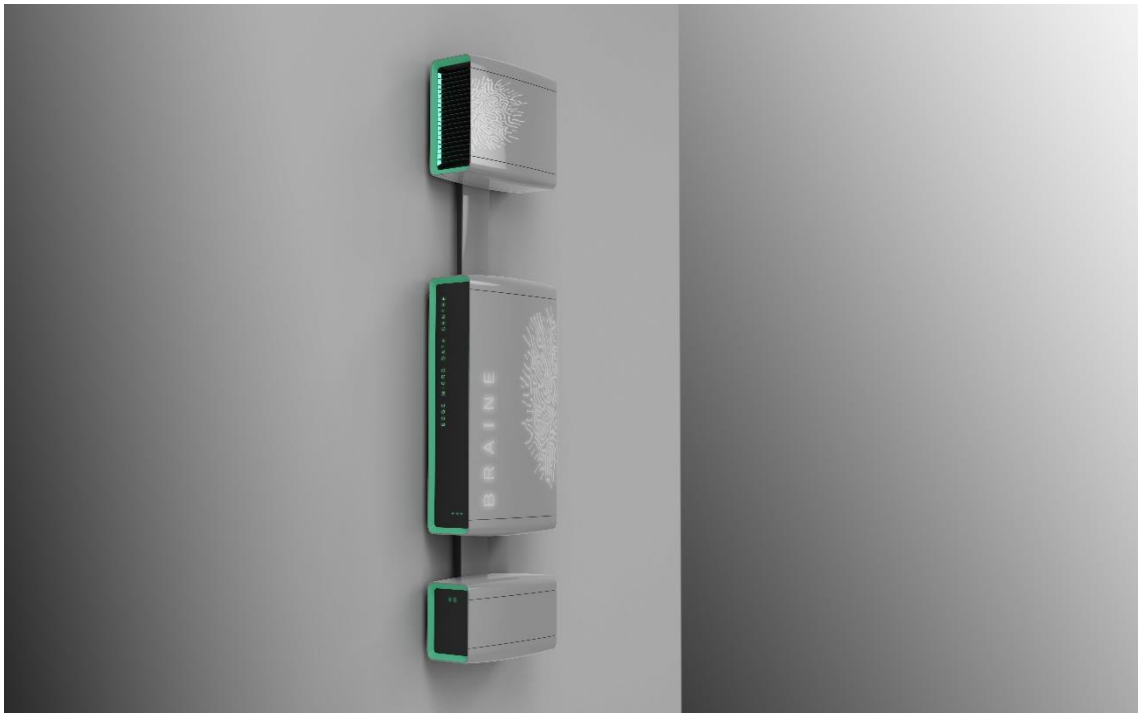


Figure 4.47d Concept design housing of the 24 node EMDC system as envisioned when placed inside (e.g. use case Hospital).

For the BRAINE prototypes it has been decided among consortium partners that not the envisioned enclosures mentioned above will be used, but that these prototypes will be placed in standard 19” rack housings. Short term practical reasons have driven this decision. This decision is taken based on a proof-of-concept approach and a consideration that the above designs are higher TRL levels than the ones targeted in BRAINE.

4.4.2 Prototype enclosures

The prototypes of the EMDC will be housed in 19” rack enclosures. These are standard enclosures that are customized.

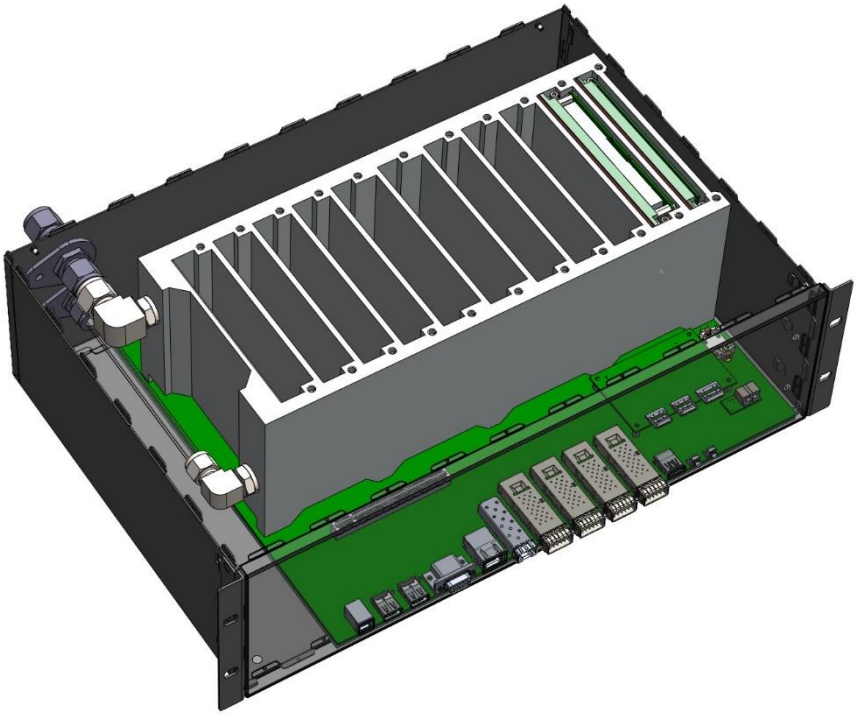


Figure 4.48a The BRAINE EMDC prototype design in a 19" 3U housing.

Figure 4.48a shows how the EMDC prototype will be placed in its enclosure. In conjunction with partners HIRO, JJC and PCB, HID has tackled different design aspects to enable the hardware to operate reliably. Some of these aspects are:

- Sufficient support of and fixation of the backplane pcb, thus enabling inserting and ejecting pcb assemblies (nodes).
- Assembly of the hardware and the coolant piping in the enclosure.
- Fixation design that leaves room for manufacturing tolerances, whilst still guaranteeing accuracy where needed.

Besides the EMDC prototype enclosure, a second enclosure will be made that houses the condenser. Figure 4.48b below shows a customized 19" 2U housing in which a custom (by JJC) condenser is placed. This condenser is made in close cooperation by JJC, PCB and HID.

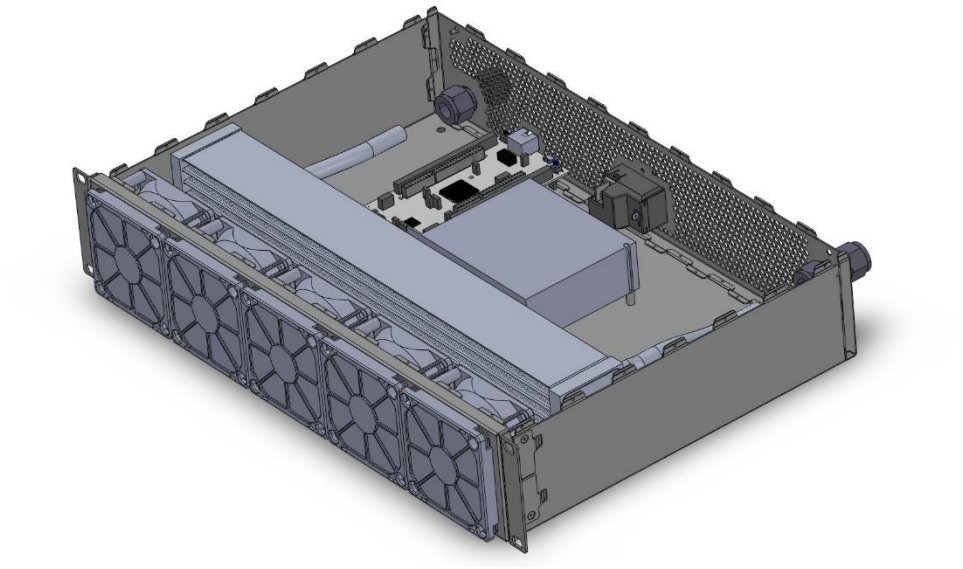


Figure 4.48b The BRAINE prototype condenser in a 19" 2U housing.

4.4.3 Eject Tool

The ejection of slots from the EMDC may become challenging as it is a tight enclosure (in order to improve cooling performance). Manual ejection may lead to damages. As a result, HID has developed an Eject Tool in order to securely remove the pcb assemblies (nodes) from the monoblock. A first test model (P1) was made to test the mechanics.

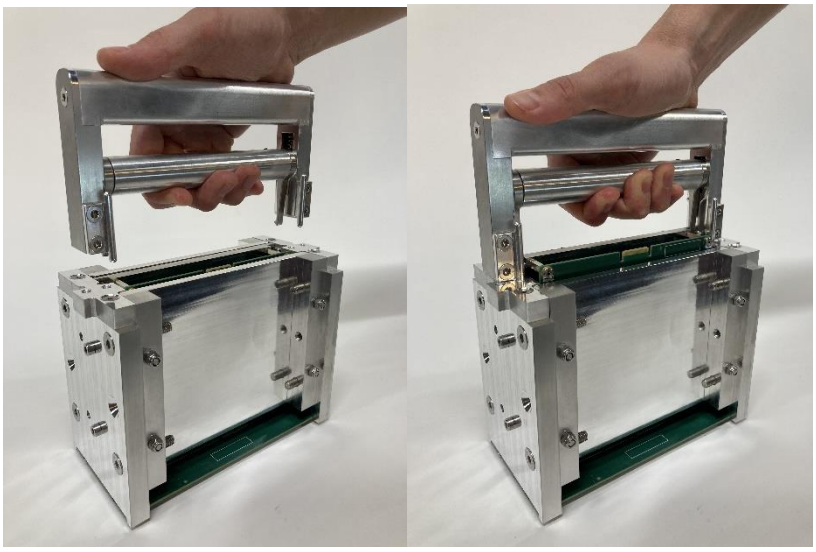


Figure 4.49a The BRAINE Eject Tool P1.

After testing mechanical changed were made, and an optimized model was made (P2).

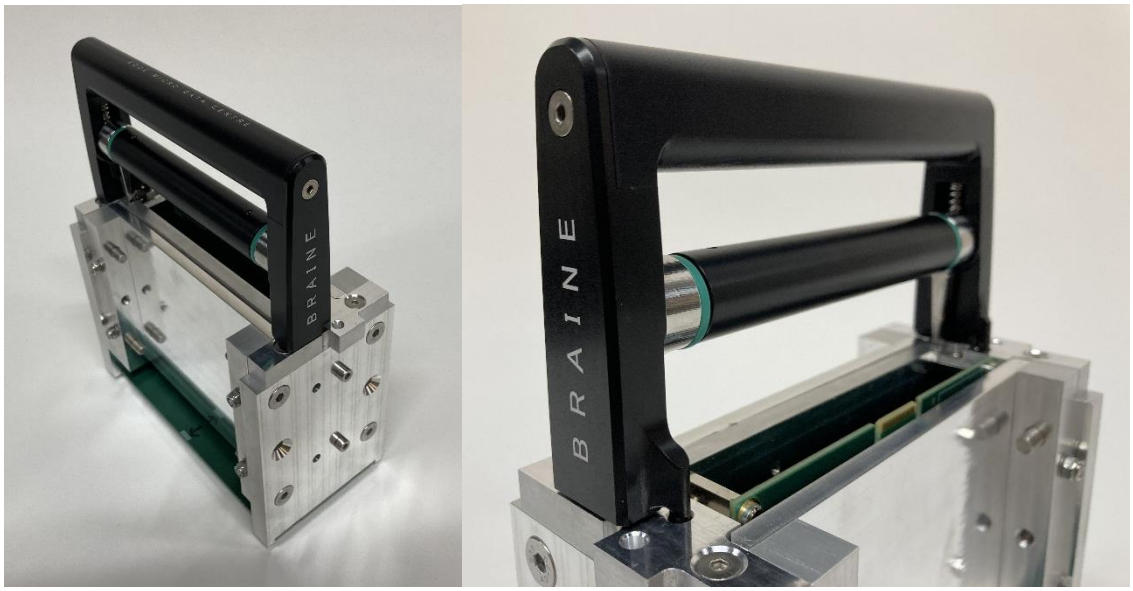


Figure 4.49b: The optimized and final version of the BRAINE Eject Tool (P2).

4.4.4 PCB assemblies – nodes physical support

In close cooperation with PCB and JJC, HID has been developing the non-computing hardware for every pcb assembly (node). Custom milled cooling parts are designed in order to transfer the heat of the computing hardware to the cooling system designed by JJC. In practice this means using a 'heatspreader' on which the computing hardware is mounted, which in turn is tightly wedged to the cooling bridge of the monobock. This concept inherently means there will be some movement of components at a right angle to the assembly direction of the node in the monoblock. This means the tolerance chain of the system must be closely monitored. HID has designed the concept of the pcb assemblies with this challenge in mind, minimizing risk by the mechanical design. During development and detailing of the components, this topic is constantly kept in mind and monitored.

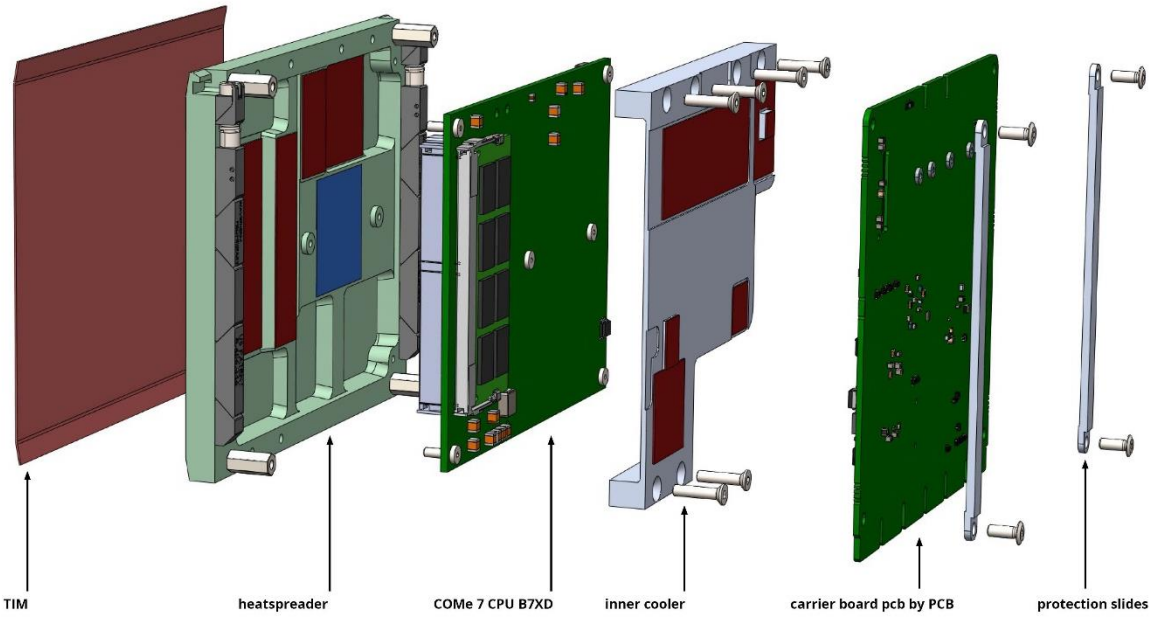


Figure 4.50a: An exploded view of the pcb assembly (node) for the CPU B7XD.

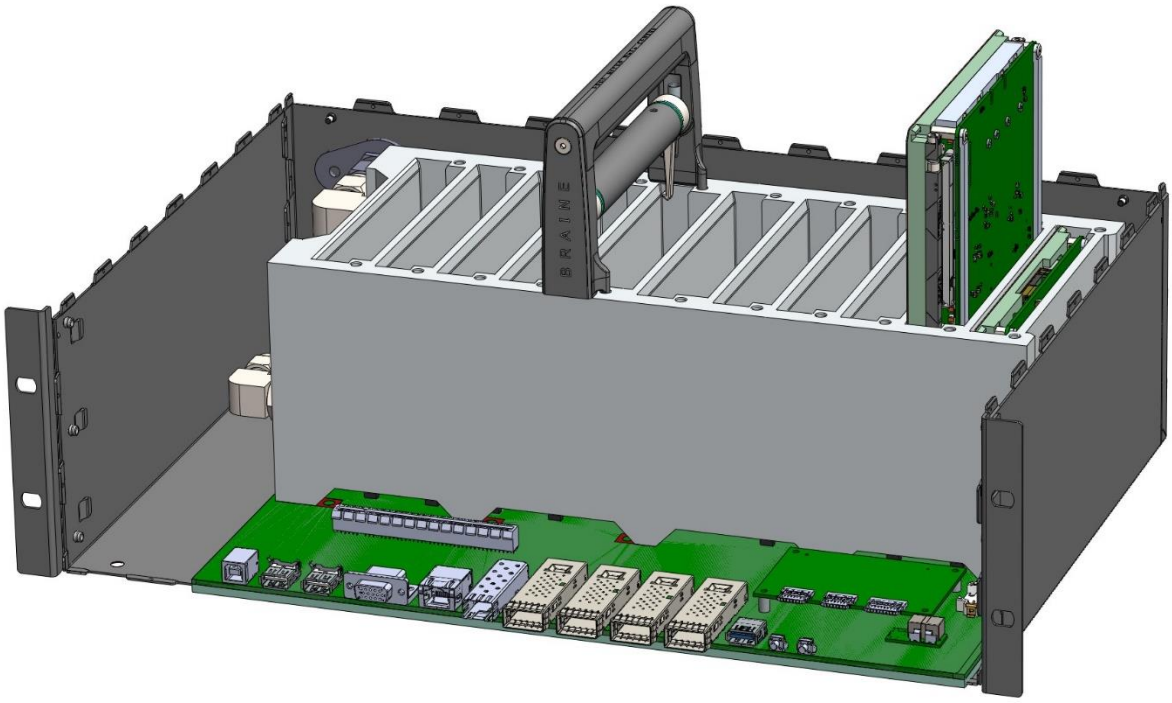


Figure 4.50b: pcb assembly being inserted into the monoblock.

The process of designing all non-computing parts is ongoing and in fruitful and close co-operation with partners PCB, JJC and HIRO.

4.5 C2.16 - Quantum safe fiber link

In this section updates on the proposed edge-to-edge quantum safe optical links based on quantum key distribution (QKD) are reported. The schematic of the QKD link is showed in Figure 4.51 and it includes three optical channels: the data channel for the encrypted cyphertext, the quantum channel for the qubits and the service channel for any supplementary information required for the key distillation, which includes processing Alice and Bob must do to retrieve the same quantum key from the exchanged qubits. QKD protocols are designed to have no critical information shared through the service channel, thus maintaining the security of the link.

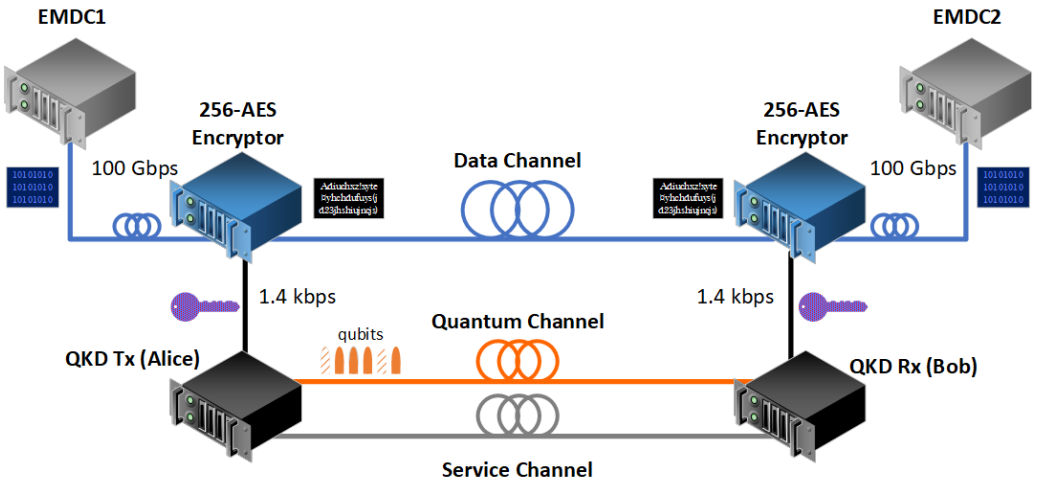


Figure 4.51: Quantum safe edge-to-edge optical link with symmetric encryption supporting QKD.

The setting of the proposed link can be split in the following steps:

1) Quantum layer that covers the QKD protocol (service and quantum channels)
2) Data plane encryption that includes the data channel
3) Single fiber QKD link with WDM
4) Upgrade setup with EMDCs and use case demonstrations

4.5.1 Quantum layer

The quantum layer covers the QDK protocol, which describes how the qubits are physically implemented and the HW at the Tx and Rx of the quantum channel. For the proposed QKD link, we selected the IDQ Clavis3 modules that supports the coherent-one-way (COW) protocol. The main advantages of COW compared to other protocols operating with single photons, such as BB84, are the more relax requirements of the photonic components. For the COW protocol the information is encoded in time. Alice sends coherent pulses that are either empty or filled with few photons (the typically mean photon number is 0.5). At the receiver side, Bob measures the power of the pulses, and he converts them into bits according to an established encoding scheme. The encoding scheme also includes decoy states, which are used by Bob to check the coherence between two consecutive pulses. In case of eavesdropping, the coherence between consecutive pulses of a decoy state is broken; therefore, Bob by randomly measuring the coherence of two consecutive pulses can identify the presence of an eavesdropper.

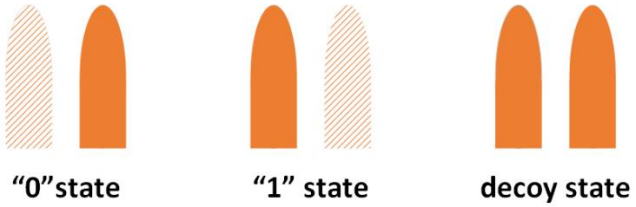
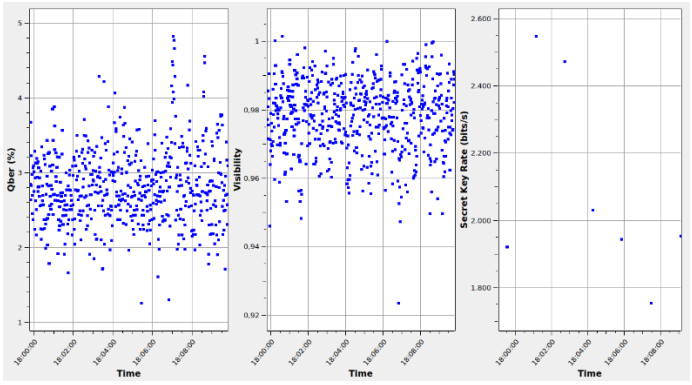


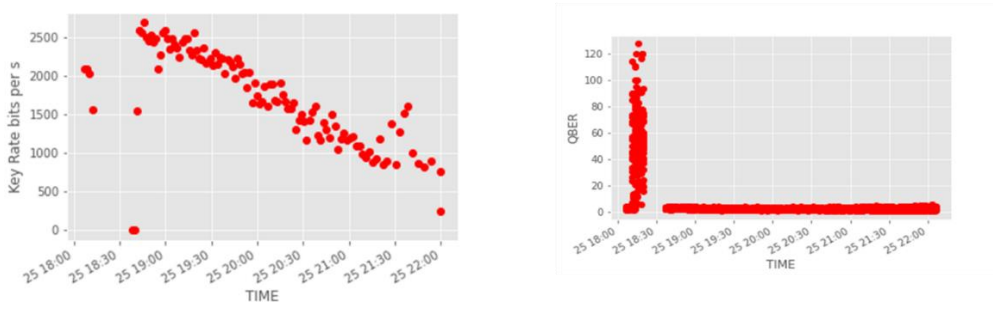
Figure 4.52: COW protocol qubits (consecutive coherent pulses) encoding into bits and the decoy state. Pulses can be empty or full (mean photon number 0.5).

After the key material, i.e. the qubits, have been exchanged and decoded into bits, it is post-processed by both parts in order to extract the same secret key. The post-processing is referred as code distillation. Key distillation requires Alice and Bob to share a portion of the key material, which is then discarded. The remaining key material will finally be used as secret key by the two parties. The key distillation has also the function to calculate the quantum bit error rate (QBER), which describe the noise in the quantum channel; more specifically the QBER is used to detect the presence of an eavesdropper, which theoretically increases the QBER above 25%. When the QBER exceeds this threshold the QKD modules continues sending qubits, but no key is provided to the encryptors until the QBER decreases.

