Energy Efficiency of Parallel File Systems on an ARM Cluster

Timm Leon Erxleben*, Kira Duwe*, Jens Saak†, Martin Köhler† and Michael Kuhn*
*Otto von Guericke University Magdeburg
Magdeburg, Germany
E-mail: timm.erxleben@ovgu.de, kira.duwe@ovgu.de, michael.kuhn@ovgu.de
†Max Planck Institute for Dynamics of Complex Technical Systems
Magdeburg, Germany
E-mail: saak@mpi-magdeburg.mpg.de, koehlerm@mpi-magdeburg.mpg.de

***Abstract*—Parallel distributed file systems are typically run on dedicated storage servers that clients connect to via the network. Regular x86 servers provide high computational power, often not required for storage management and handling I/O requests. Therefore, storage servers often use low core counts but still have a relatively high idle power consumption. This leads to high energy consumption, even for mostly idle file systems. Advanced Reduced Instruction Set Computer Machines (ARM) systems are very energy-efficient but still provide adequate performance for file system use cases. Leveraging this fact, we built an ARM-based storage system, on which we tested both CephFS and OrangeFS. We compare the performance and energy efficiency of x86 and ARM systems using several metrics. Results show that while our ARM-based approach currently provides less throughput per Watt for reads, it achieves an approximately 121 % higher write efficiency when compared to a traditional x86 Ceph cluster.**

***Keywords*—energy efficiency, CephFS, OrangeFS, x86, ARM**

I. Introduction

Storage systems are scaled up steadily to satisfy increasing storage demands, leading to growing energy consumption [1]. High-Performance Computing (HPC) storage systems are currently built from regular x86 servers, whose computing power is not fully utilized by storage applications. Traditional x86 servers feature a relatively high power consumption even when idle: It is not uncommon to measure idle consumption of more than 100 W for just the processor, main memory, and mainboard. In comparison, low-power ARM computers are often required to stay below 5–10 W maximum consumption by design. To offset the high idle consumption of x86 servers, they have to be equipped with large amounts of storage devices, such as hard disk drives (HDDs) and solid-state disk (SSDs). However, depending on the used network interconnect, only a limited number of devices can be saturated. For instance, on a 100 Gbit/s network, two to three NVMe SSDs are enough to provide the necessary throughput. This proportion gets even worse on slower networks.

Therefore, we evaluate the use of low-energy ARM-based single-board computers as a replacement for traditional servers in storage systems. To assess the feasibility of an ARM-based storage system, we evaluated the ARM cluster using CephFS and OrangeFS. Furthermore, we compared it to a productive CephFS cluster running at the computer science faculty of the Otto von Guericke University, using different metrics.

The contributions of our paper are:

1) We propose to apply the energy-delay product, typically used to evaluate the energy efficiency of computations, as a metric for storage systems as well to measure energy efficiency while still accounting for the performance needed by HPC applications.
2) We show that low-power ARM-based storage clusters can achieve throughput efficiencies comparable to or even exceeding traditional x86 systems.

The remainder of the paper is organized as follows. In section II, CephFS and OrangeFS are briefly described followed by a summary of related works. Section III describes the benchmarks which were done and discusses metrics that can be derived from the measurement data. Next, in section IV both cluster setups, ARM and x86, are described, followed by the presentation of the results. Results and setups are discussed in section V. Finally, section VI concludes the paper.

II. Background and Related Work

This section introduces background on used technologies, such as Ceph and OrangeFS, and related work.

a) Ceph: Ceph is a popular, clustered object store, which is highly scalable due to its Controlled Replication Under Scalable Hashing (CRUSH) placement algorithm, which enables all participating services, that can access the cluster map to locate and place objects [2]. A typical Ceph cluster is made of Object Storage Devices (OSDs), monitoring and management services. All components may be redundant to enable automatic failover. Apart from access through the library `librados`, many interfaces might be used. The POSIX access via CephFS, realized by additional Metadata Services (MDSs) interacting with Ceph storage pools, is particularly interesting for HPC systems. CephFS has a rich feature set, including replication, multiple storage pools, file systems, snapshots, and high control over data placement [3].

b) OrangeFS: OrangeFS is a traditional parallel file system designed for HPC [4]. Only one type of server is needed, which can handle both data and metadata, though it can be configured to handle only one type. In OrangeFS, data is striped according to a distribution function that can be specified for each file. The default is to start at a random server and use all servers in a round-robin fashion with a stripe size of 64 KiB. Unlike Ceph, which uses its own object store *Bluestore* [5], OrangeFS relies on a separate local file system. As of the current version, 2.9.8, there are no redundancy

