Benefits of staleness and asynchrony in machine learning algorithms

CS744: Big Data Systems – Group 8

Hongyi Wang, Xin Jin, Yuzhe Ma

UNIVERSITY OF WISCONSIN–MADISON

Department of Computer Sciences, University of Wisconsin–Madison

Motivation

- Scaling **synchronous** distributed machine learning is challenging because of **straggler effect**
- **Back-up worker** setups mitigate straggler effect but still suffering from losing data for each epoch since stale gradient will be dropped by master
- **Staleness** of gradient has reported that has partially equivalent effect as adding **momentum** in iterative-style optimization method
- Our approach is motivated by foregoing points, we're trying to **use stale gradients** to improve model **convergence rate** while maintain **speedup gains** under backup worker setups

Background - Overview

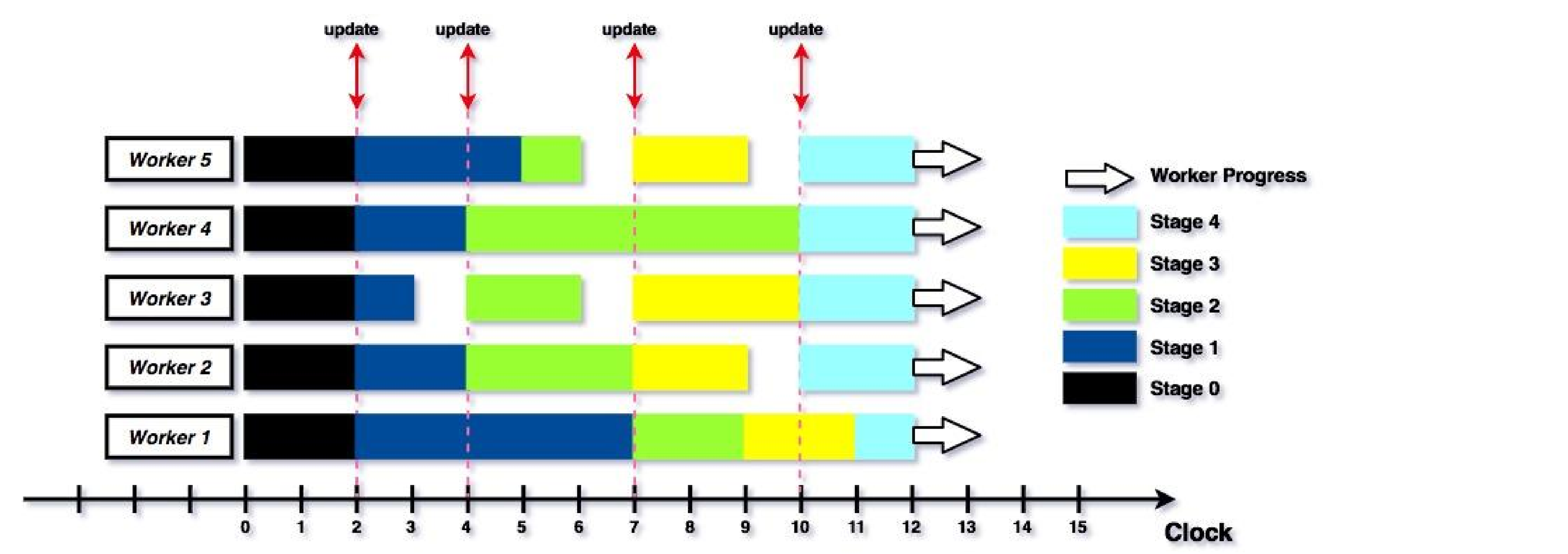


Fig. 1 Setting and Approach in this work

- For each iteration, master only wait for **k** faster workers out of **n**
- When gradients from slow workers (t-1, t-2, ...) are received, master cache and use them for next model update (for step t)

