Novelty Assessment Report

Paper: DES-LOC: Desynced Low Communication Adaptive Optimizers for Foundation Models
PDF URL: https://openreview.net/pdf?id=6N2qFixxYZ
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-07

Abstract

Scaling foundation model training with Distributed Data Parallel~(DDP) methods is bandwidth-limited. Existing infrequent communication methods like Local SGD were designed to synchronize model parameters only and cannot be trivially applied to adaptive optimizers due to additional optimizer states. Heuristic approaches that keep states local or reset them lack guarantees and can be unstable in compute-efficient batch regimes; conversely, Local Adam synchronizes all states uniformly and is provably convergent but triples communication costs. We propose Desynced Low Communication Adaptive Optimizers (DES-LOC), a family of optimizers assigning independent synchronization periods to parameters and momenta, enabling lower communication costs while preserving convergence. Our theoretical analysis shows that while parameter synchronization dominates the asymptotic rate in-expectation, high-probability convergence guarantees require at least infrequent synchronization of the second momentum. Furthermore, we prove that more frequent momentum sync permits larger stable step sizes. Experiments on language models of up to 1.7B show that DES-LOC can communicate 170x less than DDP and 2x less than the previous state-of-the-art Local Adam, enabling 1.3x–2.1x wall-clock speedups over DDP for 1-13B models on 100Gb/s links. Furthermore, unlike previous heuristic methods, DES-LOC is robust to worker failures offering a scalable, efficient, and fault-tolerant solution for foundation model training.

Core Task Landscape

This paper addresses: **Distributed Training of Foundation Models with Reduced Communication Overhead**

A total of **50 papers** were analyzed and organized into a taxonomy with **22 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Communication Compression and Gradient Reduction Techniques**
- **Communication-Computation Overlap and Scheduling**
- **Parallelism Strategy Design and Optimization**
- **Production-Scale System Implementations**
- **Federated Foundation Model Training**
- **General Distributed Learning Frameworks and Theory**
- **Surveys and Characterization Studies**

Complete Taxonomy Tree

- Distributed Training of Foundation Models with Reduced Communication Overhead Survey Taxonomy
- Communication Compression and Gradient Reduction Techniques
  - Gradient and Activation Compression (4 papers)
  - [17] Optimus-CC: Efficient Large NLP Model Training with 3D Parallelism Aware Communication Compression (Songï¼ Jaeyong, 2023) View paper
  - [18] Protocol Models: Scaling Decentralized Training with Communication-Efficient Model Parallelism (Ramasinghe, 2025) View paper
  - [23] Accelerating Distributed Deep Learning Training with Compression Assisted Allgather and Reduce-Scatter Communication (Qinghua Zhou, 2023) View paper
  - [25] Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training (Yujun Lin, 2022) View paper
  - Weight and Optimizer State Compression (1 papers)
  - [31] WeiPipe: Weight Pipeline Parallelism for Communication-Effective Long-Context Large Model Training (Junfeng Lin, 2025) View paper
  - Theoretical Foundations of Communication Compression (1 papers)
  - [34] Lower bounds and nearly optimal algorithms in distributed learning with communication compression (Huang, 2022) View paper
- Communication-Computation Overlap and Scheduling
  - Fine-Grained Overlap and Kernel Fusion (3 papers)
  - [13] Centauri: Enabling efficient scheduling for communication-computation overlap in large model training via communication partitioning (Chang Chen, 2024) View paper
  - [14] Co2: Efficient distributed training with full communication-computation overlap (Qin Zhen, 2024) View paper
  - [48] Concerto: Automatic Communication Optimization and Scheduling for Large-Scale Deep Learning (Shenggan Cheng, 2025) View paper
  - Asynchronous and Local Update Methods ★ (3 papers)
  - [0] DES-LOC: Desynced Low Communication Adaptive Optimizers for Foundation Models (Anon et al., 2026) View paper
  - [32] SlowMo: Improving Communication-Efficient Distributed SGD with Slow Momentum (Wang Jian-yu, 2022) View paper

- [38] Communication-Efficient Language Model Training Scales Reliably and Robustly: Scaling Laws for DiLoCo (Charles, 2025) View paper
- Cross-Region and Wide-Area Network Training (1 papers)
- [29] Cross-region Model Training with Communication-Computation Overlapping and Delay Compensation (Zhu, 2025) View paper

