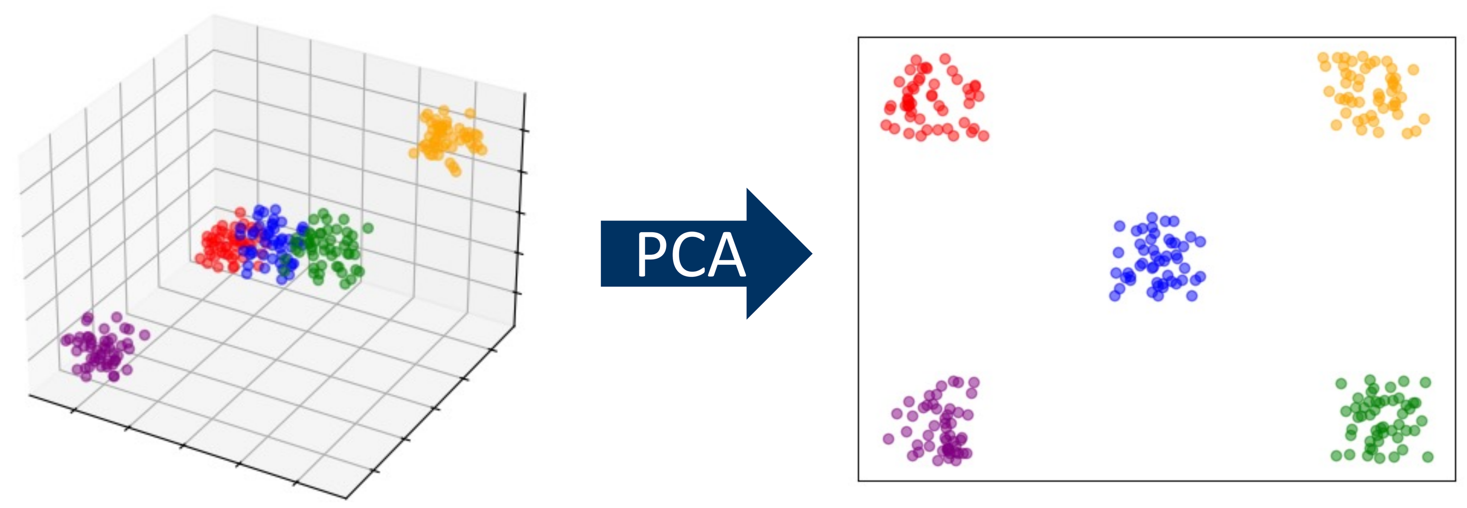
Randomized SVD for Serverless Systems

Gabriel Raulet Yen-Hsiang Chang

Berkeley UNIVERSITY OF CALIFORNIA

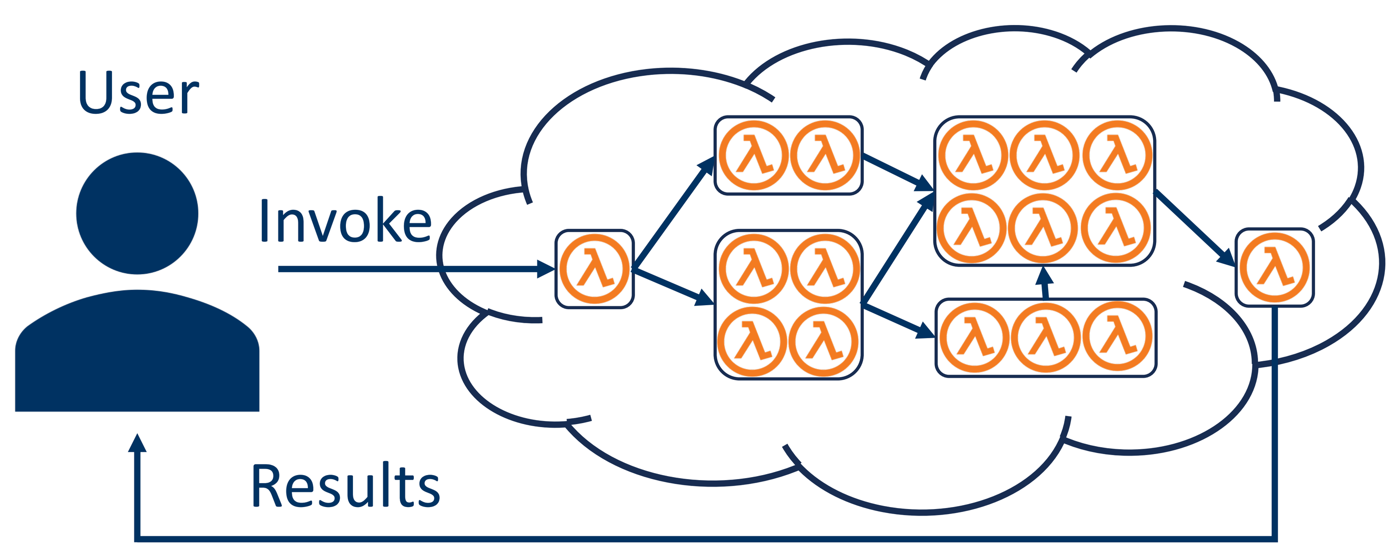
Randomized SVD for Serverless Systems

Motivating Application: Given a columnwise standardized data matrix $X \in \mathbb{R}^{n \times p}$ with n samples in p dimensional space, principal component analysis (PCA) finds a lower dimensional space that maximizes the variance of the projected data among subspaces of a given dimension $k \ll p$:



To overcome large-scale matrices, PCA utilizes randomized singular value decomposition (SVD) in a distributed manner to approximate $X_k = U_k S_k V_k^\top$, the truncated rank-k SVD of X. Unfortunately, high performance clusters are not available to everyone. Therefore, serverless systems come into play.

Serverless Systems: Serverless computing is a cloud computing paradigm that abstracts away the need for maintaining servers. Through the concept of Function as a Service (FaaS), computation is performed through stateless functions that scale elastically to the demand of applications. The recent growth of serverless computing offers the potential to close the accessibility gap.