System Design

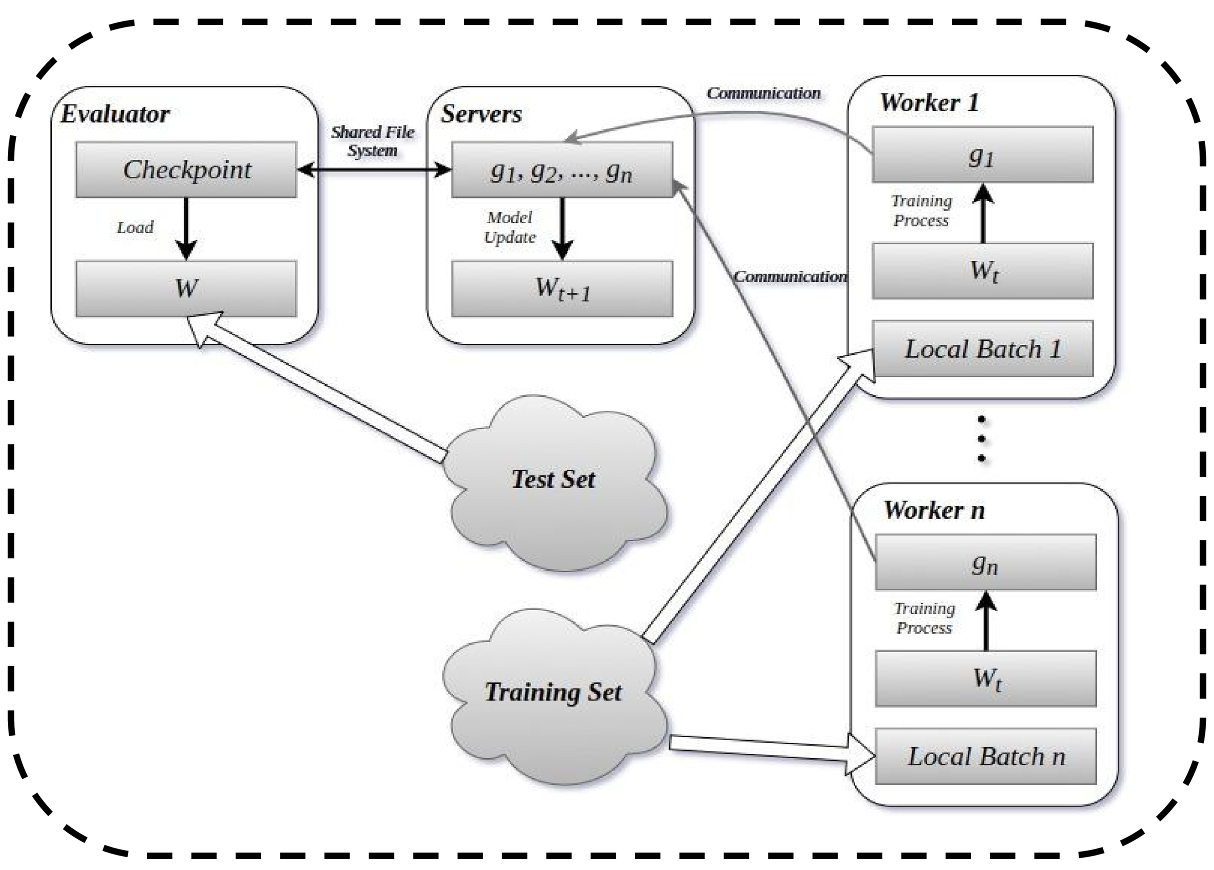


Fig. 2 Distributed Setup and System Design

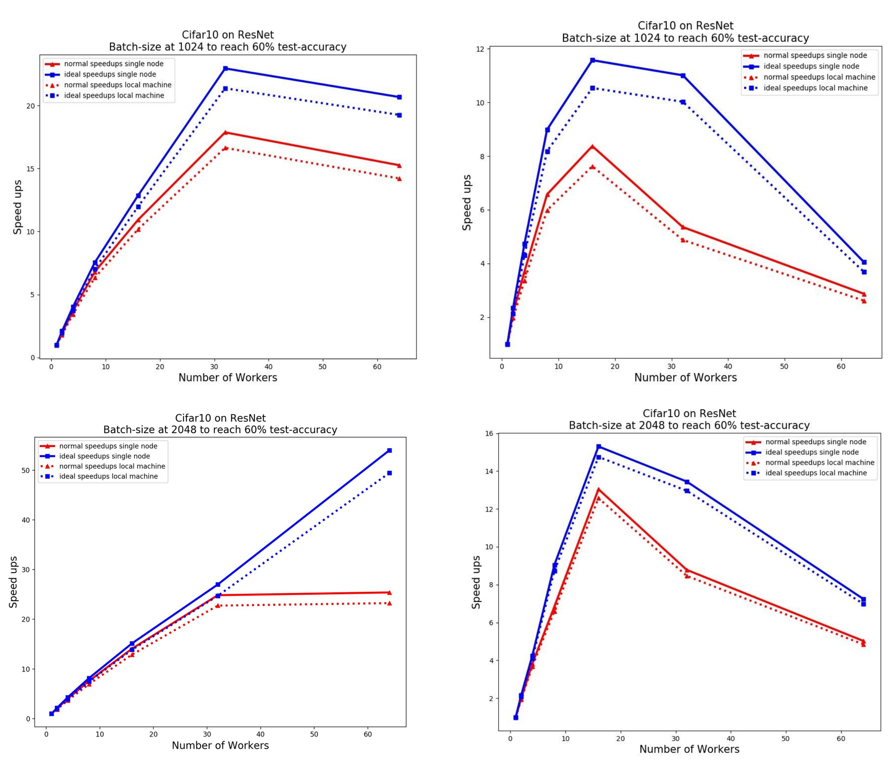


Fig. 3 Speedup Performances

- We implement **Parameter Server** distributed setting and train deep network model in **synchronous** manner
- Our Distributed Algorithm is implemented in **PyTorch** + **MPI**, model training process are handled by PyTorch while communication is achieved through MPI
- **Gradient compression** is implemented for reducing communication overhead
- Our system gain good **speedups** as number of nodes scales up

Theoretical Analysis

Mathematical Model: $w_{t+1} = w_t - \sum_{i=t-k+1}^{t} \alpha_i \nabla f(w_i)$

One Dimensional Case

Definition 1. *Generalized Curvature. The derivative of $f(x) : \mathbb{R} \to \mathbb{R}$, can be written as*

$$f'(x) = h(x)(x - x^*)$$

for some $h(x) \in \mathbb{R}$, where x^* is the global minimum of $f(x)$. We call $h(x)$ the generalized curvature.

Assumption: $h(x) \in [a, b], 0 < a \leq b.$ (bounded curvature)