 ISBN: 978-1-61208-967-6

features for data that is not marked as read-only, though this is planned for OrangeFS version 3 [6]. Many interfaces may be used to interact with OrangeFS. Most popular choices include access via the OrangeFS Linux kernel module or direct access using the library `libpvfs2`. Noteworthy is the direct Message Passing Interface I/O (MPI-IO) support by using ROMIO's [7] Abstract-Device Interface for I/O (ADIO), for which OrangeFS provides an implementation [8].

c) State of the Art and Related Work: There have been various endeavors to measure and increase the energy efficiency of large systems as energy consumption is becoming a possible constraint on HPC systems in the future. Many different aspects have to be considered, ranging from the system's energy efficiency to the scalability of the applications. As ARM processors aim to offer better energy efficiency, they have been heavily studied across the years [9–11]. Deployments, such as Fugaku [12], show that they can provide competitive performance and even work in exascale systems. Earlier research on systems like Tibidabo at Barcelona Supercomputing Center indicated that single instruction multiple, data stream (SIMD) instructions limited to single precision were a severe bottleneck for the performance [10][11][13].

Energy efficiency is also a relevant aspect in distributed systems, as examined for Peer-to-Peer systems. A survey by Brienza et al. showed that often simple energy models were used, disregarding other hardware components like intermediate routers [14]. An early approach, and still very prominent solution to energy savings in storage, is sending idle peers to sleep [15]. However, it introduces problems when the load varies. To have systems benefit from the increased energy efficiency, in the long run, applications have to be considered as well. The optimization towards energy efficiency comes indeed with its challenges for applications [13][16–18]. Reducing the performance of a single core in order to cap the power consumption means that scalability is of increased importance [13].

Gudu and Hardt evaluated the use of an ARM-based Ceph cluster, made of Cubieboards, as a replacement for traditional network-attached storage (NAS) controllers [19]. They measured the throughput of their cluster via Ceph's Reliable Autonomic Distributed Object store (RADOS) and RADOS Block Device (RBD) access and found that the Cubieboard cluster is a viable alternative to NAS controllers. However, the limited network capabilities were the bottleneck of the system.

Apart from using low-power hardware [20], there have been efforts to reduce the power consumption of existing HPC storage clusters [21][22]. For example, it was proposed to assign subsets of storage clusters to specific users and only run them at full power when said user uses the compute-cluster [23].

Considering that local file systems are often part of the storage stack, their influence on energy efficiency and performance were analyzed in [24] using simulated workloads of web, database, and file servers. It was found that the choice of file system and its configuration greatly influence performance and energy efficiency. However, no file system performed best for all workloads.

In contrast to Gudu and Hardt, we measure data throughput at the CephFS level and evaluate ARM-based clusters as a replacement for HPC storage clusters.

III. Benchmark and Metrics

We measured the performance of the clusters for sequential, independent accesses from one to four clients using `IOR` v3.3 [25] with the POSIX backend, individual files per client and five iterations for each data point. The transfer size was set to 4 MiB, which corresponds to the default stripe size of CephFS and is aligned to the stripe size of 64 KiB on OrangeFS. On the x86 Ceph cluster, 96 GiB were written and read. The amount of data was reduced to 36 GiB for the ARM setup to keep run-times manageable.

For every iteration, the power consumption of the storage cluster was measured using the methods as described in Section IV. As a result, several energy efficiency metrics can be derived from the collected data. However, choosing a specific metric is not trivial, as there is no single optimal metric indicating energy efficiency [26].

We decided to compare the results obtained by using the **energy-delay product** (EDP) [27], **throughput per Watt** and **capacity per Watt** [28].

Throughput per Watt is a commonly used metric for evaluating and comparing storage energy efficiency. The transferred data may differ between systems, so it is well suited to compare systems that greatly vary in their performance. However, this metric alone is insufficient when analyzing and optimizing storage systems, as no insight into performance is given. Geveler et al. [16] found that for simulations, in some cases, energy savings might lead to performance drops. In such cases, they motivated using the EDP as a fused metric describing energy efficiency and performance at once. The EDP is computed as the product of the total energy E consumed while performing a task and the time t needed to complete the task (Equation (1)). Depending on the performance requirements, the time may be weighted [29]. As we want to focus on energy consumption, we set $w = 1$.

$$\mathrm{EDP} = E \cdot t^w, \quad w \in \mathbb{N} \tag{1}$$

Though the energy-delay product was initially developed for hardware design, it is also useful when evaluating software, as done by Georgiou et al. [30]. Nevertheless, the amount of work needs to stay constant to compare different systems, so only the two ARM setups are compared using the EDP. Because its unit is hard to interpret and even changes with different weights, we normalized the EDP using the lowest value per comparison.