- Parallelism Strategy Design and Optimization
  - 3D Parallelism and Hybrid Strategies (2 papers)
  - [21] Merak: An Efficient Distributed DNN Training Framework With Automated 3D Parallelism for Giant Foundation Models (Zhiquan Lai, 2023) View paper
  - [40] On optimizing the communication of model parallelism (Zhuang, 2023) View paper
  - Model Parallelism Communication Optimization (2 papers)
  - [3] SWARM Parallelism: Training Large Models Can Be Surprisingly Communication-Efficient (Ryabinin, 2023) View paper
  - [46] Megatron-lm: Training multi-billion parameter language models using model parallelism (Shoeybi, 2019) View paper
  - Mixture-of-Experts Training Optimization (2 papers)
  - [10] Megascale-moe: Large-scale communication-efficient training of mixture-of-experts models in production (Jin Chao, 2025) View paper
  - [45] FasterMoE: modeling and optimizing training of large-scale dynamic pre-trained models (Jiaao He, 2022) View paper
- Production-Scale System Implementations
  - Datacenter-Scale Training Systems (2 papers)
  - [4] Boosting large-scale parallel training efficiency with c4: A communication-driven approach (Jianbo Dong, 2024) View paper
  - [8] {MegaScale}: Scaling large language model training to more than 10,000 {GPUs} (Jiang Ziheng, 2024) View paper
  - Performance Modeling and Simulation (2 papers)
  - [30] {SimAI}: Unifying Architecture Design and Performance Tuning for {Large-Scale} Large Language Model Training with Scalability and Precision (X Wang, 2025) View paper
  - [37] Mad-max beyond single-node: Enabling large machine learning model acceleration on distributed systems (Samuel Hsia, 2024) View paper
  - Network Infrastructure and Hardware Co-Design (3 papers)
  - [7] Accelerating large language model training with in-package optical links for scale-out systems (Aakash Patel, 2024) View paper
  - [41] RailX: a flexible, scalable, and low-cost network architecture for hyper-scale LLM training systems (Feng, 2025) View paper
  - [50] Vela: A Virtualized LLM Training System with GPU Direct RoCE (Apoorve Mohan, 2025) View paper
- Federated Foundation Model Training
  - Parameter-Efficient Federated Fine-Tuning (2 papers)
  - [6] Feddat: An approach for foundation model finetuning in multi-modal heterogeneous federated learning (Haokun Chen, 2024) View paper
  - [11] PromptFL: Let Federated Participants Cooperatively Learn Prompts Instead of Models â Federated Learning in Age of Foundation Model (Tao Guo, 2023) View paper
  - Heterogeneous Device Adaptation (2 papers)
  - [28] Distributed Fine-Tuning of Foundation Models Over Heterogeneous Edge Devices (Xueting Han, 2025) View paper
  - [42] Cluster Based Heterogeneous Federated Foundation Model Adaptation and Fine-Tuning (Qiao Ya-qi, 2025) View paper
  - Wireless Network Integration (5 papers)
  - [2] Distributed foundation models for multi-modal learning in 6G wireless networks (Jun Du, 2024) View paper
  - [5] Distributed large models training optimization with real-time wireless channel feedback (J Pei, 2025) View paper
  - [16] Hierarchical federated foundation models over wireless networks for multi-modal multi-task intelligence: Integration of edge learning with D2D/P2P-enabled fog â¦ (P Abdisarabshali, 2025) View paper
  - [22] Resource management for low-latency cooperative fine-tuning of foundation models at the network edge (Hai Wu, 2025) View paper
  - [36] Accelerating Distributed Model Training through Intelligent Node Selection and Data Allocation Strategies in 6G network (Yuhao Chai, 2024) View paper
- General Distributed Learning Frameworks and Theory
  - Distributed SGD Variants and Convergence (3 papers)
  - [24] Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning (Yu Hao, 2019) View paper
  - [35] Communication-Efficient Distributed Training for Collaborative Flat Optima Recovery in Deep Learning (Dimlioglu, 2025) View paper
  - [49] Large scale distributed neural network training through online distillation (Anil, 2018) View paper
  - General Communication-Efficient Learning (3 papers)
  - [33] Communication-efficient and distributed learning over wireless networks: Principles and applications (J Park, 2021) View paper
  - [43] Communication-efficient distributed learning: An overview (Xuanyu Cao, 2023) View paper
  - [47] Distributed learning in wireless networks: Recent progress and future challenges (Mingzhe Chen, 2021) View paper
  - Federated Learning Fundamentals (2 papers)
  - [12] Two-stream federated learning: Reduce the communication costs (Xin Yao, 2018) View paper
  - [27] The cost of training machine learning models over distributed data sources (Elia Guerra, 2023) View paper
  - Domain-Specific Distributed Training (1 papers)
  - [19] MS-DINO: Efficient Distributed Training of Vision Transformer Foundation Model in Medical Domain through Masked Sampling (Park Sangjoon, 2023) View paper
- Surveys and Characterization Studies
  - Foundation Model Training Surveys (3 papers)
  - [1] Efficient training of large language models on distributed infrastructures: a survey (Duan, 2024) View paper
  - [9] Cocktailsgd: Fine-tuning foundation models over 500mbps networks (Wang Jue, 2023) View paper
  - [26] Scaling Large Language Model Training on Frontier with Low-Bandwidth Partitioning (Xu Lang, 2024) View paper
  - Federated Learning Surveys (2 papers)
  - [15] Advances and open challenges in federated foundation models (Chao Ren, 2025) View paper
  - [20] Towards federated large language models: Motivations, methods, and future directions (Yujun Cheng, 2024) View paper
  - Communication Pattern Characterization (2 papers)