Randomized SVD for Serverless Systems: Directly making use of distributed randomized SVD kernels for serverless systems is not a trivial problem however. Previous work has pointed out that serverless linear algebra kernels suffer from high communication overheads due to the lack of efficient collective communication primitives. Nevertheless, linear algebra kernels nowadays have already been optimized for minimizing communication costs.

Central Challenge

How to **reduce communication overheads** in randomized SVD for serverless systems if the distributed design is **already optimized for minimizing the amount of data moved**?

Related Work

NumPyWren [1]: It provides a serverless linear algebra programming model. High communication overheads due to the lack of efficient collective communication primitives.

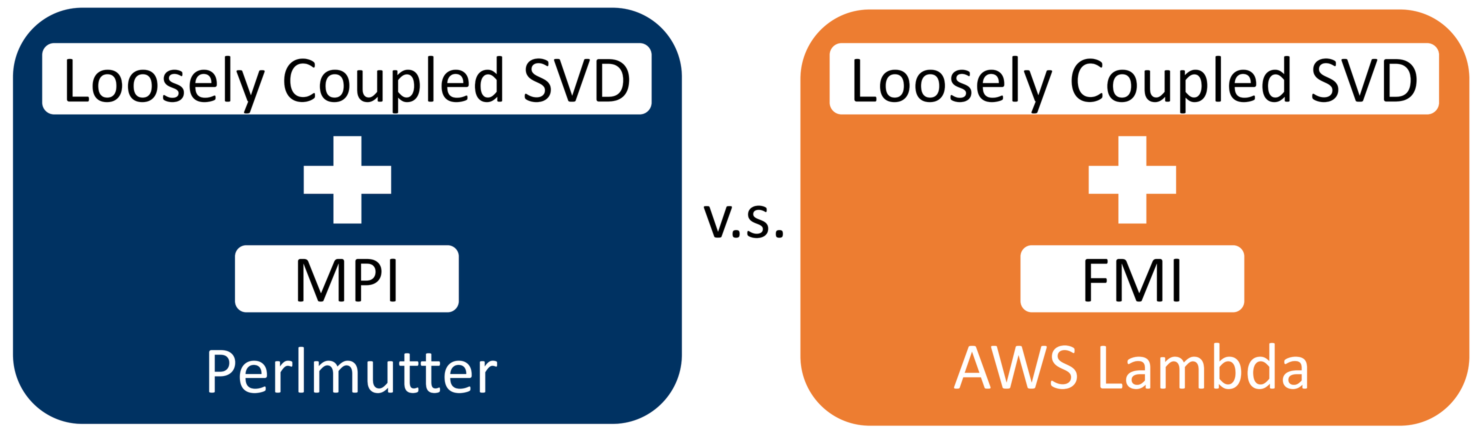
FaaS Message Interface (FMI) [2]: It provides efficient collective communication primitives for FaaS applications by establishing direct communication between services.

