Wei Hu, Guang-ming Liu, Qiong Li, Yan-huang Jiang, Gui-lin Cai, 2016. The storage wall for exascale supercomputing. *Frontiers of Information Technology and Electronic Engineering*, **17**(11):1154-1175.
http://dx.doi.org/10.1631/FITEE.1601336

The storage wall for exascale supercomputing

Key words: Storage-bounded speedup, Storage wall, High performance computing, Exascale computing

Contact: Wei Hu
E-mail: huwei@nscc-tj.gov.cn

ORCID: http://orcid.org/0000-0002-8839-7748

Introduction

- The mismatch between compute performance and I/O performance has long been a stumbling block as supercomputers evolve from Petaflops to Exaflops.

- Some important and basic issues related to the storage bottleneck, e.g., quantitative description, inherent laws, and system scalability, have not been solved yet.

- To quantify the I/O performance bottleneck and highlight the significance of achieving scalable performance in peta/exascale supercomputing, we introduce the 'storage wall' theory.

Storage-bounded speedup

$$S_{\mathrm{Sto}}^{\mathrm{P}} = S_P \frac{1}{1 + O(P)}$$ $O(P)$: storage workload factor

Storage wall

$\mathrm{sup}\mathrm{S}_{\mathrm{Sto}}^{\mathrm{P}}$ the supremum of the storage-bounded speedup

Constant and incremental systems

$O(P) \preceq \Theta(1)$ constant system

$O(P) \succ \Theta(1)$ incremental system

Write and read performance variation trends along with the processes scaling

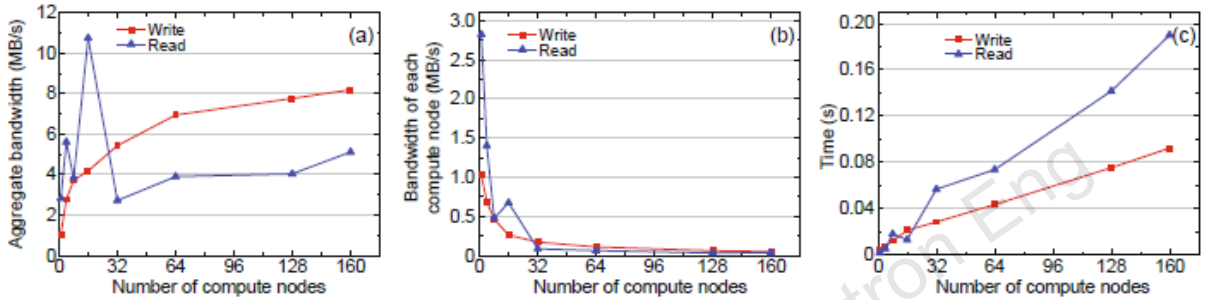

Fig. 11 Write and read operation performances when I/O block size is 4 KB: (a) aggregate bandwidth of the Lustre pool; (b) bandwidth per compute node; (c) time overhead per operation

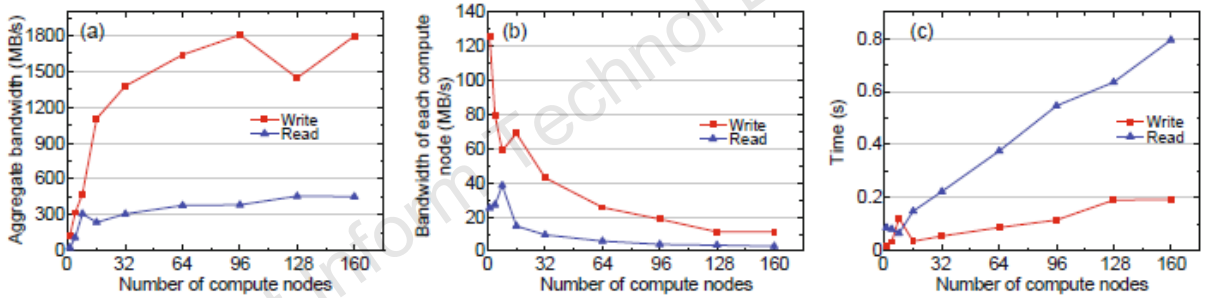

Fig. 12 Write and read operation performances when I/O block size is 2 MB: (a) aggregate bandwidth of the Lustre pool; (b) bandwidth per compute node; (c) time overhead per operation