The evaluation of the quantum layer is described by two KPIs: the average key rate and the QBER. Figure 4.53a shows the achieved key rate and QBER by the QKD modules. For these measurements, the quantum and service channels were in different fibers of 1 m. We can notice that the key rate ranges between 1.8 and 2.5 kbps approximately, which is more than enough to feed a symmetric encryptor operating at 100 Gbps. The measured QBER is below 5%, which is significantly below the threshold.



(a)



(b)

Figure 4.53: Key rate and QBER achieved by the QKD modules and encryptors with 1 m fiber (a) and with attenuation up to 21 dB over time (b)

Figure 4.53b shows the two KPIs when we gradually applied over time up to 21 dB attenuation with a variable optical attenuator (VOA) to the quantum channel. We can notice that the QBER rises it is still significantly below 25%; therefore, the QKD are still able to successfully provide keys. However, the key rate drops below 1 kbps because the QKD protocol requires to discard more qubits to estimate the QBER. In conclusion, experimental results show that the QKD modules can successfully provide secret keys with an attenuation of 21 dB in the channel, which is more than enough for edge use cases with 1-30 km fiber connections between edge nodes.

4.5.2 Data plane encryption

This step consists in the integration of the data plane components: security application entities (SAE), clients and key exchange management system (KEMS).

Currently the two clients are edge nodes implemented as two commercial servers.

The two SAEs are layer-2 symmetric encryptors. QKD protocols are independent from the symmetric encryption applied to the data; thus, the selection of the QKD protocol and the symmetric encryption are uncorrelated. Therefore, we selected the standard symmetric encryption 256-bit AES for the SAEs. Each SAE must be authenticated and receive a unique ID to be registered in a QKD environment. Currently, the SAEs in the testbed can operate up to 40 GbE. The next upgrade will include two 100 GbE encryptors as SAEs.

However, they must support the ETSI GS QKD 014 protocol, which regulates the key exchange between two or pair of SAES and their QKD modules (see Figure 4.54).

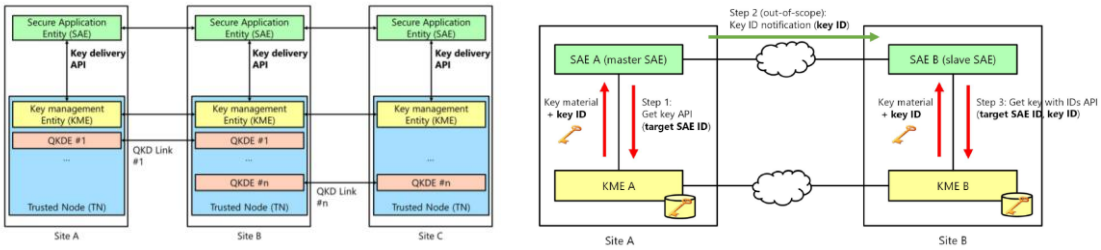


Figure 4.54: Key exchange interface between QKD modules and SAEs.

A running QKD systems enables an API from which keys can be obtained by an authenticated SAE via HTTPS requests. Once a key is extracted from the qubits, the QKD modules store it in their key management entity (KME) until they are supplied to a cryptographic application via API calls. There are three types of API calls: get status, get key and get key with IDs (see Figure 4.55). Key IDs are used by the Alice's SAE (master) to inform Bob's SAE (slave) that it successfully retrieves a key from its QKD module.

Figure 4.55: Sample response from KMS to SAE "Get Key" API call in JSON format.

The authentication of the SAEs and the QKD modules into a QKD environment is managed by an external server that is referred as key exchange management system (KEMS), as shown in Figure 4.56. The KEMS does not share any information regarding keys; however, it handles the authentication between the two parties (Alice and Bob). The KEMS runs in Docker containers; the main containers are the KEMS Server and KEMS

Database. KEMS Server container give access to the central GUI for configuring the QKD environment. It manages the configuration of QKD, KMS and SAEs, as well as the links among them for the secure transmission and use of the key. This information is collected and stored in the KEMS Database container, which uses a MariaDB.

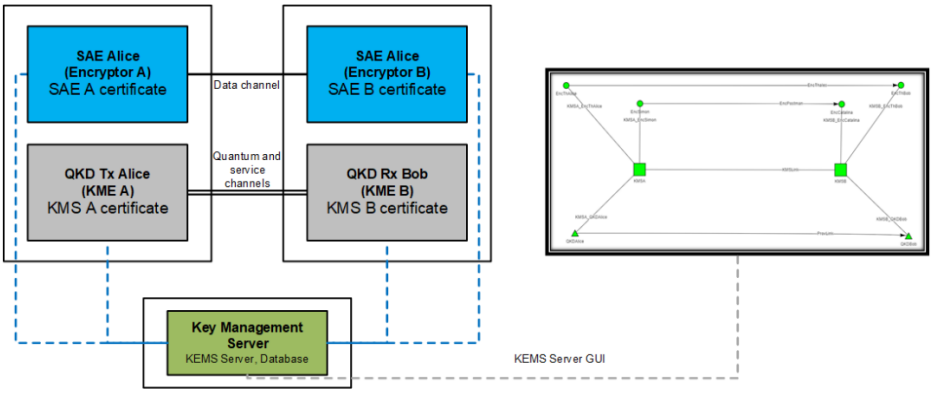


Figure 4.56: Key exchange system between QKD modules and SAEs.

The main KPI related to the data plain encryption is the latency introduced by the SAEs when external QKD keys are provided. When the QKD modules operate regularly no impact is expected to the SAEs performance. We measured the round trip time (RTT) latency introduced by the encryptor with key sharing based on QKD or no encryption (directly sending keys). The encrypted data was at 10 Gbps and the quantum channel has a fiber length of 1 m. As expected, experimental results confirmed that the QKD had no impact in the overall latency, which had a mode of 222 ms approximately. Measurements with different fiber lengths are still undergoing; we expect to complete the measurements within the following months.

4.5.3 WDM system

A typical drawback of QKD solutions is the requirement of dark fibers for the quantum channel. Dark fibers are rarely available when dealing with shared networks, thus, multiplexing the quantum channel with the other two and possibly regular traffic is a target of BRAINE solution. The schematic of the proposed single fiber QKD link is showed in Figure 4.57 that includes a quantum channel and the bidirectional service and data channels. The three channels share the same fiber by wavelength division multiplexing (WDM) to be compatible with fiber transport networks or at least consume the minimum number of fibers. Five wavelengths are used as WDM channels. A coarse (CWDM) and dense (DWDM) configurations are considered:

1) Alice data channel (Ch1): 1270 nm (CWDM) or 1550.92 nm (DWDM)
2) Alice service channel (Ch2): 1310 nm (CWDM) or 1552.52 nm (DWDM)
3) Bob data channel (Ch3): 1290 nm (CWDM) or 1551.72 nm (DWDM)
4) Bob service channel (Ch4): 1330 nm (CWDM) or 1553.32 nm (DWDM)
5) Quantum channel (Q-ch): 1550 nm (CWDM) or 1310 nm (DWDM)

The quantum channel lays in a different bandwidth to minimize the crosstalk. The choice between CWDM and DWDM solutions mainly depends on the isolation of the quantum channel; more specifically if the fiber is not dark, thus the QKD link is sharing the fiber with other services or applications. Currently in the testbed only the CWDM configuration is available; however, a possible demonstration with the DWDM configuration has not yet been discarded.

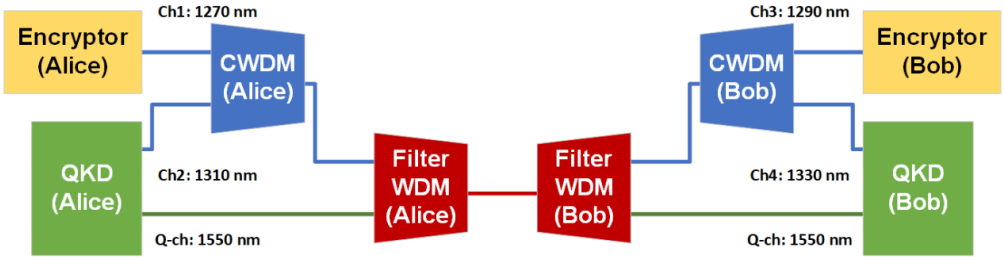


Figure 4.57: Proposed QKD link schematic with WDM and 40 Gbps links. To be upgraded with EMDC prototypes and 100 Gbps encryptors.

The separation between the 1550 and 1310 nm channels is done by the WDM filter, which is coupler based. The CWDM mux/demux is used to separate the service and data channels in both directions. The isolation provided by the WDM filters are the most significant factor and the main KPI of this development step. Currently measurements with the coupler based filters are undergoing. Moreover, different filter configurations and technologies will be considered. We expect to provide the first results in the following months.

4.5.4 EMDC upgrade and use cases validation

The QKD modules, encryptors and KEMS have been integrated and are operational. Currently the quantum channel is still in a separated fiber; however, we expect to soon complete the integration of the quantum channel into the WDM system.

Figure 4.58 shows TUe testbed after the expected upgrades. The testbed will be extended to a second rack. Two BRAINE EMDC prototypes will be added as users of the QKD link and for evaluating the SW developed for use cases 1 and 3 and their KPIs. The two current encryptors are expected to be replaced with models capable to operate up to 100 GbE.

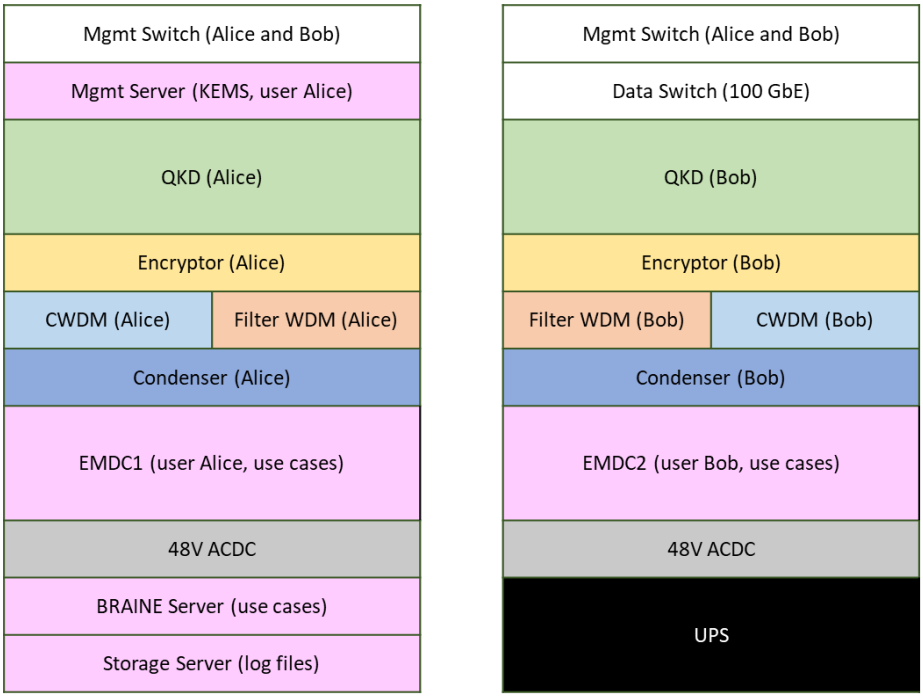


Figure 4.58: TUe testbed and QKD link after the upgrades (EMDCs and 100 GbE encryptors).

5. EMDC embedded firmware

This section describes the main elements of embedded firmware: board management related software components, FPGA node hardware architecture and software components, and EMDC switch/networking software components.

Board management functions are distributed among the main Board Management Controller and the Board Management Microcontrollers present on all BRAINE nodes. The physical layer of the communication uses USB, while the software layers communicate via Ethernet frames, allowing the implementation of transparent, application-independent extensible protocol and integration with OpenBMC. The main novelty of this approach is to use high level TCP socket-based communication through the CDC-ECM standardized USB class interface. This approach provides much higher communication bandwidth, greater flexibility and extensibility compared to traditional UART or CDC Virtual Com Port based solutions.

The FPGA node is based on the latest generation of FPGAs from Xilinx (Xilinx Versal AI). In addition to traditional FPGA logic resources, Versal AI also contains execution units which can perform DSP operations efficiently making it particularly suitable for AI centric systems, like BRAINE. Unlike traditional host - accelerator architectures found in typical HPC systems; the BRAINE FPGA node is a stand-alone unit which does not require closely coupled host management. Its firmware provides the necessary components required to execute, deploy and manage FPGA accelerated, containerized applications.

This section describes the software components developed within BRAINE to be supported by the EMDC to enable advanced networking and AI processes at the edge. This includes the

- P4 programs supporting (i) P4-based layer 2-3-4 forwarding, (ii) Traffic steering, (iii) In-band telemetry, (iv) Post-card telemetry, (v) 5G acceleration for UPF, (vi) In-network stateful feature extraction for cyber security.
- The operating system of the P4 programmable switch included within the EMDC
- The architecture and solutions supporting long reach connectivity from the switch through pluggable modules (I.e., supporting direct edge to cloud optical transparent connectivity).

5.1 Board Management Microcontroller firmware

5.1.1 Board Management Microcontroller microcontroller selection

The main criteria for the microcontroller selection for the BMMC (Board Management Micro Controller) functionality was the following:

- Well known and widespread architecture
- Well known manufacture with long term support
- High level development environment support
- Moderate calculation capability (at least 100DMIPS)
- USB device support
- Low power operation
- Appropriate basic peripheral support: UART, I2C, SPI, analog I/O

Based on the above criteria the ARM Cortex M core was selected as the microcontroller architecture. The ARM Cortex M core is currently the leading microcontroller core in the 32-bit microcontroller field based on the Embedded Market Study report.

The ARM Cortex M core is used in many manufacturer's microcontrollers. Among these manufacturers we have selected the STMicroelectronics. The reason for selecting ST is because the ST has the widest range of ARM Cortex M core microcontrollers, and it is also one of the leaders on the microcontroller market based on the Embedded Market Study report.

The ST also provides an integrated development environment support for its microcontrollers this IDE is called the Cube IDE. The Cube IDE (shown on Figure 5.1) helps to accelerate the development by providing graphically aided code generation for basic peripheral configurations (GPIO setups, low level peripheral setups: UART, SPI, etc), and by integrating the HAL firmware library and middleware functionality like RTOS support and USB device driver library.

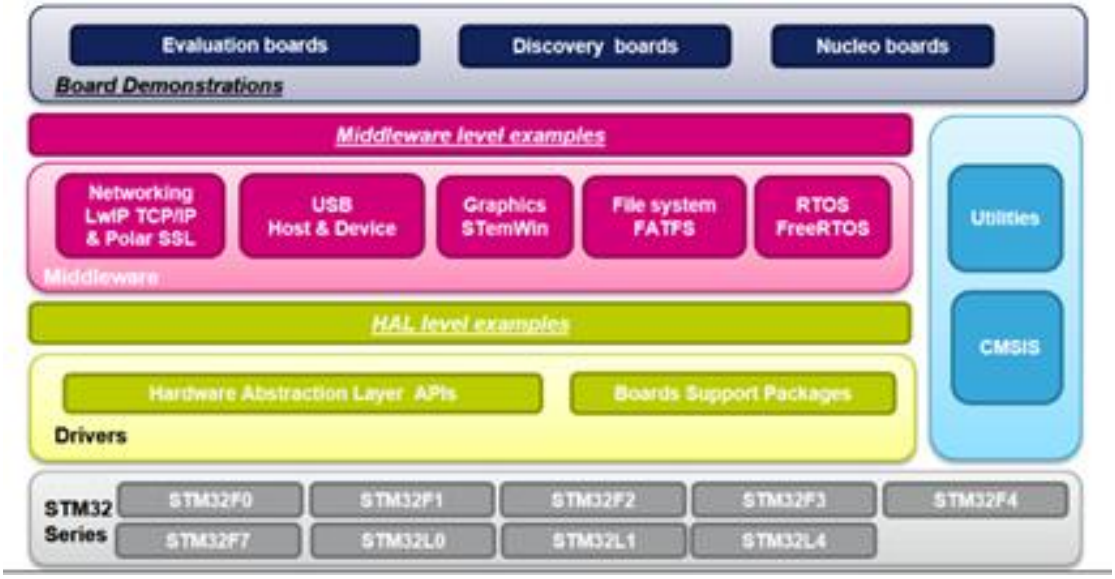


Figure 5.1: Layered model of ST CUBE

Among the ST's microcontrollers the STM32L4R5VGT6 was selected. This microcontroller is part of the recent generations of the STM32 series. Therefore, it has an improved performance at a significantly lower power as previous generations.

The internal structure is shown on the below figure.

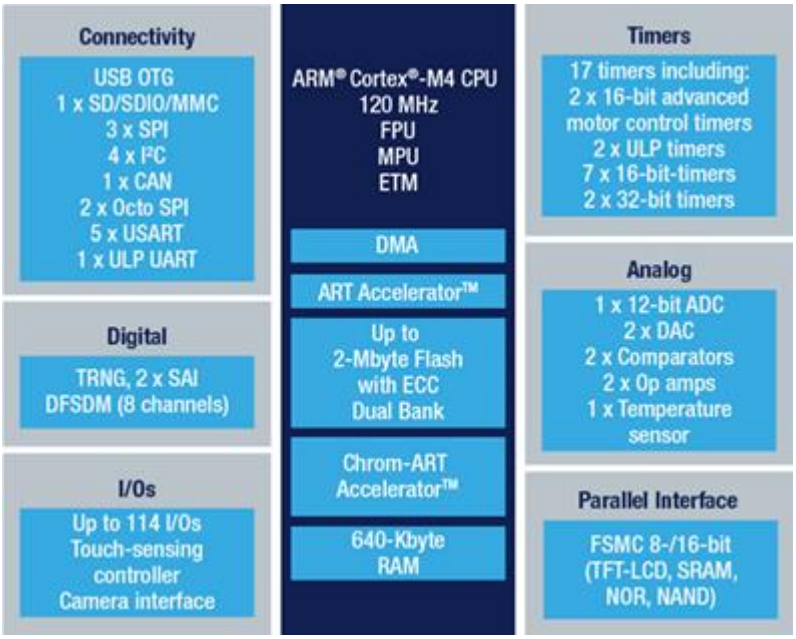


Figure 5.2: Internal structure of STM32L4R5VGT6 microcontroller

The STM32L4R5VGT6 has briefly the following main attributes:

- 32-bit ARM® Cortex®-M4 core
- Maximum 120 MHz CPU clock frequency
- 110 µA/MHz power consumption in “Run mode”
- 1MB FLASH
- 640 Kbyre RAM
- Complex and wide set of peripherals, which makes it capable for the BMMC functionality:
    - I²C, SPI, UART/USART, USB OTG
- Many analog input, and digital I/O-s
    - maximum 16 analóg, vagy 81 digitális láb

From the above list, it can be seen that the controller meets the peripheral requirements required by the BMMC: USB Device functionality, support for I2C, SPI communication for ICs implementing thermometer, power management IC, TPM chip and other board functionality. In addition, the selected controller has more than enough data and program memory, and its consumption can be described as extremely low, thus helping to keep critical thermal problems low.

An important consideration when choosing a controller was that a development board could be purchased for this controller family, if not for the specific controller, which allows you to try out some software features in parallel with the hardware design. Software development started on the development board shown in Figure 5.3.

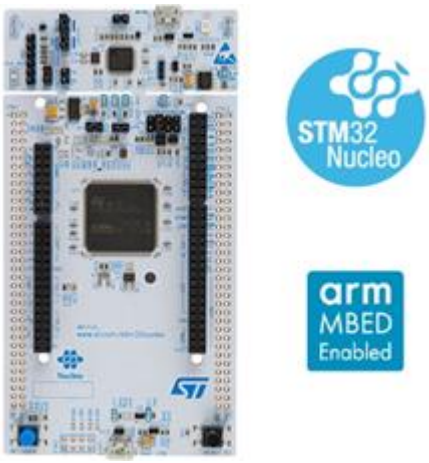


Figure 5.3: STMicroelectronics NUCLEO-L4R5ZI Developer Card

Although the STM32L4R5ZI microcontroller on the NUCLEO-L4R5ZI card differs from the controller we will use in both pin count and program memory, but the essential peripheral functionality like USB programing can be tried out using this board.

5.1.2 BMMC firmware

The Board Management Microcontroller firmware is under development. The development is based on an ST Nucleo L4R5ZI development board that contains the same series microcontroller as the final hardware. Since PCBs containing the real BMMC are not yet available, so the low-level hardware features only exist in an experimental form, and they have been tested in a development board environment. All of the functions need to be ported, integrated and tested into a real hardware setup soon.

The BMMC firmware consists of the following components:

- Basic card independent management functionality
  - Providing management communication via USB interface Communication between the MBMC and the
  - Turn the card's power on and off
  - Monitoring the current consumption of the card
  - Monitor card temperature
  - Provide low level debug functionality
  - Auxiliary option is to support for the functions provided by the TPM chip, which may include, for example, card authentication (verification of the card's origin from an authentic source)
- Card dependent management functionality
  - These functionalities are completely card dependent and will be finalized during the development of that card.
  - Example functionality: providing firmware upgrade capability for the card's main processor via AST 2500 BMC

Currently supported functionalities (BRAINE 1.0)

- CDC ECM communication, and LwIP TCP/IP stack is implemented into the ST Nucleo L4R5ZI development board. Linux based test system identify the development board as USB Ethernet interface and can communicate with it. TCP communication from the Nucleo L4R5ZI development board to the Linux test

system is reliable, but communication from the Linux to the Nucleo L4R5ZI has some instability which needs to be fixed.
- CDC VCP functionality is implemented to the L4R5ZI solution to provide a backup in case of long term CDC ECM instability.
- TMP112 chip based temperature monitoring API is developed but cannot be tested yet
- PMBus Power System Management Protocol is integrated to the BMMC software architecture to support power unit control and monitoring. Functionality needs to be tested in real hardware
- Analog measurement test has been done to simulate the current consumption management, but cannot be tested and integrated into a real hardware yet.

Development roadmap towards BRAINE 2.0

- CDC ECM communication instability needs to be solved: Mid of October
- As soon as hardware available PMBus based Power switching and monitoring should be tested: depends on hardware availability: Mid of October – End of October
- TMP112 based temperature monitoring should be tested: : depends on hardware availability: Mid of October – End of October
- Analog measurement based current monitoring should be tested: depends on hardware availability: Mid of October – End of October
- Card specific function support for CPU board: depends on hardware availability: End of October, mid of November

5.1.3 Board Management Controller firmware

Requirements, device selection and architecture decision.

We have analyzed the common BMC chip types, which offer dedicated system controller type of peripherals, during the early design.

Based on preliminary analysis we have created the following high level specification for the BMC module.

- Shall be capable to run Linux operating system
- Should have at least 1x USB 2.0 Host functionality (to connect with BMMC microcontrollers)
- Should have at least 1x USB 2.0 Device functionality to connect with CPUs, and act as remote keyboard and mouse
- Should have x1 PCIe lanes for remote video device functionality
- May support NCSI ans LPC interface
- Should have at lest 2x 1 G interfaces, both fiber ang 1 G BASE-T support preferred.
- Should have at least 6 x I2C peripherals to connect on board devices

Based on this ASPEED was selected as a primary manufacturer.

ASPEED devices are commonly used in server infrastructures, it is supported both by mainline e.g AMI BIOS (megaRAC) and open source OpenBMC firmwares.

Congatec evaluation board also contains AST2500 system supervisor, but it's firmware is unavailable due to special AMI licensing agreement.