Loosely Coupled SVD [3]: It redesigns randomized SVD to reduce its reliance on collective communication primitives by sacrificing the error bound with a factor related to the amount of parallelism.

Methodology

Solution to Challenge: We integrate FMI into loosely coupled SVD to derive a randomized SVD kernel for serverless systems that mitigates communication overheads.

Particularly, we are interested in the following questions.



Direct Comparison: Is the performance comparable between distributed systems and serverless systems? How easy is it to modify the distributed code into the serverless one?

Runtime Breakdown: How much time is spent in function startups, computations, and communications, respectively?

Scalability: Is there any modification needed in the code to exploit auto-scaling in serverless systems?

Error Analysis: Is the error propagated from randomized SVD to PCA tolerable when we increase the amount of parallelism?

Preliminary Benchmarks

Performance Comparison: The serverless implementation is under development. The distributed version on Perlmutter indicates potential underutilization of resources, which might make serverless computing a more attractive choice.

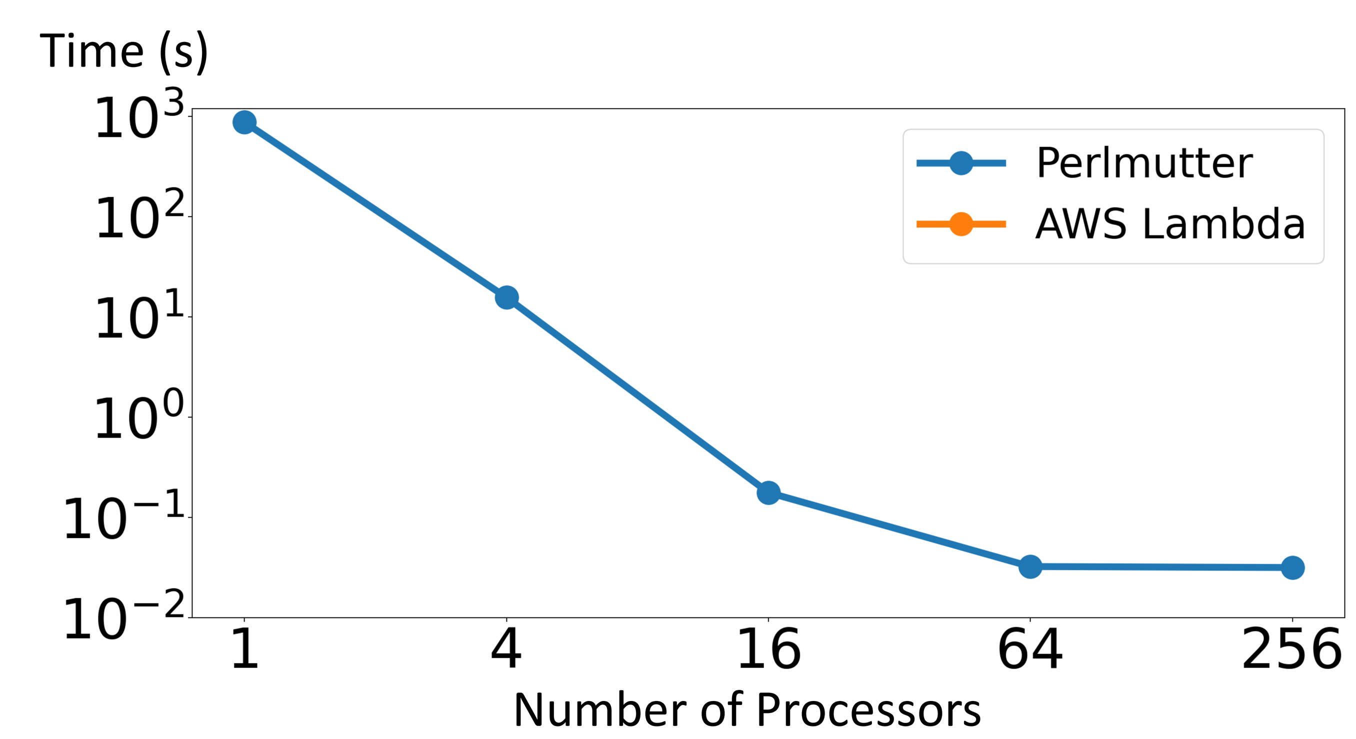


Figure 1. Strong scalability experiments for loosely coupled SVD on Perlmutter and AWS Lambda with $n = 4096$ and $p = 4096$

Error Analysis: The relative errors of explained variance for the first few principal components stay tiny when scaling up. However, the errors become nonnegligible for further principal components, which might be a concern.