- [39] Characterizing the Efficiency of Distributed Training: A Power, Performance, and Thermal Perspective (Park, 2025) View paper
- [44] Understanding and Characterizing Communication Characteristics for Distributed Transformer Models (Quentin Anthony, 2025) View paper

Narrative

Core task: distributed training of foundation models with reduced communication overhead. The field addresses the challenge of scaling large-scale model training across multiple devices while minimizing the communication bottleneck that often dominates training time. The taxonomy reveals several complementary strategies: Communication Compression and Gradient Reduction Techniques focus on reducing the volume of data exchanged through methods like gradient quantization and sparsification, as seen in Deep Gradient Compression[25] and Compression Assisted Allgather[23]. Communication-Computation Overlap and Scheduling aims to hide communication latency by overlapping it with computation, exemplified by Co2 Overlap[14] and Cross-region Overlapping[29]. Parallelism Strategy Design explores hybrid parallelism configurations that balance computation and communication trade-offs, with works like Megatron-lm[46] and SWARM Parallelism[3] demonstrating different partitioning approaches. Production-Scale System Implementations such as MegaScale[8] and Megascale MoE[10] integrate multiple techniques for real-world deployments, while Federated Foundation Model Training addresses decentralized scenarios with heterogeneous devices, as in PromptFL[11] and Edge Foundation Finetuning[22]. General frameworks and surveys like Efficient LLM Training Survey[1] provide broader perspectives on the landscape.

A particularly active line of work explores asynchronous and local update methods that reduce synchronization frequency, trading off some coordination for substantial communication savings. DES-LOC[0] falls within this branch alongside SlowMo[32] and DiLoCo Scaling Laws[38], which investigate how infrequent global synchronization with local SGD steps affects convergence and scalability. While SlowMo[32] introduced momentum-based slow updates to stabilize training, DiLoCo Scaling Laws[38] examines how these methods scale with model size and cluster topology. DES-LOC[0] emphasizes a different angle by focusing on decentralized strategies that adapt synchronization patterns dynamically, contrasting with the more structured periodic updates in DiLoCo[38]. This cluster of methods represents a shift from traditional synchronous data parallelism toward more flexible, communication-efficient paradigms that are especially valuable when training across geographically distributed or bandwidth-constrained environments.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. SlowMo: Improving Communication-Efficient Distributed SGD with Slow Momentum

Authors: Wang Jian-yu, Jianyu Wang, Tantia, Vinayak, Vinayak Tantia, et al. (10 authors total) | **Year/Venue**: 2022 | **URL**: View paper

Abstract

Distributed optimization is essential for training large models on large datasets. Multiple approaches have been proposed to reduce the communication overhead in distributed training, such as synchronizing only after performing multiple local SGD steps, and decentralized methods (e.g., using gossip algorithms) to decouple communications among workers. Although these methods run faster than AllReduce-based methods, which use blocking communication before every update, the resulting models may be ...

Relationship Analysis

Both papers belong to the Asynchronous and Local Update Methods category, focusing on reducing communication overhead through infrequent synchronization in distributed training. While DES-LOC independently synchronizes model parameters and optimizer states (first and second momenta) at different frequencies to minimize communication costs for adaptive optimizers like Adam, SlowMo applies a slow momentum update after periodic synchronization of parameters following multiple local steps of a base optimizer. The key difference is that DES-LOC targets adaptive optimizers with state-specific synchronization schedules, whereas SlowMo adds a momentum layer on top of various base optimizers (including SGD and decentralized methods) with uniform synchronization periods.

