Measurement of Stragglers in Distributed Machine Learning Model Training

1 Alibaba PAI Production Cluster Data

1.1 Data from April 2, 2022

Coefficient of variation of worker iteration time: Coefficient of variation (CoV) is the ratio of the standard deviation to the mean. Figure 1 shows the CoV of five randomly selected machine learning (ML) jobs.

It can be observed that the CoV of the iteration time of job_13 and job_74 varies considerably between 0.1 and 2.4 over time and that of job_15 and job_81 ranges between 0.05 and 1. The CoV of job_27 keeps stable from the beginning to the end. During the selected timeline, job_13, job_74 and job_81 suffer from stragglers obviously all the way. The CoV of job_27 is relatively stable, which means that it may not have stragglers. job_15 experiences stragglers in the end of the timeline, while the straggling degree is not so significant compared to other jobs (except job_27) for most of the time.

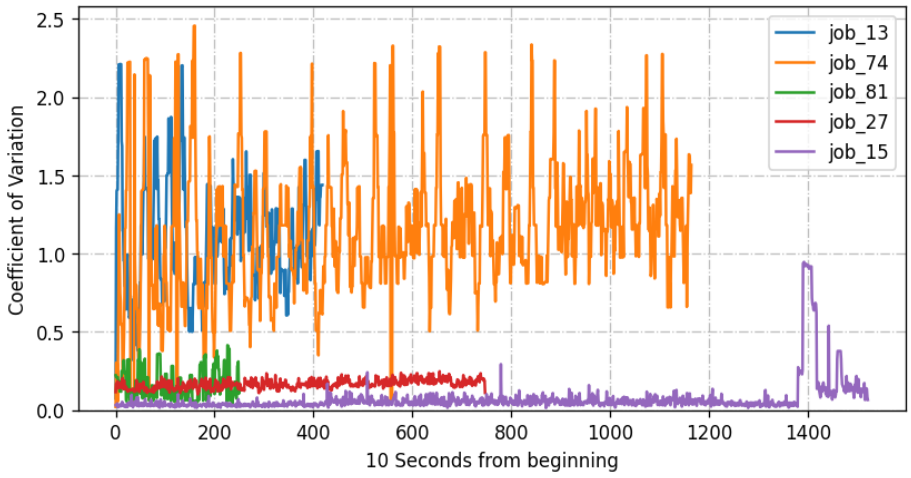


Figure 1: Coefficient of variation of the iteration times of a job.

Next, we show the iteration time of different workers of other 4 randomly selected jobs.

Time series of job 1: Figure 2 shows the worker iteration times of tracked job_1 over time. The plan ratio of requested GPUs of each worker is 25%. The iteration time of workers except worker_6 keeps almost the same from the beginning to the end. During the time between approximately 1800s-2000s, worker_6's iteration time drastically by 150%, becoming a straggler.

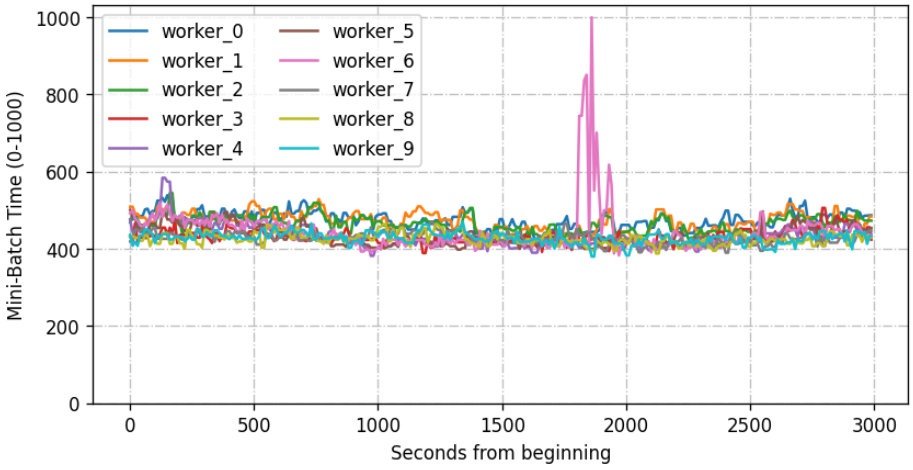


Figure 2: Time series of iteration times of the workers of job 1.

Time series of job 2: Figure 3 shows the worker iteration times of job_2 over time. The plan ratio of requested GPUs of each worker is 20%. Even though the iteration time of each worker of job_2 keeps relatively stable, different workers inherently have different iteration times. The job completion time of job_2 is limited by the slowest worker, i.e., worker_5. The relatively slow workers, worker_3 - worker_8, can be regarded as stragglers during the training.

Time series of job 3: Figure 4 shows the worker iteration times of job_3 over time. The plan ratio of requested GPUs of each worker is 25%. Different workers of job_3 have different iteration times. worker_0 and worker_1 are the two slowest workers during the training. From the beginning to approximately 2100s, the iteration times of all workers keep stable. After 2100s, the iteration time of worker_6 - worker_9 increase by approximately 50% and the iteration time variation among all workers becomes larger.