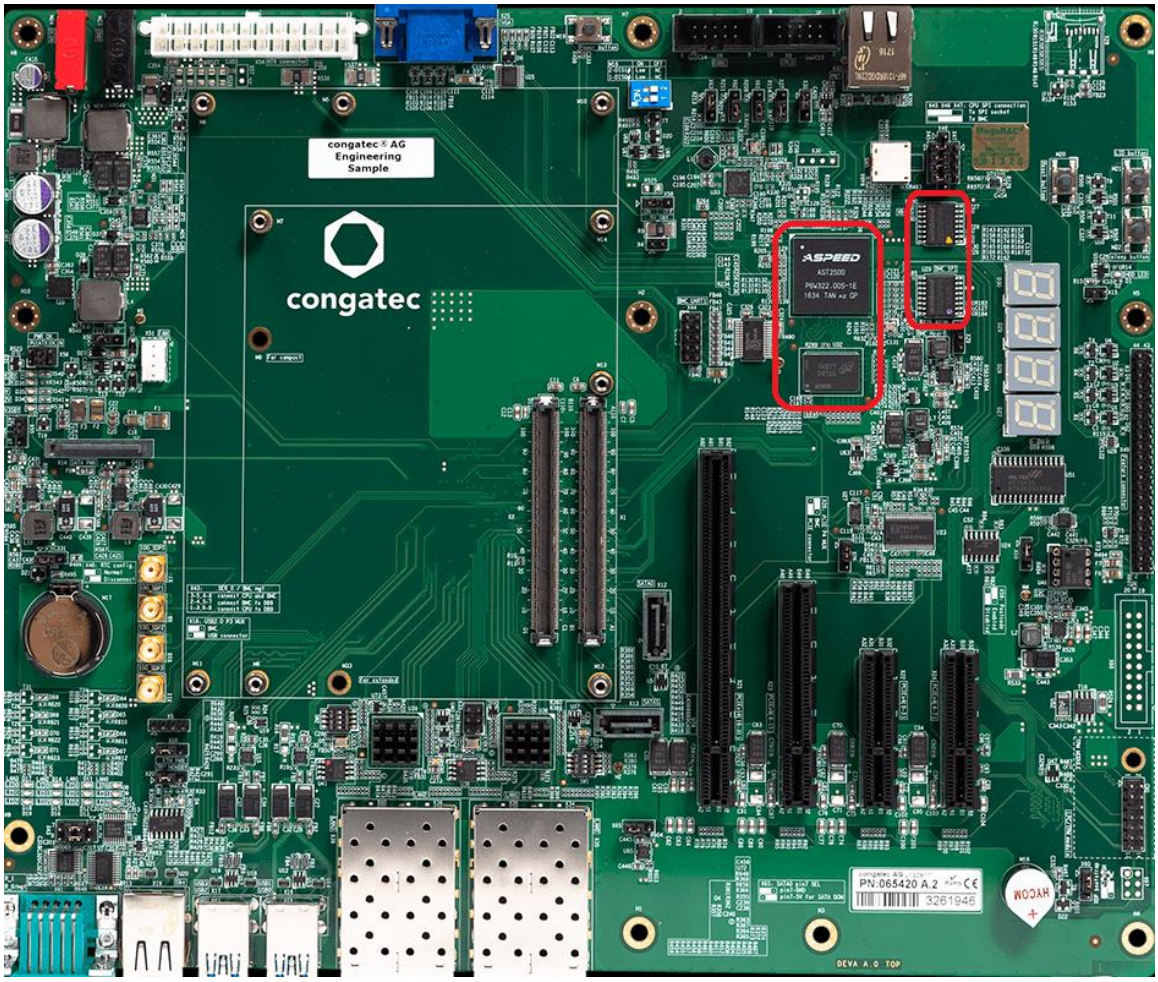


Figure 5.4: Congatec (X7AVAL) evaluation board,

AST2500 BMC and major peripherals marked red.

During the first year of the project Conagtec evaluation board was being used to develop the EMDC 1.0 BMC firmware functionalities, while interfaces towards the slots are quite similar to the one used on this evaluation platform.

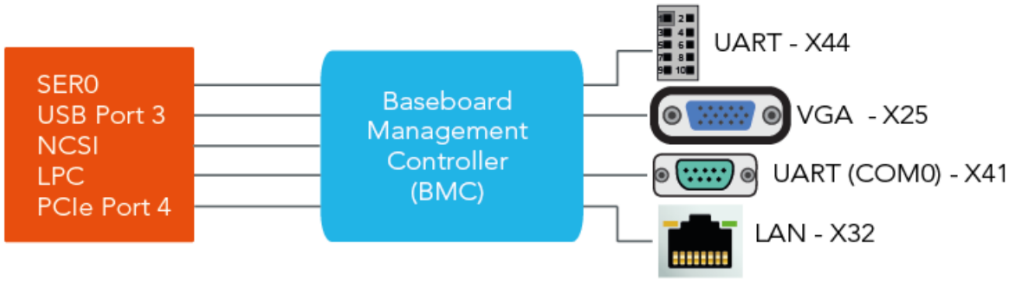


Figure 5.5: Congatec evaluation board BMC connection.

OpenBMC was selected which is widely used in several Open Compute Projects, including e.g. Facebook servers.

Special build target with custom device tree and device drivers were built to support Congatec evaluation platform connection. Development was performed in 3 major steps.

- Booting Linux image, allowing console connection
  - Definition of device tree / GPIO based on evaluation board schematics.
  - Method to flash Main and Video SPI Flash devices.
- iKVM (Keyboard Video Mouse connection) over Ethernet
  - Setting up Ethernet connection, peripheral USB (gadget) driver
  - Enabling embedded VGA BIOS and PCIe video target
- Host CPU enable, SOL (Serial over LAN), BIOS setup remotely.

OpenBMC features include the following:

- Host management: Power, Cooling, LEDs, Inventory, Events, Watchdog
- D-Bus based interfaces

- Web-based user interface
- REST interfaces
- Host SSH based SOL
- Host Remote KVM

Moreover,Redfish Compliance is under development too.

5.2 FPGA node firmware

5.2.1 Device selection

At the time of writing there are six major FPGA manufacturers: Xilinx, Intel, Achronix, Microchip, Lattice and QuickLogic. The first three offer high-end devices which are candidates to be used as high-performance hardware accelerators, as they offer a huge amount of logic and high bandwidth internal and external memory.

In the BRAINE architecture each node is a standalone node which has its own operating system and virtualized environment to execute workloads. Consequently, each node should have a CPU subsystem which executes the OS. In case of the FPGA node, this can be realized as a multi-chip solution using discrete CPU and FPGA, or as a single-chip solution where the FPGA also contains CPU core(s). The latter architecture has definite advantages regarding the connection between the FPGA logic and the CPU: it offers more flexibility and higher bandwidth. CPUs can be implemented in two ways: as the FPGA contains general digital logic, it can be used to implement soft-core processor; or the processor can be implemented as a hard-core during the manufacturing of the silicon (system on chip). Hard-core CPU cores not only offer better performance but also higher performance/watt, so they are the preferred solution for an energy efficient architecture.

Xilinx

Xilinx offers several FPGA families, including traditional FPGA devices (like Virtex UltraScale+) and SoC devices with hard-core CPU subsystem (Zynq UltraScale+ MPSoC and Versal ACAP). Although there are slight differences in the general FPGA logic architecture between these devices, these are not significant enough to be a decision factor: all Xilinx FPGAs are based on 6-input LUTs, flip-flops and adder logic.

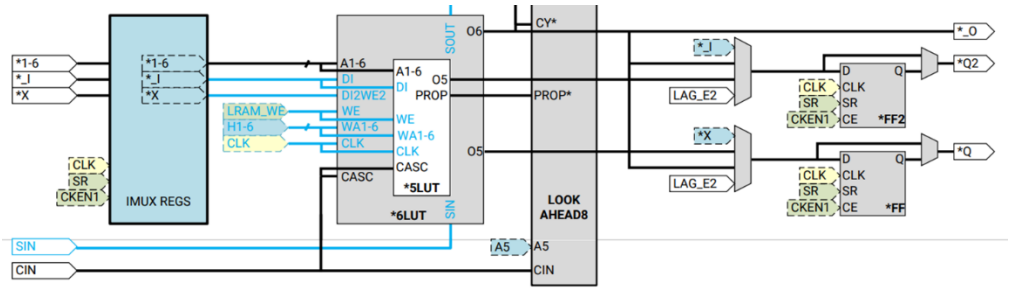


Figure 5.6: Internal architecture.

The difference is more significant in case of the integrated signal processing (DSP) blocks. While each FPGA family can implement multiply-add operation, the Versal DSP block considerably extends the possibilities by offering 32-bit floating point support and 3-dimensional vector dot product. The added functionality is useful in high dynamic range computations and low-precision multiply-accumulate operations like convolution and matrix operations.

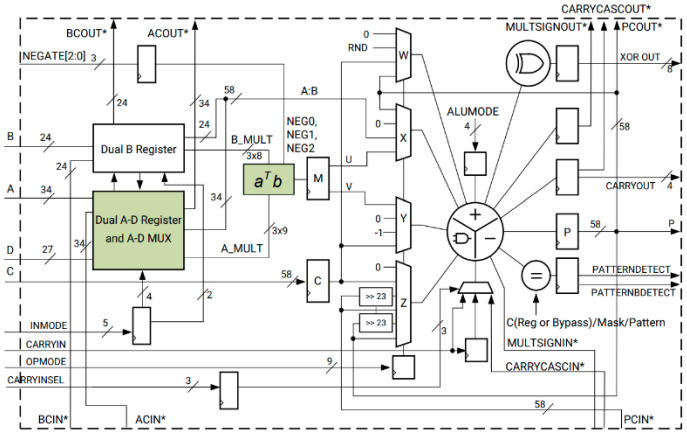


Figure 5.7: Matrix operations flow.

The Versa AI series takes heterogeneity one step further with the integration of the AI Engine Array.

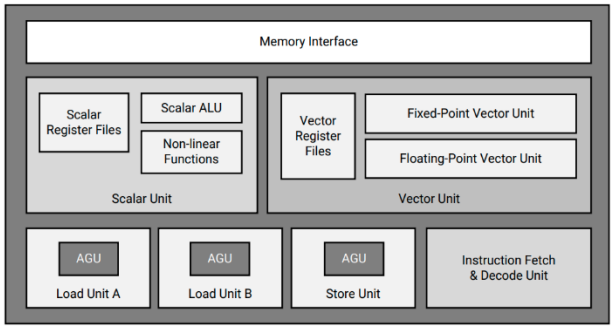


Figure 5.8: Versa AI blocks.

The AI Engine Array consists of several AI Engines arranged into a 2D matrix. Each engine contains a scalar RISC processor and a wide vector unit which can execute 128 8-bit MAC operations per clock cycle, making it ideal for executing convolution-based algorithms, like convolutional neural networks. Versal AI Edge devices also supports double-speed 4-bit MAC operations.

Internal memory resources are quite similar for the more recent families, consisting of small (18 kb/36 kb) BlockRAMs and 288 kb UltraRAMs. External memory interfaces are implemented differently. MPSoC devices offer one hard-core DDR3/DDR4 memory controller inside the CPU subsystem and additional memory controller can be implemented using the general FPGA logic. Versal devices, on the other hand, offers 1-4 hard-core DDR4 controllers.

On the connectivity side, most MPSoC devices offer PCIe Gen3 interface, 25G Ethernet and some devices also integrates 100G Ethrnet MAC. On the other hand, all Versal devices supports PCIe Gen4 and 100G Ethernet.

Intel

Similarly to Xilinx, Intel offers a wide range of FPGA devices, starting with the low-cost Cyclone series to the high-performance Startix and Agilex families. Higher-end SoC devices contain hard-core CPU subsystem with four ARM Cortex A-53 cores.

Architecturally the Stratix 10 and Agilex FPGAs are a bit different from Xilinx. They are based on adaptive logic modules (ALMs) which has an adaptive LUT, consisting of several smaller LUTs, and therefore allows more versatile mapping of various complexity logic functions. Depending on the actual design this may translate into better usage of the resources, but most of the time it does not make large differences, as both architectures are fine-grained.

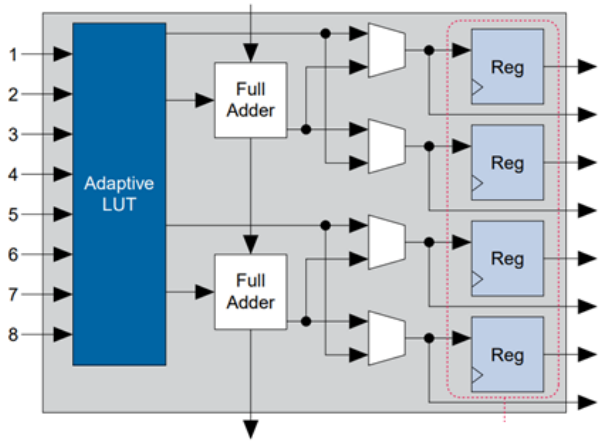


Figure 5.9: Adaptive LUT architecture.

While addition and substraction can be efficiently implemented using the ALMs, Intel also employs dedicated functional elements to execute more complex arithmetic functions. The Stratix DSP blocks can compute two 18x19 bit or one 27x27 bit or one 32-bit floating point MAD operation per clock cycle, while Agilex devices expands the possibilities with computing two 16-bit floating point MADs.

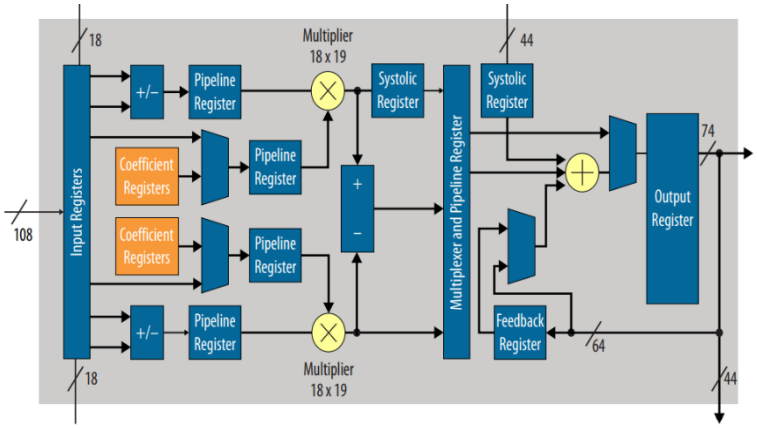


Figure 5.10: Agilex expanded computing.

Beyond the arithmetic elements Intel DSP blocks also contain small memory blocks which allows in-block storage of one of the operands, reducing the additional resource usage for certain applications.

The granularity of the on-chip memories is quite similar to Xilinx: all devices contain a large number of 20 kb blocks and some of them also has a smaller number of large, 18 Mb SRAM blocks. The fastest external memory interfaces in Startix 10 devices supports DDR4 memories at 2666 Mb/s. Agilex devices increase DDR4 data rate to 3200 Mb/s and adds support of 4400 Mb/s DDR5 memories. The highest end devices have stacked HBM memory which considerably increase the achievable memory bandwidth.

The speed of the transceivers ranges from 16 to 57.8 Gb/s; most Startix 10 devices support PCIe Gen3 x16 interface, while Stratix 10 DX and Agilex devices offer PCIe Gen4 x16. On the Ethernet side, 100G MAC is present in most devices.

To compete with the Versal AI series, Intel announced the Startix 10 NX family, which includes an additional AI Tensor Block, which is design to perform multiple MAC operations in parallel, using 4 or 8-bit integer or 12-16-bit shared exponent float or 32-bit float operands. Unlike the Xilinx AI cores these are dedicated hardware blocks within the programmable logic.

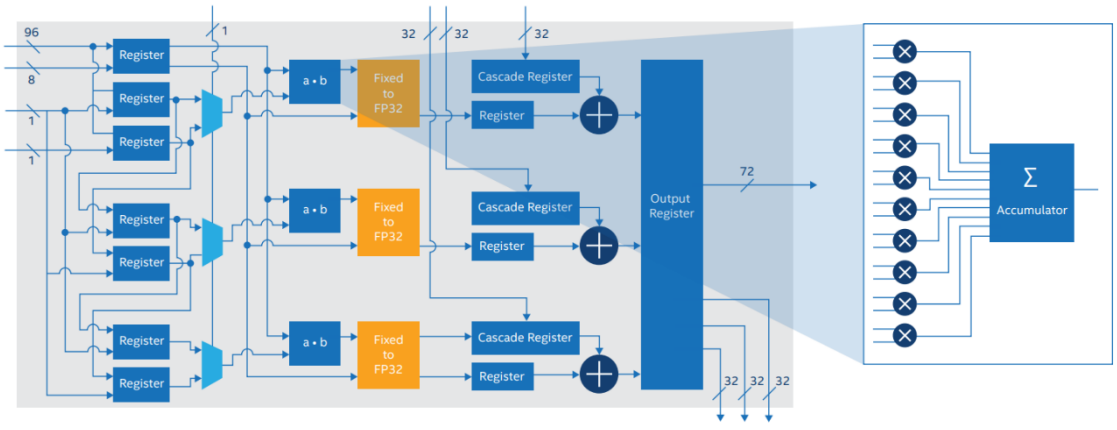


Figure 5.11: Dedicated hardware blocks within the programmable logic architecture.

The performance of the highest-end Startix 10 NX FPGA matches the performance of the Versal AI devices.

Achronix

Achronix offers only one, high performance FPGA family, the Speedster7t. The basic architecture of the programmable logic is quite similar to Xilinx: it is based on 6-input LUTs with dedicated arithmetic logic which allows efficient implementation of addition and multiplication.

The dedicated arithmetic units in the Speedster7t FPGAs – Machine Learning Processor (MLP) – are designed to perform MAC operations on various data types.

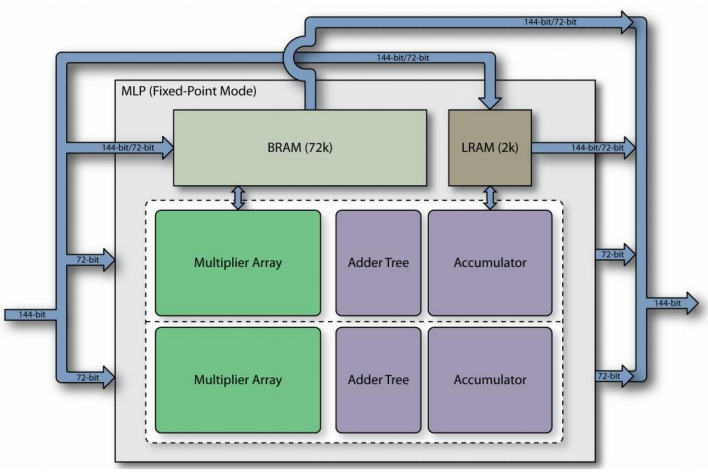


Figure 5.12: Architecture of the machine learning processor from Speedster7t.

Each of these blocks can perform 32/16/4 MAC operations using 4/8/16 bit integers or 2 MAC operations using 16 or 24-bit floats – offering similar capabilities to AI engines and Tensor Blocks. The blocks also contain a relatively large amount of storage, in the form of 72k and 2k memory blocks. The maximum performance is 30 TOPs using 8-bit operands, which is approximately one third of the competition.

One area where the Speedster7t FPGAs shine is external memory bandwidth: depending on the device the 1-2 DDR5 memory interfaces are complemented with 8-16 GDDR6 interfaces which offer 16 Gb/s data rate.

On the interface side the devices offer maximum of two PCIe Gen5 x16 interfaces and 8-32 100G Ethernet with 400G MAC, thus making it outstanding in this respect.

As the Speedster7t devices do not contain dedicated CPU subsystem, they are more suitable as a hardware accelerator attached to a CPU. Although a soft-core CPU could be used as an alternative, the performance of such cores is not sufficient to implement the complex software stack required in BRAINE.

5.2.2 FPGA firmware

BRAINE 1.0 does not include the FPGA node; therefore, all related developments are in progress and will be available in BRAINE 2.0. The pre-development is done on Xilinx Zynq UltraScale+ MPSoC (ZCU106) and Versal AI development boards (VCK190).

The FPGA firmware consists of the following components:

- Base FPGA HW architecture, which defines the external components connected to the FPGA and the internal structure.
- Petalinux operating system for the given HW architecture.
- Higher-level software:
    - Application-specific acceleration kernel.

    - Docker runtime which allows execution of Docker images with access to the FPGA accelerator.
    - Kubernetes device plugin to advertise FPGA resources.

Although the FPGA firmware is not part of BRAINE 1.0, the current development phase encompasses:

- Xilinx Zynq UltraScale+ MPSoC based system
    - Base FPGA design for ZCU106.
    - Petalinux operating system
        - Network boot.
        - Docker runtime.
        - Basic Kubernetes device plugin.
    - Acceleration function: neural network inference for pedestrian detection.
- Xilinx Versal AI based system
    - Limited base FPGA design for VCK190.
    - Base Petalinux operating system.
    - Acceleration function: accelerated matrix operation using general FPGA resources.

Development roadmap towards BRAINE 2.0:

- Docker runtime integration on VCK190 by Nov 2021.
- Kubernetes device plugin integration on VCK190 by Nov 2021.
- 10G/40G Ethernet integration into base FPGA design on VCK190 by Jan 2022.
- Acceleration function on VCK190: neural network inference for pedestrian detection by Feb 2022.
- Dynamic function exchange trial on VCK190 by March 2022.

5.3 Software components for the EMDC programmable switch

This section describes the software components developed within BRAINE to be supported by the EMDC to enable advanced networking and AI processes at the edge.

In particular, the networking components descriptions are complemented with the research studies that designed and employed specific extensions published as advance with respect with the existing state-of-the-art

5.3.1 C2.17 - P4 programs

The BRAINE edge solution leverages on a P4-based programmable switch for advanced networking functionalities.

The P4 technology is still in its infancy and as of today only limited P4 functions are actually implemented in ASIC. For this reason, two versions of P4 programs have been developed for BRAINE 1.0.

The first version targets the Mellanox Spectrum1 ASIC that will be included in the BRAINE EMDC HW developed by Task T2.1. This version encompasses:

- P4-based layer 2-3-4 forwarding
- Traffic steering

The second version targets P4 nodes (e.g., bmv2) that support the full list of standard P4 capabilities.

This version encompasses:

- P4-based layer 2-3-4 forwarding
- Traffic steering
- In-band telemetry,
- Post-card telemetry,
- 5G acceleration for UPF,
- In-network stateful feature extraction for cyber security

Development roadmap towards BRAINE 2.0:

The following functionalities will be added within the BRAINE P4 programs for BRAINE 2.0:

- Stateful P4-based telemetry aggregation by Nov 2021
- Inter-EMDC P4 postcard telemetry exchange by Jan 2022
- P4 INT/postcard telemetry collector by March 2022
- Specific support for the applications of use case 2

P4 Postcard Telemetry

The P4 switch has the capability to extract selected packet metadata conveying monitoring information, such as the packet timestamp, the packet hop latency and other switch parameter states. In this P4 program, we target the hop latency as a key monitoring parameter. The latency experienced at each switch is a key metric for EMDC network infrastructures SLA. Thus, instead of using external generator/analyzers, we rely on extended P4 switches to extract latency directly. The implementation is based on postcard-based telemetry, a method used to extract features at each network node and provide results to a collector using a dedicated monitoring interface. The postcard-based telemetry has a simple implementation effort, since no complex in-band telemetry (INT) solutions are required. The program is composed of two main pipeline controls: ingress and egress. In the ingress pipeline, standard forwarding table is implemented. Moreover, a further flow table is deployed to match the traffic flows to be monitored. Upon match, the executed action is the cloning of the packet towards the monitoring interface. In the egress table, a match on the cloned packet triggers the Report creation action. In this action, the old packet is replaced with a new Ethernet-IP-UDP stack. Moreover, the Postcard Telemetry Report, as defined by the P4.org INT specifications, is added. With respect to the format specified in the INT document, the *flow_id* field has been extended and processed. This value is added as flow entry parameter in the match performed in the ingress pipeline, conveniently set by the SDN controller in charge of enforcing switch flow entries. The value is mapped in the Report packet. Thus, each Report conveys the information related to the monitored flow (*flow_id*) and the originating switch (*switch_id*). The format of the Report packet is shown in Figure 5.13.

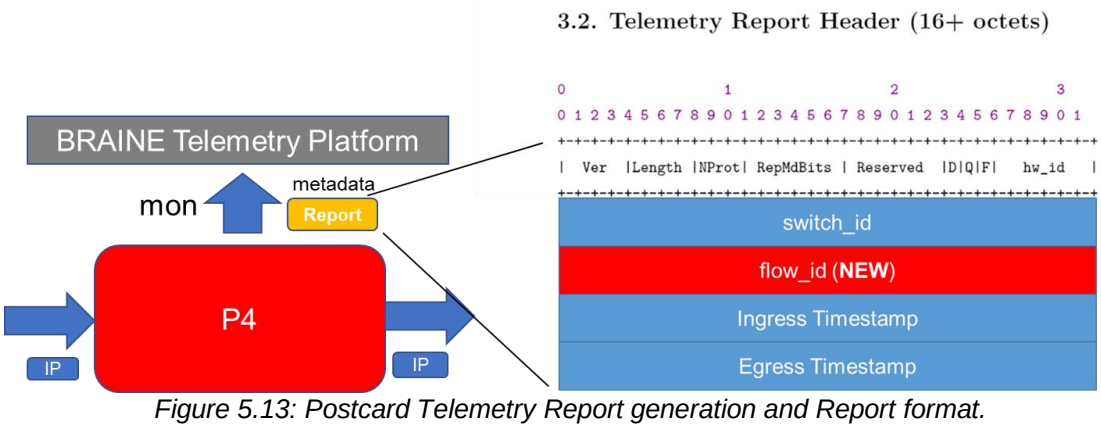


Figure 5.13: Postcard Telemetry Report generation and Report format.