The third metric considered measures the capacity of the storage system per Watt. Because of growing storage demands and, therefore, growing storage systems, optimizing systems regarding this metric is critical for the cost-efficient and environmentally friendly operation of data centers.

 ISBN: 978-1-61208-967-6

IV. Evaluation

In this section, the hardware and software setup is described, followed by an analysis of the respective clusters' theoretical peak performance and the presentation of the results.

a) Reference Cluster: The reference cluster is a four-node subset of the productive Ceph cluster running at the computer science faculty at the Otto von Guericke University using Ceph 16.2.7 deployed as containers. Three nodes of the subset are part of the Supermicro AS 2124BT-HNTR [31] multi-node system, each of which is equipped with four Intel P4510 NVMe SSDs [32]. The fourth server is a Gigabyte R282-Z94 [33] equipped with one Intel P4510 NVMe SSD and eight Samsung MZQL23T8HCJS-00A07 NVMe SSDs [34]. All nodes are connected by 100 Gbit Ethernet, with a separate 100 Gbit network for communication between Ceph OSDs. Though Ceph does not exclusively use the nodes, they are idle most of the time. The average idle power consumption of the four nodes was measured to be **699.3 W**. This power measurement was done on a Sunday since the servers are mostly idle on the weekend. It lasted for one hour, starting at 14:00, and had a standard deviation of 13.98 W. While running, the benchmark power consumption peaked at 1,057 W. The existing monitoring solution, gathering power samples over IPMI every 15 seconds, was used to collect power samples.

For each SSD, two Ceph OSDs are deployed. The Ceph monitor and a standby metadata service are located at the Gigabyte server, while the active metadata service runs on one of the Supermicro servers. Ceph pools use the default replication settings and, therefore, produce three replicas of the data and return to the client after two replicas are written. The clients used for the benchmark were four servers equipped with an AMD Epyc 7443, with 24 cores at 2.85 GHz, 128 GB RAM, and 100 Gbit Ethernet.

b) ARM cluster nodes: The low-power cluster is built of six Odroid HC4 nodes featuring the Amlogic S905X3 SoC, with four cores at 1.8 GHz, 4 GiB DDR4 RAM, two SATA-3 ports, and a 1 Gbit NIC [35]. We used Armbian Buster [36], and Ceph version 14.2.21, which is available in the Buster backports repository. We built OrangeFS version 2.9.8 with GCC version 8.3.0 and LMDB 0.9.22 from the Buster repository. Four of the nodes are equipped with two 1 TB WD Black HDDs [37] and one is equipped with two 512 GB Samsung V-NAND SSD 860 PRO SSDs [38]. All nodes are connected to a Netgear GS110EMX switch [39].

One OSD is deployed for each storage device. The node which is equipped with SSDs additionally runs one MDS. The Ceph monitor and management daemon run on the sixth node, which has no disks attached. The two storage pools needed for CephFS use different CRUSH rules to distribute objects. While the data pool uses all HDDs and manages replicas on the node level, the metadata pool uses the two SSDs and manages replicas on the OSD level. Both pools are configured to use 64 placement groups. Ceph is configured to produce two replicas and return immediately after one replica is written, allowing a fairer comparison with OrangeFS.

As explained above, OrangeFS has only a single type of daemon, which is running on all nodes with disks. Metadata is stored by the daemon, which is deployed on the SSD node, while the other nodes store the data. As OrangeFS offers no data redundancy for data that is not read-only, ZFS version 2.0.3 was used to mirror disks locally.

The complete cluster, including the switch, is powered by an MW HRP450-15 PSU [40] and consumes 56.36 W, measured over one hour with a standard deviation of 0.14 W, in idle state, with HDDs spun up. For comparison with the reference cluster, which does not include the switch in the power measurements, we subtracted the average idle power of the switch, which was measured to be 15.46 W, with a standard deviation of 1.13 W over one hour. The adjusted idle power consumption of the ARM cluster, therefore, is **40.9 W**. The highest peak in power consumption measured while running the benchmark was 58.9 W.