Theorem 1. *Let $f(w)$ be strictly convex, and assume the generalized curvature $h(w) \in [a, b]$, where $0 < a \leq b$. If $c \leq \alpha_t \leq \frac{1}{b}$ for some $c > 0$ and $\sum_{i=t-k+1}^{t-1} \alpha_i \leq \frac{a}{2b}\alpha_t$, then $\lim_{t\to\infty} |w_t - w^*| = 0$.*

High Dimensional Case

Definition 2. *High-dimensional Generalized Curvature. The derivative of a strictly convex function $f(x) : \mathbb{R}^d \to \mathbb{R}$, can be written as*

$$\nabla f(x) = H(x)(x - x^*)$$

for some $\nabla f(x) \in \mathbb{R}^d$, where x^* is the global minimum of $f(x)$. Let $\lambda_i, i \in [d]$ be the eigenvalues of $H(x)$ and also use $v_i, i \in [d]$ to denote the corresponding eigenvectors. We call λ_i the generalized curvature along direction v_i.

Assumption: all the eigenvalues satisfy $\lambda_i \in [a, b], 0 < a \leq b.$

Theorem 2. *Let $f(w)$ be strictly convex. Assume the generalized curvature $\lambda_i \in [a, b]$ for all i at any w, where $0 < a \leq b$. If $c \leq \alpha_t \leq \frac{1}{b}$ for some $c > 0$ and $\sum_{i=t-k+1}^{t-1} \alpha_i \leq \frac{a}{2b}\alpha_t$, then $\lim_{t\to\infty} \|w_t - w^*\| = 0$.*