2. Communication-Efficient Language Model Training Scales Reliably and Robustly: Scaling Laws for DiLoCo

Authors: Charles, Zachary, Zachary Charles, Gabriel Teston, Rush, et al. (16 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

As we scale to more massive machine learning models, the frequent synchronization demands inherent in data-parallel approaches create significant slowdowns, posing a critical challenge to further scaling. Recent work develops an approach (DiLoCo) that relaxes synchronization demands without compromising model quality. However, these works do not carefully analyze how DiLoCo's behavior changes with model size. In this work, we study the scaling law behavior of DiLoCo when training LLMs under a fi...

Relationship Analysis

Both papers belong to the asynchronous and local update methods category, using infrequent synchronization to reduce communication overhead in distributed training. The candidate paper (DiLoCo scaling laws) focuses on empirically characterizing the scaling behavior and hyperparameter tuning of DiLoCo across model sizes, while the original paper (DES-LOC) proposes a novel optimizer family that independently synchronizes parameters and optimizer states at different frequencies with theoretical convergence guarantees. The key difference is that DES-LOC introduces desynchronized state updates with provable convergence, whereas DiLoCo maintains uniform synchronization periods and focuses on empirical scaling law analysis.

Contributions Analysis

Overall novelty summary. The paper proposes DES-LOC, a family of adaptive optimizers that assign independent synchronization periods to parameters and momenta to reduce communication costs in distributed training. It sits within the Asynchronous and Local Update Methods leaf, which contains only three papers including this one. This is a relatively sparse research direction compared to more crowded areas like Gradient and Activation Compression (four papers) or Wireless Network Integration (five papers), suggesting the specific problem of desynchronized optimizer state updates remains underexplored within the broader distributed training landscape.

The taxonomy reveals that neighboring leaves pursue complementary strategies: Fine-Grained Overlap and Kernel Fusion (three papers) focuses on hiding latency through scheduling, while Cross-Region Training (one paper) addresses wide-area network challenges. The parent branch Communication-Computation Overlap and Scheduling excludes pure compression methods, which are handled separately under Communication Compression. DES-LOC's approach of varying synchronization frequencies for different optimizer components bridges asynchronous methods and optimizer state management, connecting to Weight and Optimizer State Compression but diverging by focusing on scheduling rather than compression.

Among 25 candidates examined, the first contribution (independent synchronization periods) shows two refutable candidates from six examined, indicating some prior work on related optimizer synchronization strategies. The convergence theory contribution examined nine candidates with none clearly refuting it, suggesting theoretical novelty within the limited search scope. The empirical validation examined ten candidates without refutation, though this reflects the specific 170× reduction claim rather than exhaustive comparison with all communication-reduction methods. The analysis covers top-K semantic matches and citation expansion, not the entire field.

Given the sparse taxonomy leaf and limited refutation across contributions, the work appears to occupy a relatively novel position within the examined literature. However, the two refutable candidates for the core optimizer design suggest some conceptual overlap exists. The scope limitations mean adjacent research directions or recent preprints may contain additional relevant prior work not captured in this 25-candidate analysis.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: DES-LOC optimizer family with independent synchronization periods

Description: The authors introduce DES-LOC, a new family of adaptive optimizers that assigns different synchronization frequencies to model parameters and optimizer momentum states. This design reduces communication overhead compared to existing methods while maintaining theoretical convergence guarantees.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. DES-LOC: Desynced Low Communication Adaptive Optimizers for Training Foundation Models

URL: View paper

Brief Assessment

DES-LOC Foundation[61] appears to be the same work as the original paper (identical authors, title, and content). This is the original paper itself, not prior work that could refute novelty claims.

---

2. FlexDeMo: Decoupled Momentum Optimization for Hybrid Sharded Data Parallel Training

URL: View paper

Brief Assessment

FlexDeMo[62] focuses on hybrid sharded data parallel training combining FSDP with decoupled momentum, whereas DES-LOC addresses independent synchronization of parameters and momentum states in distributed adaptive optimizers. The technical approaches and problem formulations differ substantially.

---

3. MT-DAO: Multi-Timescale Distributed Adaptive Optimizers with Local Updates

URL: View paper

Prior Art Analysis

MT-DAO[64] demonstrates that prior work exists on assigning independent synchronization frequencies to different optimizer states in distributed adaptive optimization. Both papers propose frameworks where model parameters and optimizer momentum states can be synchronized at different intervals to reduce communication overhead while maintaining convergence guarantees. The candidate paper explicitly describes synchronizing parameters at frequency kx and multiple momentum states at different frequencies (k1, k2, etc.), which directly overlaps with the original paper's core contribution of independent synchronization periods for parameters versus momenta.

