Energy Efficiency of Parallel File Systems on an ARM Cluster

Timm Leon Erxleben, Kira Duwe, Jens Saak, Martin Köhler and Michael Kuhn
timm.erxleben@ovgu.de

May 16, 2022

Faculty of Computer Science
Otto von Guericke University Magdeburg

MAX PLANCK INSTITUTE
FOR DYNAMICS OF COMPLEX
TECHNICAL SYSTEMS
MAGDEBURG

Timm Leon Erxleben

Timm Leon Erxleben is currently a bachelor student at the computer science faculty of the Otto von Guericke University Magdeburg.

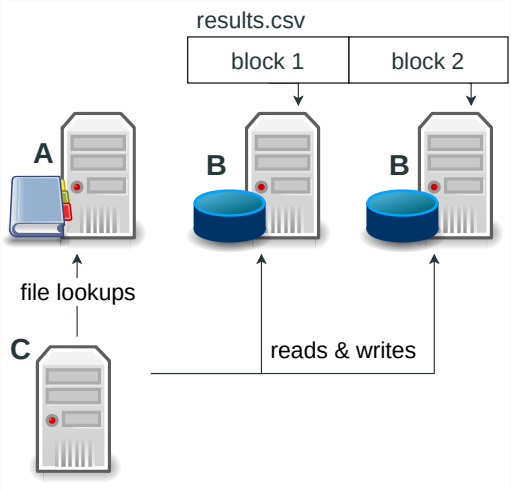
Recently, he gained interest in scientific computing and GreenHPC.

Motivation

- growing storage systems and growing energy consumption
- storage systems typically build from regular x86 servers with low energy proportionality [1]
- ARM-based single-board computers are designed for low energy consumption

Distributed File Systems

- Typically metadata is split and stored on a dedicated metadata server (A)
- Actual file contents are striped and distributed to multiple storage servers (B)
- Clients (C) directly communicate with storage servers after metadata lookups



CephFS

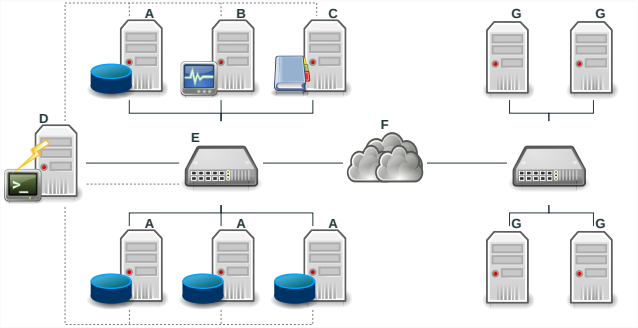
- built on the Ceph Clustered Object Store [2]
  - made of monitor, management and object storage services
- distribution of data via CRUSH
  - takes cluster map and some admin defined rules to calculate data positions
- fail-over possible due to replication
- additional metadata services implement access as file system

OrangeFS

- OrangeFS [3] is a traditional HPC file system
- only one service type that handles metadata and data
- metadata management via LMDB
- no replication or built-in fail-over
- file striping with round-robin and 64 KiB stripe size

ARM-based Cluster Setup

- A,B,C: Odroid HC4 [4]
  - Quad-Core 1.8GHz ARM 64-bit CPU
  - 4 GB DDR4 memory
  - 1x Gbit/s Ethernet
- A: 8x 1 TB 2.5" HDD [5]
- C: 2x 512 GB SSD [6]
- D: ZES Zimmer LMG 450 [7]
- E: Netgear Switch [8]
- F: Network infrastructure of the MPI Magdeburg
- G: Dell Clients [9]



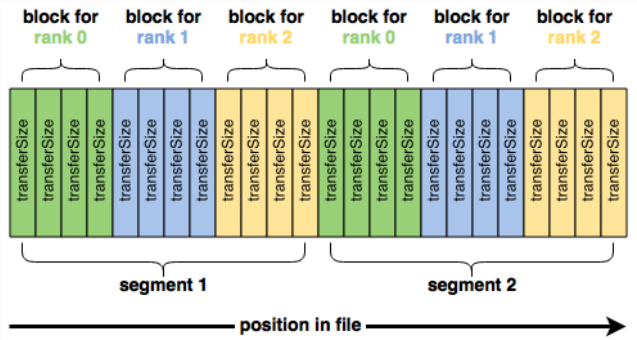
Reference-Setup

- four node subset of a productive Ceph cluster
- 3x Supermicro AS 2124BT-HNTR [10]
- 1x Gigabyte R282-Z94 [11]
- 2x 100 Gbit/s Ethernet
- 13x Intel P4510 NVMe SSD [12]
- 8x Samsung MZQL23T8HCJS-00A07 NVMe SSDs [13]
- power measurement over the existing monitoring solution via IPMI every 15s