This implementation allows to activate the postcard telemetry on demand on the desired flow, as specified by the controller. This way, the association between flows is orchestrated by the SDN controller. The results obtained on the monitoring interface is reported in the Wireshark capture of Figure 5.14. The considered Report packet is related to switch_id 0xb (10), to flow-id 1 and conveys the input and output timestamps. Similar implementations include also the direct value of hop latency, computed as the difference between the two timestamps.

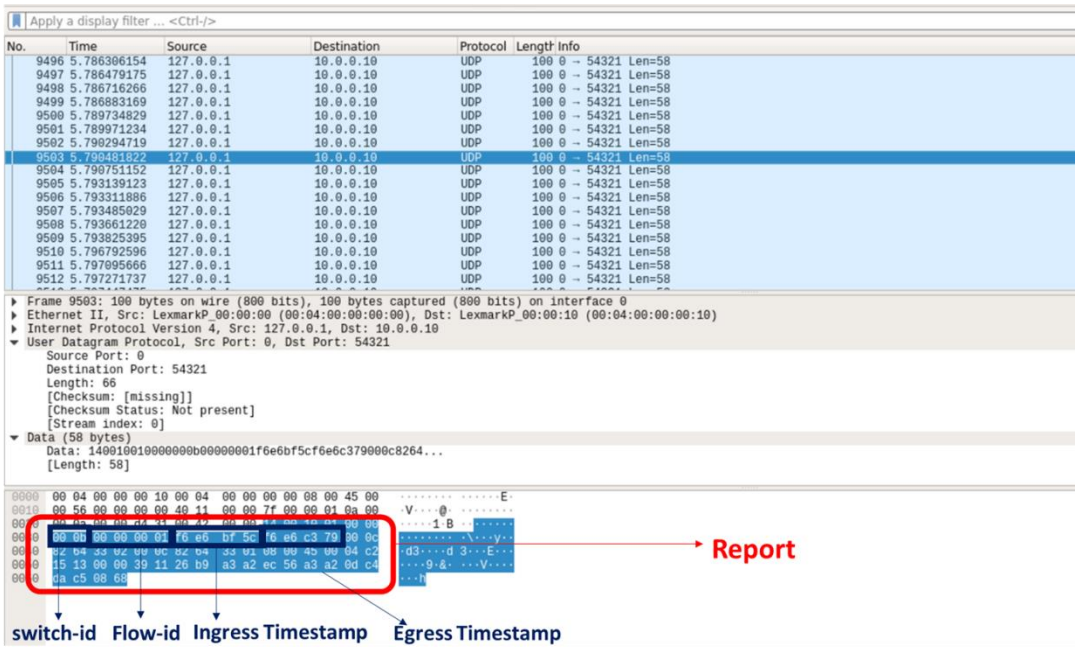


Figure 5.14: Postcard Telemetry Report generation and Report format.

It is worthwhile to note that the Postcard Telemetry granularity is at the packet level, meaning that each considered flow packet is mapped to a single Postcard Telemetry Report, with no data aggregation. Aggregation is performed at the Telemetry Collector level, that implements a Metric Exporter tool pushing data to the BRAINE Telemetry and Monitoring Platform, developed by WP4. Aggregated data are the average and the maximum latency experienced by each flow at each considered switch in a time window composed by a configurable number of packets. Figure 5.15 reports the relationship between Reports and Monitoring data exported to the BRAINE platform. The P4 Postcard Telemetry implementation has been published in the DRCN21 conference [PAO-DRCN21] and used in other works referenced by the project.

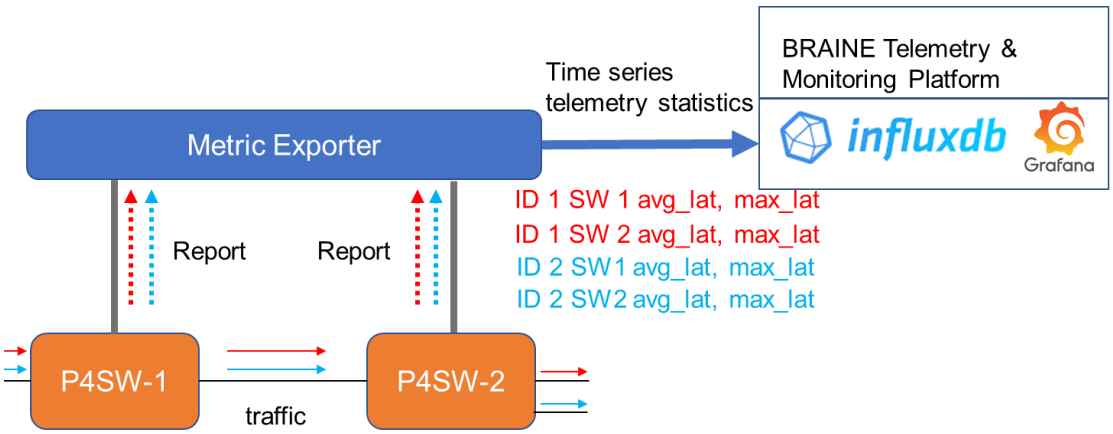


Figure 5.15: Postcard Telemetry Report generation and Report format.

P4 In-band Network Telemetry (INT)

INT solutions employ the metadata communication directly inside the traffic packet, resorting to additional and dedicated extra-headers.

The first P4 INT implementation has been developed in the context of serverless edge steering. Latency-critical traffic are monitored using a per-app per-node segments. Ultra-low latency serverless app is dynamically deployed and migrated between two EMDC in less than 10ms. Switches run the INT 1.0 P4 code. INT is initiated by a source node, updated by each transit node and terminated by the sink edge NIC. The INT extra header, added to each matched packet over the TCP layer, includes a shim header (SH), inserted only by the source node, and a variable header (VH), added by each involved INT node. The SH carries global INT information (e.g, INT type, next protocol type, length), needed by the sink NIC during the INT header decapsulation. The VH is inserted by each INT node and encloses per-node metadata (i.e., switch id, hop latency, egress timestamp). The sink NIC is responsible for removing the whole INT extra header thus restoring the original packet. Moreover, it enables the generation of an INT Report over UDP, to be consumed by the TS. The INT Report structure is reported in Figure 5.16. Besides general header fields, the switch id identifies the Report node generator (i.e., the INT sink NIC) and the ingress timestamp according to the local generation time. In addition, it includes all the VH collected at each INT-enabled node. This way, the Telemetry System (TS) may receive Reports from different sink node and, inside each sink, from different application flows with the detailed metadata list of each crossed node, including the sink NIC. This allows the TS to store and provide actual per-node and per-application latencies.

Figure 5.16 shows the testbed used to assess INT. The packet-optical switch performs P4-based forwarding, while S0 (switch id 10) and the NICs (switch id 12 and 14) implement P4-based INT over the BMv2. INT Reports, shown in the captures collected at N1, are generated by NICs and provided to dedicated TS collector servers.

Figure 5.16 shows the application end-to-end latency variation measured by TS resorting to INT data for traffic reaching E1. Upon latency increase, the system moves the application to E2 and latency is reduced. The placement takes place in around 10ms without experiencing service outage.

Full results have been published in a dedicated OFC paper [PEL-OFC21].

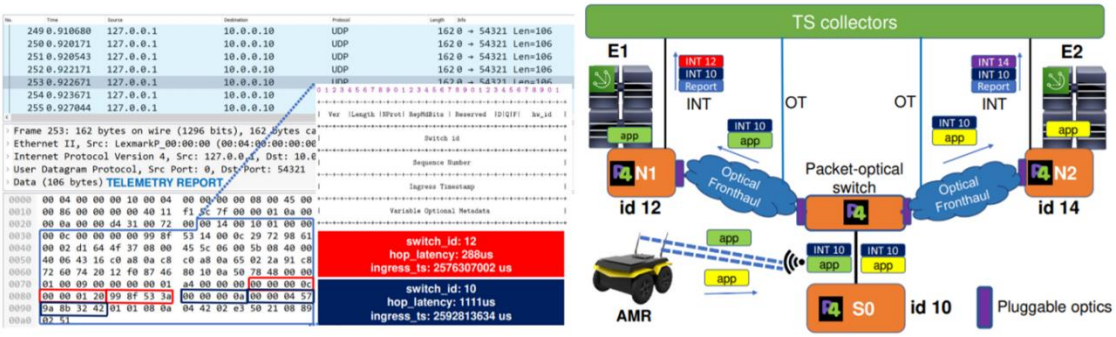


Figure 5.16: INT Report format and serverless application testbed.

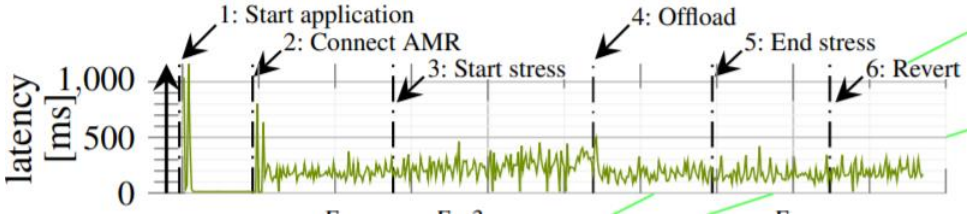


Figure 5.16: INT-based latency monitored at the application layer.

INT extensions to the User Equipment for decentralized steering

Current INT implementations are able to provide detailed intra-switch latency information. Such information is sufficient to compute end-to-end latency only in specific network scenarios. For example, in a single-layer packet-switched wired network, latency variations typically occur inside the switch due to packet processing, packet queuing, congestion, contention and buffering. However, end-to-end latency is subject to significant variations in complex scenarios (i.e., multi-segment, multi-layer, hybrid wired-wireless) that include a variation of link latency. Link latency computations are not computed by INT implementations and may impact end-to-end latency from the User Equipment up to the EMDC or the cloud data center.

To enable telemetry-driven in-network steering without the intervention of the SDN controller, we propose an overall extension of the INT architecture that includes the following novelties:

1. the inclusion of the UE in the INT domain;

2. the extension of the INT mechanism providing backward INT information with actual end-to-end latency and geolocation metadata;

3. automatic decentralized steering strategies triggered by extended INT awareness.

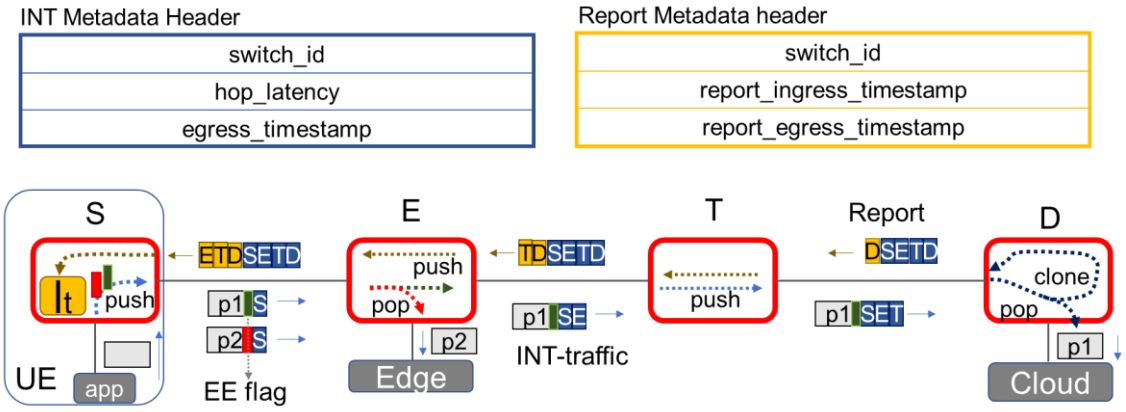


Figure 5.17: Extended INT metadata and UE-triggered automatic steering.

The first innovation relates to the INT domain enlargement, involving the UE. The inclusion of the UE is performed by implementing an SDN P4 programmable switch inside the UE acting as a service app. A real implementation may rely on a lightweight virtual container programmed to process only data generated and received by the considered application. The switch, co-located at the UE, is programmed to act as telemetry source node (i.e., responsible for pushing the shim header and include the UE metadata). The extended INT details are shown in Figure 5.17 and Figure 5.18, respectively. In both solutions, the INT domain starts at the UE acting as source node (node S in the figures) and terminates at sink node D, co-located with the data center or the edge node, responsible for removing all the INT headers and route the traffic transparently to the application server. Along the end-to-end path, transit nodes T simply add their own INT headers. Among these nodes, special steering switches E are extended to support steering functions to the attached local edge node and, in the case of local edge steering, similarly as node D, are programmed to pop INT headers to provide traffic transparently to the local edge server.

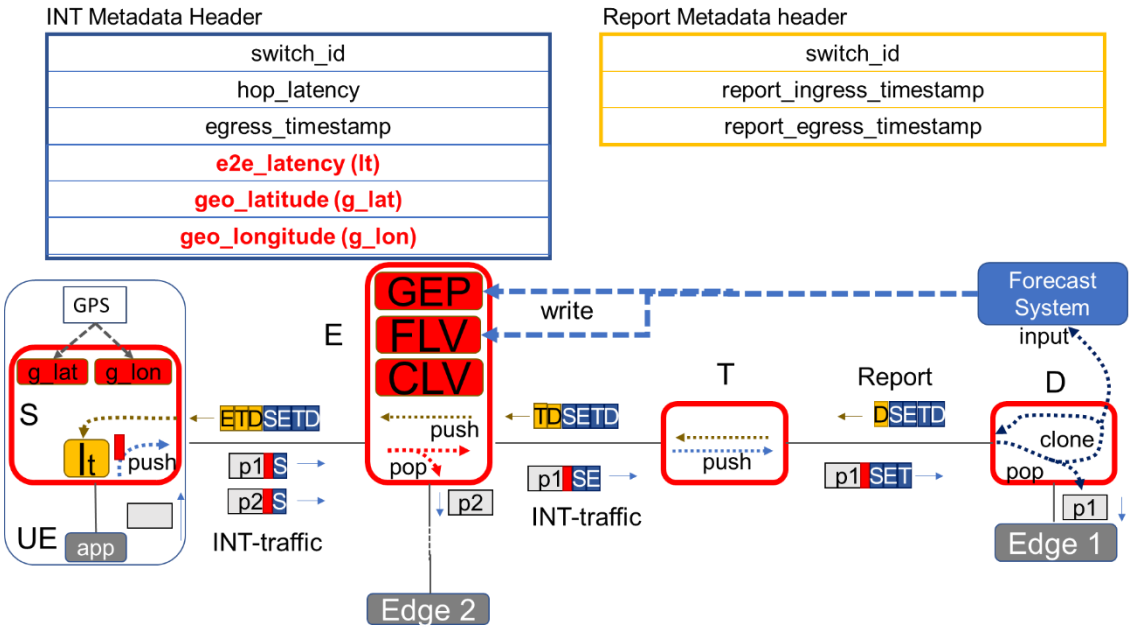


Figure 5.18: Extended INT metadata including UE geolocation information.

The second innovation, allowing the end-to-end latency computation, envisions an extended INT Report packet processing. Typically, the Report packet is generated as out-of-band packet destined to monitoring plane analysis, reporting the hop latency and egress timestamp information retrieved along the standard INT path forward direction (i.e., S, E, T, D blue headers of Figure 5.17). First, node D is programmed to generate and recirculate a Report packet in the backward direction to the source (i.e., the UE). Second, all the switches are programmed to push the ingress and the egress timestamps related to the Report transit in the backward direction up to S. Both forward and backward timestamps stored in the Report packet (i.e., blue and yellow headers referred to S, E, T, D nodes in Figure 5.18) allow all the switches to compute the link latency experienced by the packet. Thus, the proposed method computes the link latency without the need of having all the nodes synchronized, since all the latency contributions are computed as differences between timestamps collected at the same switch. In the edge-to-edge scenario, following Figure 5.18, besides local timestamps and hop latencies, the UE adds to packet p2 its INT header pushing its current position (i.e., the latitude and longitude pair, *geo_latitude* and *geo_longitude)* and the latency experienced by the previous packet p1, stored in the local UE P4 switch instance. The current space position information is retrieved from internal UE sensors (e.g., GPS) and stored inside two dedicated registers (*g_lat* and *g_lon*). This allows to store in the same traffic packet the combined synchronized information of latency and geolocation, that is utilized for decentralized steering decisions.

The third innovation, driven by the extended INT, resides in the awareness acquired at the different INT nodes, utilized to take forwarding/steering decisions based on the INT information. Two decentralized steering are proposed.

In the UE-suggested steering (UES), the end-to-end latency is stored at the UE and application traffic Application traffic is generated by the UE and INT-tagged by S with a novel INT shim header flag, named Enable Edge (EE). The flag reports the state of the experienced latency l_t , compared with a pre-defined target bound latency threshold. The flag is initially set to zero (packet p1 with green EE flag) triggering E to steer traffic towards the cloud, and is turned to 1 when l_t > ITH. The EE flag is inspected only by switch E. In the case EE is set to 1 (packet p2 with red EE flag), the node automatically steers traffic to the local edge node while popping INT headers. In order to preserve network stability and avoid unexpected steering events, the UE S node is programmed to set the EE flag

when a certain threshold number of packets experience l_t > lTH. This is realized in the switch by resorting to a dedicated P4 counter. Figure 5.19 shows the results obtained in the CNIT testbed. In particular, the extended INT carrying out the source hop latency and the EE flag is shown and the experimental results of the automatic steering at the local EMDC, enforced after 5 violations of the end-to-end latency threshold, occurring after 500ms elapsing from the first violation. The five violations hysteresis have been programmed in the switch to avoid steering instability.

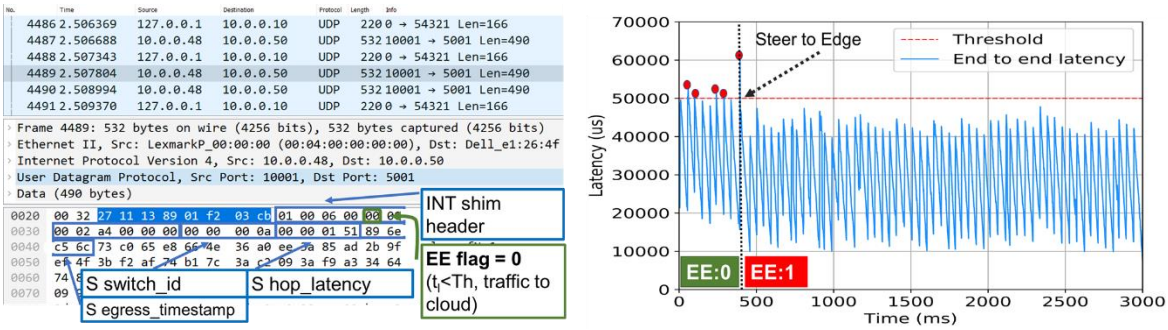


Figure 5.19: Results: extended INT packets and automatic decentralized steering at the edge switch.

In the forecast switch-triggered steering (FSS), the E node resorts to a forecast system that computes the future positions of the UE, and based on such information written in specific P4 registers, it automatically steers the traffic to the closest edge node. The forecast system is implemented in WP3 and details are reported in D3.2 deliverable.

The full details and results of this implementation have been published in a JOCN paper [SCANO-INT-JOCN21].

5.3.2 C2.18 Programmable switch operating system

The operating system of the P4-based programmable switch part of the BRAINE edge solution depends on the underlying hardware. For this reason, multiple OS versions have been considered for BRAINE 1.0.

The first version is designed for the Mellanox spectrum1 ASIC as standalone, off-the-shelf switch currently implemented in the BRAINE testbed at CNIT Lab. This version is based on ONYX over ONIE, augmented with a specifically designed P4 docker container released by Mellanox.

The second version targets the Mellanox spectrum2 ASIC as standalone, off-the-shelf switch. This version is based on SONiC over ONIE, augmented with a different P4 docker container released by Mellanox. So far, this version has been tested on Spectrum 1 switch. Given the different underlying ASIC and APIs, only limited P4 capabilities and pluggable module configurations have been implemented over SONiC.

SONiC is an open-source network operating system (NOS) that can run on switches from multiple vendors and ASICs with the support of the Switch Abstraction Interface (SAI), SONiC NOS architecture is utilizing multiple containerized components that enable NOS extensions like programable P4 applications.

SONiC's modular and extensible design can accelerate innovation and support customers adding new network applications that will be fully integrated with the SONiC system, P4 applications can use this extensible design to integrate into the SONiC NOS. SONiC NOS has a rapidly growing ecosystem and was adopted in huge data centers and cloud providers.

Development roadmap towards BRAINE 2.0:

- Apply SONiC over ONIE over Mellanox spectrum2 standalone for full support of both pluggable modules and P4 programs (by Feb 2022)
- Mellanox spectrum1 for EMDC: SONiC (preferred, ONIX as alternative OS) over ONIE

ONYX

The first OS solution is based on the proprietary Mellanox Onyx operating system. Mellanox has released, for the scope of the project, a dedicated P4 container to be run inside Onyx. The container includes the environment exposing the compiled P4 program to the internal P4 SDK, responsible of mapping the compiled P4 in a set of C++ files that are used to deploy the desired pipeline at the ASIC level. The capabilities provided by the compiler and by the SDK are related to the Spectrum-1 capabilities, that include all the OpenFlow-based packet processing (e.g., routing, forwarding, address translation). Additional features reside in the flexible parser and deparser sections, allowing the dissection and the processing of extra headers.

While supporting flexible parsing/deparsing capabilities and having ability to estimate flow bandwidth, Spectrum-1 platforms have a limited support of the rest needed for network telemetry metadata like flow latency and ability to modify packets inline based on flexible network format in order to generate in-band telemetry format. Thus, Spectrum-1 platforms are not supporting full required network telemetry. Required capabilities are supported on Spectrum-2 and above platforms running SONiC NOS.

Figure 5.20: Spectrum-1 Onyx over ONIE.

SONiC over ONIE

SONiC (Software for Open Networking in the Cloud) is an open-source network operating system already deployed in production intra-DC networks and it is also considered a strong candidate to control packet and packet-optical nodes although some operational extensions are needed to fill the existing architectural gaps.

The packet-optical node Operating System architecture based on SONiC is better detailed in Figure 5.21. It consists of a Mellanox/NVIDIA SN2010 Ethernet switch running SONiC operating system over ONIE. Besides the basic components (i.e., soniccfggen, syncd, swss, pmon and the redis database) SONiC also includes the P4/P4 Runtime docker container and the NETCONF docker container. Using these two containers two parallel communication channels are established between the packet-optical device and the packet-layer SDN controller to enable configuration of resources. Specifically, the P4/P4 runtime docker container is an experimental P4 implementation developed and provided

by Mellanox/NVIDIA, allowing the SDN controller to configure the packet resources using P4 Runtime. The full implementation has been described in two published OFC and ECOC papers, including a demo session showing the hierarchical coordination between the packet and optical controllers [SGAMBELLURI-OFC21], [SCANO-ECOC21].