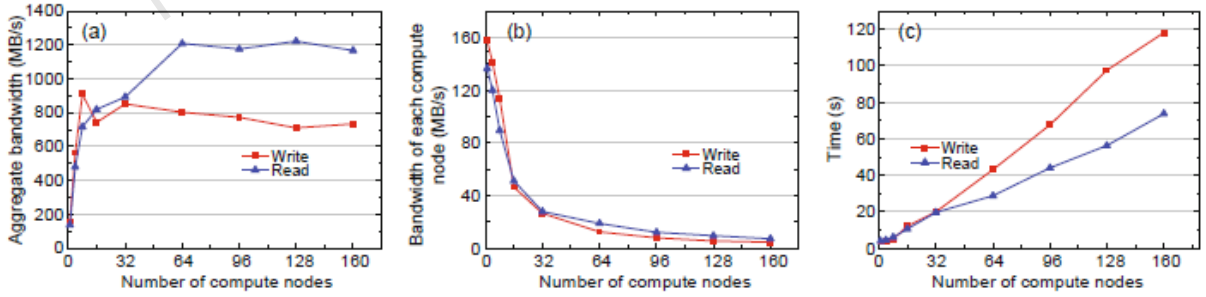

Fig. 13 Write and read operation performances when I/O block size is 512 MB: (a) aggregate bandwidth of the Lustre pool; (b) bandwidth per compute node; (c) time overhead per operation

The presence of the storage wall of the Tianhe-1A

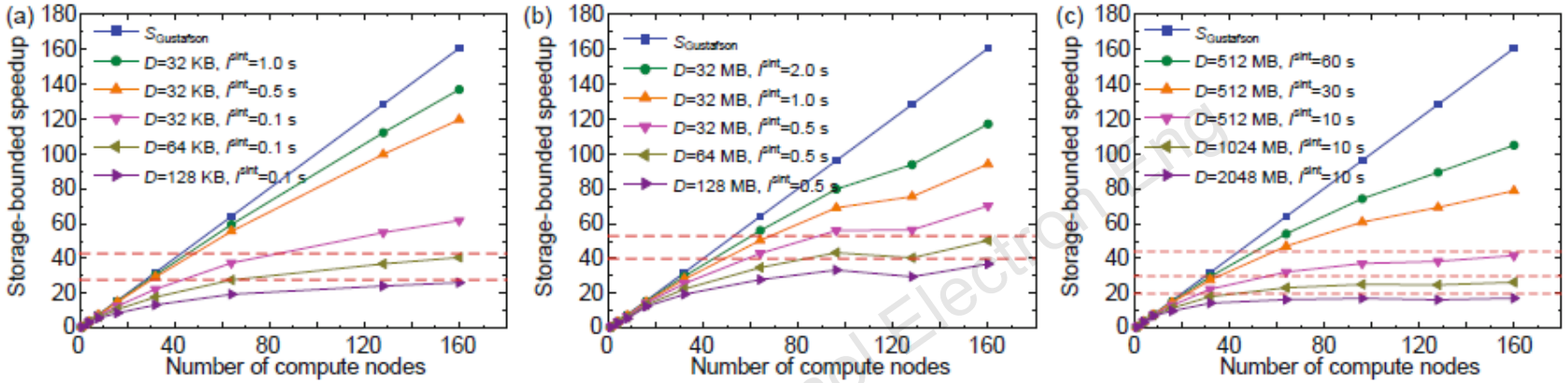

Fig. 14 Cases for the storage wall in centralized, distributed, and parallel (CDP) architecture in different block sizes, I^{sint}, and D: (a) block size = 4 KB; (b) block size = 2 MB; (c) block size = 512 MB

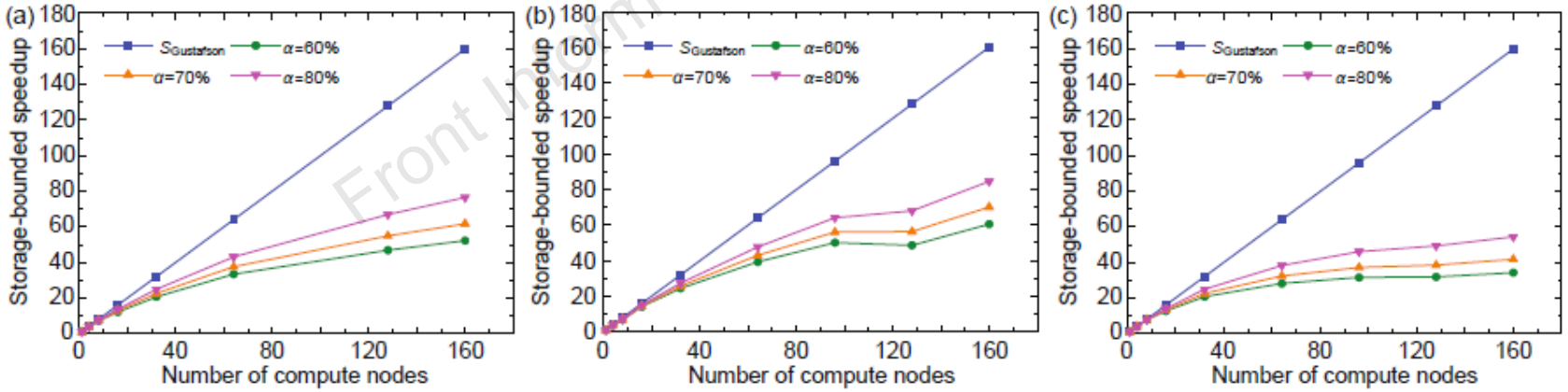

Fig. 15 The impact of α on the storage wall: (a) D=32 KB, I^{sint}=0.1 s, block size=4 KB; (b) D=2 MB, I^{sint}=0.5 s, block size=32 MB; (c) D=512 MB, I^{sint}=10.0 s, block size=512 MB