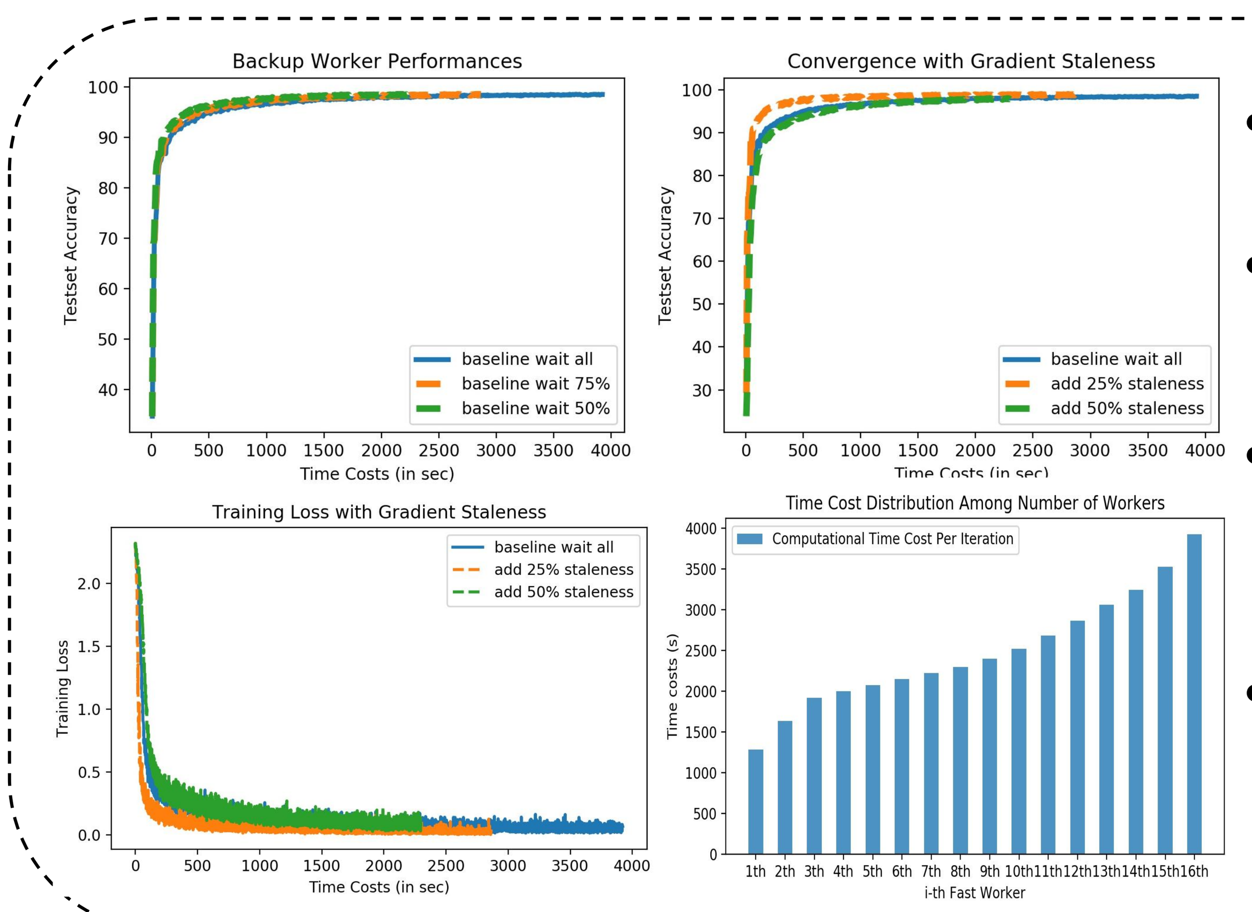
Example 1. *Ridge Regression. Let $f(w) = \|Xw - y\|^2 + \eta\|w\|^2$. The global minimum $w^* = (X^\top X + \eta I)^{-1} X^\top y$. $f(w) = (Xw - y)^\top (Xw - y) + \eta w^\top w$, thus $\nabla f(w) = 2X^\top X w - 2X^\top y + 2\eta w = 2(X^\top X + \eta I)(w - w^*)$ and $H(w) = 2(X^\top X + \eta I)$, which is a constant with respect to w. Now consider the generalized curvature of $f(w)$, which are the eigenvalues of $H(w) = 2(X^\top X + \eta I)$. Without loss of generality we assume single instances satisfy $\|x\| \leq 1$. Then we have the following claim:*

Theorem 3. *All the eigenvalues of $H(w) = 2(X^\top X + \eta I)$ lies between $[2\eta, 2n + 2\eta]$, where n is the training set size.*