Evidence

Evidence 1 - **Rationale**: Both papers describe frameworks that assign independent synchronization periods to different optimizer states. The candidate explicitly mentions synchronization at different frequencies (kx for parameters, kj for momenta) with convergence guarantees, which is the same core concept as DES-LOC's independent synchronization periods. - **Original**: we propose desynced low communication adaptive optimizers (des-loc), a family of optimizers assigning independent synchronization periods to parameters and momenta, enabling lower communication costs while preserving convergence. - **Candidate**: mt-dao reduces communication costs by (1 kx + pn j=1 1 kj + 1 kv)-1 over ddp. this generalized framework recovers previous distributed adaptive optimizers [51, 12, 10, 19]. it also introducesthe first-ever formulations for provably convergent distributed variants of multi-momentum optimizers[36, ...

Evidence 2 - **Rationale**: Both algorithms explicitly define independent synchronization periods: DES-LOC uses kx for parameters and {kj} for optimizer states, while MT-DAO uses the identical notation kx for parameters and {kj} for momentum states, demonstrating the same architectural approach to independent synchronization. - **Original**: as shown in algorithm 1, des-loc synchronizes parameters x∈r d and optimizer states {sj}n j=1 at state-specific intervals kx,{k j}n j=1 ∈n +. - **Candidate**: require: model tensors, hyper-parameters 1:x 0 ∈r d,{¯uj -1}n j=1 ∈(r d)n ,¯v-1 ∈r d - initial params,nfirst momenta, second momentum... 6:k x,{k j}n j=1, kv ∈(n +)n+2 - communication periods for parameters and states

Evidence 3 - **Rationale**: Both papers provide theoretical analysis showing that parameter synchronization frequency is more critical than momentum synchronization frequency, with the candidate demonstrating that momentum sync frequency impact is modulated by decay rates—establishing prior theoretical understanding of independent synchronization effects. - **Original**: our theoretical analysis shows that while parameter synchronization dominates the asymptotic rate in-expectation, high-probability convergence guarantees require at least infrequent synchronization of the second momentum. - **Candidate**: the step size η is constrained by βω and ψ. the dependence ψ=o(1/p 2 x) shows that model synchronization frequency px is critical. the impact of momentum synchronization is nuanced: reducing a momentum's sync frequency pj increases its contribution to ψ, but this is modulated by its decay rate βj.

---

4. DeMo: Decoupled Momentum Optimization

URL: View paper

Prior Art Analysis

DeMo[53] demonstrates prior work on decoupling synchronization frequencies for different optimizer components. The candidate paper explicitly describes synchronizing fast-moving momentum components at every step while allowing slow-moving components to remain decoupled across accelerators. This directly challenges the novelty claim of DES-LOC being the first to assign independent synchronization frequencies to parameters and momentum states, as DeMo[53] already implements differential synchronization based on component velocity.

Evidence

Evidence 1 - **Rationale**: DeMo[53] describes a method that decouples momentum across accelerators while synchronizing only extracted fast components, demonstrating independent synchronization periods for different optimizer components before DES-LOC. - **Original**: we propose desynced low communication adaptive optimizers (des-loc), a family of optimizers assigning independent synchronization periods to parameters and momenta - **Candidate**: Starting from sgd with momentum, we make two key modifications: first, we remove the all-reduce operation on gradients ˜gk, decoupling momentum m across the accelerators. Second, after updating the momentum, we extract and remove its fast components q, which can be efficiently synchronized with mini...

---

5. Distributed Low-Communication Training with Decoupled Momentum Optimization

URL: View paper

Brief Assessment

Distributed Decoupled Momentum[63] focuses on momentum compression via DCT for communication reduction, not on assigning independent synchronization periods to parameters and momentum states as a core design principle.

---

6. Accelerated federated learning with decoupled adaptive optimization

URL: View paper

Brief Assessment

Federated Decoupled Adaptive[54] focuses on federated learning with momentum decoupling for client-server settings, not general distributed data-parallel training with independent synchronization periods for parameters and optimizer states.

Contribution 2: Convergence theory for desynchronized parameter and momentum updates

Description: The authors provide theoretical convergence guarantees for DES-LOC under non-convex objectives for SGDM and weakly convex objectives for Adam. Their analysis shows that parameter synchronization dominates the asymptotic convergence rate, while momentum synchronization frequency affects stable step sizes and high-probability bounds.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. A Bias Correction Mechanism for Distributed Asynchronous Optimization

URL: View paper

Brief Assessment

Bias Correction Asynchronous[59] focuses on asynchronous distributed optimization with bias correction mechanisms for handling data heterogeneity, not on desynchronized parameter and momentum updates with independent synchronization periods as in the original paper.