Theoretical Peak Performance

Cluster	Network	Throughput Storage Devices	TPP
ARM	4 × 124.1 MB/s	8 × 115 MB/s	496.4 MB/s
Supermicro	3 × 12.5 GB/s	12 × 2.9 GB/s	34.8 GB/s
Gigabyte	1× 12.5 GB/s	1 × 2.9 GB/s / 8 × 4 GB/s	12.5 GB/s

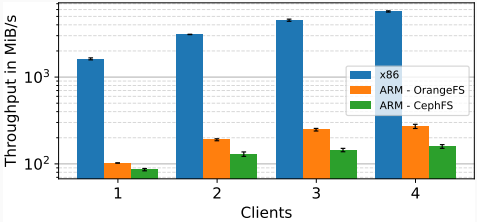
Benchmark



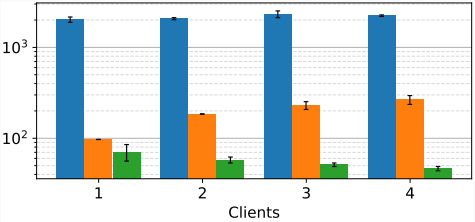
Layout of an IOR test file [14]

- IOR [15] v3.3 with POSIX backend and a transfer size of 4 MiB
- aggregated size of 96 GiB on the reference and 36 GiB on the ARM cluster

Data Throughput



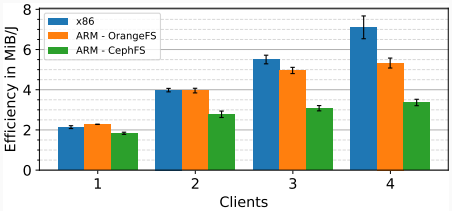
reads



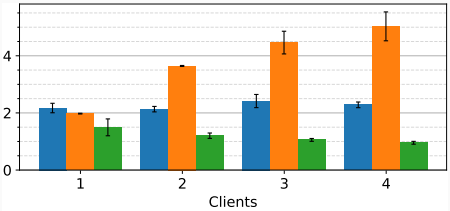
writes

System	Write in MiB/s / % TPP	Read in MiB/s / % 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

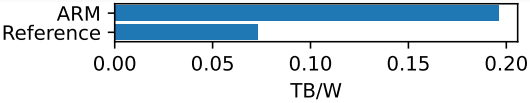
Energy Efficiency



(a) read efficiency



(b) write efficiency

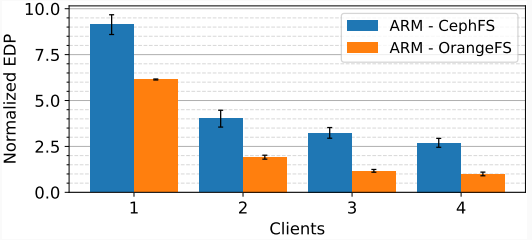


(c) capacity provided per Watt

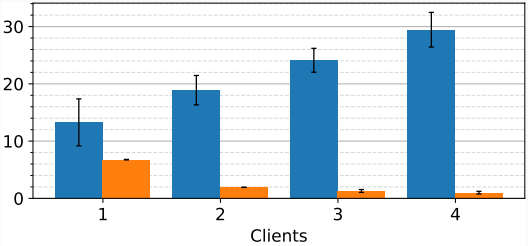
Throughput and capacity efficiency metrics [16]

Energy-Delay Product

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



reads



writes

Normalized Energy-Delay Product (EDP) with a weight of one [17, 18]

Conclusion and future work

- promising energy efficiency values for capacity and throughput
- lightweight distributed file system achieved better performance and thus efficiency
- Future Work:
  - test more distributed file systems
  - use a more similar cluster as reference
  - evaluate larger clusters

References

[1] L. A. Barroso and U. Hölzle, "The case for energy-proportional computing," *Computer*, vol. 40, no. 12, pp. 33–37, 2007.

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

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

References ...

[5] 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].

[6] 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].

[7] 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].

References ...

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

[9] 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].

[10] 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].

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

References ...

[12] 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].

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

[14] I. D. Team, “IOR 3.3.1 Documentation,” https://ior.readthedocs.io/en/3.3/, [retrieved: 03.05.2022].

[15] 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.

References ...

[16] 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

[17] 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.

[18] 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