For power measurements, the ZES Zimmer LMG 450 [41] is used to measure the power consumption of the PSU for the whole cluster. The power meter is connected to one of the clients via USB, which collects samples with 20 Hz. The clients used to perform the benchmark were four Dell Precision 3650 Tower workstations [42] each with an Intel Core i7-11700 CPU with 8 cores at 2.5 GHz, 8 GB RAM, and a 1 Gbit NIC. They were connected via the network infrastructure of the Max Planck Institute Magdeburg.

c) Theoretical Peak Performance: As can be seen in Table I the theoretical peak performance (TPP) of the ARM cluster is limited by the network throughput of each node which is not as high as the aggregated throughput of all storage devices of the node. As no measurements could be made in the productive reference cluster, the maximum throughput of the components is taken from the respective datasheets. Adding together the TPP of the two-node types, the reference cluster's TPP is **47.3 GB/s**.

This analysis neglects metadata operations which are reasonably assumed not to limit the data throughput of the cluster for a few files in use. Furthermore, the table only presents the performance for writes. However, as the network already limits peak performance for the ARM cluster and aggregated throughput of the SSDs in Supermicro nodes of the reference cluster is close to the network speed, the same applies approximately to reads.

d) Results: The results of the performance efficiency metrics are shown in Figure 1. Each value of the throughput per Watt metric is computed as the mean of five samples, each divided by the mean power consumption of their iteration. Error bars on the plots depict the standard deviation. As explained above, the EDP (see Figure 2) is normalized by the lowest value per comparison. The capacity metric was computed using the idle power consumption of the clusters and the raw storage capacity. The usable storage capacity depends on the respective setup. The ARM cluster achieved **0.196 TB/W** and the reference cluster **0.073 TB/W**, see Figure 3.

 ISBN: 978-1-61208-967-6

TABLE I. THROUGHPUT OF COMPONENTS RELEVANT FOR THEORETICAL PEAK PERFORMANCE (TPP) THROUGHPUT

Cluster	Network	Throughput Storage Devices	Storage Devices per Node	# Nodes	TPP
ARM	124.1 MB/s	115 MB/s	2	4	496.4 MB/s
Supermicro	12.5 GB/s	2.9 GB/s	4	3	34.8 GB/s
Gigabyte	12.5 GB/s	2.9 GB/s / 4 GB/s	1+8	1	12.5 GB/s

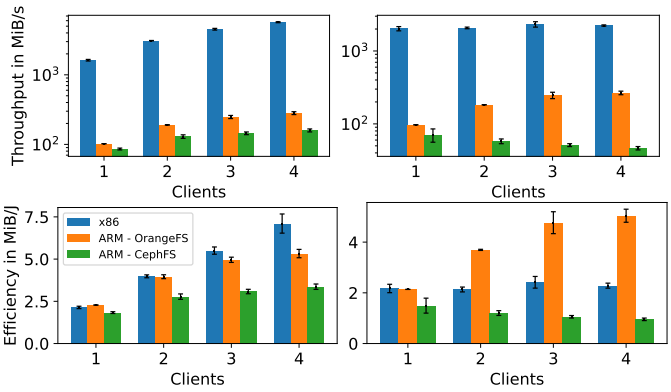

Figure 1. Throughput (top), Throughput per Watt (bottom) for reading (left) and writing (right)

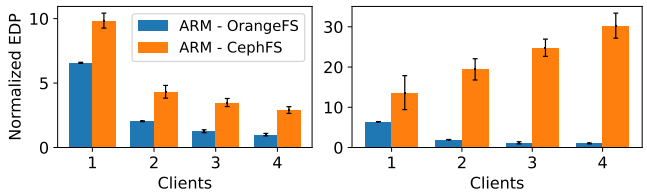

Figure 2. Normalized energy-delay product for reading (left) and writing (right)

V. DISCUSSION

All results need to be seen in relation to the respective systems' cost, as the ARM cluster nodes and disks cost only about € 1,350, while the reference cluster nodes and disks cost around € 40,000. In addition, the reference cluster only uses NVMe SSDs, while the ARM cluster uses HDDs for data object storage. Due to the low sampling rate of the power measurements for the reference cluster, some spikes in the energy consumption are possibly missed, resulting in an underestimation. In contrast, power measurements on the ARM cluster can be expected to overestimate the actual power consumption of the nodes and disks, as only the average idle power consumption of the switch is subtracted.

