

SC16


Submit

To-Do List 2

My Account

Logout

Jay Lofstead



SC16

My Submissions

Make a New Submission

My People Conflicts

My To-Do List

hide

Submit

an IEEE copyright form for *DAOS and Friends: A Proposal for an Exascale Storage System* (pap633).

Submit

the stage 4 (Paper: Upload of Final Paper) of pap633 *DAOS and Friends: A Pr...* by 11:59pm (UTC-12)

hide item

Aug 10, 2016

You can always view your current and completed to-do items on the To-Do List page (link in the top bar).

Decision: Accept with Shepherd

Submission: Lessons Learned from the Fast Forward Storage and I/O Project

Contributors: Barton, Bent, Jimenez, Koziol, Lofstead, Maltzahn

Key for the below column headings: [show](#)

Summary of Reviews of pap633s2: Lessons Learned from the Fast Forward Storage and I/O Project

Reviewer	rel	sound	import	orig	pres	rec	conf	exp	award
Reviewer 1	HIGH (4)	HIGH (4)	MODERATE (3)	MODERATE (3)	MODERATE (3)	WEAK ACCEPT (3)	MODERATE (3)	MODERATE (3)	No, it is not award quality (o)
Reviewer 2	HIGH (4)	VERY HIGH (5)	HIGH (4)	MODERATE (3)	HIGH (4)	ACCEPT (4)	MODERATE (3)	MODERATE (3)	No, it is not award quality (o)
Reviewer 3	HIGH (4)	VERY HIGH (5)	HIGH (4)	MODERATE (3)	MODERATE (3)	ACCEPT (4)	HIGH (4)	HIGH (4)	No, it is not award quality (o)
Reviewer 4	VERY HIGH (5)	HIGH (4)	HIGH (4)	MODERATE (3)	MODERATE (3)	ACCEPT (4)	HIGH (4)	MODERATE (3)	No, it is not award quality (o)
Averages:	4.3	4.5	3.8	3.0	3.3	3.8	3.5	3.3	0.0

Committee Comments

[jump](#)

Author Rebuttal

[jump](#)

Review of pap633s2 by Reviewer 1

top

Summary and High Level Discussion

This paper describes a holistic approach to storage and I/O problems arising for extreme scale computing and Big Data challenges. The approach described in this work is comprehensive and based on a user interface derived from an abstraction of the HDF5 library. The work is of sure interest and the description of the project is exhaustive.

Nevertheless most of the information reported in the paper are related to the description of the FFSIO project already publicly available, hindering the results

and the lessons learned. Most of the descriptive sections could be reduced referring to the FFSIO documentation, without losing technical quality and giving more relevance and space to the report of the results and lessons learned.

► Comments for Rebuttal

- reduce the amount of information already available from FFSIO project to give more relevance to lessons learned
- replace the figures 5 and 6 with readable fonts

► Detailed Comments for Authors

Only a few minor remarks:

- 1) there are some conceptual "jumps" that make the reading of the paper sometimes difficult. For example, in section III.B containers, array and datatypes for the FFSIO are only introduced without any further information. This concept is continued in V.A. Probably, it makes more sense (and less confusion) to introduce these concept directly in the V.A because the information in III.B just adds confusion.
- 2) the discussion about the "Related work" seems to be really long and it should be shortened, and in enumerating other FS, it will fair to cite BeeGFS as well.
- 3) the caption in Fig.3,4,5 and 6 do not permit to understand the figures. Expand.
- 4) The section about demonstration seems to be the weakest in the paper. The discussion of the results is missing and the presented plots are not meaningful without a proper discussion (why there are many different lines? what is the difference between each dataset? are those differences expected? etc.). There are also some typos that should be checked (ES-2670 should be E5-2670).

Review of pap633s2 by Reviewer 2[top](#)**► Summary and High Level Discussion**

The paper presents the Fast Forward Storage and IO Stack project, addressing IO performance and features needed for future extreme-scale requirements. The authors give an excellent overview of the motivation and challenges of the project and the design considerations for the now starting Phase 2.

The concepts and considerations for the IO stack are clearly explained. The paper is generally well structured, readable and will be interesting for a broader HPC audience, in particular as the concepts are likely to be relevant for future HPC systems.