---

2. FADAS: Towards federated adaptive asynchronous optimization

URL: View paper

Brief Assessment

FADAS[58] focuses on asynchronous federated optimization with adaptive methods (Adam/AMSGrad), not on desynchronized parameter and momentum updates with independent synchronization periods as in the original paper. The candidate analyzes gradient delays in federated settings, while the original analyzes independent sync frequencies for parameters vs. momenta in distributed training.

---

3. Privacy-Preserving Asynchronous Federated Learning Framework in Distributed IoT

URL: View paper

Brief Assessment

Privacy Preserving Asynchronous[56] focuses on blockchain-based federated learning in IoT with differential privacy, not on convergence theory for desynchronized parameter and momentum updates in distributed optimization.

---

4. ADMM-tracking gradient for distributed optimization over asynchronous and unreliable networks

URL: View paper

Brief Assessment

ADMM-tracking[60] focuses on distributed optimization using ADMM-based consensus protocols for asynchronous networks. The candidate addresses consensus optimization over networks with packet losses, not the specific problem of desynchronized parameter and momentum updates in adaptive optimizers for foundation model training that the original paper tackles.

---

5. DeMo: Decoupled Momentum Optimization

URL: View paper

Brief Assessment

DeMo[53] does not provide formal convergence theory. The paper explicitly states 'We leave formal proofs of these conjectures to a later work' and bases its method on empirical conjectures rather than theoretical convergence guarantees.

---

6. Advances in asynchronous parallel and distributed optimization

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

---

7. Sharper Convergence Guarantees for Asynchronous SGD for Distributed and Federated Learning

URL: View paper

Brief Assessment

Asynchronous SGD Convergence[55] analyzes asynchronous SGD with delayed gradients in distributed settings, focusing on gradient delays rather than desynchronized parameter and momentum synchronization frequencies. The candidate does not address the specific problem of independently setting synchronization periods for parameters versus momentum states in adaptive optimizers.

---

8. Communication Efficient Asynchronous Stochastic Gradient Descent

URL: View paper

Brief Assessment

Asynchronous SGD Communication[57] focuses on asynchronous gradient descent with stale gradients in distributed settings, not on desynchronized parameter and momentum updates with independent synchronization periods as in DES-LOC.

---

9. Accelerated federated learning with decoupled adaptive optimization

URL: View paper

Brief Assessment

The candidate paper analyzes convergence in federated settings with momentum decoupling but does not provide convergence theory specifically for desynchronized parameter and momentum updates with independent synchronization frequencies as in DES-LOC.

Contribution 3: Empirical validation showing 170× communication reduction over DDP

Description: The authors demonstrate through experiments on language models up to 1.7B parameters that DES-LOC achieves substantial communication reductions: 170× compared to standard DDP and 2× compared to Local Adam, resulting in significant wall-clock speedups while maintaining competitive performance on in-context learning benchmarks.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Taming Momentum in a Distributed Asynchronous Environment

URL: View paper

Brief Assessment

Taming Momentum Distributed[67] focuses on asynchronous distributed training with momentum in parameter servers, not communication-efficient synchronous data-parallel methods. The candidate addresses gradient staleness in asynchronous SGD rather than reducing synchronization frequency in adaptive optimizers.

2. Distributed Low-Communication Training with Decoupled Momentum Optimization

URL: View paper

Brief Assessment

Distributed Decoupled Momentum[63] reports 3000× reduction over DDP and 16× over DiLoCo, but uses a different technical approach (DCT-based momentum compression) rather than the desynchronized state updates of the original paper.

3. Delayed Momentum Aggregation: Communication-efficient Byzantine-robust Federated Learning with Partial Participation

URL: View paper

Brief Assessment

Delayed Momentum Byzantine[65] focuses on Byzantine-robust federated learning with partial client participation, not on general distributed training communication efficiency. The 170× reduction claim in the original paper pertains to asynchronous momentum synchronization in foundation model training, while the candidate addresses robustness against malicious clients in federated settings.

4. A2CiD2: Accelerating Asynchronous Communication in Decentralized Deep Learning

URL: View paper

Brief Assessment

A2CiD2[72] focuses on asynchronous decentralized training with gossip-based communication acceleration, not on adaptive optimizers with desynchronized momentum synchronization as in the original paper. The communication reduction mechanisms are fundamentally different.

5. Efficient asynchronous federated learning with prospective momentum aggregation and fine-grained correction

URL: View paper

Brief Assessment

Prospective Momentum Aggregation[71] focuses on asynchronous federated learning with client-server communication in distributed settings, not on distributed data parallel training for foundation models. The communication reduction mechanisms and experimental contexts differ fundamentally.