Time series of job 4: Figure 5 shows the worker iteration times of job_4 over time. The plan ratio of requested GPUs of each worker is 50%. For most of the time, the iteration time of each worker is relatively stable. Different workers have different iteration times. worker_1, worker_3, worker_4 and worker_9 are fast workers. Workers of job_4 occasionally become stragglers during the training. There are more than one straggler and the straggling situation becomes better or worse over time. The iteration time of worker_5 and worker_6 significantly increases by 400% in the end of selected timeline and the straggling degree becomes extremely severe.

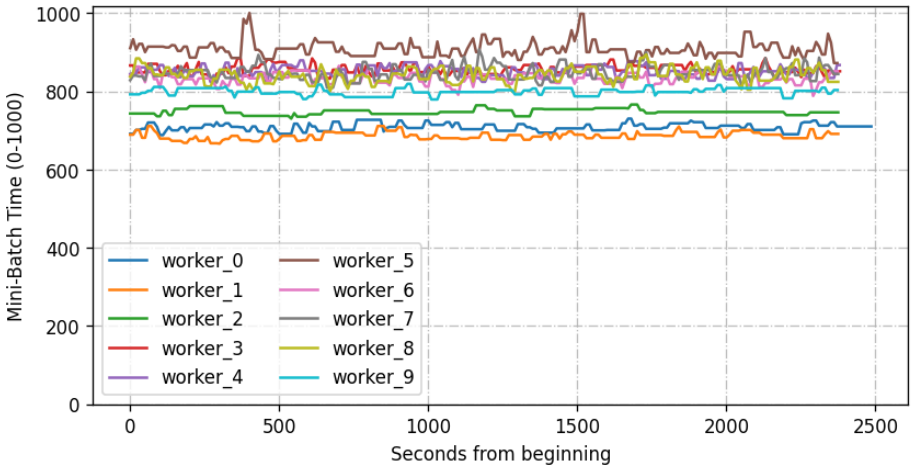


Figure 3: Time Series of iteration times of the workers of job 2.

Figure 4: Time series of iteration times of the workers of job 3.

1.2 Trace Data from July to August of 2020

In the production cluster, the workers of a job are placed in different machines [3, 4]. We also conducted experiments to observe how the workers are placed in different machines using the Alibaba trace data [3]. Figure 6 shows the cumulative distribution function (CDF) of the number of servers that host the workers of a job of al the jobs occurs from July to August of 2020. Figure 6(a) demonstrates the CDF for all the jobs. We can see that the workers are placed across 100 servers for around 98% jobs. Figure 6(b) shows the CDF for the jobs whose workers are hosted by no more than 100 servers. Here, we can observe

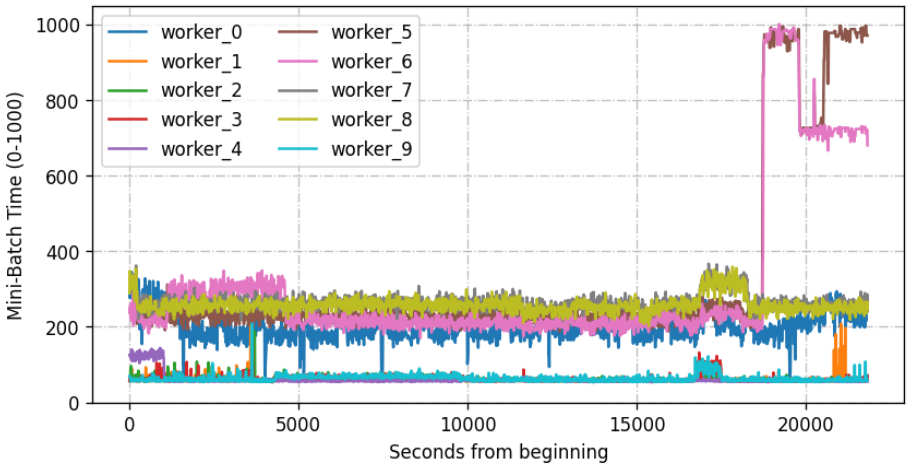


Figure 5: Time series of iteration times of the workers of job 4.

that the workers are placed across 20 servers for around 80% jobs. This confirms that almost for all jobs, the workers are placed in multiple servers. Since the workers are placed in different servers, they tend to experience different resource contention, which causes some workers to run slower than other workers and thus result in stragglers.