The case study to verify and analyze the storage wall of Jaguar supercomputer

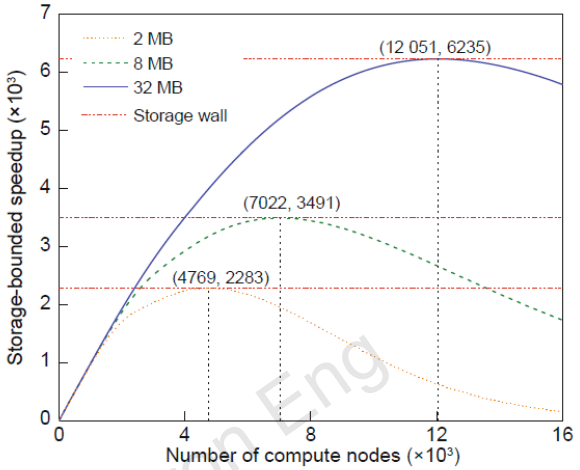

Fig. 18 $S^{\mathrm{P}}_{\mathrm{Sto}}$ **variation trends in parallel applications with full-memory checkpointing on Jaguar for write aggregate bandwidths with buffer size per core at 2, 8, and 32 MB, respectively (checkpointing interval is 3 h)**

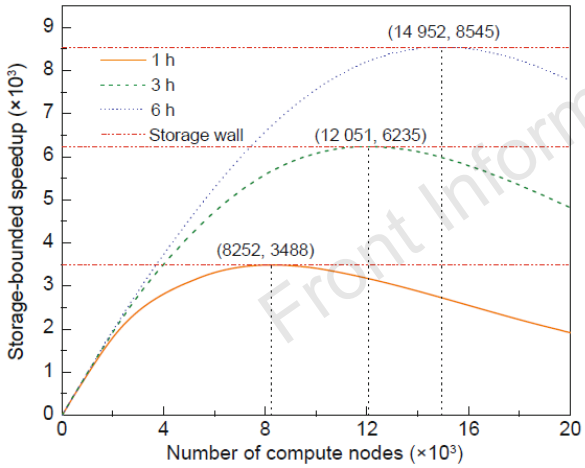

Fig. 19 $S^{\mathrm{P}}_{\mathrm{Sto}}$ **variation trends in parallel applications with full-memory checkpointing on Jaguar, for checkpointing intervals of 1 h, 3 h, and 6 h, respectively (write aggregate bandwidth of buffer size per core at 32 MB)**

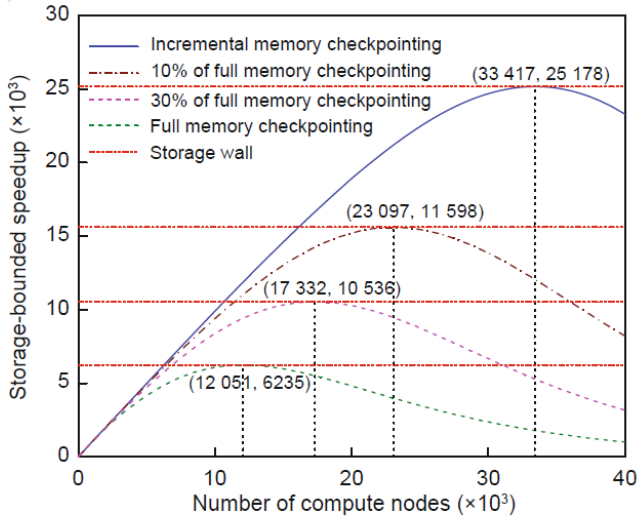

Fig. 20 Comparison of $S^{\mathrm{P}}_{\mathrm{Sto}}$ **among applications with full-, partial-, and incremental-memory checkpointing with the write aggregate bandwidth data of buffer size per core at 32 MB (checkpointing interval is 3 h)**

Conclusions

- The storage-bounded speedup and storage wall allow for the effects of the storage bottleneck on the scalability of parallel applications for large-scale parallel computing systems.
- The experiment results verify the existence of the storage wall, and reveal the key factors that affect the storage wall.
- Our work enables researchers to push the storage wall forward by designing or improving storage architectures, applying I/O resource-oriented programming models, and so on.