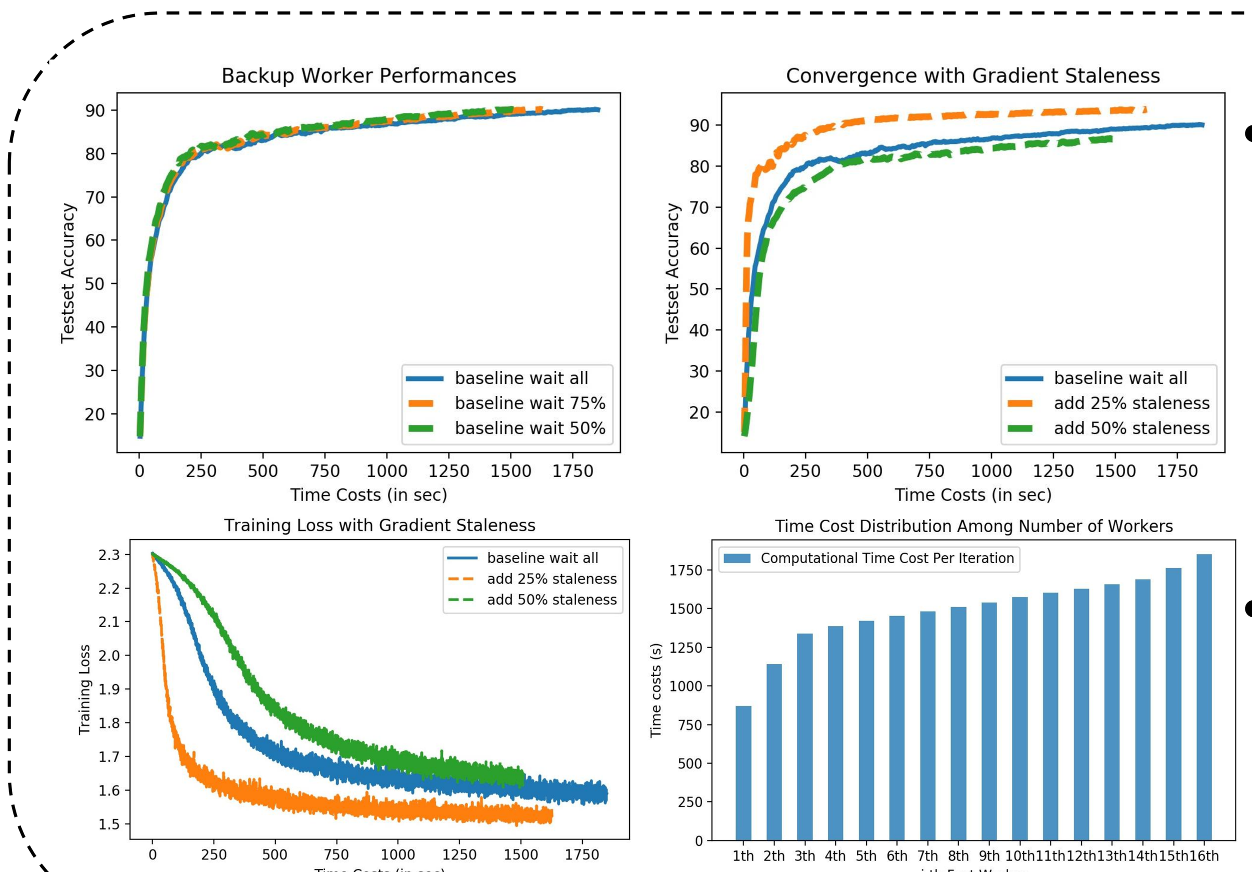
Extension to SGD

Update Rule: $w_{t+1} = w_t - \sum_{i=t-k+1}^{t} \alpha_i X_i, \mathbf{E}X_i = \nabla f(w_i).$

Theorem 4. *Let $f(w)$ be strictly convex. Assume the generalized curvature $\lambda_i \in [a, b]$ for all i at any w, where $0 < a \leq b$. If $c \leq \alpha_t \leq \frac{1}{b}$ for some $c > 0$ and $\sum_{i=t-k+1}^{t-1} \alpha_i \leq \frac{a}{2b}\alpha_t$, then $\lim_{t\to\infty} \|\mathbf{E}[w_t - w^*]\| = 0$.*



- Experiments are running on **m4.2xlarge** instances on AWS EC2
- The Deep Network **LeNet** and hand-written image dataset **MNIST** are used for these results
- **mini-batch SGD** is implemented for this experiment global batch size B=256
- global batch are **splitted among workers** each worker shares local batch size at B/n



- Following the same settings these experiments are running on multi-layer **fully connected** neural network with **MNIST** dataset
- 3 hidden layer are used with number of hidden units at 800, 500, 10 respectively

References

[1] J. Zhang, I. Mitliagkas, and C. Ŕe. "Yellowfin and the art of momentum tuning." arXiv preprint arXiv:1706.03471, 2017.
[2] Li, Mu, et al. "Scaling Distributed Machine Learning with the Parameter Server." *OSDI*. Vol. 1. No. 10.4. 2014.