During previous experiments on a BananaPi M1 single-board computer cluster, the deployment of traditional parallel file systems proved difficult. Tested file systems were CephFS, OrangeFS and BeeGFS. Both CephFS and BeeGFS needed small patches to run on the unusual setup. OrangeFS could not run the client on ARM 32-bit using the upstream kernel module. Additionally, we observed low read throughput if no direct I/O was used. For four clients reading a 2 GiB file each, only 12.41 MiB/s could be achieved. Consequently, measurements on OrangeFS are done with direct I/O.

Our prototype cannot compete with the throughput of the reference cluster. For real world HPC applications, more storage nodes need to be added to achieve higher throughput. This cluster was built as a proof-of-concept for throughput efficiency and to gain insight in ARM single-board computer storage clusters.

The different read and write sizes on both setups were chosen to achieve reasonable run-times of the benchmarks on both settings. Neither throughput nor throughput efficiency are influenced by the different amounts of transferred data if run-times are long enough.

Both clusters show good scaling behavior in all metrics. Exceptions occur for writes. On the reference cluster, one client achieves close to the observed maximum performance, and no further improvement can be seen when adding more clients. In addition, both Ceph-based systems only reached a fraction of the theoretical peak performance, as can be seen in Table II.

For the ARM cluster, this is most likely related to data replication over the public network. Ceph OSDs reported slow operation warnings due to waiting times for sub-operations. As pointed out by Just [43], the Ceph OSD service utilizes many threads, leading to performance issues for a few cores as context switches introduce additional overhead. Ceph's behaviour is strongly influenced by the number of placement groups per OSD [3]. While a higher ratio of placement groups to OSDs ensures a balanced data distribution, management of each placement group consumes memory and CPU time. To minimize overhead we set both pools to 64 placement groups. The number of placement groups per OSD also influences recovery behavior for larger clusters as more placement groups need to be replicated in case of a server crash. Further experiments are needed to evaluate different placement group counts and placement group to OSD ratios for productive usage of Ceph on large ARM clusters.

Nevertheless, replication cannot explain the performance drop for the reference cluster, which needs further investigation. One impacting factor for reads was that only one process

TABLE II. MAXIMUM THROUGHPUT ACHIEVED IN MIB/S AND PERCENT OF TPP.

System	Write / % TPP	Read / % TPP
ARM - CephFS	95.22 / 20.11	172.12 / 36.36
ARM - OrangeFS	289.23 / 61.10	296.82 / 62.70
Reference	2322.47 / 5.15	5705.0 / 12.65

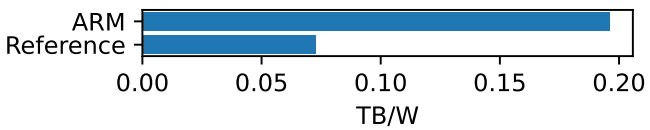

Figure 3. Storage capacity per Watt

 ISBN: 978-1-61208-967-6

per client was used, resulting in only one network stream, insufficient to saturate the network. This decision was made for comparability with the ARM cluster.

Both systems might be impacted by CephFS' lazy deletes [3], which are done asynchronously by an MDS and probably overlapped with reads and writes, resulting in lower throughput.

OrangeFS performs better than CephFS on ARM in nearly all measurements. In contrast to CephFS, the OrangeFS daemon is lightweight and does not use many threads. As a consequence, context switches introduce less overhead on low core counts. Because no replication is done between nodes, less data needs to be transferred via the network, and the management of replicas does not consume resources. The downside is that faults of nodes can lead to data loss. Even though performance is higher compared to CephFS, only about 60 % of the TPP (see Table II) can be achieved. This can certainly be improved by tuning the stripe size of OrangeFS and the record size of ZFS. Compared to the defaults of other parallel file systems, OrangeFS has a low default stripe size of 64 KiB. Further benchmarks should be done to evaluate bigger stripes which could result in larger disk accesses depending on server-side cache size and cache times. As shown by traces of MPI-IO calls and OrangeFS' internal Trove layer, which does the actual disk I/O, single client-side write calls can result in multiple server-side Trove write calls [44]. Those should align to ZFS record sizes, if possible, to minimize read-modify-write cycles.

Compared to the other metrics, the EDP, as shown in Figure 2 is a fused metric that measures performance and energy efficiency at once. The use of this metric for tuning storage systems enforces that balanced configurations are found. Neither performance nor energy-saving efforts are neglected in favor of the other one. Considering that OrangeFS achieves both higher performance and energy efficiency, the EDP of CephFS is up to 30 times higher.