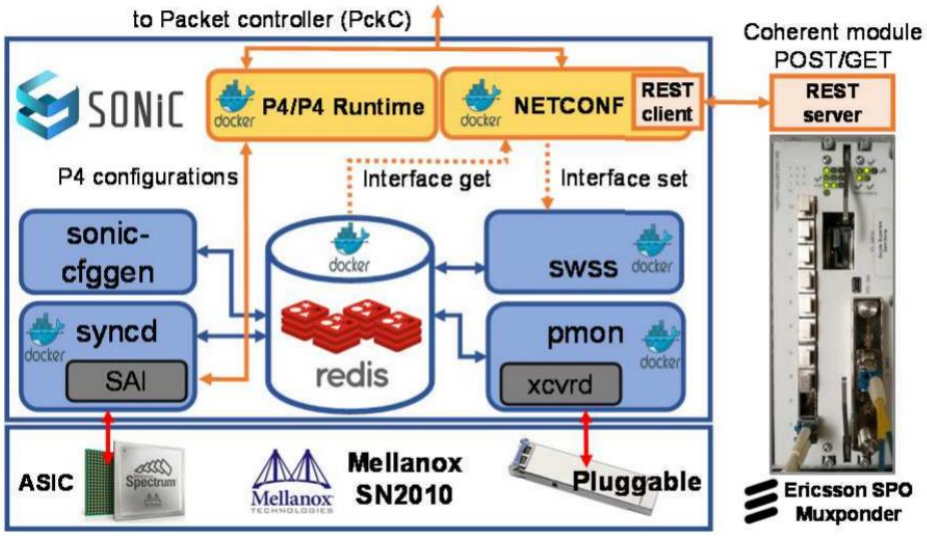


Figure 5.21: Implementation of SONiC over Spectrum-1 and P4/P4Runtime support.

5.3.3 C2.19 - Switch connectivity and monitoring through pluggable modules

The operating system of the programmable switch part of the BRAINE edge solution, besides supporting P4 programs, requires supporting the interconnection to other EMDCs as well as to the remote cloud (medium/long reach transmission). To guarantee unprecedented network awareness at the control/orchestration level, new monitoring capabilities of port interfaces and transceivers have been implemented. In particular, for BRAINE 1.0 an open-source NETCONF Agent supporting OpenConfig YANG model has been enhanced to support:

- Added the support of coherent pluggable modules within the agent data store
- Updated OpenConfig YANG model to 2021 release
- Implementation as docker container within SONiC operating system
- Ownership segregation. Implemented exploiting the Network Configuration Access Control Model (NCACM) solution as detailed in RFC 8341. This way, the Optical Controller in charge of the metro optical network is provided with writing rights on the optical parameters and read-only rights (including enabling notifications upon subscriptions) on packet parameters. Similarly, EMDC Controller is provided with writing rights on packet parameters and read-only rights on optical parameters.
- Monitoring through SONiC

Development roadmap towards BRAINE 2.0:

- Validation in metro scenario of the NETCONF Agent supporting OpenConfig model updated for pluggable modules (by Oct 2022)
- Peer-to-peer optical telemetry for pluggable modules including AI embedded processing of collected data (by Apr 2022)

Integration of pluggable modules in packet optical nodes

Replacement of standalone transponders with pluggables modules in the packet devices directly connected to the optical network drives significant benefits in terms of CapEx, power consumption and occupied space in central offices. Furthermore, it enables a tight integration between packet and optical networks, which is of special interest as transport is dominated by Ethernet and IP traffic. For example, a single packet switch can provide both intra-data center (DC) traffic aggregation and, thanks to coherent pluggables, effective DC-to-DC interconnection. However, controlling packet-optical solutions requires a complete operating system that is much more complex than traditional NETCONF/YANG software agents employed in standalone transponders.

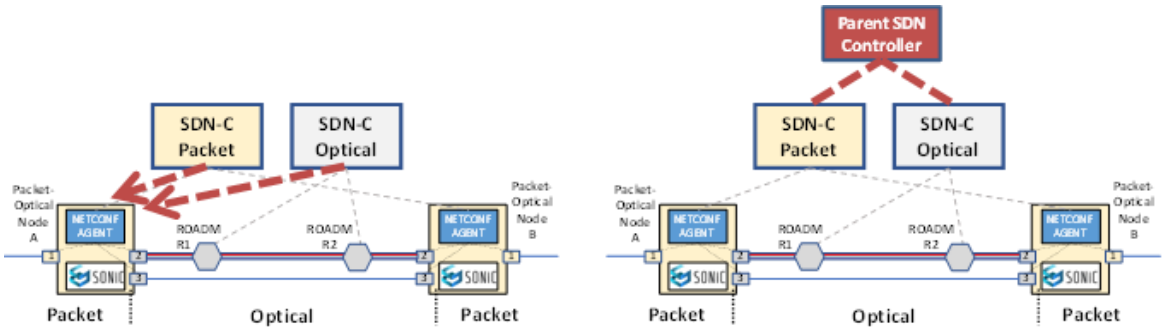


Figure 5.22: Coordinated packet-optical control (left); Hierarchical packet-optical control (right)

Two solutions are implemented. The former exploits coordination at the agent connected to both controllers. The latter connects only to PckC and a hierarchical controller is introduced.

The first reference scenario is illustrated in Figure 5.22 (left). Packet-optical nodes are equipped with pluggable transceivers. In large metro networks, a single controller with visibility on both packet and optical resources is hardly implementable due to scalability issues. Two controllers are then typically considered: an Optical SDN Controller (OptC) in charge of the optical transport network and a Packet Controller (PckC) supporting Layer 2-7 configurations. Traditionally, each SDN controller has full visibility on all components and software modules of every controlled network element. For example, OptC, besides configuring the ROADMs, is responsible for the configuration of the line interfaces of transponders. However, in the considered scenario, transponders are replaced by pluggable modules equipped within the packet-optical nodes, traditionally controlled only by PckC. That is, two different controllers need to concurrently operate on packet-optical nodes. Thus, a proper workflow needs to be defined to enable the SDN agent of the packet-optical node to coordinate the operations imposed by each controller. Indeed, without proper coordination, complex multi-layer operations, such as recovery upon soft failure, would lead to management conflicts on the packet-optical nodes as well as to potential traffic disruptions. Both controllers are connected to the agent. In particular, the agent exposes the whole YANG model including packet and optical descriptors. To avoid non-standard and complex peer/hierarchical operations, the two controllers do not communicate each other to coordinate their actions. Instead, they leverage on the proposed workflow to avoid conflicts and guarantee segregation of control. Ownership segregation has been implemented exploiting the NCACM solution as detailed in IETF RFC 8341. In particular, the OptC is provided with writing rights on the optical parameters and read-only rights (including enabling notifications upon subscriptions) on packet parameters. Similarly, PckC is provided with writing rights on packet parameters and read-only rights on optical parameters.

An alternative approach is based on inter-Controller communication. With this solution, packet-optical nodes only interact with the PckC, as shown in Figure 5.22 (right). In turn, the PckC is enabled to configure optical parameters by proper interaction with the OptC. The solution includes the extension of the T-API connectivity service request to include the pluggables relevant parameters. Specifically, the T-API interface has been extended to include the pluggable optical description that may have an impact on the physical impairment aware path computation performed by the optical controller. The whole inter-controller workflow has been designed in WP3 and includes a parent controller in charge of coordinating provisioning and recovery steps of a packet-over-optical intent. To effectively perform impairment-aware optical path computation, the OptC must be aware of pluggable supported features (e.g., supported modulation formats, FECs, operational modes). To this goal, two cases can be considered. In the first case, the type and features of the pluggable modules are included in the set of static information related to inter-layer connectivity, that are manually loaded at the OptC. In the second case, the type and features of the pluggable modules are discovered by the PckC. In both cases the PckC forwards the pluggable details to the Parent Controller, that, in turn, forwards them as part of the T-API based connectivity request to the OptC. The latter case is adopted in the solution, it enables automatic discovery of pluggables but requires additional parameters to be exchanged among controllers for each request, even if the information is quasi static. When the PckC applies the connectivity configuration two different configuration messages are received by each packet-optical node. The first is processed by the NETCONF agent that results in the activation/configuration of the pluggable; the second is processed by the P4/P4 Runtime agent that configure the P4 pipeline to enable a bidirectional connectivity between tributary ports and the pluggable.

The two solutions for the packet-optical node agent have been detailed in two published papers at OFC [SGAMBELLURI-OFC21] and ECOC [SCANO-ECOC21] conferences, including a demo session.

Peer-to-peer optical telemetry

Optical telemetry services, defined as augmented real-time monitoring streaming services, are designed on the degree of optical network disaggregation. In full disaggregation, telemetry is enabled between the SDN controller, the whitebox controllers and the controlled device agents, allowing hierarchical architectures. For example, a disaggregated node (e.g., ROADM) may be delegated to run local telemetry to controlled device agents and report the results to the control/monitoring plane.

Figure. 5.23 shows a telemetry service for a disaggregated network. Telemetry subscriptions are performed by the control plane (i.e., the SDN controller) to monitor a set of devices crossed by a lightpath. In the figure, lightpath La quality of transmission is monitored through telemetry monitor streams generated at La xPonder (e.g., OSNR, Pin, BER) and at intermediate ROADM2 (e.g., Pin and Pout). The subscription set may be enlarged, e.g., including xPonder of a co-routed adjacent Lb that may interfere with La. Streams are collected by central collectors (step 1), processed by AI-based Analytics Handler and, in case of deviations, feedback is provided to the controller (step 2) in order to react properly (step 3) with recovery procedures (e.g., elastic operation, re-routing), also including fine tuning (e.g., launch power, signal adaptations, central frequency offset).

Typically, if compared with traditional 15-minutes averaged statistics, telemetry with sampling period in the order of 1sec is enough to monitor effectively malfunctioning, soft failures, even though a number of open ROADMs platforms support up to 50ms telemetry update interval. This means that, with respect to packet-switched networks running in-

band telemetry where the service rate scales with data plane traffic rate, in optical telemetry a large number of streaming sessions may be easily handled even by a central collector.

The proposed Peer-to-Peer Telemetry (P2PT) is conceived to bring network awareness at the next-generation (pluggable) optical card, possibly equipped with lightweight storage, processing, and AI (e.g., GPU) resources. P2PT may be activated upon card configuration (i.e., at provisioning) or on-demand upon notifications during soft failures. The purpose of P2PT is twofold: 1) enable an additional hierarchical degree in the telemetry architecture, thus improving scalability; 2) enable the xPonder to take local lightpath-based decisions within the SDN controller configuration margins, based on telemetry data processed at the card itself. In both cases, the SDN controller selects a set of devices and optical parameters to be monitored, configuring telemetry subscriptions having the La transmitter card as telemetry stream collector. Telemetry data streamed to the transmitter (step 1) are consumed locally to identify and trigger possible soft failure recovery at the card directly (step 2), within the constraints imposed by the current SDN controller configurations. For example, fine central frequency tuning within the allocated frequency slot, proprietary FEC or constellation shaping adaptation, launch power variation are possible countermeasures that may enforced directly by the card agent and subject to subsequent monitor via P2PT.

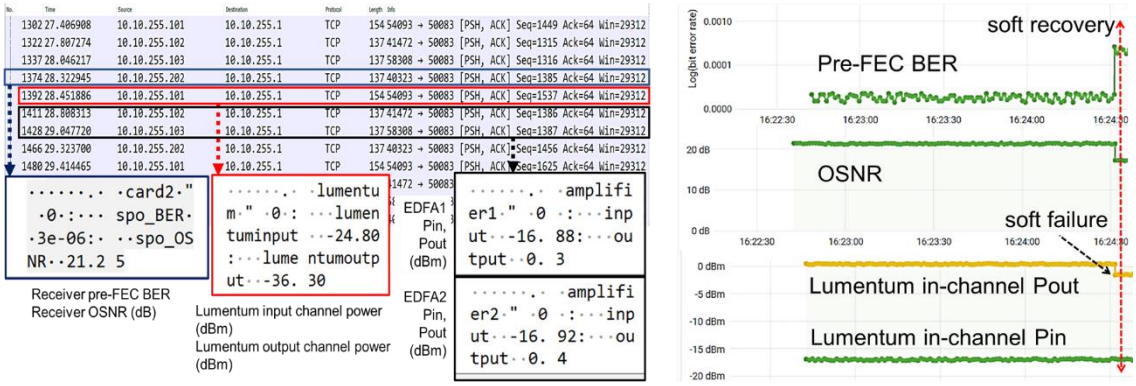


Figure 5.23: Peer-to-peer optical telemetry results: gRPC streams and soft failure test.

Results obtained by implementing the P2PT in a metro disaggregated optical network are reported in Figure 5.23. The figure shows P2PT streams captures at the transmitter (IP 10.10.255.1) with 4 subscriptions at 1s sampling rate (card receiver 202, Lumentum ROADM 101, EDFA1 102, EDFA2 103). The aggregated streaming rate is 4kbit/s average and 250kbit/s peak, including TCP overhead. This means that a 10G control plane interface is able to support up to 40k cards receiving the same amount of data. The figure shows the P2PT data at the TX. A filter shift soft failure at the intermediate ROADM WSS induces a degradation of the in-channel output power (-1.5 dB), impacting receiver OSNR (-2dB) and BER degradation (1.5x10-4). The card, using a Linux script, identifies the type of soft failure and, after 3 degraded samples (3s), triggers adaptation to FEC2, more robust to narrow filtering. The whole ONOS controller intervention would have required at least 3s more due to recomputation and control plane message exchanges.

The full P2PT implementation has been published at the OFC conference [PAO-P2PT-OFC21][1] and a journal version is currently under submission, extending the scope to AI-based soft failure detection at the optical pluggable level.

[1] [PAO-P2PT-OFC21] F. Paolucci, A. Sgambelluri, P. Castoldi and F. Cugini, "Telemetry Solutions in Disaggregated Optical Networks: an Experimental View," 2021 Optical Fiber Communications Conference and Exhibition (OFC), 2021, pp. 1-3.

6. WP2 Software Components for AI acceleration

6.1 Introduction

This section provides an overview of the different components developed within Workpackage 2 for AI acceleration in all their aspects; this means that BRAINE considers AI acceleration not defined only by the acceleration of the process itself, but also the communication and systems supporting the AI end-to-end. The rationale is that the overall user experience benefits from an end-to-end accelerated system, rather than dedicated AI-processes only surrounded by system bottlenecks.

6.2 C2.20 - 5G NR model

6.2.1 Brief description and motivations for main architectural choices (PHY high model for 5G RU)

The language chosen to build NR model simulation is MATLAB due to the fact that testing algorithms and implementing without the need of compilation is possible in MATLAB, as well as, no need to define the variables such as int, double, etc. for each input to be used, and many other reasons. And since we started to build the tool upon the 3GPP release 16, so we decided to follow that in the tool. Despite the advantages of using MATLAB, there are consequences when planning to use it in a real-life environment. Firstly, MATLAB is not suitable for online (On-air) implementation so for real demonstration, the files need to be captured and saved then run on MATLAB tool. Secondly, It is hard to configure MATLAB to interface with real HW transport interfaces such as eCPRI since MATLAB is running on a server computer and doesn't have a dedicated HW. Moreover, it is tricky to interface it with upper stack protocols as well in an online mode manner due to synchronization issues.

6.2.2 Current development status in BRAINE 1.0

SW design

The model is done completely from the scratch based on 3GPP release 16 to support at least some of the functionalities of the latest release initially and to be upgraded in the future to newer releases of NR for BRAINE use, the NR model needs to include the gNB transmission and reception blocks to provide the basic functionality which can be seen in the figure below:

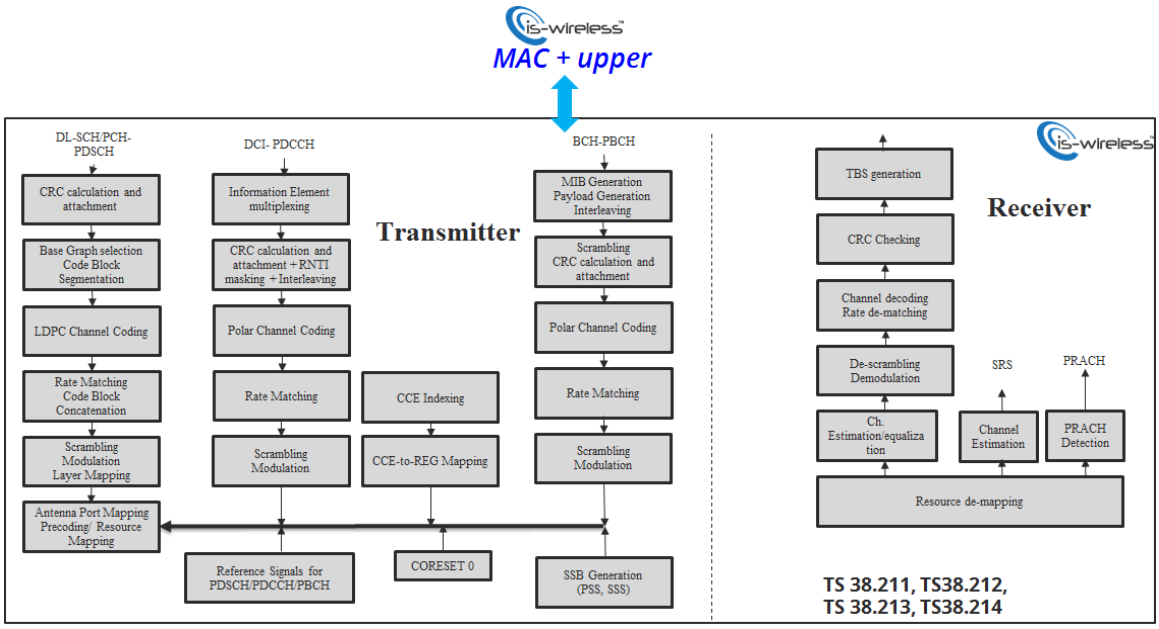


Figure 6.1 functionalities

90% of the transmitter is done, till the antenna port mapping, some kind of precoding is implemented but it is not yet integrated with the code.

From the receiver side, only around 5% done, there will be more concentration on receiver implementation in the incoming months.

Some kind of different block testing is done to confirm functionality and in the meantime, a separate mini RRC is implemented in the code that configures higher layer parameters randomly each time the code is run to confirm different modes of operations.

The model still needs to be upgraded in a manner to be able to transmit/receive data from upper protocol stack layers.

So far, we are targeting to implement the complete NR model to support the complete PHY (high and low PHY) but at the same time, we are investigating if we can interface the MATLAB with eCPRI so in that case, we can provide 7-2 split where eCPRI will be the contact point towards lower PHY. In case we successfully integrate the eCPRI, then the output will be the layer mapped modulated symbols towards the low PHY, and for sure, there must be some control info from the upper layer (RRC) that needs to be delivered as well.

In case we implemented the whole PHY, then investigation of communicating with USRP is needed (read the captured data by USRP and sending the data towards USRP), and this part still needs to be investigated as well since so far, we are aware that MATLAB can connect with USRP using MATLAB communication toolbox.

Development roadmap towards BRAINE 2.0

The target for BRAINE 2.0 is to finalize 5G NR model of High PHY for both gNB transmitter and receiver. In addition, eCPRI will be configured in MATLAB and based on the completion of both model and eCPRI, an offline testing low PHY RU will be implemented (Q2-Q3/2022).

6.3 C2.21 - vRAN adjustments prototypes

6.3.1 Brief description and motivations for main architectural choices

The accelerators include Graphics Processing Unit (GPU), Application Specific Integrated Circuit (ASIC) and Field Programmable Gate Array (FPGA). FPGAs with its reconfigurability and high energy efficiency are widely used in many edge computing scenarios. Recently, FPGA becomes a promising solution for algorithm acceleration with high energy efficiency and reconfigurability attributes compared to other platforms such as CPU and DSP for many edge computing scenarios. FPGA facilitates developers to implement only the necessary logic in hardware according to the target algorithm. Nonetheless, the system performance depends also on the scheduling efficiency between software tasks on CPUs and hardware tasks on FPGAs.

IS-Wireless virtual RAN architecture includes disaggregated 3GPP stack as well as RAN intelligent controller (RIC). For real time operation, RIC executes various protocols such as PHY-High, MAC, and RLC inside the stack. In the most of cases of the SOTA examples, the FPGA is usually installed with network interface card (NIC) in the DU MAC layer. The MAC layer functionalities such as PRB allocation, power control, user clustering need rapid CSI exchange between gNB and UEs. Thus, each TTI should be less than 1 ms to perform MAC functionalities in real time for highly dense 5G HetNets. The MAC scheduler collects the CSI from the users, and it copies the collected data to the central processing unit (CPU) main memory. This generates a significant overhead because the CPU needs to repeat the step of copying the collected data to the internal memory of the HW. FPGA based Time-Shared Optical Network (TSON)-CPRI interface protocol in MAC is used for providing fronthaul (FH) services. As the heavily loaded functionalities are performed in the MAC layer of the DU unit, therefore, MAC layer, is one of the top candidates to get more priority for acceleration from the virtual RAN stack. From the vRAN perspective, our proposal was to perform acceleration in radio protocol stack layers, i.e., MAC layer or PDCP layer. The initial conception of acceleration can be performed in MAC layer by using FPGA. Then we discuss the initial results about the acceleration of PDCP layer. We also provide solution proposal for vRAN DU/CU acceleration in BRAINE 1.0 and next steps towards the final version for BRAINE 2.0.

6.3.2 Design choices Virtual RAN (5G): Proposed FPGA based MAC layer accelerator

The physical resource block (PRB) allocation and power allocation (PA) allocation require time series data, such as channel state information (CSI) between the eNB and the mobile terminals (MTs) to allocate the resources in downlink and uplink transmissions. The CSI needs to be copied in every scheduling period of 1 ms and it is also required in each transmission time interval (TTI) to allocate resources to the users.

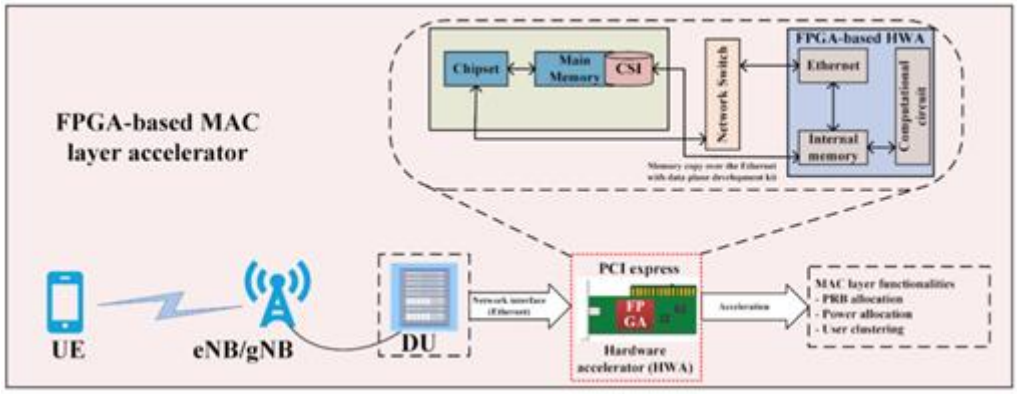


Figure 6.2 FPGA based MAC layer Acceleration.

Therefore, the MAC layer functionalities need to be accelerated to ensure the fairness of users in a highly dense 4G/5G and beyond 5G networks. The MAC layer functionalities e.g., PRB allocation along with PA, and CSI-assisted user clustering in a network function virtualization (NFV) environment can be accelerated through field-array programmable gate array (FPGA) based hardware accelerator (HWA) in distributed unit (DU) by implementing central unit (CU) with a standard server. The HWA will mainly handle the heavy processing of the MAC layer and the rest of functionalities related to L2 and L3 will run on the standard server. The MAC scheduler collects the CSI from the users, and it copies the collected data to the central processing unit (CPU) main memory which will generate a significant overhead because the CPU needs to repeat the step of copying the collected data to the internal memory of the HWA. The proposed tightly coupled FPGA-based HWA will collect the CSI and send it to the FPGA's internal memory. Then, the scheduler and the user clustering algorithm will schedule the resources and create clusters based on the reported CSI which is stored in the internal memory of the HWA.

6.3.3 Current development status: solution proposal for vRAN DU/CU acceleration in BRAINE 1.0

The acceleration needs targeted through special purpose hardware accelerators and artificial intelligence model training (or exploitation) enhancements are one of the key target solutions for the BRAINE project. The software functionalities from 5G vRAN stacks with acceleration attributes envisages the enhanced performances of the EMDC. To facilitate the acceleration capabilities in the EMDC, several investigations have been performed for the suitability of the protocol stacks. In particular, for BRAINE 1.0, the suitability of the packet data convergence protocol (PDCP) layer has been selected for acceleration. The selected layer could be accelerated through the hardware implementation of the (a) PCIe based FPGA development board or (b) a smart NIC.

The results achieved by applying the first configuration have proven to be limiting its usefulness for the PDCP instrumentation. In the case of PCIe based FPGA card that hosts the encryption algorithm (OpenAES) from PDCP, the results proven to increase latency for single packet encryption from the level of 50 microsecond to the few-fold increase by 20 milliseconds. This is only beneficial in case of enormous amount of data (at the level of hundred MBs per packet) which is definitely not feasible for a 3GPP-compliant 5G radio stack.