6. Ordered momentum for asynchronous SGD

URL: View paper

Brief Assessment

Ordered Momentum[68] focuses on asynchronous SGD with momentum for distributed training, not on adaptive optimizers or communication-efficient training of foundation models. The paper addresses gradient delay and momentum incorporation in parameter servers, which is a different technical approach from DES-LOC's desynchronized state synchronization.

7. SlowMo: Improving Communication-Efficient Distributed SGD with Slow Momentum

URL: View paper

Brief Assessment

SlowMo[32] focuses on reducing communication frequency through slow momentum updates in distributed SGD, not on adaptive optimizers like Adam/AdamW which are central to DES-LOC's approach and the 170× reduction claim.

8. Asynchronous Distributed Bilevel Optimization

URL: View paper

Brief Assessment

Asynchronous Bilevel[69] focuses on asynchronous distributed bilevel optimization for hyperparameter tuning, not on communication-efficient training of foundation models through momentum desynchronization. The methods, problem settings, and technical approaches are fundamentally different.

9. ASMAFL: Adaptive staleness-aware momentum asynchronous federated learning in edge computing

URL: View paper

Brief Assessment

ASMAFL[70] focuses on asynchronous federated learning in edge computing with staleness-aware momentum, addressing wireless communication constraints and non-IID data. This is fundamentally different from the original paper's synchronous distributed training framework with desynchronized optimizer state updates for foundation models.

10. {eSGD}: Communication efficient distributed deep learning on the edge

URL: View paper

Brief Assessment

eSGD[66] focuses on edge-based distributed training with gradient sparsification techniques, achieving different compression ratios (50-87.5% gradient dropping) on MNIST. This is a distinct technical approach from the original paper's asynchronous momentum synchronization for foundation models, operating in different deployment contexts (edge vs. data center) and model scales.

Appendix: Text Similarity Detection

Textual similarity detection checked 24 papers and found 5 similarity segment(s) across 2 paper(s).

The following **2 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. DES-LOC: Desynced Low Communication Adaptive Optimizers for Training Foundation Models

