



FPGA-Based Network-Attached Accelerators – An Environmental Life Cycle Perspective

Fritjof Steinert^{1,2} and Benno Stabernack^{1,2}

¹ Fraunhofer for Telecommunications, Heinrich Hertz Institute (HHI),
Einsteinufer 37, 10587 Berlin, Germany
fritjof.steinert@hhi-extern.fraunhofer.de,
benno.stabernack@hhi.fraunhofer.de

² Embedded Systems Architectures for Signalprocessing, University of Potsdam,
August-Bebel Straße 89, 14469 Potsdam, Germany

Abstract. Homogeneous computing systems are reaching their limits with the growing demands of current applications. Accelerating compute-intensive applications ensures manageable computing times and boosts energy efficiency, which is an important lever as part of ongoing efforts to tackle global climate change. Field Programmable Gate Array (FPGA) accelerators are well-known for increasing throughput and, in particular, energy efficiency for many applications. FPGA accelerators connected directly to the data center high-speed network are ideal for integration into a heterogeneous data center, avoiding the energy and resource overhead of a carrier system. The standalone Network-attached Accelerators (NAAs) further benefits from low latency and predictable line-rate network throughput, as well as an interoperable communications interface. For selected use cases, we compare a heterogeneous computing cluster extended by NAAs with a homogeneous CPU-based cluster not only in terms of computing performance and energy efficiency, but also considering resource efficiency. For this purpose, we perform a Life Cycle Assessment (LCA) for both systems based on the Key Performance Indicators for Data Center Efficiency (KPI4DCE) indicator set, which takes into account the manufacturing phase in addition to the usage phase. The KPI4DCE tool has been extended to include modeling of NAAs. This allows us to show that NAAs are not only more energy-efficient, but also more resource-efficient for the selected applications, leading to a strong improvement of the environmental impact of the manufacturing phase.

Keywords: FPGA · Network-attached Accelerator · Data Center · Life Cycle Assessment · Reconfigurable Computing · Heterogeneous Computing

This work is part of the NAAICE project, which is funded by the German Federal Ministry of Education and Research under reference 16ME0624.

© The Author(s) 2023

G. Goumas et al. (Eds.): ARCS 2023, LNCS 13949, pp. 248–263, 2023.

https://doi.org/10.1007/978-3-031-42785-5_17

1 Introduction

With ever-increasing compute requirements of applications such as Machine Learning (ML), image and video processing (e.g. video transcoding for social media), distributed databases and the like, the need for energy-efficient acceleration in Data Centers (DCs) is growing. Classical processor-based architectures are reaching their limits, especially after the end of Dennard scaling. Homogeneous computing systems are accompanied by Graphics Processing Units (GPUs), Field Programmable Gate Arrays (FPGAs) and Application Specific Instruction Processors (ASIPs) like a Tensor Processing Unit (TPU) to provide the required processing capabilities in an energy-efficient manner. In this heterogeneous landscape, the challenge is to find the appropriate compute node for a workload. Energy efficiency in the usage phase has become the most important design parameter in this regard, as it helps lessen the enormous greenhouse impact of global DCs. Although ASIPs are to be preferred for energy efficiency and performance reasons, long development times and in particular the lack of flexibility outside a specific domain are an obstacle to their deployment in DCs.

Escobar et al. conducted an extensive study to determine which application groups are suitable for which kinds of accelerators and are consequently the most efficient [6]. They distinguish 4 groups:

1. High arithmetic demand and relatively regular memory access patterns → on GPU.
2. High arithmetic demand and irregular regular memory access patterns → on FPGA.
3. Low arithmetic demand and sophisticated memory management → on multi-core processors.
4. Low arithmetic demand and operators are mapped directly to hardware → on FPGA.

While the study provides an initial guide to selecting an accelerator, it does not consider the communication interface. Microbenchmarks indicate that hardware-based network implementations offer line-rate throughput as well as low variance and deterministic latencies, unlike typical software stacks [3,23]. The implementation of Network Interface Controller (NIC) directly in FPGA also saves the communication detour via the host when the data is delivered via network to the compute node. In a DC architecture with compute nodes and storage nodes, this is always the case resulting in a latency reduction. Therefore, network-coupled FPGAs are well suited e.g. for latency-critical tasks.

Connecting the FPGA directly to the DC network degrades the host to a power-only enclosure. By using a standalone, network-attached FPGA accelerator, called Network-attached Accelerator (NAA), as proposed in [20,24], the baseline energy requirements of the system can be greatly reduced without sacrificing performance, thereby boosting the energy efficiency. NAAs are treated as distinct and fully equal nodes in the DC.

However, in addition to the operational phase, the environmental aspects of the manufacturing phase, transportation, installation, and disposal should also

be included in a Life Cycle Assessment (LCA), as these phases can have a significant impact. Not just a single component (as one NAA), but the entire system (whole DC) should be considered in order to exclude undesirable interactions. It is evident that the NAA approach also performs well in the extensive environmental analysis as the reduced number of components in an NAA architecture decreases the impact of the manufacturing phase on the environmental footprint, in tandem to energy consumption.

Our main contribution is to perform and evaluate an environmental life cycle analysis for a heterogeneous cluster accelerated by standalone NAA nodes and its comparison with a homogeneous CPU-based cluster.