The next target considers the FPGA based Smart NIC (e.g. Napatech) that will be included in the BRAINE EMDC HW in the later stage. The current BRAINE 1.0 version encompasses:

- Containerized 5G vRAN with split 7.2 and able to deliver e2e connectivity (when integrated with EMDC)

6.3.4 Development roadmap towards BRAINE 2.0:

The following functionalities will be considered for the hardware-based network acceleration for AI for BRAINE 2.0:

- Second approach to PDCP acceleration with the use of SmartNIC (H2/2022)
- GPU or FPGA based HW acceleration of the AI/ML models from T3.3 (e.g. for training or exploitation) (Q3/2022)
- Specific adjustments to be able to support use-case UC2 (Q2-Q3/2022)

6.4 C2.25 - OpenCL program for Low-PHY

SSSA experimentally evaluates the performance of offloading some vDU functions onto an FPGA. Specifically, the offloaded functions are the Inverse Fast Fourier Transform (IFFT) and the Cyclic Prefix (CP) addition in Orthogonal Frequency Division Multiplexing (OFDM) signals. FPGAs can be a very good accelerator for these functions since they offer almost deterministic latency and high processing capacity per Watt.

6.4.1 Current Development Status

The Low-PHY function to be implemented in FPGA is shown in Figure 6.3 for both downlink and uplink direction. In the downlink direction, IQ samples in the frequency domain consisting of 32 bits (16 bits for the real part and 16 bits for the complex part) are received in the FPGA, which then performs the IFFT to convert the samples to the time domain. The cyclic prefix (CP) is then inserted as a guard interval to avoid inter-symbol interference (ISI).

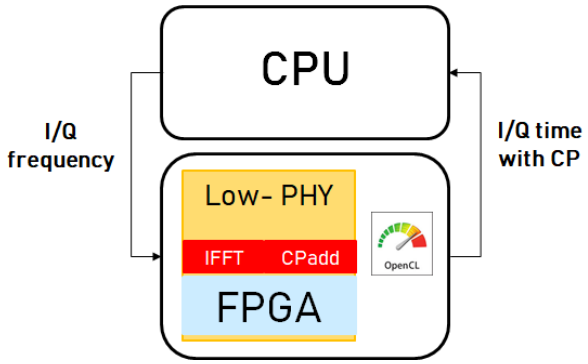


Figure 6.3: Low-PHY layer protocol implemented in hardware

The Low-PHY layer function implementation is based on the Open Computing Language (OpenCL). The OpenCL is a parallel computing API that provides flexibility to run a single program in multiple platforms (e.g., FPGA, GPU, CPU, and DPU). Thus, it is suitable for the implementation of the considered vDU, where some functions are implemented in the FPGA and others are implemented in the CPU. The considered FPGA hardware is a DE10-Pro Terasic FPGA board equipped with Stratix 10 GX Intel FPGA and 2 banks of DDR4 memory.

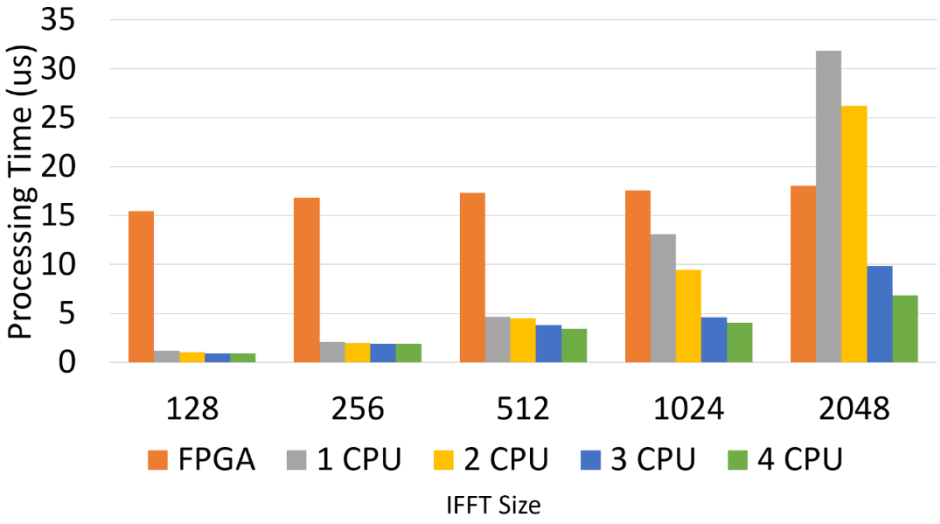


Figure 6.4: FPGA vs CPU Processing Time

Figure 6.4 shows the processing time as a function of the IFFT size of 5G low-PHY in OpenCL exploiting an FPGA, a CPU with a different number of processing cores (from one to four). The figure shows that the CPU-based processing time decreases as a function of increasing CPU cores. This is expected, since multiple cores allow to run multiple processes at the same time with greater ease, increasing the performance when multitasking or with demanding computations. Such processing time performance is not noticeable in lower IFFT sizes (i.e., 128 and 256 IFFT points) due to the lower complexity, which can be handled by a single CPU core. However, a notable performance difference can be highlighted for 2048 IFFT points, where it only took around 6.38 µs to run the implementation with 4 CPU cores, while it needs 31.84 µs with 1 CPU core. This is because the computational effort for IFFT and CP addition is shared among 4 CPU cores, resulting in an approximately four times lower processing time.

Moreover, the processing time of both the CPU-based and the FPGA-based implementations increases as a function of the IFFT size. However, the slope of the increase experienced by the FPGA-based implementation (approximately 1 µs if the IFFT size doubles) is much lower than the one experienced by the single CPU core implementation. At 2048 OFDM symbol size, the FPGA-based implementation has a shorter processing time compared to single and dual-core CPUs. This means that implementing 5G LOW-PHY function in FPGA can free up to 2 CPU cores.

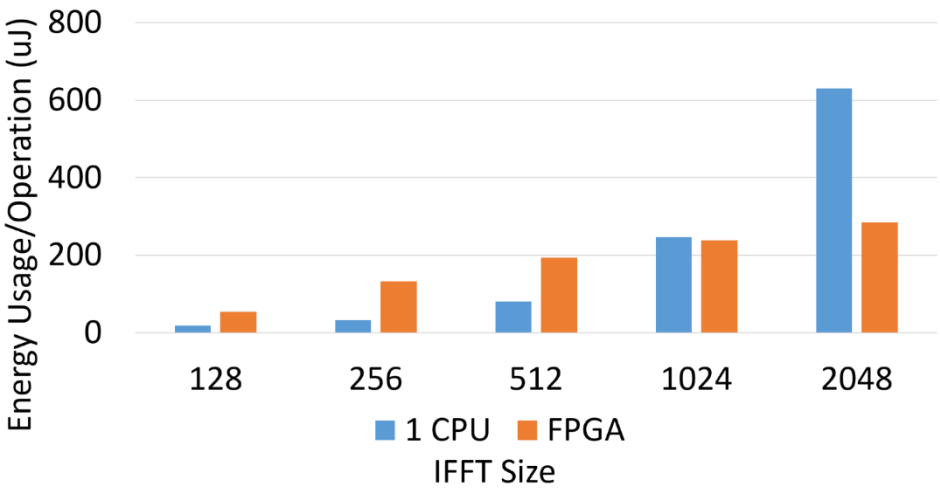


Figure 6.5: Energy Usage per Low-PHY Operation in FPGA and CPU

Figure 6.5 shows the energy consumption per Low-PHY operation in FPGA and a single CPU core. It is measured as the power consumed during the processing time of each Low-PHY operation. The s-tui tool is used to measure the CPU power, while quartus_pow (included in the DE10_pro board support package) is used to estimate the power dissipated in the FPGA.

As shown, the energy usage per operation increases as a function of the considered IFFT points when using either FPGA or CPU. Results also show that CPU-based implementation of Low-PHY on the has the least energy consumption up to 512 IFFT size. However, at higher IFFT sizes, FPGA-based implementation outperforms CPU –based implementation (*1.03x* lower than single-cored CPU for *1024* IFFT size and *2.22x* lower than single-cored CPU for *2048* IFFT size). This means that as the IFFT size increases, offloading the Low-PHY function into the FPGA becomes more energy efficient compared to processing them into the CPU.

6.4.2 Roadmap Towards BRAINE 2.0

Although the FPGA-based implementation of the 5G Low-PHY layer provides a faster processing for bigger IFFT sizes with lower energy consumption, the data transfer and synchronization between the host and the device memory becomes a bottleneck in this implementation scenario due to the contribution of the host-to-device transfer latency. Aside from the host-to-device memory transfer, data transfer between FPGAs' global and local memory also adds to the FPGA-based Low-PHY processing time. Compared to GPUs where the memory bandwidth offered by GDDR5X or HBM2 is in the order of hundreds of GB/s, FPGA boards usually offer much less memory bandwidth (e.g., DDR4 with around 32 GB/s).

To further improve the FPGA-based implementation of Low-PHY, SSSA plans to implement an FPGA-accelerated martNIC. This implementation reduces the Low-PHY processing time since data are transferred from the CPU to the FPGA through the PCIe interface only once. This is possible by exploiting auto-run kernels and OpenCL channels. The auto-run kernels allow to execute the processing in hardware without interaction with the host and the global memory. Indeed, the host starts the auto-run kernel that forwards the data to the NIC interface (e.g., QSFP) of the FPGA after the Low-PHY implementation

by means of the I/O OpenCL channels. Finally, the FPGA-based Low-PHY will be integrated with the other 5G protocol stack implemented by OpenAirInterface to analyze the overall gNB performance.

6.5 C2.26 - VTU – virtual transcoding unit

The i-EVS, developed by Italtel, is a framework that provides video processing and data storage functionalities. It also manages the information related to the users (organized in groups) accessing i-EVS services, by storing and maintaining updated the related data in the i-EVS user and group database.

The i-EVS consists of three components, the Virtual Transcoding Unit (VTU), the User and Group Database Manager (UGDM) and the sTORE,

The VTU provides three main functionalities:

- video and audio transcoding;
- real time and on-demand media streaming;
- video and audio file upload/download leveraging a local storage service.

The UGDM handles and stores all the information about crowded event participants, which have registered to the I-EVS. It provides Operation and Maintenance (O&M) functionalities, and manages the local storage system where audio and video data are stored (sTORE), as shown in Figure 6.6.

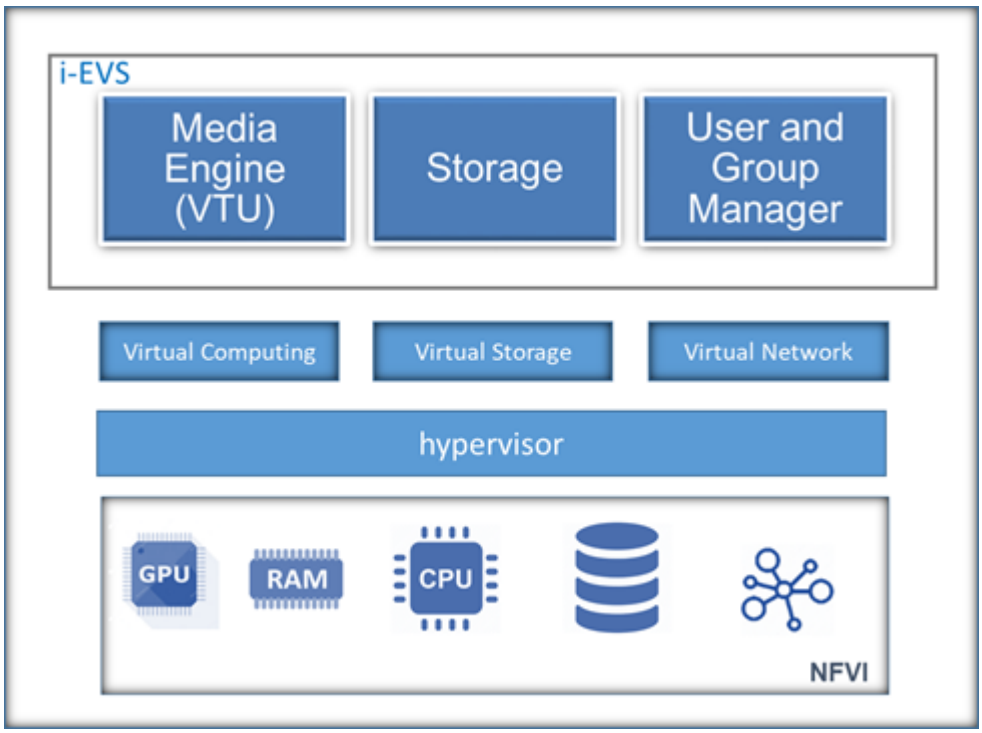


Figure 6.6: i-EVS Architecture

The i-EVS is being enhanced and adapted in order to be used in UC2 for supporting the multi-tenancy requirements at application level in case of video stream sharing. Furthermore, the VTU engine has been adapted to the BRAINE environment in order to be used as workload for testing the EMDC performance leveraging the novel architecture proposed in BRAINE (GPU HW acceleration).

VTU

The VTU can convert audio and video streams from one format to another. The source stream can originate from a file within the local storage system, or maybe a packetized network stream. The requested transcoding service can be monodirectional, as in video streaming, or bi-directional, like in videoconferencing (see Figure 6.7). The transcoding capabilities of the VTU are provided by Libav[1]. It is an open-source library, which can handle a wide variety of audio and video coding standards. For the most computationally intensive video encoding tasks, the VTU relies on Graphical Processing Unit (GPU) resources.

[1] Available at: https://libav.org/

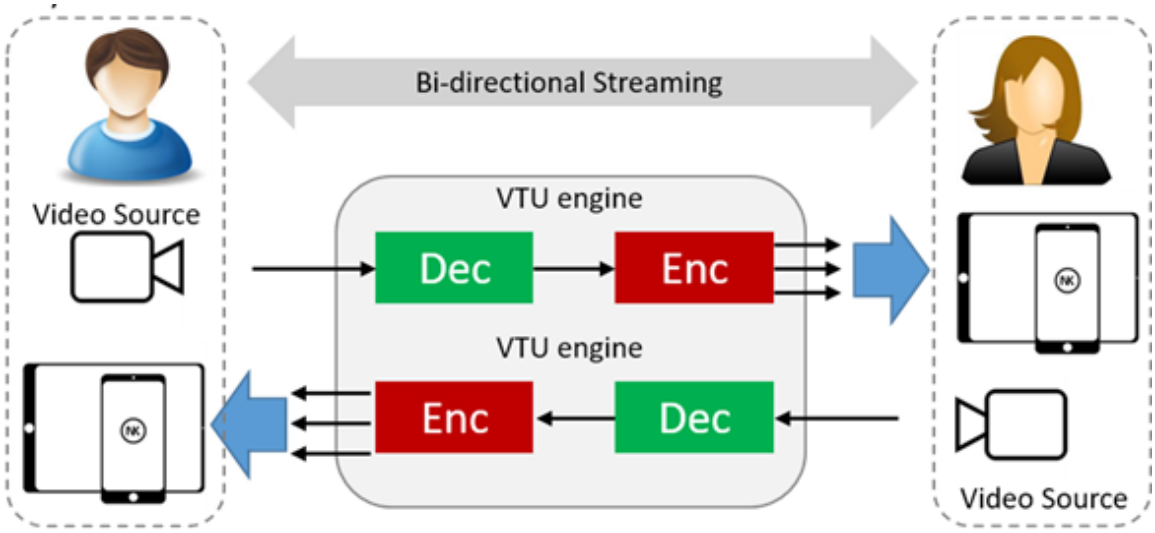


Figure 6.7: VTU engine

Two versions of the VTU are currently available:

- Single stream environment (one POD can manage a single camera, selection is made via an environment variable).
- Multi stream environment (a POD can manage multi cameras, camera run time selection/association via REST API).

Docker container is built on:

- Alpine (suggested for the single stream case) – 150MB image size
- Ubuntu (suggested for the multi stream case) – 750MB image size

The VTU performance obtained when running on the EMDC can be compared with VTU performance obtained using other HW solution[2].

As further development, the VTU will be integrated with the telemetry system to demonstrate its ability to adapt to the network traffic conditions.

6.6 C2.27 - N2Net

With N2Net we propose a new approach that efficiently leverages programmable NICs' hardware to enable high throughput and low latency network traffic analysis, while maintaining comparable accuracy with respect to existing machine learning-based traffic-

[2] Please see: https://www.researchgate.net/publication/316665777_GPU-accelerated_Video_Transcoding_Unit_for_Multi-access_Edge_Computing_Scenarios

analysis solutions implemented in software. The key insight is to exploit binary neural networks (BNNs), a recently-proposed machine learning (ML) model targeting battery-powered edge devices. Our tests show that BNNs can provide better classification accuracy than Decision Trees and Random Forests. Importantly, BNNs use single bits to represent inputs and weights, which provides two critical properties: (i) they exhibit a very compact memory footprint even for larger models; (ii) unlike mainstream Deep Neural Networks (DNNs), BNNs require only simple operations such as XOR and population count. This enables the implementation of efficient BNNs executors in a NIC's data plane.

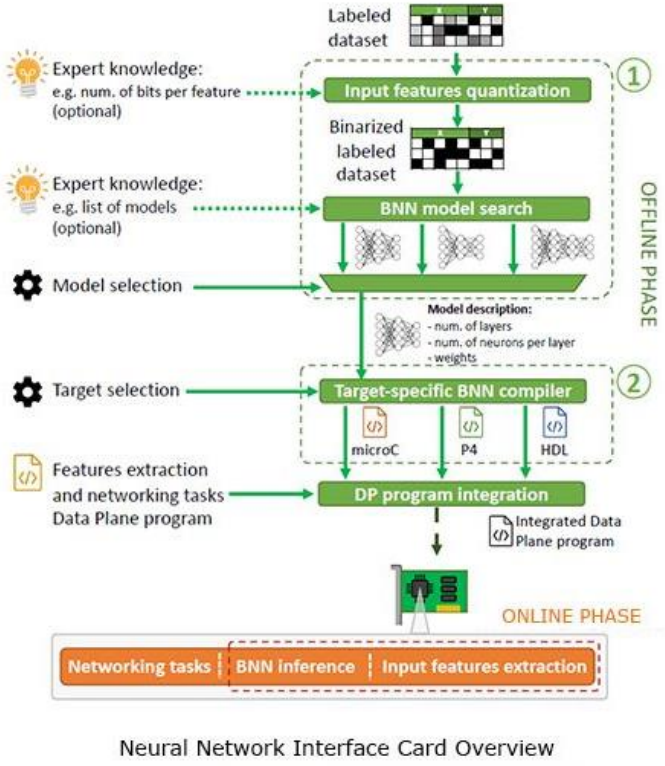


Figure 6.8: N2Net concept overview

Building on this insight, we developed N2Net to automate the implementation of binary neural network tasks in a network data plane. N2Net includes a framework to train a BNN using a labeled dataset provided by the user, and a compiler that translates the trained model into target-specific executable code. N2Net is currently able to target P4 enabled devices and is modular to enable the implementation of additional hardware targets.

For future work towards Braine 2.0, we will prove the potential of using N2Net-based traffic analysis for use cases that are relevant to the edge infrastructure, e.g., IoT traffic classification (helpful for instance for QoS management), and anomaly detection for network security.

6.7 C2.28 - Quantum-safe readiness

(I) Brief description and motivations for the task strategies

The Braine consortium has a clear understanding of the threat on cryptographic security that can appear with de development of quantum computing. Research on quantum-safe cryptography, also called Post-Quantum Cryptography (PQC), that is to say cryptographic

algorithms that cannot be broken by quantum computers, has begun worldwide a few years ago. Time scale on cryptography fundamentals is decade rather than months. Nevertheless, considering the importance of the change, standardization processes on PQC has begun worldwide. A prominent actor on the field for western country is the NorthAmerican NIST (National Institute of Standards and Technology).

It was then understood by the consortium of the primary importance for the Braine project to follow the evolution of PQC and to prepare readiness to PQC standardization, both for security of cryptography in use in the project and for acceptability by end-users requiring compliance to standard.

(II) Current development status

We have followed the NIST competition for PQC standardization and analyzed the outcome of round 3 of the process. In summer 2021, the third PQC standardization conference announced 7 finalists algorithms and 8 alternates candidates.

In the same time, NIST and also some essential industry standard related organization such as ETSI (European Telecommunications Standards Institute) have published first studies on challenges for adoption and use of PQC algorithm. Securities agency such as the German BSI as also published Theses literatures (with focus on NIST publication 10.6028/NIST.CSWP.05262020-draft and ETSI Technical Report TR 103 619) have been studied by cryptographers at SIC and were reported to the partners in a dedicated meeting and discussion on the expected outcome of the NIST round 3. Alternates candidates of Round 3 will be considered in a Round 4 phase.

Round 3 results are planned by NIST to be published by spring 2022. Current trend is that a Lattice-based algorithm will more likely be the final result of the contest. A prototype of TLS software library implementing TLS protocol thanks to quantum-safe cryptographic algorithms instead of the classical one has been realized and tested.

(III) Next steps towards the final version for Braine 2.0

The quantum-safe TLS implementation of SIC will be integrated on the QKD testbed of TU/e located in Eindhoven. This join SIC - TU/e works is ongoing.

In the same time, another activity in the sub-task is to analyze, for migration preparation to PQC, jointly between SIC and ISW, thanks to the work done on compilation of guideline previously performed. This work has started, and is made easier by the former collaboration of the two partners on WP4 on another cyber security topic (microarchitecture vulnerabilities detection in T4.1).

6.8 C2.29 - Low-bit-rate blockchain protocols

(I) brief description and motivations for main architectural choices.

We are bound by safety-critical IoT applications (smart hospital, assisted living) where the application pull is strong but it is predicated on ***both*** the technology push and regulatory compliance. The key technology is the *digital twin*, which processes messages from IoT platforms and uses AI to recognise and respond to situations. However, the compliance can only be assured by cyber security measures that guarantee non-repudiation. The reason is that multiple principals are involved in a healthcare organisation, each with their own compliance requirements; in the event of a significant malfunction or malpractice the

technology provider requires the ability to trace any actions to their origin, authenticate them and prove that they indeed took place irrespective of any principals' actions, including group actions that involve collusion. This is what non-repudiation means in practice, and the best way to achieve it is to use a zero-trust environment.

However, all known zero-trust environments are blockchains based on a distributed consensus mechanism, whose implementation is resource hungry to the extent that it is beyond the budget of an IoT platform. We have proposed a blockchain system architecture where the only required trust is in a single sealed, non-reprogrammable, non-internet-connected device (the *Sequencer*) operating on a radio network (LoRa). Our solution is based on the trust in the Sequencer's ability to hold a secret inside for roughly 15min, after which time it is published and a new secret is created, etc.

Due to the so-called Guy-Fawkes Protocols (GFP), this ability is sufficient to guarantee nonrepudiation of messages published on a blockchain by an untrusted Fog/Edge server, leaving DoS as the only security threat to worry about. DoS is a general threat; it is well understood and we have nothing further to propose in that area.

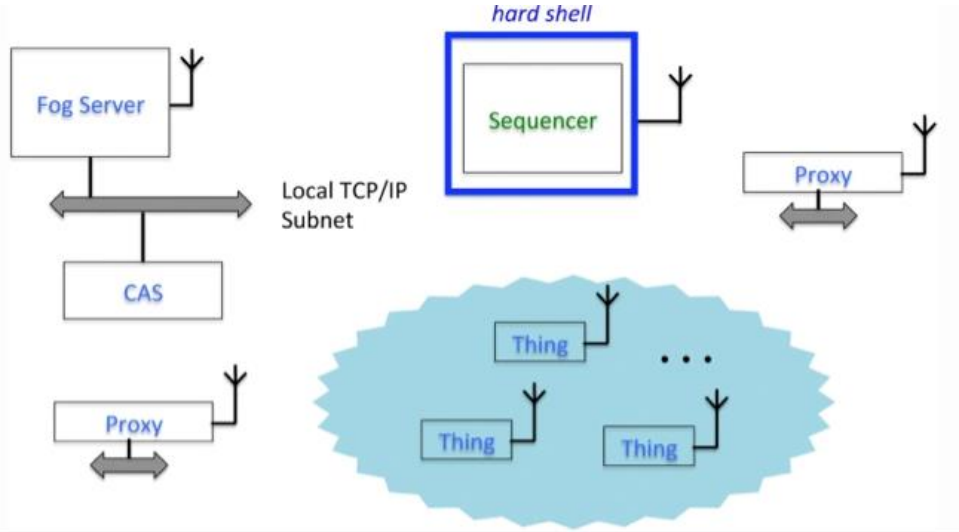


Figure 6.9: reference scenario

GFPs make it possible to post hashes of IoT platforms' messages (or the messages themselves, whichever shorter) on our special blockchain *directly* and also for the IoT platform to hold every existing block's Root of Trust (RoT) on the platform itself. A self-certified cloud storage subsystem, the Content-Addressible Store (CAS), holds all data authenticated by these roots. It can provide parts of the blocks relevant to an individual IoT platform at request via untrusted public networks, with the platform itself authenticating the data using solely the relevant RoTs. The protocols permit the Server to delegate its communication functions to Proxys, with the latter introducing no additional trust requirements.