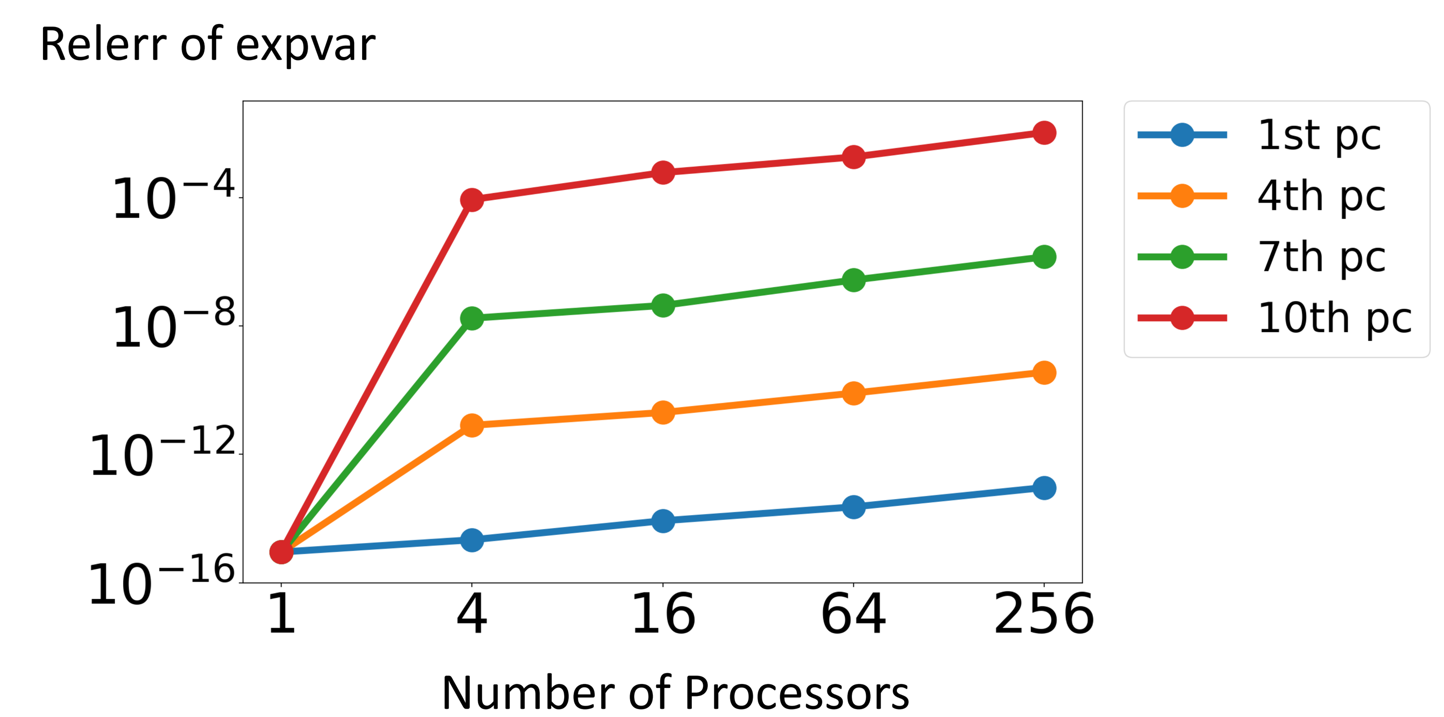


Figure 2. Relative errors of explained variance (Relerr of expvar) for principal components (pc) in PCA v.s. number of processors used with $n = 4096$ and $p = 4096$

References

[1] V. Shankar et al., Serverless linear algebra. SoCC, Oct. 2020.

[2] M. Copik, R. Böhringer, A. Calotoiu, and T. Hoefler, FMI: Fast and cheap message passing for serverless functions. ICS, June. 2023.

[3] S. Fang and R. Hauser, Distributed Computing of Large-Scale Singular Value Decompositions. March. 2018