The following paper is structured as follows: In Sect. 2, the background and related work regarding an optimal NAA communication model, LCA for DCs and NAA use cases is presented. Section 3 describes our FPGA hardware framework for NAA, which incorporates one-sided Remote Direct Memory Access (RDMA) communication. It also introduces a flexible and scalable energy measurement system for DCs. A review of the environmental impacts over the complete life cycle of an NAA cluster compared to a homogeneous cluster is conducted in Sect. 4. Section 5 summarizes and gives an outlook on our future developments.

2 Related Work

2.1 Communication Model for Network-Attached Accelerators

In the past, high-speed interfaces were usually used for coupling the appropriate accelerators, such as Peripheral Component Interconnect Express (PCIe). For coupling in a multiprocessor configuration, even more tightly coupled processor interfaces such as QuickPath Interconnect (QPI), which provide cache coherence, have been used. Due to their high data rates, these interfaces allow very close coupling with the program flow of the main process and are suited for communication-bound compute problems. However, the decisive factor for selecting an adequate compute accelerator is not only the question of the available bandwidth of the interfaces, but the speedup including the communication time over the compute time on the host. It can be observed that there is a class of computing problems that require only a small amount of data, and thus a small bandwidth for actual communication, which very quickly become compute-bound instead of communication-bound.

This relationship can be visualized by the so-called Roofline model [14]. Figure 1 shows the relationship between achieved computational complexity and required computational complexity. In the figure, applications that are in the right area of the graph of the Roofline model are particularly well suited, e.g. the MobileNetV2 inference kernel used in the analysis.

2.2 Environmental Life Cycle Assessment for Data Centers

To evaluate the environmental impact of a DC, numerous indicators based on a literature review are presented in [17]. The indicators describe the impacts at

different system levels and are classified according to the objectives: energy consumption, Global Warming Potential (GWP), raw materials and others such as water consumption. Some indicators include more than one objective. The most common efficiency indicator is Power Usage Effectiveness (PUE), which describes the efficiency of building services as the ratio between the used energy of the whole DC and the consumed energy of Information Technology (IT) devices like compute nodes, storage servers and switches. Whether IT systems perform relevant tasks with the energy consumed is not part of the PUE and thus a weak point of this indicator.

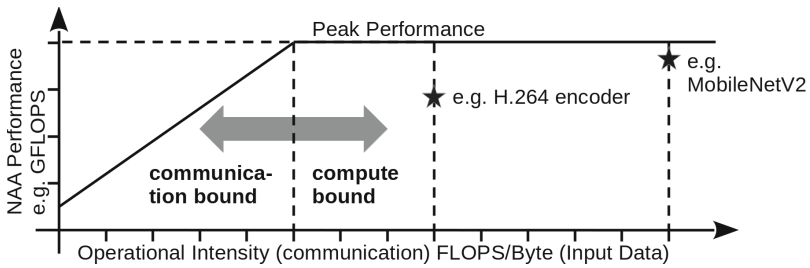


Fig. 1. NAA communication Roofline model.

At the level of the entire DC, which includes building services and IT equipment, only 8 indicators are applicable, of which only Data Center Performance Per Energy (DPPE) takes into account energy demand, GWP and raw material demand. Moreover, almost all indicators including DPPE consider only the usage phase, which is insufficient for a comprehensive LCA.

$$\text{KPI} = \frac{\text{benefit [e.g. ops]}}{\text{effort [e.g. kg Sb.eq./a]}} \quad (1)$$

Therefore, the study [17] developed the indicator set Key Performance Indicators for Data Center Efficiency (KPI4DCE) as a quotient of benefit to effort (cf. Eq. (1)), where the benefit metric is throughput in operations per second (ops). It relates the generic benefit of the DC to the environmental effort in the sub-areas computing power of the nodes, utilized storage capacity, external data traffic as well as infrastructure of the DC for Cumulative Energy Demand (CED), GWP, Abiotic Depletion Potential (ADP) and water consumption. Hence, 16 sub-indicators exist. As a simplification, only the manufacturing and usage phases are considered for KPI4DCE, since the influence of the transportation and disposal phases is marginal according to the case studies in [8, 17]. In [8], the indicator set was developed further, and the database was updated. The KPI4DCE effort indicators in detail are:

- **ADP**: usage of non-renewable raw materials and minerals in kg of antimony equivalents per year [kg Sb.eq./a].

- **CED**: consumption of non-renewable and renewable energy resources in megajoules per year [MJ/a].
- **GWP**: effect on global warming in kilograms of carbon dioxide equivalents per year [$kg\ CO_2\ eq./a$].
- **Water**: Water consumption in cubic meters per year [m^3/a].

2.3 Use Cases for Network-Attached Accelerators

The survey by Kachris et al. shows significant performance and energy efficiency gains for FPGA-based Map Reduce and FPGA-based Key-value Database (KVD) applications, which both are part of Group 3 (cf. Sect. 1) [13]. This highlights that instead of processors, NAAs are the best accelerators for KVD applications.