(ii) current development status

Our proposed solution necessitated a great deal of research, which has now been completed. The interaction between the Sequencer, Server, CAS and an IoT platform requires better protocols (two more advanced versions of GFP than were known before the start of BRAINE: PLS and SLVP) and novel indexing structures. Both are needed to minimise IoT communication given the stringent constraints on LoRa traffic in the EU. The results have been published in two large journal papers.

Work in progress includes the implementation of the Server and Sequencer technology demonstrators, which will show the feasibility of our approach and its low cost.

Components produced:

Component	Description
Blockchain DB (block, CAS, mt, node)	Code for basic blockchain mechanisms
fog_server	Logic for receiving/validating blocks using SLVP for transmission via PLS
sequencer	A PLS handler (active side)
thing	A PLS client/SLVP handler

All components have been produced with their associated test suites.

The next step is porting the last two components to our chosen IoT platform: LoRa/ESP32 board.

(iii) next steps towards the final version for Braine 2.0

Final steps will include integration of the Server component of the blockchain software with the Digital Twin demonstrating the IoT- and Cloud user-traffic on our special blockchain that guarantees nonrepudiation.

6.9 C2.30 - Transport Layer Security accelerator

BRAINE architecture and data place supporting AI require secure connections in which security is not increased at cost of latency, as AI processes themselves are very latency sensitive. IPsec and TLS provide both full data-path encryption and transport security. In BRAINE, we tackled autonomous offloads for the TLS layer (standalone encryption offloads) through dedicated accelerators (which can be implemented in both CPUs and silicon).

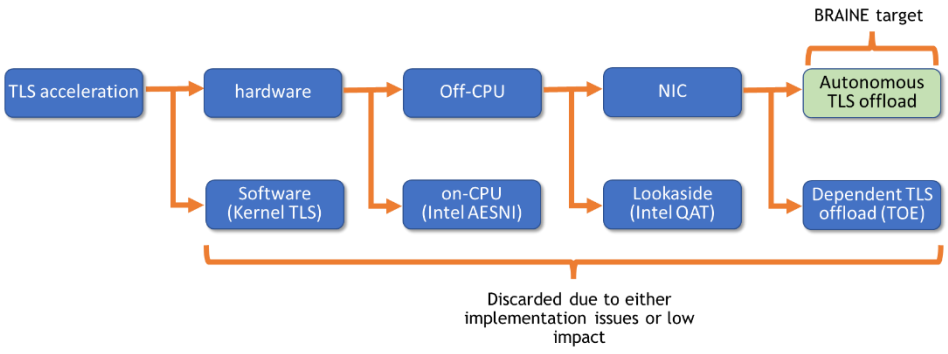


Figure 6.10: TLS acceleration roadmap

TLS accelerations can be conducted at both hardware and software level. At software level (through kernels), CPU clocks are used intensively, and hence, the scalability of the system becomes highly impacted (an initial investigation indicates around 50% of the CPU cycles would be spent on the encryption/decryption).

TLS acceleration on hardware systems may be conducted off-CPU or on-CPU. On-CPU offloading may be a bit more efficient than off-CPU accelerations, mainly due to the

overhead introduced in the IO set of operations. In addition, off-CPU accelerators require more parallelization and hence are a bit more intensive in terms of programmability and coding.

However, off-CPU accelerators, despite the increase of complexity in the system architecture and implementation, offer the possibility to employ highly powerful dedicated processors. In particular, some vendors propose to use dedicated lookaside processors which have in the fabric the encryption and decryption process. However, this option becomes very cost intensive, as a dedicated CPU is needed for each system CPU.

In BRAINE, the direction was decided to be employing network interface cards, thereby offloading networking operations to a fully dedicated subsystem. In this scenario, dependent TLS offloads can be implemented, which require to conduct the encryption using CPU cycles of the NIC. In a way, this leads to the same bottlenecks are conducting the offloading at CPU level of the machine itself. Hence, BRAINE took the decission to work on autonomous TLS offloads, where the encryption is conducted at NIC level by a dedicated processor (ARM processor), hence achieving perhaps the best performance in terms of latency.

Autonomous offloads have been implemented during the beginning of BRAINE and are now being transferred to NIC cards that can be employed in BRAINE.

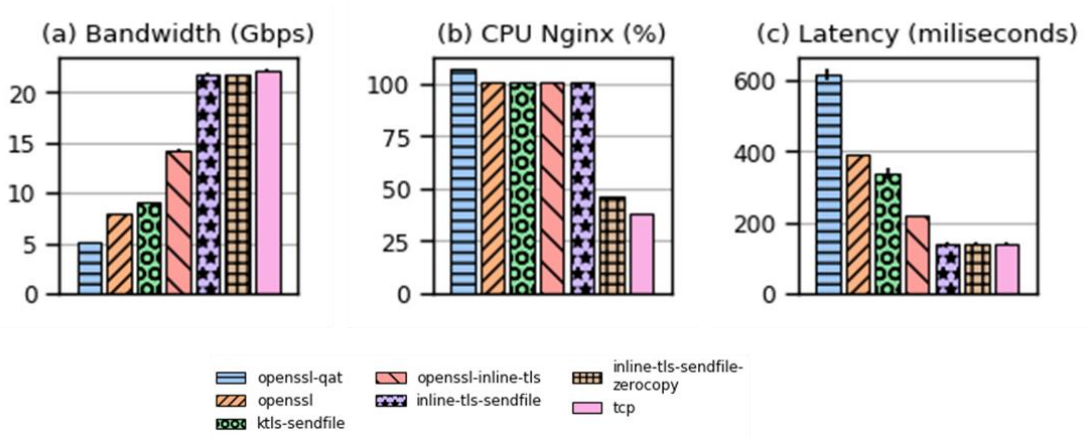


Figure 6.11: Preliminary performance evaluation – TLS offloads provide lower latency, lower CPU utilization, and enable higher bandwidth utilization.

Towards BRAINE 2.0, we will integrate the offloads to the NIC and these NIC tested at TUE testbed.

6.10 C2.31 - Smart sensors integration into the EMDC

In this section, we demonstrate the contribution with smart sensors and chips that allow smooth hardware multilevel integration. We do so by developing two demonstrator components that include both hardware (the sensors), and intelligent data processing & communication on the edge (AI algorithms and cloud/edge communication). Overall, we are following the notion of a set of interconnected systems (of systems).

The initial component environment is composed of the following two demonstrators: (1) the “smart wristband for voting” platforms are physically distributed and represent the sensor nodes, while a display presentation of the wristbands height, retrieved from the barometric pressure, represents the actuator side. Beyond that, (2) we are investigating (in another set-up) possibilities to design a biometric identification device that captures Photoplethysmograph (PPG) signal via an infra-red sensitive PALS sensor (PALS2).

Both demonstrators have in common that we are establishing wireless communication – first, on cloud level integrating the Arrowhead framework as service-oriented architecture; and second, by moving towards edge communication using EMDC in BRAINE. This way, by adding advanced data processing through AI algorithms on the edge, we demonstrate how an intermediate wireless communication gateway serves as the computing node. We present the principal notion in Figure 6.12. Basically, we differentiate two general system components: the hardware devices, i.e., the sensor systems, and the data processing unit that is running in some sort of infrastructure, which is communicating to the devices in a wireless fashion. Methodologically, we establish a secure and stable connection to a local cloud layer first (i.e., the service-oriented Architecture as reflected in the Arrowhead Framework), before establishing a connection to EMDCs. The main motivation is to demonstrate embedded IoT devices with micro data storage, micro data processing (smart algorithms), and (near) run-time capabilities. We will follow this method for both demonstrators, i.e., smart wirstband and PPG sensor.

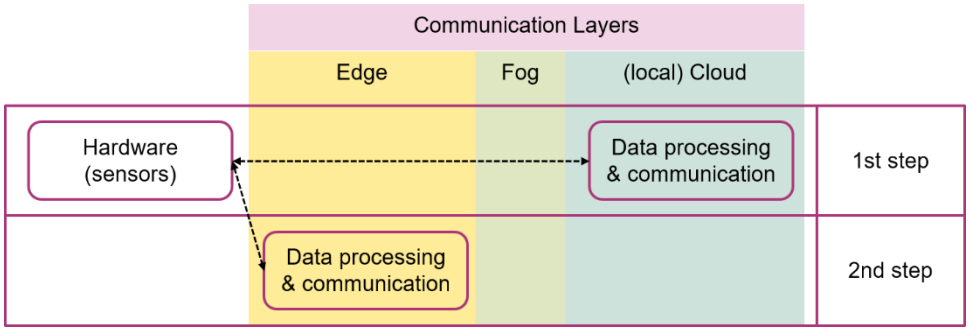


Figure 6.12: Methodological approach for smart sensor integration.

First, with the barometric pressure sensor, we follow our vision of a distributed, secure, democratic decision-making system. Democratic decision-making processes play a crucial role in any society, but also in organizations, especially those containing large interest groups and distributed stakeholders. Likewise, digitalization is at the forefront among companies in recent years. We developed a digital voting system with convincing performance, with special emphasis on the response time, user-friendliness and scalability. We achieve this goal by deploying the service-oriented architecture described in the Arrowhead Framework for a wearable device. Using an open-source environment, we follow the conception of the Internet of Things and, by inherited features of the IEC 61499 standard, such as interoperability and scalability of automated industrial systems, we meet the aspirations of Industry 4.0 paradigms. With this validated set-up is driven further in BRAINE using EMDC.

We present the conceptual communication processes in Figure 6.13. With this set-up, we are adopting our demonstrator to the service-oriented architecture as present in the Arrowhead Framework. Hence, we can establish late-binding in run-time, security in communication, and scalability of devices. After comprehensive evaluation of the demonstrators, we move forward to transferring data processing and communication to the edge layer, where we aim to enable real-time processing.

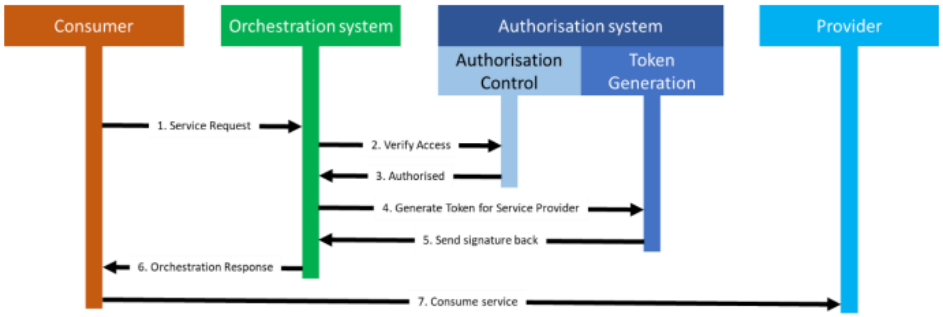


Figure 6.13: The process of consuming a service.

Second, Photoplethysmograph signal captured via PALS2 sensors, while being a non-invasive biosignal detection method, could provide very detailed information about the heart rate, oxygen saturation in blood, and even blood pressure could be estimated as well. We are currently developing circuit design for it. Communication is also running via ESP8266 and integration with Arrowhead Framework is already conceptualized. Further advanced signal processing via Artificial Intelligence Algorithms (on the edge) is planned for future periods. This contributes to our vision of enabling biometric identification via PPG. For both demonstrator ideas, an intermediate wireless communication gateway serves as the computing node. We are specifically focusing on security and data plausibility aspects.

6.11 C2.32 Distributed sound sensors

MAI is developing distributed sound sensors for the Smart City use case. The architecture of the integration of the audio analysis to the EMDC is as follows:

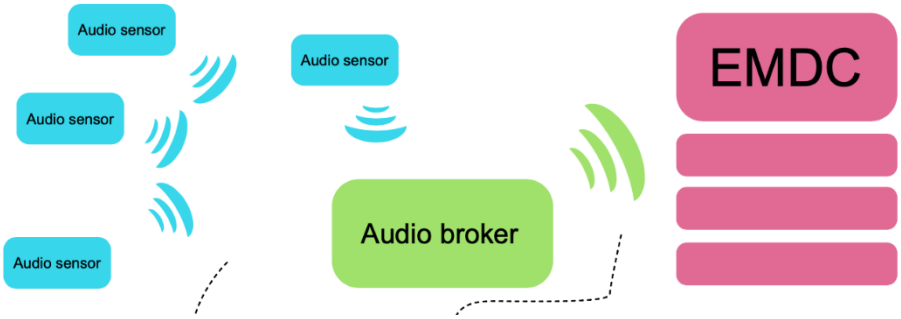


Figure 6.14: Audio sensor communications architecture.

Audio sensor	Current prototype is a Arduino BLE Sense based device, that will be replaced by nRF5340 based BLE Audio solution during the project. The current plan is to run Zephyr RTOS and the LC3 audio codec for lower latency and better transmitted audio per wattage ratio.
Audio broker	A lazy proxy-connection between the audio sensors and the EMDC. The current prototype is based on a simple Raspberry Pi device.
EMDC	The EMDC runs the MarshallAI end-to-end signal analytics platform with audio inference.

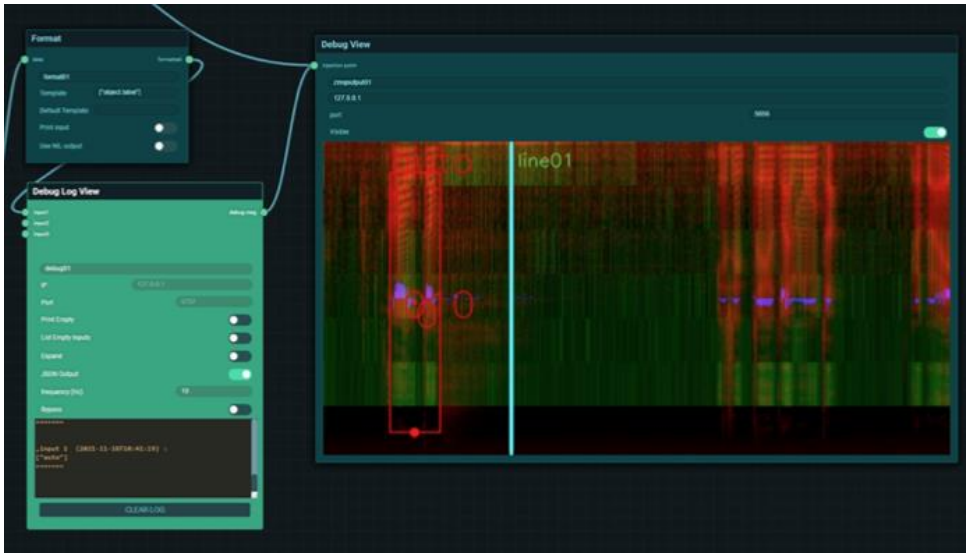


Figure 6.15: MarshallAI end-to-end signal analytics platform running audio inference.

The MarshallAI platform combines time domain amplitude information with spectral features extracted from the frequency domain. The result is a synchronized feature rich representation of audio that can be used for training CNN based object detectors. Audio detection & classification can then benefit from various tools that have been developed for temporospatial signal analytics, for example tracking and spatial filtering.

Using scale adaptive deep convolutional neural networks enables the optimization and balancing between feature resolution and prediction accuracy easily. The method enables learning features that will adapt to various sample rates and noise patterns from a variety of signal sources (microphones).

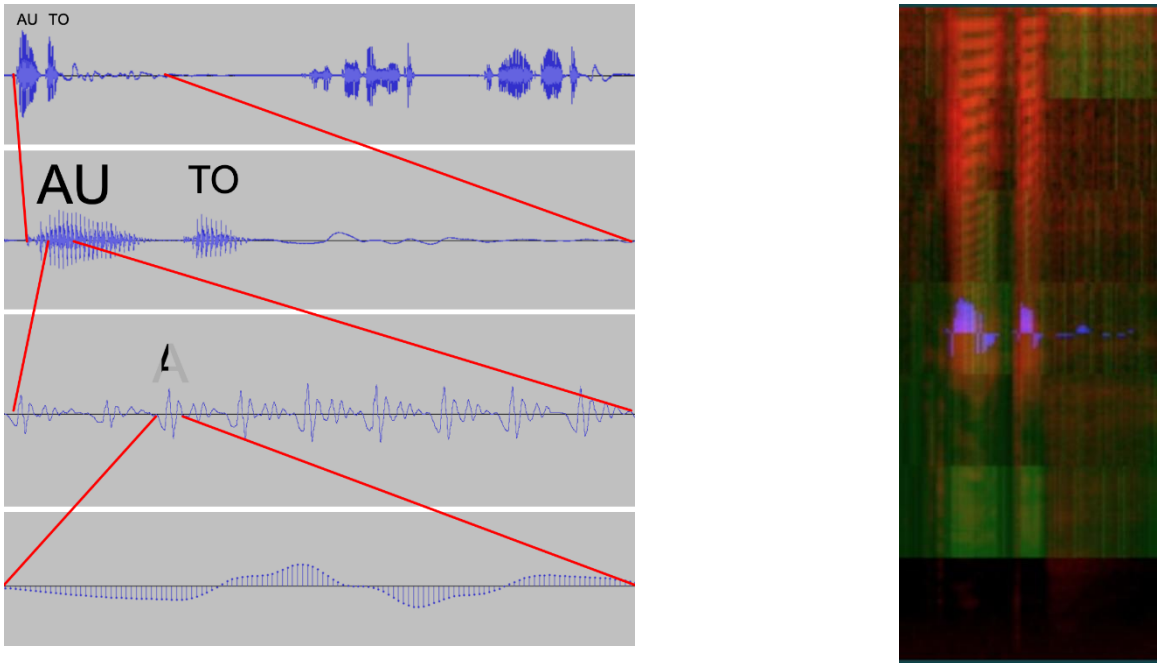


Figure 6.16: Audio analysis of the word AUTO.

RED The mel-frequency cepstrum frequency bin coefficients mapped between 0-255.

GREEN Spectral contrast, the greener the clearer the sound as in a pure string instrument and noisier on the blacker side. Divided into 7 frequency bands.

BLUE Down sampled time-domain amplitude representation.

6.12 C2.33 Quantum-Safe Cryptographic Solutions Evaluation

This evaluation is a sub-task within T2.3. The objective is to produce an informed recommendation based on the evaluation. The key focus will be to highlight the tradeoff between security, speed, and efficiency. It is unlikely that one algorithm will be the standout in all three of these areas, thus, several of algorithms may be recommended with each suiting a differing purpose. A secondary component of this evaluation will examine the potential impact of replacing contemporary cryptography with Quantum-Safe cryptography in the BRAINE EMDC. The impacts which are of interest to task partners are; latency overheads, required bandwidth, additional load on the CPU. One point regarding key sizes is that: where key sizes in a given source are presented as a value which has many decimal places, this value is rounded to the nearest accurate value of fewer decimal places in order to maintain consistency in tables and figure. The outcome of the NIST post-quantum cryptography standardization competition should take precedence over the material produced in this evaluation.

6.12.1 Motivation for Evaluation

In an ideal scenario there would already have been an accepted standard for quantum-safe cryptography that the BRAINE project could have utilized from inception. The National Institute of Standards is currently in the process for developing such a standard for both ‘public-key encryption and key establishment algorithms’ and ‘digital signatures’. The NIST standardization process also includes an ‘alternate candidate’ for each of the aforementioned. The process is currently in its third, and final, round which is due to complete sometime in mid-2022. Since it is not realistic within BRAINE’s timeline to wait for this standardization; the evaluation will be carried out on those algorithms selected for the NIST final round and will act as supplementary material to the standard itself once it is released.

6.12.2 Comparisons and Differing Sources

To ensure the evaluation compares like with like, direct comparisons cannot be made between the data presented in certain tables of this report. The reason for this is that different sources and experiments underpin the data in some tables. When comparing data such as key sizes this is not so much of an issue, however when comparing the data produced by different power usage experiments it could lead to incorrect assumptions due to different underlying hardware and other factors. Where several data tables are assembled from the same source direct comparisons can be made. Conclusions drawn from comparisons between data from different sources will take the underlying experiments into account. Information will be stated in the relevant sections to highlight which sources different elements are derived from.

6.12.3 Key and Signature Sizes

Type	Public Key Size	Private Key / Signature Size

Algorithm		Low	Mid	High	V-High	Low	Mid	High	V-High
RSA	Asymmetric key (Bit Size)	256b	512b	1024b	2048b	512b	1024b	2048b	4096b
RSA	Asymmetric Key (Byte Size)	32B	64B	128B	256B	64B	128B	256B	512B

Figure 6.17: RSA Asymmetric-Key & Signature Sizes

Algorithm	Type	Key Size		
		Low	Mid	High
AES	Symmetric Key (Bit Size)	128b	192b	256b
AES	Symmetric Key (Byte Size)	16B	24B	32B

Figure 6.18: AES Symmetric-Key Conventional Data Encryption Algorithm

Algorithm	Type	Public Key Size		Private Key Size	
		NIST1	NIST5	NIST1	NIST5
Classic McEliece	Code-Based	250KB	1300KB	11.5KB	...
CRYSTALS-KYBER	Lattice-Based	800B	1.5KB	1.6KB	3.1KB
NTRU	Lattice-Based	6.1KB	...	6.7KB	...
SABER	Lattice-MLWR	6B	1.3KB	1.5KB	3KB

Figure 6.19: NIST Round 3 PKE/KEM Finalists Size

Algorithm	Type	Public Key Size		Private Key Size	
		NIST1	NIST5	NIST1	NIST5
BIKE	Code-Based	2.5KB	8.1KB	248B	514B
FrodoKEM	Lattice-Based	9.6KB	21.5KB	19.8KB	43KB

HQC	Code-Based	3KB	8KB	6KB	17.3KB
NTRU Prime	Lattice-Based	…	…	…	...
SIKE	Super-Singular Isogeny-Based	330B	564B	128B	256B

Figure 6.20: NIST Round 3 PKE/KEM Alternates Size

Algorithm	**Type**	**Public Key Size**		**Private Key Size**		**Signature Size**	
		NIST1	NIST5	NIST1	NIST5	NIST1	NIST5
CRYSTALS-DILITHIUM	Lattice-Based	1.18KB	1.76KB	…	…	2.7KB	3.4KB
FALCON	Lattice-Based	897B	1.8KB	...	...	618B	1.2KB
Rainbow	Multi-Variate Based	149KB	1.7MB	93KB	1.2MB	48B	184B

Figure 6.21: NIST Round 3 Signature Finalists Size

Algorithm	**Type**	**Public Key Size**								
		NIST1	NIST3	NIST5	NIST1	NIST3	NIST5	NIST1	NIST3	NIST5
GeMSS	Multi-Variate Based	352 KB	1.237MB	N/A	13.4 KB	34KB	N/A	258 B	576 B	N/A
Picnic	Hash-Based	32B	48B	64B	16B	48B	64B	34 KB	76 KB	132KB
SPHINCS+	Hash-Based	32B	...	64B	64B	...	128B	8 KB	...	29.7 KB

Figure 6.22: NIST Round 3 Signature Alternates Size

6.12.4 Power & Energy Consumption

The data presented in this section is collated from sources [Esl], [Imr] and [Roma].