The paper is more an overview of the project concepts and ideas than lessons-learned from Phase 1.

► Comments for Rebuttal

None

► Detailed Comments for Authors

The paper presents a good overview of the motivation of the project, the concepts and the design considerations for Phase 2. It
The whole paper seems to be focusing on the concepts, however the title suggests

"lessons learnt" from the project. Such lessons-learnt do not feature strongly in the paper, and could be strengthened.
Also, the demonstration section (IX) about the prototype implementation could be expanded and strengthened.

Minor issues:

Figure 1 – there is some inconsistency between the figure and the explanatory text on the following page and this should be reviewed: e.g. purple box ("lustre server") in the figure vs. VOSD in the text; explanation of the "dark pink and yellow" is not quite clear; the references to the coloured boxes in the text are possible in the wrong place in the sentence. Eg. "Below the user API is an IO forwarding layer (in black) ..." may be better understandable than having it at the end of the sentence " ... IO despatching layer (in black)".

Figure 4 is not referenced in the text

Figures 5 and 6 are rather small and difficult to read.

Also, there are some typos in the text.

Review of pap633s2 by Reviewer 3

[top](#)

► Summary and High Level Discussion

The authors present their design for a new extreme-scale I/O architecture developed as part of the Fast Forward Storage and IO project. It improves scalability by including I/O forwarding and dispatching, and makes it possible to transparently use new storage technologies such as node-local NVRAM. Additionally, backwards compatibility for existing applications is provided by adapting HDF5 to support this new framework.

The paper is important as it makes available design considerations etc. of what is likely going to be the next I/O stack for HPC. The overall design is sound: Backwards compatibility is important and leveraging information about data structures from the application layer should allow more sophisticated optimizations to be performed within the storage layer. It also makes sense to keep parts of the stack optional to ease integration in smaller clusters.

► Comments for Rebuttal

Some restructuring might be appropriate to introduce the necessary terminology earlier in the paper. Currently, the information is scattered across the paper and then (re)introduced quite late in the Broader Design section.

Additionally, the performance demonstration is quite shallow and does not really allow judging the implementation. More details about the tests and existing implementation should be added.

For more information, see the detailed comments.

► Detailed Comments for Authors

Introduction

- Page 2: "Becuase" - Should be "Because."
- Page 2, figure 1: Depending on whether the proceedings will be printed in

black/white or not, it might make sense to modify the description of the different layers as it directly refers to colors that might not be easily distinguishable in b/w.

- Page 2: "... and a more complex API for accessing the lower level components." - Which API is this? The figure only mentions HDF5, MPI-IO and POSIX.

- Page 2: "... no dependence on any technologies specified above it (in dark pink and yellow)." - This description (or the figure) should be made more clear. The Lustre Client (yellow) mentions DAOS and it is therefore not clear how it can be a layer above DAOS. Also, the NVRAM component (dark pink) is not explained at all.

- Page 2: "At the bottom is the Versioning Object Storage Device (VOSD) (in purple)." - The purple layer is called Lustre Server in the figure.

IO Dispatcher Layer

- Page 6: "... are capable of better regulating ... storage array better than ..." - Drop one "better."

DAOS Layer

- Page 7: "The FFSIO stack does not support multiple applications from the same or different platforms using a shared DAOS layer to write to the same container at the same time. This functionality is not supported by popular existing parallel file systems either." - What exactly is a platform here, different clusters accessing the same storage area? Also, since a container is basically a file, what does the last sentence mean? Surely multiple applications can access a shared file at the same time in parallel file systems. Or do you mean something else here?

- Page 7: "... may thin outsourcing ..." - Add a comma after "thin."

Broader Design

- Page 8: "Transactions: ..." - The explanation of terminology is a bit late here since most of it has been used before. Maybe this explanation could be moved towards the beginning of the paper to give the reader a rough idea of the concepts. Currently, most of it seems like a repetition (and is not really about transactions).

- Page 8: "Epochs: ..." - This has also been mentioned before. Maybe reorganizing (parts of) the broader design section towards the beginning of the paper makes sense.

- Page 9, figure 4: This figure is not referenced anywhere in the text.