In [24], NAAs are presented for speeding up text tokenization with regular expressions that transfers natural language into a structured form as a prerequisite for subsequent text analyzes. A control node forwards through a 10 Gigabit Ethernet (GbE) interface a text document to be analyzed to a process pipeline consisting of 2 NAAs, with the last NAA sending the results back to the server. Compared to a software solution with 2 servers as processing units instead of NAAs, the throughput was increased by 14–18 times depending on the document size, the latency was reduced 12–40-fold and the response time variance was reduced to 0.5 ms from 3–4 ms. Compared to employing tightly-coupled FPGAs in the nodes, the NAAs were able to increase the throughput by 10.8–14.8, the latency was reduced up to 1.11x, and the variance of the response times was reduced to the same extent as in the software solution. The server system consumes a total of 600 W and the tightly-coupled FPGA solution requires overall 650 W. The NAA architecture, on the other hand, requires only 250 W with increased throughput, which improves energy efficiency by 33.6x–43.2x compared to the software solution. Tightly-coupled FPGA increase energy efficiency by only 10x–14x.

The acceleration of a Jacobi 2D Stencil is shown for an upgrade of the NAA approach of [24] in a cluster with up to 31 NAAs (each with 10 GbE) and one CPU [15]. The application is automatically synthesized for the FPGA using Message Passing Interface (MPI) transpilation and communication is based on a MPI/User Datagram Protocol (UDP) stack. For data sizes from 16×16 to 1024×1024 , speedups over a homogeneous CPU cluster of 1.96–5.55 are achieved for different cluster sizes. The energy efficiency increases by 5.74–31.31.

In [5], the authors investigate the acceleration of Monte-Carlo European Option Pricing (MCE), which is applied in the financial community for the pricing of an option with multiple uncertainties, for NAAs among others. The high-end GPU A100 with 7 nm node technology achieves the highest throughput, however, the end-to-end response time of the NAA (mid-range FPGA) with 20 nm node technology for cold runs, which is particularly important in the financial industry, is 3 orders of magnitude better. A more recent and performant FPGA can certainly improve the throughput significantly. This is underlined by the fact that a 3 times larger tightly-coupled FPGA with 16 nm achieves 7–8

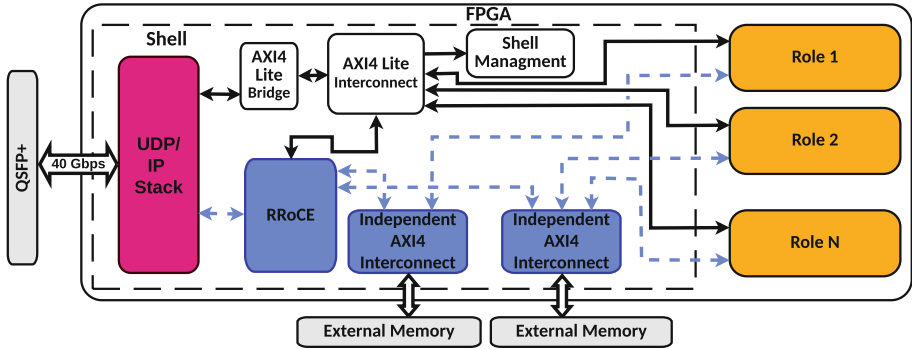


Fig. 2. Exemplary NAA hardware framework.

times the throughput of the NAA due to more parallel MCE cores. However, the cold start time of the tightly-coupled FPGA is worse, so the authors do not investigate this solution further.

In [2], a comparison of a network protocol load balancer between Virtual Machine (VM)-based software and NAA, which needs an extra serial port for control purposes, is shown. The VM already experiences packet drops and latency variance starting at 25 MBps. In contrast, the NAA can operate up to 100 MBps without losses and constant latency variance.

In addition, there are some applications that have not been implemented on standalone NAAs, but can probably be easily adapted. Due to space limitations, they are not presented in detail such as tightly-coupled accelerators with direct network access [4, 12] or FPGA-based switches for In-Network data Processing (INP) [9].

The presented use cases for the NAAs reveal the throughput, latency, and energy efficiency advantages of the distributed NAA architecture in a heterogeneous DC for certain problems, which motivates us to also investigate this architecture in terms of resource efficiency and manufacturing phase impact.

3 Exemplary NAA Framework

For the NAAs, we adopt the hardware abstraction layer described in [20] as an exemplary hardware framework that divides the FPGA into a static shell and up to N roles (cf. Fig. 2). The shell provides a 40 GbE with UDP/IP stack as communication interface and Routable RDMA over Converged Ethernet (RRoCE) protocol stack based on it, which is introduced in more detail in [16]. RRoCE is intended for reliable, connection-oriented RDMA communication via one-sided WRITE transfers, especially between servers and NAAs. This permits scalable and interoperable communication in a heterogeneous DC with a low processor load on the server side. A 512-bit Advanced eXtensible Interface Bus (AXI)-4 interconnect is used to access external memory. The roles are managed by a

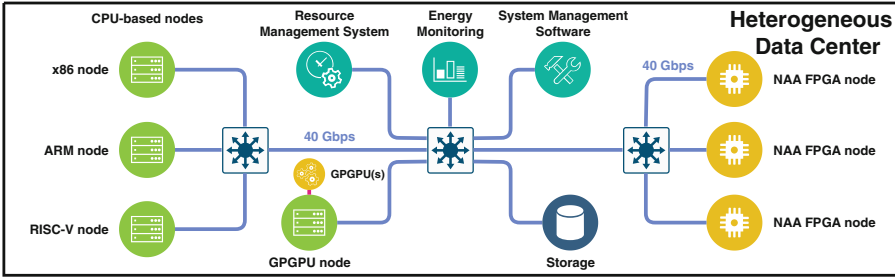


Fig. 3. Heterogeneous DC with NAA nodes.

lightweight AXI-4-Lite system, that can only be controlled by a manager from within the shell to prevent unwanted control of one role by another. Further possibilities of the framework such as partial reconfiguration or streaming communication of the network stack with the roles are not used for performance reasons, even though these functionalities are important in a real DC environment.