6.12.5 Conventional Cryptography

Algorithm	**Type**	**Key generation**	**Encapsulation**	**Decapsulation**

		Power	Energy	Power	Energy	Power	Energy
Round5 NI 0d AES	Symmetric	16.53 W	26.19mJ	15.73W	25.24m J	15.60W	1.41mJ
Round5 NI 0d SHAKE	Asymmetric	16.60 W	26.52mJ	15.82W	25.61m J	15.65W	1.50mJ

Figure 6.23: Conventional Cryptography Power & Energy Consumption

Intel Core i7-6700 CPU (PAPI) [Roma]

Round5 - library, NI - new instruction set, 0d - no forward error correction, PAPI – performance API

6.12.6 RSA Power Consumption

Algorithm	Type	Key generation	
		Power	Energy
RSA-1024 Software	Asymmetric	0.73W	…
RSA-2048 Software	Asymmetric	5.5W	…
RSA-1024 Hardware	Asymmetric	0.03W	…
RSA-2048 Hardware	Asymmetric	0.08W	…

Figure 6.24: Power Consumption of RSA [Abdu]

6.12.7 Comparing AES and AES New Instructions

Algorithm	Type	Key generation	
		Power	Energy
AES	Symmetric	3.1W	28.90J/GB
AES-NI	Symmetric	2.95W	2.7J/GB

Figure 6.25: AES Power & Energy Consumption

ARMv8 Processor rev 1 v81 OS: Ubuntu 18.04.2 LTS [Esl]

Algorithm	Type	Key generation	
		Power	Energy
AES	Symmetric	7.37W	20J/GB
AES-NI	Symmetric	7.31W	2.2J/GB

Figure 6.26: AES Power & Energy Consumption

Intel Core i5-8250U CPU@1.6GHZ OS: Ubuntu 18.04.2 LTS [Esl]

6.12.8 Quantum-Safe Cryptography Power & Energy Consumption

Algorithm	Type	Key generation		Encapsulation		Decapsulation	
		Power	Energy	Power	Energy	Power	Energy
Classic McEliece	Code-Based	16.79 W	2672.43mJ	16.37W	1.02mJ	14.35W	273.22mJ
CRYSTALS-KYBER	Lattice-Based	16.20 W	0.64mJ	16.22W	0.83mJ	16.21W	0.97mJ
NTRU (HPS)	Lattice-Based	15.77 W	55.70mJ	16.19W	3.68mJ	16.40W	8.03mJ
SABER	Lattice-MLWR	16.15 W	0.50mJ	15.92W	0.63mJ	16.05W	0.70mJ
Round5 NI 0d AES	Symmetric	16.53 W	26.19mJ	15.73W	25.24mJ	15.60W	1.41mJ
Round5 NI 0d SHAKE	Asymmetric	16.60 W	26.52mJ	15.82W	25.61mJ	15.65W	1.50mJ

Figure 6.27: NIST Round 3 PKE/KEM Finalists Power & Energy Consumption [Roma]

Algorithm	Type	Key generation		Encapsulation		Decapsulation	
		Power	Energy	Power	Energy	Power	Energy
BIKE (1-CCA)	Code-Based	16.37W	3.97mJ	16.44W	4.96mJ	16.34W	30.70mJ
FrodoKEM (SHAKE)	Lattice-Based	15.82W	47.88mJ	15.93W	52.21mJ	15.94W	51.84mJ
FrodoKEM (AES)	Lattice-Based	15.32W	210.663mJ	15.23W	213.74mJ	15.23W	213.49 mJ
HQC (1)	Code-Based	16.13W	5.71mJ	16.30W	10.91mJ	16.25W	17.13mJ
NTRU Prime (L)	Lattice-Based	14.71W	100.82mJ	14.53W	178.36mJ	14.51W	266.83 mJ
SIKE	Super-Singular Isogeny-Based	14.98W	263.93mJ	14.96W	430.651mJ	14.97W	459.94 mJ
Round5 NI 0d AES	Symmetric	16.53W	26.19mJ	15.73W	25.24mJ	15.60W	1.41mJ
Round5 NI 0d SHAKE	Asymmetric	16.60W	26.52mJ	15.82W	25.61mJ	15.65W	1.50mJ

Figure 6.28: NIST Round 3 PKE/KEM Alternates Power & Energy Consumption [Roma]

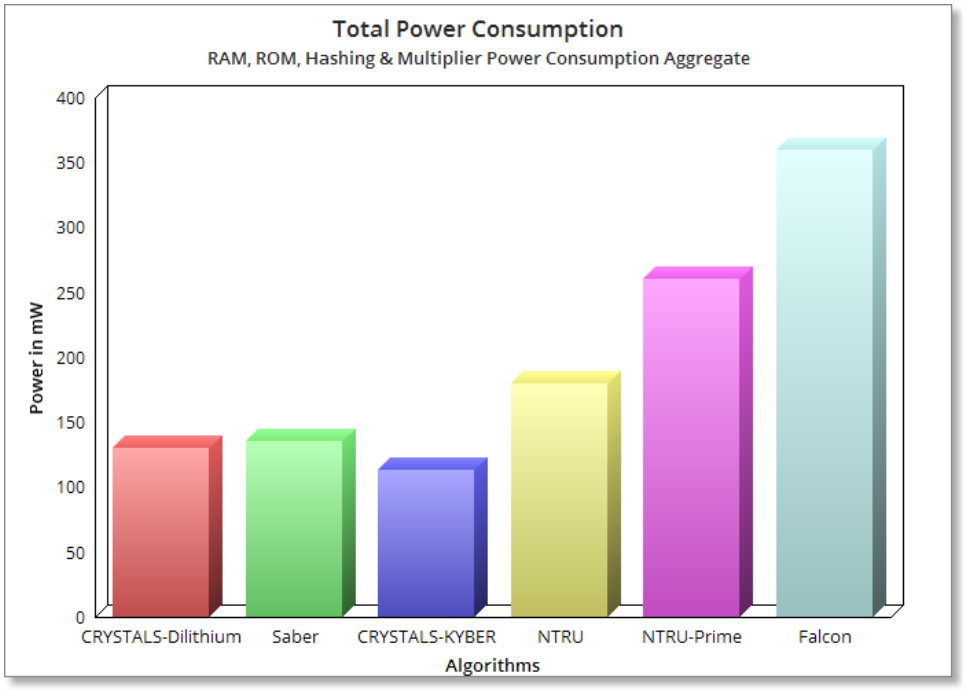


Figure 6.29: Lattice-Based Algorithms - Total Power Aggregate [Imr]

The above Figure shows the aggregate power consumption of ROM, RAM, Hashing and Cryptographic Multipliers for a set of ASIC-synthesis experiments investigating performances of lattice-base quantum-safe algorithms as hardware accelerators. The original paper [Imr] omits peripheral power consumption costs such as glue logic in an attempt to provide results on only the algorithms. Twelve algorithms were selected for the original work but only those which are present in the NIST competition's final round were selected for the BRAINE evaluation.

6.12.9 Other Power Consumption Analyses

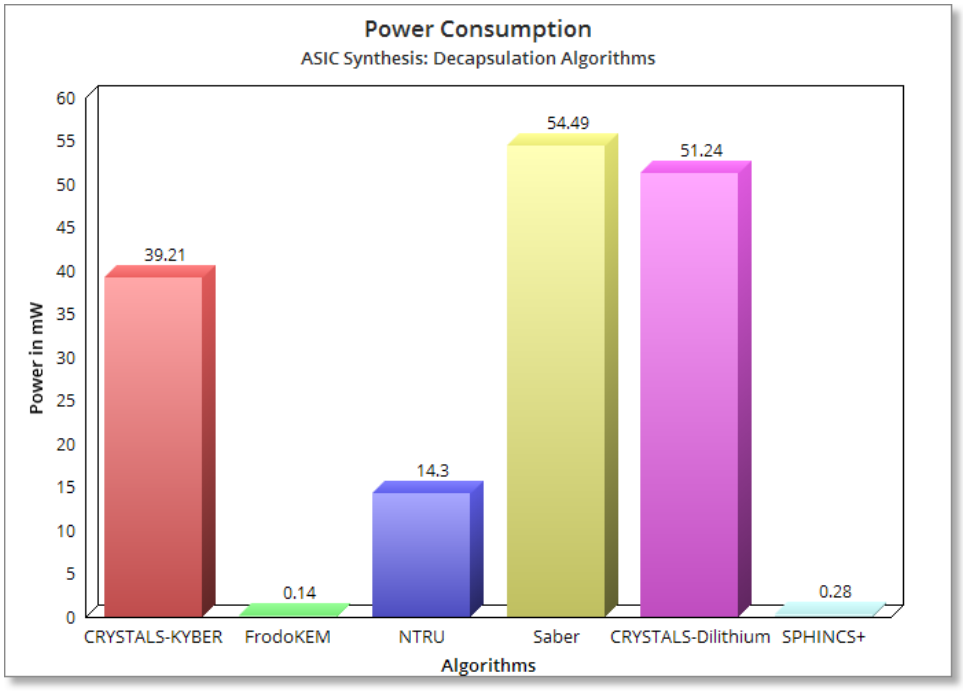


Figure 6.30: Select Decapsulation Algorithms - Power Consumption [Imr]

The above Figure gives the results of a hardware-based comparative study of NIST candidates. with this figure specifically covering the power consumption of the decapsulation algorithm component of some of the NIST contestants, the 'NewHope' algorithm is included in the original work but due to it not proceeding from NIST round 2 to the final round it has been excluded from this evaluation.

6.12.10 CPU Latency & RAM

6.12.11 CPU & Latency

The data in this section was gathered form sources [Basu] and [Gry].

Algorithm	Type	Flip-Flops	Lookup Tables	Encrypt Clock (nanosec)	Decrypt Clock (msec)	Latency (cycles)	Device
RSA-mMMM42-128	Asymmetric	3246	8294	50	19	…	Atrix-7
RSA-mMMM42-128	Asymmetric	3514	8202	50	19	…	Virtex-5
RSA-RSACIPHER-128	Asymmetric	2591	6850	30	19	…	Atrix-7
RSA-RSACIPHER-128	Asymmetric	2850	7108	30	19	…	Virtex-5

Figure 6.31: Area and Latency of Conventional Asymmetric Key Cryptography [Gowd]

Algorithm	Type	Flip-Flops	Lookup Tables	Clock (nanosec)	Latency (cycles)	Device
AES-EncDec	Symmetric	…	15919	21	55	Virtex 6
AES-Modes	Symmetric	…	…	72	188	Artix 7
AES-Efficient	Symmetric	665	393	32	84	Virtex 5

Figure 6.32: Area and Latency of Conventional Symmetric Key Cryptography [Gry]

Algorithm	Type	NIST Level	Flip-Flops	Lookup Tables	Clock (nanosec)	Latency (cycles)
KEM Algorithms						
CRYSTALS-Kyber	Lattice-Based	1	40720	230540	15	56345
FrodoKEM	Lattice-Based	1	14516	82265	10	469217
NTRU	Lattice-Based	1	6633	33845	15	1496914
Saber	Lattice-MLWR	3	38495	214764	15	499812
Classic McEliece	Code-Based	5	60264	840384	10	5128978
Signature Algorithms						
CRYSTALS-Dilithium	Lattice-Based	1	25926	133461	10	609828
SPHINCS+	Hash-Based	1	8461	31147	10	628778326

Figure 6.33: Area and Latency of Encapsulation Algorithms (unoptimized) [Basu]

Algorithm	Type	NIST Level	Flip-Flops	Lookup Tables	Clock (nanosec)	Latency (cycles)
KEM Algorithms						
CRYSTALS-Kyber	Lattice-Based	1	33030	186244	15	53553
FrodoKEM	Lattice-Based	1	14461	82307	10	220344
NTRU	Lattice-Based	1	5292	29532	15	1003222
Saber	Lattice-MLWR	3	33751	189597	15	89392
Classic McEliece	Code-Based	5	70112	847949	10	146126996
Signature Algorithms						
CRYSTALS-Dilithium	Lattice-Based	1	20865	108878	10	5380
SPHINCS+	Hash-Based	1	3335	11438	10	937975

Figure 6.34: Area and Latency of Decapsulation Algorithms (unoptimized) [Basu]

Lit review

Defining the comparison parameters / optimizations

6.12.12 RAM

6.12.12.1 Required memory instance size of RAM

In this section information is presented in a concise format compared to that of the source [Imr]. In some instances, in source [Imr] multiple rows of values are provided for a given algorithm, the tables in this section include data from two rows from the source [Imr] per algorithm, using the rows which produce the highest and lowest total memory instance size figures to represent the algorithm. The memory instance aggregate column has been included, which represents the sum of all memory instance total sizes for all rows (even those excluded from tables in this section for brevity) in the source [Imr]. The total number of rows, per the source [Imr], which form the aggregate is displayed in the 'number of aggregated rows' column.

Algorithm	Type	Required Memory Instances		Memory Addresses Per Instance		Bits Stored Per Address		Memory Instance Size		Memory Instance Total Size		Memory Instance Aggregate	Number of Aggregated Rows
		Low	High	Low	High	Low	High	Low	High	Low	High		
CrystalS-KYBER	Lattice-Based	1	5	128	256	16b	16b	0.256 KB	0.512 KB	0.256 KB	2.560KB	**2.816KB**	2
Classic McEliece	Code-Based	…		…		…		…		…	…	…	
Saber-firesaber	Lattice-MLWR	1	1	32	4	8b	1024b	0.032 KB	0.512 KB	0.032 KB	0.512KB	**1.888KB**	10
NTRU	Lattice-Based	…	14	…	821	…	16b	…	1.644 KB	…	22.988 KB	**22.988KB**	1

RAM Usage PKE / KEM Finalists [Imr]

Algorithm	Type	Required Memory Instances		Memory Addresses Per Instance		Bits Stored Per Address		Memory Instance Size		Memory Instance Total Size		Memory Instance Aggregate	Number of Aggregated Rows
		Low	High	Low	High	Low	High	Low	High	Low	High		
BIKE	Code-Based	…											

FrodoKEM	Lattice-Based	5	3	64	10752	16b	16b	0.128KB	21.504 KB	0.640KB	64.512 KB	**65.152KB**	2
SIKE	Super-Singular Isogeny-Based	…											
NTRU-Prime	Lattice-Based	1	1	24	256	64	8	0.192 KB	0.256 KB	0.192 KB	0.256 KB	**0.448KB**	2
HQC	Code-Based	…											

RAM Usage PKE / KEM Alternates [Imr]

Algorithm	**Type**	**Required Memory Instances**		**Memory Addresses Per Instance**		**Bits Stored Per Address**		**Memory Instance Size**		**Memory Instance Total Size**		**Memory Instance Aggregate**		**Number of Aggregated Rows**
		Low	High	Low	High	Low	High	Low	High	Low	High			
Crystals-Dilithium	Lattice-Based	…	3	…	256	…	32	…	1.024KB	…	3.072KB	**3.072KB**		1
Falcon	Lattice-Based	5	6	1024	521	16	32	2.048KB	2.084KB	10.240KB	12.504KB	**22.744**		2
Rainbow	Multi-Variate Based	…	…	…	…	…	…	…	…	…	…	…	…	…

RAM Usage Signature Finalists [Imr]

6.12.12.2 RAM - Stack Memory for key Generation, Encapsulation and Decapsulation

The tables below include data regarding the implementation of cryptographic scheme, there are four types of implementation described in the source [Kann]:

- **Clean**: a clean *pqclean* implementation
- **Ref**: A NIST reference submission *C language* implementation from the *mupq* repository
- **Opt**: *C language* optimized implementation. Optimized implementations submitted to NIST are in the mupq repository
- **Opt-CT**: opt with a constant-time variant added to the emulation

- **M4**: A Cortex-*M4* optimized implementation. Usually, assembly code is present. Such implementations are held on the *pqm4* repository

Algorithm & Cryptographic Scheme Implementation	Type	Cryptographic Scheme Implementation Type	Key Generation	Encapsulation	Decapsulation	Experiment Exclusion Reason
Crystals-KYBER kyber768	Lattice-Based	m4	3.848KB	3.128KB	3.072KB	N/A
Crystals-KYBER kyber1024	Lattice-Based	m4	4.360KB	3.584KB	3.592KB	N/A
Saber	Lattice-MLWR	ref	13.256KB	15.544KB	16.640KB	N/A
NTRU ntruhps2048509	Lattice-Based	m4	21.412KB	15.452KB	14.828KB	N/A
NTRU ntruhps4096821	Lattice-Based	m4	34.532KB	24.924KB	23.980KB	N/A
Classic McEliece	Code-Based	X	X	X	X	Key too large for platform

PKE/KEM Finalists Key Encapsulation Schemes Stack Memory Usage [Kann]

Algorithm & Cryptographic Scheme Implementation	Type	Cryptographic Scheme Implementation Type	Key Generation	Encapsulation	Decapsulation	Experiment Exclusion Reason
FrodoKEM frodokem640shake	Lattice-Based	m4	26.528KB	51.904KB	72.600KB	N/A
FrodoKEM frodokem640shake	Lattice-Based	opt	36.672KB	58.312KB	78.944KB	N/A
FrodoKEM frodokem640aes	Lattice-Based	m4	31.992KB	62.488KB	83.112KB	N/A

SIKE sikep434	Super-Singular Isogeny-Based	opt	6.776KB	7.088KB	7.424KB	N/A
SIKE sikep751	Super-Singular Isogeny-Based	opt	11.624KB	11.768KB	12.328KB	N/A
NTRU-Prime sntrup653	Lattice-Based	ref	13.976KB	14.004KB	16.700KB	N/A
NTRU-prime sntrup857	Lattice-Based	ref	18.264KB	17.908KB	21.444KB	N/A
BIKE	Code-Based	X	X	X	X	Integration difficulty
HQC	Code-Based	X	X	X	X	Integration difficulty

PKE/KEM Alternates Key Encapsulation Schemes Stack Memory Usage [Kann]

Algorithm & Cryptographic Scheme Implementation	**Type**	**Cryptographic Scheme Implementation Type**	**Key Generation**	**Sign**	**Verify**	**Experiment Exclusion Reason**
Crystals-Dilithium dilithium2	Lattice-Based	clean, m4	36.424KB	61.312KB	40.664KB	N/A
Crystals-Dilithium dilithium4	Lattice-Based	Clean, m4	67 136KB	104.408KB	71.472KB	N/A
Falcon falcon512	Lattice-Based	opt-ct	1.680KB	2.524KB	0.512KB	N/A
Falcon falcon1024	Lattice-Based	opt-ct	1.640KB	2.728KB	0.512KB	N/A
Rainbow	Multi-Variate Based	X	X	X	X	Key too large for platform

Signature Schemes Finalists Stack Memory Usage [Kann]

Algorithm & Cryptographic Scheme Implementation	Type	Cryptographic Scheme Implementation Type	Key Generation	Sign	Verify	Experiment Exclusion Reason
SPHINCS+ sphincs-sha256-128f-simple	Hash-Based	clean	2.192KB	2.248KB	2.544KB	N/A
SPHINCS+ sphincs-sha256-256s-robust	Hash-Based	clean	6.048KB	5.880KB	5.360KB	N/A
SPHINCS+ sphincs-shake256-128f-simple	Hash-Based	clean	2.208KB	2.368KB	2.664KB	N/A
SPHINCS+ sphincs-shake256-256s-robust	Hash-Based	clean	6.008KB	5.840KB	5.320KB	N/A
GemSS	Multi-Variate Based	X	X	X	X	Key too large for platform
Picnic	Hash-Based	X	X	X	X	Memory requirement too high for platform

7. Conclusions

This deliverable provides the description of the hardware subsystems, system and embedded software components of BRAINE release 1. The components, to be integrated in WP5 with the software components built by WP3 and WP4, will constitute a powerful edge micro data center (EMDC) able to support intensive artificial intelligence (AI).

The components of release 1.0 have been designed and implemented. The refined design and updated implementation towards the final release BRAINE 2.0 are in progress.

This document reported on each single component as well as its current status development.

The first technical section of this document reported on the HW components on BRAINE 1.0, providing its overall architecture, and all detailed technological aspects, including the enclosure which encompasses the cooling system, ejection system, and power system. In addition, it reports on the activities at the HW level to provide a quantum-safe fiber layer.

The document then provides details on the embedded, low-level software components required to operate the EMDC hardware. This includes the EMDC firmware, the firmware of the board management controller, the firmware of the FPGA node, and the software components for the EMDC embedded programmable switch (P4 programs, operating system, interconnectivity).

Furthermore, the deliverable includes the description of all additional building blocks for the whole BRAINE solution. This includes the 5G components, the virtual transcoding unit, the acceleration for network interfaces, security components (e.g., quantum-safe readiness and performance assessment, TLS acceleration), and smart sensors.

8. References

[Abdu] Abdullah Said Alkalbani, T. Mantoro and A. O. M. Tap. (2010). "Comparison between RSA hardware and software implementation for WSNs security schemes," Proceeding of the 3rd International Conference on Information and Communication Technology for the Moslem World (ICT4M) 2010, pp. E84-E89, doi: 10.1109/ICT4M.2010.5971920.

[Basu] Basu, K., Soni, D., Nabeel, M.M., & Karri, R. (2019). NIST Post-Quantum Cryptography- A Hardware Evaluation Study. IACR Cryptol. ePrint Arch., 2019, 47.

[Esl] Eslam G. AbdAllah, Yu Rang Kuang, and Changcheng Huang. 2020. *Advanced Encryption Standard New Instructions (AES-NI) Analysis: Security, Performance, and Power Consumption*. Proceedings of the 2020 12th International Conference on Computer and Automation Engineering (ICCAE 2020). Association for Computing Machinery, New York, NY, USA, 167–172.

[Gowd] Gowda, Leelavathi & Shaila, K. & K R, Venugopal. (2021). Hardware performance analysis of RSA cryptosystems on FPGA for wireless sensor nodes. International Journal of Intelligent Networks. 2. 184-194. 10.1016/j.ijin.2021.09.008.

[Gry] Grycel, J.T., & Walls, R.J. (2020). DRAB-LOCUS: An Area-Efficient AES Architecture for Hardware Accelerator Co-Location on FPGAs. 2020 IEEE International Symposium on Circuits and Systems (ISCAS), 1-5.

[Imr] Imran, M., Abideen, Z., & Pagliarini, S.N. (2020). An Experimental Study of Building Blocks of Lattice-Based NIST Post-Quantum Cryptographic Algorithms. Electronics, 9, 1953.

[Kann] Kannwischer, M.J., Rijneveld, J., Schwabe, P., & Stoffelen, K. (2019). pqm4: Testing and Benchmarking NIST PQC on ARM Cortex-M4. IACR Cryptol. ePrint Arch., 2019, 844.

[PAO-DRCN21] F. Paolucci et al., "User Plane Function Offloading in P4 switches for enhanced 5G Mobile Edge Computing," 2021 17th International Conference on the Design of Reliable Communication Networks (DRCN), 2021, pp. 1-3, doi: 10.1109/DRCN51631.2021.9477338.

[PEL-OFC21] I. Pelle, F. Paolucci, B. Sonkoly and F. Cugini, "Fast Edge-to-Edge Serverless Migration in 5G Programmable Packet-Optical Networks," 2021 Optical Fiber Communications Conference and Exhibition (OFC), 2021, pp. 1-3

[Roma] C. A. Roma, C. -E. A. Tai and M. A. Hasan, Energy Efficiency Analysis of Post-Quantum Cryptographic Algorithms, in IEEE Access, vol. 9, pp. 71295-71317, 2021, doi: 10.1109/ACCESS.2021.3077843.

[SCANO-ECOC21] D. Scano et al., "Hierarchical Control of SONiC-based Packet-Optical Nodes encompassing Coherent Pluggable Modules," 2021 European Conference on Optical Communication (ECOC), 2021, pp. 1-3, doi: 10.1109/ECOC52684.2021.9605850.

[SCANO-INT-JOCN21] D. Scano, F. Paolucci, K. Kondepu, A. Sgambelluri, L. Valcarenghi and F. Cugini, "Extending P4 in-band telemetry to user equipment for latency- and localization-aware autonomous networking with AI forecasting," in Journal of Optical Communications and Networking, vol. 13, no. 9, pp. D103-D114, September 2021, doi: 10.1364/JOCN.425891.

[SGAMBELLURI-OFC21] A. Sgambelluri et al., "Coordinating Pluggable Transceiver Control in SONiC-based Disaggregated Packet-Optical Networks," 2021 Optical Fiber Communications Conference and Exhibition (OFC), 2021, pp. 1-3.

[SGAMBELLURI-OFC21] A. Sgambelluri et al., "Coordinating Pluggable Transceiver Control in SONiC-based Disaggregated Packet-Optical Networks," 2021 Optical Fiber Communications Conference and Exhibition (OFC), 2021, pp. 1-3.