In terms of capacity per Watt, the ARM cluster is superior to the reference cluster, achieving 2.68 more TB per Watt. The ARM cluster's low idle power and maximum power consumption allow for usage of the cluster in places or situations where power restrictions apply, enabling the usage as a mobile storage solution.

VI. Conclusion and Future Work

We evaluated CephFS for HPC workloads on a productive cluster based on traditional x86 servers and an ARM-based low-power cluster. We compared the results in terms of throughput and efficiency. The ARM cluster is able to provide more than twice as much TB per Watt as the reference cluster and can achieve comparable throughput efficiency. OrangeFS has been shown to perform better than CephFS on the ARM cluster. Due to the low idle power consumption and low power peaks, ARM-based storage solutions are helpful in situations where power restrictions apply, for example, when used as a mobile storage cluster. In summary, we have shown that the energy efficiency of storage solutions depends significantly on both the used architecture and the file system. Lightweight solutions can reduce energy consumption and thus cost, which is becoming increasingly important due to the exponentially growing volumes of data.

As a next step, we will evaluate the use of other parallel file systems, such as MooseFS, and compare the results with an x86 setup, which is more similar in terms of network and disks compared to the ARM-based cluster. Throughput scaling of the ARM cluster while adding more storage nodes needs to be measured, so that the use in real world applications can be evaluated. In addition to sequential throughput other workloads, such as metadata-focused or mixed workloads, are of interest.

Acknowledgment

This work is partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 417705296. More information about the CoSEMoS (Coupled Storage System for Efficient Management of Self-Describing Data Formats) project can be found at https://cosemos.de.

References

[1] J. G. Koomey, "Worldwide electricity used in data centers," *Environmental Research Letters*, vol. 3, no. 3, jul 2008. [Online]. Available: https://doi.org/10.1088/1748-9326/3/3/034008

[2] S. A. Weil, S. A. Brandt, E. L. Miller, D. D. E. Long, and C. Maltzahn, "Ceph: A Scalable, High-Performance Distributed File System," in *7th Symposium on Operating Systems Design and Implementation (OSDI '06), November 6-8, Seattle, WA, USA*, B. N. Bershad and J. C. Mogul, Eds. USENIX Association, 2006, pp. 307–320. [Online]. Available: http://www.usenix.org/events/osdi06/tech/weil.html

[3] Ceph authors and contributors, "Ceph Documentation," https://docs.ceph.com/en/latest, 2021, [retrieved: 04, 2022].

[4] M. M. D. Bonnie *et al.*, "OrangeFS: Advancing PVFS," in *USENIX Conference on File and Storage Technologies (FAST)*, 2011.

[5] K. Duwe and M. Kuhn, "Using Ceph's BlueStore as Object Storage in HPC Storage Framework," in *CHEOPS@EuroSys'21*. ACM, 2021, pp. 3:1–3:6.

[6] J. Edge, "The OrangeFS distributed filesystem," https://lwn.net/Articles/643165/, 2015, [retrieved: 04, 2022].

[7] R. Thakur, W. Gropp, and E. Lusk, "A Case for Using MPI's Derived Datatypes to Improve I/O Performance," in *Proceedings of SC98: High Performance Networking and Computing*. ACM Press, November 1998. [Online]. Available: http://www.mcs.anl.gov/~thakur/dtype/

[8] M. Vilayannur, R. Ross, P. Carns, R. Thakur, A. Sivasubramaniam, and M. Kandemir, "On the performance of the POSIX I/O interface to PVFS," in *12th Euromicro Conference on Parallel, Distributed and Network-Based Processing, 2004. Proceedings.*, 2004, pp. 332–339.

 ISBN: 978-1-61208-967-6

[9] Z. Ou, B. Pang, Y. Deng, J. K. Nurminen, A. Ylä-Jääski, and P. Hui, "Energy- and Cost-Efficiency Analysis of ARM-Based Clusters," in *CCGRID*. IEEE Computer Society, 2012, pp. 115–123.

[10] E. L. Padoin, D. A. G. de Oliveira, P. Velho, and P. O. A. Navaux, "Evaluating Performance and Energy on ARM-based Clusters for High Performance Computing," in *41st International Conference on Parallel Processing Workshops, ICPPW 2012, Pittsburgh, PA, USA, September 10-13, 2012*. IEEE Computer Society, 2012, pp. 165–172. [Online]. Available: https://doi.org/10.1109/ICPPW.2012.21

[11] N. Rajovic, A. Rico, N. Puzovic, C. Adeniyi-Jones, and A. Ramírez, "Tibidabo: Making the case for an ARM-based HPC system," *Future Gener. Comput. Syst.*, vol. 36, pp. 322–334, 2014. [Online]. Available: https://doi.org/10.1016/j.future.2013.07.013

[12] M. Sato *et al.*, "Co-Design for A64FX Manycore Processor and "Fugaku"," in *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, 2020, pp. 1–15.