4 Environmental Life Cycle Assessment of NAA Nodes

4.1 Initial KPI4DCE Observations

For the LCA, a heterogeneous cluster accelerated by NAA nodes (cf. Fig. 3) is compared against a homogeneous cluster as baseline architecture, which relies on classical CPU-based nodes typically found in DCs. The comparison is based on the indicator set KPI4DCE and the advancements in [8]. KPI4DCE applies the integer portion of the Standard Performance Evaluation Corporation (SPEC)-2006 benchmark as a measure of *beneficial computing power*. This benchmark contains 12 individual benchmarks [19], which are not readily executable on an application-specific FPGA accelerator. Instead of abstract benchmarks, we apply real life benchmarks that represent relevant applications, as described below. According to [8], the adoption of own benchmarks instead of SPEC-2006 provides comparable results as long as the benchmarks are adopted on all compared systems.

MobileNetV2: MobileNetV2, as a current Deep Neural Network (DNN) for image classification on 224×224 images, is part of an important application category and was thus considered as benchmark. We rely on the work introduced in [16] with two MobileNetV2 roles per NAA. The images to be classified are aggregated and transmitted via RRoCE to the NAA nodes for classification. From there, the results are sent back to a server via RRoCE for further processing. For our tests we used the Imagenet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012) validation data set with 50000 images.

H.264 Encoder: Video transcoding is needed for internet video platforms or social media to adapt video resolution and quality to different devices. Encoding with for example H.264/Advanced Video Coding (AVC) is the computationally intensive part, which is an element of SPEC-2006. The employed NAA implementation with two parallel H.264 High Definition (HD) encoders running at 30 frames per second (fps) is based on the work presented in [21]. The decoded video data is transmitted via 40 GbE using RRoCE to the NAAs, where they are encoded and then sent to a server via RRoCE for further playout. The functionally identical C reference software is used as the software encoder. As test sequence (SteamLocomotive) with 1920×1080 pixels, YUV 4:2:0, 8-bit color depth and 300 frames targeting 5 Mbps was encoded.

The total manufacturing expenses in all categories (CED, GWP, ADP and water consumption) are distributed over the expected service periods in years given in Tables 1 and 2 [8, 17]. In the majority of DCs examined in both studies, the ADP is dominated by the manufacturing phase, even when excluding significant portions of building services due to their low impact in the analysis [8, 17]. The remaining ADP in the usage phase is caused by the combustion of fossil fuels to generate electricity. To increase resource efficiency, it is advisable to maximize the lifespan of IT equipment. But this creates a trade-off with energy footprint, as more efficient IT devices help reduce that consumption. However, in the other categories (CED, GWP, and water consumption), the usage phase dominates.

In an LCA for electronic products, just 10% of the components contribute 90% to the GWP, which is used as a simplified indicator in [22]. Therefore, the consideration of the main contributions is particularly relevant, which are in descending order in a DC context [22]:

1. Integrated Circuits (ICs): active semiconductors like memory (DRAM, Solid-State Drive (SSD), HDD) or logic (CPU, GPU, FPGA).
2. Printed Circuit Board (PCB): material (substrate, finish and solder).
3. Ports: power and communication interfaces such as electrical/optical connectors and cables.
4. Chassis: housing materials plus cooling.

Despite numerous uncertainties, such as the assessment of the benefit of IT operations, the usage of accelerators, the neglect of internal network traffic or, in particular, the very incomplete data basis for the resource consumption of IT components, the authors in [17] consider KPI4DCE as robust and reliable in trend. The authors of [8, 17] see further research required to improve the database, especially for the determination of ADP, since the number of data sets in the electronic area is small, complex to create, and they also quickly become outdated. For the calculation of KPI4DCE, an Excel-based tool has been published [7], which is subsequently used in version 2.5.

4.2 CPU-Based Nodes

For the examination of the homogeneous CPU-based cluster as a baseline architecture, the resources shown in Table 1 are assumed. Due to the absence of

Table 1. CPU-based cluster resources.

#	Qty.	Type	Lifetime	Description
1	80	compute nodes	5 years	Intel Xeon Silver 4114, 10C@2.2 GHz, 13.75 MiB L3 Cache, 6 × 8 GiB DDR4 SDRAM, 1 TB HDD, 40 GbE, 2×PWS-1K43F-1R power supply
2	1	control server	5 years	same as #1
3	1	storage server	6 years	same as #1 but 8 × 4 TB HDD
4	3	40 GbE switch	7 years	with 32 ports

a server cluster, the software implementations are run on a single node (running a bare metal Ubuntu 20.04.5 LTS) and the results are extrapolated. In our experience, this extrapolation leads to a negligible error for the application type used, since the applications are embarrassingly parallel. In addition, the energy requirements of the infrastructure were taken into account with PUE=1.2 according to a typical PUE of an energy-efficient DC [10]. For modeling the GWP of electricity consumption, the medium-voltage electricity mix of Germany is assumed based on environmental LCA database Ecoinvent V3.5 (published 2018) [8]. The power measurement of the servers was carried out via the Power Management Bus (PMBus) of the power supply via the Baseboard Management Controller (BMC). The measured idle power of a computing node is 100 W. A typical power consumption of 150 W with passive copper cabling is reported per switch [18]. For classical DCs without accelerators, we believe that the KPI4DCE tool can be applied well. No additional assumptions had to be taken regarding the servers, the storage servers or the network infrastructure.