Detected in: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

---

2. SlowMo: Improving Communication-Efficient Distributed SGD with Slow Momentum

Detected in: Core Task (sibling), Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] DES-LOC: Desynced Low Communication Adaptive Optimizers for Foundation Models View paper
- [1] Efficient training of large language models on distributed infrastructures: a survey View paper
- [2] Distributed foundation models for multi-modal learning in 6G wireless networks View paper
- [3] SWARM Parallelism: Training Large Models Can Be Surprisingly Communication-Efficient View paper
- [4] Boosting large-scale parallel training efficiency with c4: A communication-driven approach View paper
- [5] Distributed large models training optimization with real-time wireless channel feedback View paper
- [6] Feddat: An approach for foundation model finetuning in multi-modal heterogeneous federated learning View paper
- [7] Accelerating large language model training with in-package optical links for scale-out systems View paper
- [8] {MegaScale}: Scaling large language model training to more than 10,000 {GPUs} View paper
- [9] Cocktailsgd: Fine-tuning foundation models over 500mbps networks View paper
- [10] Megascale-moe: Large-scale communication-efficient training of mixture-of-experts models in production View paper
- [11] PromptFL: Let Federated Participants Cooperatively Learn Prompts Instead of Models â Federated Learning in Age of Foundation Model View paper
- [12] Two-stream federated learning: Reduce the communication costs View paper
- [13] Centauri: Enabling efficient scheduling for communication-computation overlap in large model training via communication partitioning View paper
- [14] Co2: Efficient distributed training with full communication-computation overlap View paper
- [15] Advances and open challenges in federated foundation models View paper
- [16] Hierarchical federated foundation models over wireless networks for multi-modal multi-task intelligence: Integration of edge learning with D2D/P2P-enabled fog â¦ View paper
- [17] Optimus-CC: Efficient Large NLP Model Training with 3D Parallelism Aware Communication Compression View paper
- [18] Protocol Models: Scaling Decentralized Training with Communication-Efficient Model Parallelism View paper
- [19] MS-DINO: Efficient Distributed Training of Vision Transformer Foundation Model in Medical Domain through Masked Sampling View paper
- [20] Towards federated large language models: Motivations, methods, and future directions View paper
- [21] Merak: An Efficient Distributed DNN Training Framework With Automated 3D Parallelism for Giant Foundation Models View paper
- [22] Resource management for low-latency cooperative fine-tuning of foundation models at the network edge View paper
- [23] Accelerating Distributed Deep Learning Training with Compression Assisted Allgather and Reduce-Scatter Communication View paper
- [24] Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning View paper
- [25] Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training View paper
- [26] Scaling Large Language Model Training on Frontier with Low-Bandwidth Partitioning View paper
- [27] The cost of training machine learning models over distributed data sources View paper
- [28] Distributed Fine-Tuning of Foundation Models Over Heterogeneous Edge Devices View paper
- [29] Cross-region Model Training with Communication-Computation Overlapping and Delay Compensation View paper
- [30] {SimAI}: Unifying Architecture Design and Performance Tuning for {Large-Scale} Large Language Model Training with Scalability and Precision View paper
- [31] WeiPipe: Weight Pipeline Parallelism for Communication-Effective Long-Context Large Model Training View paper
- [32] SlowMo: Improving Communication-Efficient Distributed SGD with Slow Momentum View paper
- [33] Communication-efficient and distributed learning over wireless networks: Principles and applications View paper
- [34] Lower bounds and nearly optimal algorithms in distributed learning with communication compression View paper
- [35] Communication-Efficient Distributed Training for Collaborative Flat Optima Recovery in Deep Learning View paper
- [36] Accelerating Distributed Model Training through Intelligent Node Selection and Data Allocation Strategies in 6G network View paper
- [37] Mad-max beyond single-node: Enabling large machine learning model acceleration on distributed systems View paper
- [38] Communication-Efficient Language Model Training Scales Reliably and Robustly: Scaling Laws for DiLoCo View paper
- [39] Characterizing the Efficiency of Distributed Training: A Power, Performance, and Thermal Perspective View paper
- [40] On optimizing the communication of model parallelism View paper
- [41] RailX: a flexible, scalable, and low-cost network architecture for hyper-scale LLM training systems View paper
- [42] Cluster Based Heterogeneous Federated Foundation Model Adaptation and Fine-Tuning View paper
- [43] Communication-efficient distributed learning: An overview View paper
- [44] Understanding and Characterizing Communication Characteristics for Distributed Transformer Models View paper
- [45] FasterMoE: modeling and optimizing training of large-scale dynamic pre-trained models View paper
- [46] Megatron-lm: Training multi-billion parameter language models using model parallelism View paper
- [47] Distributed learning in wireless networks: Recent progress and future challenges View paper
- [48] Concerto: Automatic Communication Optimization and Scheduling for Large-Scale Deep Learning View paper
- [49] Large scale distributed neural network training through online distillation View paper

- [50] Vela: A Virtualized LLM Training System with GPU Direct RoCE View paper
- [51] Advances in asynchronous parallel and distributed optimization View paper
- [52] Decoupled momentum optimization View paper
- [53] DeMo: Decoupled Momentum Optimization View paper
- [54] Accelerated federated learning with decoupled adaptive optimization View paper
- [55] Sharper Convergence Guarantees for Asynchronous SGD for Distributed and Federated Learning View paper
- [56] Privacy-Preserving Asynchronous Federated Learning Framework in Distributed IoT View paper
- [57] Communication Efficient Asynchronous Stochastic Gradient Descent View paper
- [58] FADAS: Towards federated adaptive asynchronous optimization View paper
- [59] A Bias Correction Mechanism for Distributed Asynchronous Optimization View paper
- [60] ADMM-tracking gradient for distributed optimization over asynchronous and unreliable networks View paper
- [61] DES-LOC: Desynced Low Communication Adaptive Optimizers for Training Foundation Models View paper
- [62] FlexDeMo: Decoupled Momentum Optimization for Hybrid Sharded Data Parallel Training View paper
- [63] Distributed Low-Communication Training with Decoupled Momentum Optimization View paper
- [64] MT-DAO: Multi-Timescale Distributed Adaptive Optimizers with Local Updates View paper
- [65] Delayed Momentum Aggregation: Communication-efficient Byzantine-robust Federated Learning with Partial Participation View paper
- [66] {eSGD}: Communication efficient distributed deep learning on the edge View paper
- [67] Taming Momentum in a Distributed Asynchronous Environment View paper
- [68] Ordered momentum for asynchronous SGD View paper
- [69] Asynchronous Distributed Bilevel Optimization View paper
- [70] ASMAFL: Adaptive staleness-aware momentum asynchronous federated learning in edge computing View paper
- [71] Efficient asynchronous federated learning with prospective momentum aggregation and fine-grained correction View paper
- [72] A2CiD2: Accelerating Asynchronous Communication in Decentralized Deep Learning View paper