[13] D. Göddeke *et al.*, "Energy efficiency vs. performance of the numerical solution of PDEs: An application study on a low-power ARM-based cluster," *J. Comput. Phys.*, vol. 237, pp. 132–150, 2013. [Online]. Available: https://doi.org/10.1016/j.jcp.2012.11.031

[14] S. Brienza, S. E. Cebeci, S. S. Masoumzadeh, H. Hlavacs, Ö. Özkasap, and G. Anastasi, "A Survey on Energy Efficiency in P2P Systems: File Distribution, Content Streaming, and Epidemics," *ACM Comput. Surv.*, vol. 48, no. 3, pp. 36:1–36:37, 2016. [Online]. Available: https://doi.org/10.1145/2835374

[15] G. Lefebvre and M. J. Feeley, "Energy efficient peer-to-peer storage," Technical Report TR-2003-17. Department of Computer Science, University of British Columbia, Tech. Rep., 2000.

[16] M. Geveler, B. Reuter, V. Aizinger, D. Göddeke, and S. Turek, "Energy efficiency of the simulation of three-dimensional coastal ocean circulation on modern commodity and mobile processors," *Comput. Sci. Res. Dev.*, vol. 31, no. 4, pp. 225–234, 2016. [Online]. Available: https://doi.org/10.1007/s00450-016-0324-5

[17] F. Mantovani *et al.*, "Performance and energy consumption of HPC workloads on a cluster based on Arm ThunderX2 CPU," *CoRR*, vol. abs/2007.04868, pp. 800–818, 2020. [Online]. Available: https://arxiv.org/abs/2007.04868

[18] M. Ponce *et al.*, "Deploying a Top-100 Supercomputer for Large Parallel Workloads: the Niagara Supercomputer," in *PEARC*. ACM, 2019, pp. 34:1–34:8.

[19] D. Gudu and M. Hardt, "ARM Cluster for Performant and Energy-Efficient Storage," in *Computational Sustainability*, ser. Studies in Computational Intelligence, J. Lässig, K. Kersting, and K. Morik, Eds. Springer, 2016, vol. 645, pp. 265–276. [Online]. Available: https://doi.org/10.1007/978-3-319-31858-5_12

[20] A. Kougkas, A. Fleck, and X.-H. Sun, "Towards Energy Efficient Data Management in HPC: The Open Ethernet Drive Approach," in *2016 1st Joint International Workshop on Parallel Data Storage and data Intensive Scalable Computing Systems (PDSW-DISCS)*, 2016, p. 43–48.

[21] L. Zhang, Y. Deng, W. Zhu, J. Zhou, and F. Wang, "Skewly replicating hot data to construct a power-efficient storage cluster," *Journal of Network and Computer Applications*, vol. 50, pp. 168—-179, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1084804514001362

[22] X. Ruan *et al.*, "ECOS: An energy-efficient cluster storage system," in *2009 IEEE 28th International Performance Computing and Communications Conference*, 2009, p. 79–86.

[23] C. Karakoyunlu and J. A. Chandy, "Techniques for an energy aware parallel file system," in *2012 International Green Computing Conference, IGCC 2012, San Jose, CA, USA, June 4-8, 2012*. IEEE Computer Society, 2012, pp. 1–5. [Online]. Available: https://doi.org/10.1109/IGCC.2012.6322247

[24] P. Sehgal, V. Tarasov, and E. Zadok, "Evaluating Performance and Energy in File System Server Workloads," in *8th USENIX Conference on File and Storage Technologies, San Jose, CA, USA, February 23-26, 2010*, R. C. Burns and K. Keeton, Eds. USENIX, 2010, pp. 253–266. [Online]. Available: http://www.usenix.org/events/fast10/tech/full_papers/sehgal.pdf

[25] H. Shan and J. Shalf, "Using IOR to Analyze the I/O Performance for HPC Platforms," in *In: Cray User Group Conference (CUG'07)*, 2007.

[26] S. Rivoire, M. A. Shah, P. Ranganathan, C. Kozyrakis, and J. Meza, "Models and Metrics to Enable Energy-Efficiency Optimizations," *Computer*, vol. 40, no. 12, pp. 39–48, 2007.