MobileNetV2: The CPU-based MobileNetV2 runs on 10 physical cores, using thread pinning, by means of ONNX runtime. Using more cores did not result in more throughput due to hyperthreading overhead. Per compute node, using a batch size of 20 frames, MobileNetV2 classifies 182.67 fps (measured with 10 iterations), which is the benefit. This yields a system performance of 14613.6 fps or 414.77 billion frames/year assuming a cluster utilization over the year of 90%. The energy consumption amounts to 111 MWh per year. With the DC components and infrastructure, this adds up to a demand of 142.84 MWh, resulting in an electrical energy expenditure of 1239.78 mJ per frame during the usage phase.

H.264 Encoder: Each compute node encodes 20 parallel HD video streams employing all CPU cores. The benefit is defined as the number of encoded fps. A frame rate of 4.06 fps per node is achieved and 166 W is consumed during encoding measured with 10 iterations. The 80 nodes thus encode 9.21 billion images/year at an assumed average CPU utilization of 90%, consuming 112.92 MWh during this time. The remaining components of the cluster increase the energy consumption along with an energy consumption of 24 MWh to simulate the PUE to 145.24 MWh per year. Per image, this corresponds to an electrical energy expenditure of 56718.4 mJ in usage phase.

Table 2. Resources of NAA-accelerated cluster.

#	Qty.	Type	Lifetime	Description
1	1	control server	5 years	Intel Xeon Silver 4114, 10C@2.2 GHz, 13.75 MiB L3 Cache, 6 × 8 GiB DDR4 SDRAM, 1 TB HDD, 40 GbE, 2×PWS-1K43F-1R power supply
2	1	NAA	5 years	8 NAAs (10AX115N3F40E2SG), each with 2×4GiB DDR3 SDRAM and 40 GbE, ASPOWER R2A-DV0800-N with 2 redundant power supplies
3	1	storage server	6 years	same as #1 but 8 × 4 TB HDD
4	1	40 GbE switch	7 years	with 12 ports

4.3 NAA-Based Nodes

The heterogeneous cluster accelerated with NAA-based nodes consists of the components shown in Table 2. Compared to Table 1, the compute nodes have been replaced by 8 NAAs in a chassis equipped with simple backplanes (just 2 layers) and components for power supply and cooling. The same assumptions regarding PUE and electricity supply of the DC are made as in Table 1. The idle power averages at 260 W after a few minutes of runtime. The 385A PCB from Bittware was selected as NAA [1]. The switches were substituted with a scaled-down model with a measured average power of 40 W, since the NAA cluster requires fewer ports. A more detailed description of the NAA node can be found in [16].

For the adaptation of KPI4DCE to standalone NAA chassis, some assumptions have to be stated for the application of the KPI4DCE tool. To determine the effort in the manufacturing phase for logic ICs, KPI4DCE applies the formula Eq. (2), which takes the number of CPU cores as a measure of the die size [8]. The die size is used to infer the effort using manufacturing data from an Intel factory in Ireland from 2017 and 2018.

$$\text{CPUDieSize}[\text{cm}^2] = 0.24584 \cdot \text{CoresNumberPerCPU} + 0.49157 \quad (2)$$

When modeling the NAA architecture for KPI4DCE, it should be noted that an FPGA as a spatial architecture cannot be compared with the invariant cores of a CPU. However, to enable modeling nevertheless, we have inferred an equivalent number of CPU cores based on the die size of the FPGA according to the formula Eq. (3):

$$\text{CoresNumberPerCPU} = (\text{CPUDieSize}[\text{mm}^2]/100 - 0.49157)/0.24584 \quad (3)$$

The die area of FPGAs is usually not publicly known, in contrast to the package size. However, for the 10AX115N3F40I2SGES FPGA, which is part of the Arria 10 GX family, this information was published in a forum by the manufacturer [11]. This FPGA is the equivalent of the 10AX115N3F40E2SG used in the environmental assessment except for the temperature range and that it is an engineering sample. Neither factor should affect the die area. The stated die area

is 337.9 mm^2 , which results in an equivalent number of 11.7 CPU cores, rounded up to 12, according to Eq. (3). The estimate is subject to large uncertainties due to insufficient data, since other process technologies are applied for FPGA manufacturing, so equating them with CPUs is only a rough approximation. Also, even the database for manufacturing CPUs based on only one fab from 2 years is very poor. The influence of the external memory on the FPGA PCB is modeled by specifying it as a RAM module, and the flash memory for booting the FPGA is modeled as an SSD.