- Page 9: "changes to the underlying ... is proposed." - Should be "are proposed."

Demonstration

- Page 10, figures 5 and 6: Replace the figures with high resolution versions if possible. As it is, they are very blurry and hard to read.

- Page 10, figure 6: It would be nice to have consistent colors for (a) and (b). Currently, the different file systems' colors change.

- Page 10, figure 6: PLFS is not mentioned anywhere in the text. Since it performs consistently better than the proposed IOD approach, some information about it would be appropriate.

- Page 10: "We run two different sets of tests." - It would be nice to have some additional information about the test setup. How is data read/written and which interface is used? Which I/O size is used for figure 5? Some additional information about the performance results would also be nice. Why is Lustre's performance so limited?

Conclusions

- Page 10: "With the overall stack design a prototype implementation complete, ..."
- Please rephrase this sentence, it is hard to understand.
- Page 10: Some information about how to get involved and how to get access to the existing design/implementation would be interesting.

References

- Please check the titles as they are all lower case ("io" instead of "IO"). Might be due to the style template. Maybe adding extra {}s around the title helps.

Review of pap633s2 by Reviewer 4

[top](#)

► Summary and High Level Discussion

The paper reports early results from DOE activities designed to adapt the I/O stack to exascale computing. The paper provides some evidence of consideration of BIG Data as well as traditional HPC file system state-of-the-art. It envisions that the design choices described in the paper will be reflected in future HPC systems at any scale.

This is a very timely and important topic, not just from the perspective of exascale. HPC has long suffered from a fragile support structure for file systems. Fundamental limitations of the POSIX interface have never been resolved. SSDs have come and gone multiple times without demonstrating lasting value. The notion that HPC file systems will be used to address data analytic workloads is far more than a technical issue. I wish the authors had spent more time on the overall strategic context beyond exascale. As written it is a bit inwardly focused.

This is a solid and helpful technical paper. The figures provide helpful elaboration and clarity of the paper's technical details. That said, while the topic is of broad interest and relevance, the presentation clarity and motivation could be improved. I also did not find the ACG and Big Data discussions satisfying: the topic was not given sufficient technical depth and breadth.

► Comments for Rebuttal

None

► Detailed Comments for Authors

A careful read by an external reviewer would help the flow. For example, the following sentence could be improved for readability.

"Since space is limited in expensive, in compute area storage resources, a copy-on-write approach is used for new versions of the same data. "

Committee Comments for Authors

[top](#)

The Committee agreed that the topic is really important and timely and that the paper is very stimulating. It was, however, noticed that not that many lessons learned are presented and the title was thus considered to be misleading. The Committee agreed that it would like this paper to be presented at SC16 assuming that the authors stick to their promise to follow-up the detailed comments made by the

reviewers.

===== Post-acceptance notes: Plagiarism detected =====

During the review process, we detected that your submission plagiarized noticeable amounts of text from other sources (possibly including your own prior papers, which would constitute self-plagiarism). This message is a warning that the same degree of plagiarism in future submissions could result in administrative rejection of your work without review.

Approximately 2300 words, or 21% of the text of this paper, were taken verbatim from an earlier workshop paper by the same authors without citation [<http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7103439>]. (Section VIII contains much of this text.) Presumably, omitting the citation may have been a misunderstanding of the goals of double-blind review, but even with a citation, this degree of direct copying is not considered good practice.

Authors' Rebuttal of Reviews

[top](#)

Rebuttal of Reviews

Rebuttal:

We thank the reviewers for the detailed reviews. We will incorporate the suggested definitions earlier in the paper and make the noticed minor corrections, including the figure edits, in the camera ready version.

Review Process Survey

Rate the overall quality of the reviews that you received on this paper. 4

How does the SC16 technical papers peer review process compare to other conferences in which you participate? 3

Do these reviews provide specific feedback that will allow you to improve the paper? 5 (agree)

How can we improve the SC review process? (100 words max):

Make the double blind truly double blind in that the only conflicts are organizational. The hybrid form is still preventing high conflicted papers from getting reviews from qualified reviewers that no longer collaborate with one or more authors, but did in the somewhat recent past.