[27] M. Horowitz, T. Indermaur, and R. Gonzalez, "Low-power digital design," in *Proceedings of 1994 IEEE Symposium on Low Power Electronics*, 1994, pp. 8–11.

[28] D. Chen *et al.*, "Usage centric green performance indicators," *SIGMETRICS Perform. Evaluation Rev.*, vol. 39, no. 3, pp. 92–96, 2011. [Online]. Available: https://doi.org/10.1145/2160803.2160868

[29] J. H. Laros III *et al.*, *Energy Delay Product*. London: Springer London, 2013, p. 51–55. [Online]. Available: https://doi.org/10.1007/978-1-4471-4492-2_8

[30] S. Georgiou, M. Kechagia, P. Louridas, and D. Spinellis, "What Are Your Programming Language's Energy-Delay Implications?" in *Proceedings of the 15th International Conference on Mining Software Repositories*, ser. MSR '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 303–313. [Online]. Available: https://doi.org/10.1145/3196398.3196414

[31] Super Micro Computer, Inc., "Supermicro AS 2124BT-HNTR Datasheet," https://www.supermicro.com/en/Aplus/system/2U/2124/AS-2124BT-HNTR.cfm, 2020,

[retrieved: 04, 2022].

[32] Intel Corporation, "Intel P4510 Datasheet," https://ark.intel.com/content/www/us/en/ark/products/122579/intel-ssd-dc-p4510-series-4-0tb-2-5in-pcie-3-1-x4-3d2-tlc.html, 2018, [retrieved: 04, 2022].

[33] GIGA-BYTE Technology Co., "Gigabyte R282-Z94 Datasheet," https://www.gigabyte.com/Enterprise/Rack-Server/R282-Z94-rev-100#Specifications, 2021, [retrieved: 04, 2022].

[34] Samsung, "Samsung MZQL23T8HCJS-00A07 Datasheet," https://semiconductor.samsung.com/ssd/datacenter-ssd/pm9a3/mzql23t8hcjs-00a07/, 2021, [retrieved: 04, 2022].

[35] HARDKERNEL CO., LTD., "Odroid HC4 Datasheet," https://wiki.odroid.com/odroid-hc4/hardware/hardware, 2021, [retrieved: 04, 2022].

[36] Armbian, "Armbian Odroid HC4," https://www.armbian.com/odroid-hc4/, 2022, [retrieved: 04, 2022].

[37] Western Digital Corporation, "WD Black WD10SPSX Datasheet," https://documents.westerndigital.com/content/dam/doc-library/en_us/assets/public/western-digital/product/internal-drives/wd-black-hdd/product-brief-western-digital-wd-black-mobile-hdd.pdf, 2020, [retrieved: 04, 2022].

[38] Samsung, "Samsung V-NAND SSD 860 PRO Datasheet," https://www.samsung.com/semiconductor/global.semi.static/Samsung_SSD_860_PRO_Data_Sheet_Rev1_1.pdf, 2018, [retrieved: 04, 2022].

[39] NETGEAR, Inc., "Netgear GS110EMX Datasheet," https://www.netgear.com/images/datasheet/switches/webmanagedswitches/GS110EMX_GS110MX.pdf, 2021, [retrieved: 04, 2022].

[40] MEAN WELL, "MW HRP 450-15 Datasheet," https://www.meanwell.com/webapp/product/search.aspx?prod=HRP-450, 2021, [retrieved: 04, 2022].

[41] ZES ZIMMER Electronic Systems GmbH, "ZES Zimmer LMG 450 Brochure," https://www.zes.com/en/content/download/286/2473/file/lmg450_prospekt_1002_e.pdf, 2010, [retrieved: 04, 2022].

[42] Dell Inc., "Dell Precision 3650 Tower Hardware Specification," https://www.delltechnologies.com/asset/en-us/products/workstations/technical-support/precision-3650-spec-sheet.pdf, 2021, [retrieved: 04, 2022].

[43] S. Just, "Crimson: A new ceph OSD for the age of persistent memory and fast NVMe storage," Santa Clara, CA, Feb. 2020.

[44] T. Ludwig, S. Krempel, J. Kunkel, F. Panse, and D. Withanage, "Tracing the MPI-IO Calls' Disk Accesses," in *Recent Advances in Parallel Virtual Machine and Message Passing Interface*, B. Mohr, J. L. Träff, J. Worringen, and J. Dongarra, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 322–330.

 ISBN: 978-1-61208-967-6