The influence of the FPGA PCB, each with 115.48 cm^2 and unknown number of layers [1], is not modeled by the chosen approach. However, the impact can be modeled approximately through the backplane, which is considered with the fixed area 1006.5 cm^2 and 6 layers for all server types in the KPI4DCE tool [8, 17]. Together with our backplane, which is 508.5 cm^2 in area with 2 layers, this gives a total PCB area of 1432.3 cm^2 . To accurately reflect the impact of the PCBs, a correction factor for the PCB area compared to the static PCB area of the modeled servers was added to the KPI4DCE tool and set to 1.423 for the PCBs of the NAA chassis including 8 NAAs.

The power distribution is modeled through the NAA enclosure, which corresponds to a server enclosure. Network cables are not part of the calculation tool [17], but since the required number for the NAA-accelerated DC is lower due to the smaller node count, this simplification is slightly unfair towards the heterogeneous DC. No further assumptions need to be made for the NAA chassis regarding housing and cooling compared to the modeled server chassis, since it consists of the same components as the latter.

Overall, we can only agree with the authors of [8, 17] and see major research required in the adoption of accelerator technologies such as GPUs and FPGAs.

MobileNetV2: The classification performance of the 8 NAAs is 10340 fps employing a batch size of 128 at a consumption of 420 W. Due to the higher throughput, the classification was carried out in a continuous loop and the measurement was performed over a period of 15 min after a startup phase of 5 min. This results in an annual output of 293.5 billion frames/year at 90% utilization with an electrical energy consumption of 140.6 mJ/frame including the DC overhead (classification only needs mJ/frame). Compared to the CPU-based classification, the energy consumption per frame in the utilization phase was reduced by a factor of 8.82.

H.264 Encoder: The 8 NAAs, with an assumed power consumption of 450 W, can encode 480 HD frames per second. At an expected workload of 90%, this corresponds to 13.6 billion frames/year and an electrical energy consumption of 3098,6 mJ per encoded frame, which equates to an efficiency increase of 18.31 in the usage phase.

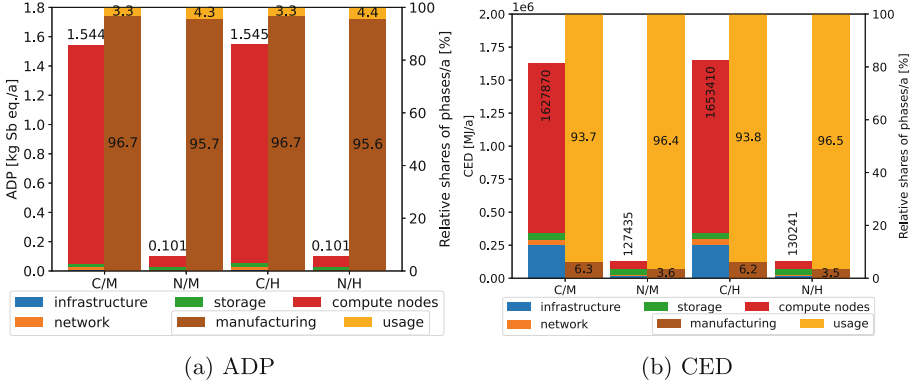


Fig. 4. Absolute ADP and CED for CPU-based cluster (C) and NAA-accelerated cluster (N) for H.264 (H) and MobileNetV2 (M). Relative shares of the manufacturing and usage phases per year.

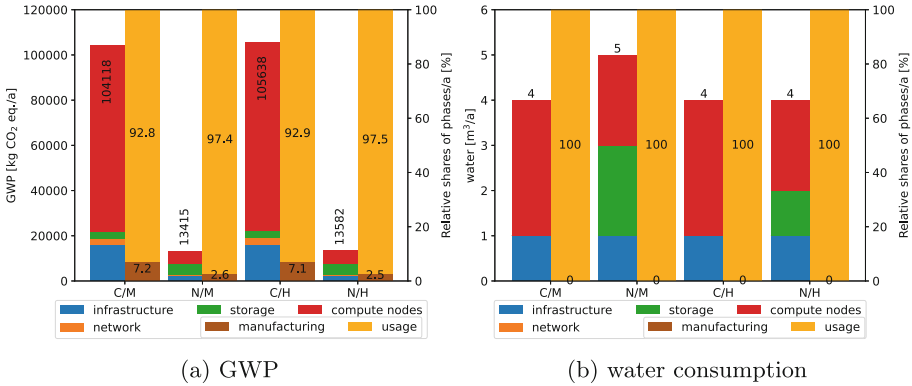
4.4 KPI4DCE Evaluation

The PUE is assumed to be constant at 1.2 for both clusters, since the infrastructure of the DC can potentially be reduced to the same extent due to the lower requirements of fewer computing nodes. The simplification is not relevant for the comparison of the two clusters with each other. The same service life is assumed for NAAs as for servers for better comparability. However, due to the typically lower energy consumption in the usage phase and the slower product cycle for FPGAs, longer lifetimes are reasonable. It is evident from Fig. 4a that the heterogeneous cluster accelerated with NAAs (cf. Table 2) has a reduced absolute ADP compared to the homogeneous CPU-based cluster (cf. Table 1). Note that the CPU-based cluster provides higher benefit for MobileNetV2, but lower benefit for H.264. Normalizing the indicators to effort following Eq. (1) restores comparability, as seen in Table 3. Thus, for MobileNetV2, the ADP of the NAA-accelerated cluster is 10.8x better and H.264 even 22.6x. Due to the high system performance of the NAAs and the high energy efficiency, a targeted performance can be achieved with a lower node number. The smaller compute node number as well as the generally smaller resource requirements of a NAA node, caused by saving the host server, are the main reasons for the improvement of the resource efficiency. The same compute resources are used for both benchmarks (H.264, MobileNetV2), which explains why the ADP is nearly identical. It differs only by the ADP part of the electrical supply. The relative allocation of the ADP to the manufacturing and usage phase for one year is also shown in Fig. 4a. It can be seen that the manufacturing phase dominates for all use cases. This is consistent with the initial considerations presented in Sect. 4.1.

In Fig. 4b, the CED of the clusters is shown for the different use cases. It is evident that the NAA-accelerated cluster consumes significantly less energy due to the much improved energy efficiency as well as the infrastructure adapted to the smaller node number. In contrast to ADP, CED is dominated by the usage

Table 3. To effort normalized KPI4DCE for CPU-based cluster (C) and NAA-accelerated cluster (N) for H.264 (H) and MobileNetV2 (M). Higher is better.

#	[effort]							
	ADP [kg Sb eq./a]	Im.	CED [MJ/a]	Im.	GWP [CO ₂ eq./a]	Im.	water [m ³]	Im.
C/M [frames]	268.6 G	1x	254793	1x	3983644	1x	103692 M	1x
N/M [frames]	2905.7 G	10.8x	2302931	9x	21876557	5.5x	73369 M	0.7x
C/H [frames]	6.0 G	1x	5576	1x	87266	1x	2305 M	1x
N/M [frames]	134.9 G	22.6x	104603	18.8x	1003059	11.5x	3406 M	1.5x

**Fig. 5.** Absolute GWP and water usage for CPU-based cluster (C) and NAA-accelerated cluster (N) for H.264 (H) and MobileNetV2 (M). Relative shares of the manufacturing and usage phases per year.

phase in all cases. The clarity of the relative distribution is due to the high utilization of the DC, which optimally exploits the fixed effort of the manufacturing phase.

In Fig. 5a the GWP of the clusters is depicted, which is fed from coolant leakages and from the fossil shares of the DC energy supply according to the applied electricity mix. Compared to the homogeneous cluster, the NAA-accelerated cluster emits fewer greenhouse gases due to lower CED. From the normalized numbers in Table 3, it is evident that the unchanged rate of coolant leakage decreases the GWP for the NAA accelerated cluster less than the CED. As expected, the usage phase dominates the GWP in all cases. In order to reduce the CED and thus the GWP, it is particularly worthwhile to optimize the usage phase, for example by taking advantage of energy-efficient sleep states for unused components both at the system level (energy-saving mode for complete nodes) and within a node (e.g., shutting down unused DRAM).

Figure 5b reports the direct water consumption of the DC. This is determined by the cooling systems, which is allocated by the KPI4DCE tool (with rounding errors) into the categories infrastructure, network, storage systems and compute

nodes on a percentage basis. Therefore, the water consumption for the NAA-accelerated cluster tends to be overestimated. The reduced number of nodes is expected to result in lower cooling requirements and thus, after adjusting the cooling capacity, in reduced water consumption.

5 Conclusion

For the given examples, we can summarize that the NAA-accelerated cluster performs significantly better than the homogeneous cluster with CPU-based compute nodes in terms of resource efficiency (improved by up to 22.6), energy efficiency (improved by up to 18.8), and greenhouse gas emissions (improved by up to 11.5), as evident in Table 3. Therefore, we consider standalone NAA as an ideal addition to a heterogeneous DC to increase energy and resource efficiency and thus reduce GWP.

Our future work will focus on the development of a software framework with hardware support for energy measurement of NAAs. In addition, we will investigate further use cases, possibly with other FPGAs, such as Agilix, using KPI4DCE.

References

1. Bittware 385a. <https://www.bittware.com/fpga/385a/>. Accessed 06 Apr 2023
2. Byma, S., Steffan, J.G., Bannazadeh, H., Leon-Garcia, A., Chow, P.: FPGAs in the cloud: booting virtualized hardware accelerators with OpenStack. In: 2014 IEEE 22nd FCCM. IEEE (2014). <https://doi.org/10.1109/fccm.2014.42>
3. Caulfield, A.M., Chung, E.S., Putnam, A., Angepat, H., et al.: A cloud-scale acceleration architecture. In: 2016 49th IEEE/ACM MICRO, Taipei, Taiwan, pp. 1–13. IEEE (2016). <https://doi.org/10.1109/MICRO.2016.7783710>
4. Chung, E., Fowers, J., Ovtcharov, K., Papamichael, M., et al.: Serving DNNs in real time at datacenter scale with project brainwave. *IEEE Micro* **38**, 8–20 (2018). <https://doi.org/10.1109/MM.2018.022071131>
5. Diamantopoulos, D., Polig, R., Ringlein, B., Purandare, M., et al.: Acceleration-as-a- μ Service: a cloud-native Monte-Carlo option pricing engine on CPUs, GPUs and disaggregated FPGAs. In: 2021 IEEE 14th CLOUD, Chicago, IL, USA, pp. 726–729. IEEE (2021). <https://doi.org/10.1109/CLOUD53861.2021.00096>
6. Escobar, F.A., Chang, X., Valderrama, C.: Suitability analysis of FPGAs for heterogeneous platforms in HPC. *IEEE TPDS* **27**, 600–612 (2016). <https://doi.org/10.1109/TPDS.2015.2407896>
7. Gröger, J., Liu, R.: Green cloud computing. <https://www.oeko.de/publikationen/p-details/green-cloud-computing>. Accessed 29 Mar 2023
8. Gröger, J., Liu, R., Stobbe, L., et al.: Green cloud computing. Technical report, UBA (2021). https://www.umweltbundesamt.de/sites/default/files/medien/5750/publikationen/2021-06-17_texte_94-2021_green-cloud-computing.pdf
9. Hartmann, M., Weber, L., Wirth, J., Sommer, L., Koch, A.: Optimizing a hardware network stack to realize an in-network ML inference application. In: 2021 IEEE/ACM H2RC, St. Louis, MO, USA, pp. 21–32. IEEE (2021). <https://doi.org/10.1109/H2RC54759.2021.00008>

10. High-Performance Computing Data Center Power Usage Effectiveness. <https://www.nrel.gov/computational-science/measuring-efficiency-pue.html>. Accessed 06 Apr 2023
11. How tall is the die for the 10ax115n3f40i2sges? <https://community.intel.com/t5/Programmable-Devices/How-tall-is-the-die-for-the-10AX115N3F40I2SGES/mp/592124>. Accessed 05 Apr 2023
12. Javaid, H., Yang, J., Santoso, N., Upadhyay, M., et al.: Blockchain machine: a network-attached hardware accelerator for hyperledger fabric. In: 2022 IEEE 42nd ICDCS, Bologna, Italy, pp. 258–268. IEEE (2022). <https://doi.org/10.1109/ICDCS54860.2022.00033>
13. Kachris, C., Soudris, D.: A survey on reconfigurable accelerators for cloud computing. In: 2016 26th IEEE FPL. IEEE (2016). <https://doi.org/10.1109/fpl.2016.7577381>
14. Ofenbeck, G., Steinmann, R., Caparros, V., Spampinato, D.G., Püschel, M.: Applying the roofline model. In: 2014 IEEE ISPASS, pp. 76–85 (2014). <https://doi.org/10.1109/ISPASS.2014.6844463>
15. Ringlein, B., Abel, F., Ditter, A., Weiss, B., et al.: Programming reconfigurable heterogeneous computing clusters using MPI with transpilation. In: 2020 IEEE/ACM H2RC, GA, USA, pp. 1–9. IEEE (2020). <https://doi.org/10.1109/H2RC51942.2020.00006>
16. Schelten, N., Steinert, F., Knapheide, J., Schulte, A., Stabernack, B.: A high-throughput, resource-efficient implementation of the RoCEv2 remote DMA protocol and its application. *ACM Trans. Reconfigurable Technol. Syst.* 1–23 (2022). <https://doi.org/10.1145/3543176>
17. Schödwel, B., Zarnekow, R., Liu, R., Gröger, J., Wilkens, M.: Kennzahlen und Indikatoren für die Beurteilung der Ressourceneffizienz von Rechenzentren und Prüfung der praktischen Anwendbarkeit. Technical report, UBA (2018). https://www.umweltbundesamt.de/sites/default/files/medien/1410/publikationen/2018-02-23_texte_19-2018_ressourceneffizienz-rechenzentren.pdf
18. Sn2700 open ethernet switch. Technical report, Mellanox Technologies (2019)
19. SPEC CINT2006: Integer component of spec CPU2006 (2006). <https://www.spec.org/cpu2006/CINT2006/>
20. Steinert, F., Schelten, N., Schulte, A., Stabernack, B.: Hardware and software components towards the integration of network-attached accelerators into data centers. In: 2020 23rd Euromicro DSD, Kranj, Slovenia, pp. 149–153. IEEE (2020). <https://doi.org/10.1109/DSD51259.2020.00033>
21. Steinert, F., Stabernack, B.: Architecture of a low latency h.264/AVC video codec for robust ML based image classification. *J. Sign. Process. Syst.* **94**(7), 693–708 (2022). <https://doi.org/10.1007/s11265-021-01727-2>
22. Stobbe, L.: Workshop: Grundlagen der Ökobilanzierung und methodisches vorgehen für die umweltpotenzialbewertung. Technical report, Fraunhofer IZM (2023). <https://owncloud.fraunhofer.de/index.php/s/IzOQHShUbTMbMX5#pdfviewer>
23. Weerasinghe, J., Abel, F., Hagleitner, C., Herkersdorf, A.: Disaggregated FPGAs: network performance comparison against bare-metal servers, virtual machines and Linux containers. In: 2016 IEEE CloudCom, Luxembourg, Luxembourg, pp. 9–17. IEEE (2016). <https://doi.org/10.1109/CloudCom.2016.0018>
24. Weerasinghe, J., Polig, R., Abel, F., Hagleitner, C.: Network-attached FPGAs for data center applications. In: 2016 IEEE FPT, Xi'an, China, pp. 36–43. IEEE (2016). <https://doi.org/10.1109/FPT.2016.7929186>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