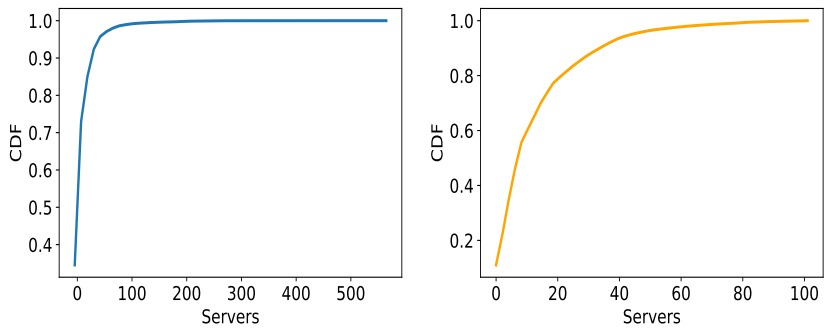
2 Measurement on Our Institute Cluster

2.1 Existence of Worker Stragglers

To observe the prevalence of stragglers, we run 100 DL training jobs in a cluster at our institute. The jobs are executed using 5 servers following the arrival of Microsoft trace [1]. We use PS architecture for 50 jobs and all-reduce architecture for 50 jobs.

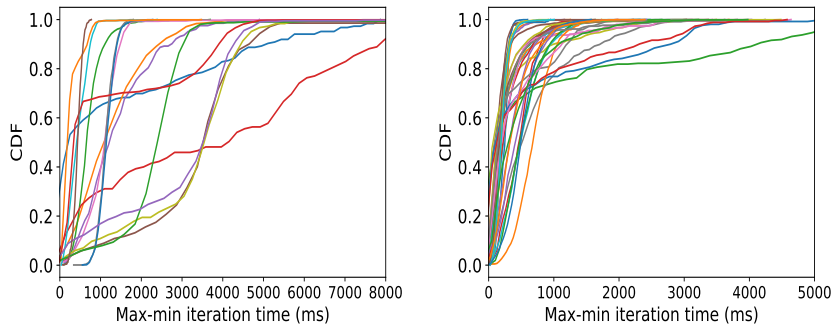
We measured the iteration time that each worker takes to complete its task in an iteration. The CDF of time difference for the fastest and slowest workers for the DL jobs are shown in Figure 7. Figure 7(a) and Figure 7(b) show time difference for DL jobs in the parameter server (PS) architecture with 8 workers and 4 workers respectively. For the jobs with 8 workers, we can see that the difference between the maximum and minimum iteration time of workers is more than 1 second for almost all jobs. For jobs with 4 workers, we can see that the difference between the maximum and minimum iteration time of workers is more 0.5 second for almost all jobs. The results indicate that the time difference is high, and could generate stragglers.

Figure 7(c) and Figure 7(d) show the time difference for DL jobs in the all-reduce architecture with 8 workers and 4 workers, respectively. Here, for the jobs with 8 workers, we can see that the difference between the maximum and

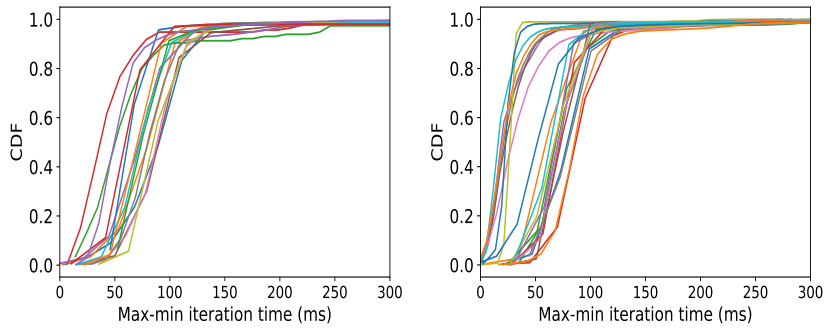


(a) Worker distribution in all servers. (b) Worker distribution upto 100 servers.

Figure 6: CDF of the number of servers that host the workers of a job of all the jobs in the Alibaba trace.



(a) DL jobs with 8 workers in PS architecture. (b) DL jobs with 4 workers in PS architecture.

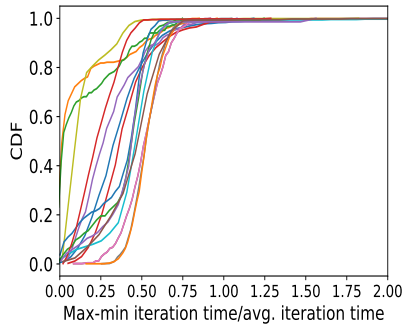


(c) DL jobs with 8 workers in all-reduce architecture. (d) DL of jobs with 4 workers in all-reduce architecture.

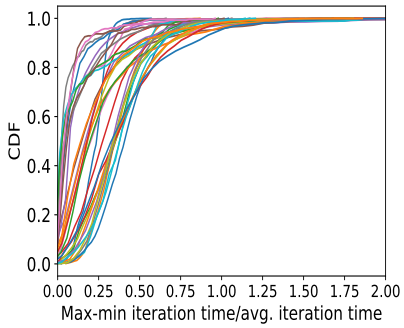
Figure 7: Time difference between the maximum and minimum iteration time of workers for different jobs.

minimum iteration time of workers is more than 50 millisecond for almost all jobs. For the jobs with 4 workers, we can observe that the difference is also more that 50 millisecond for almost all jobs. This is because the slower workers or stragglers encounter different resource contention. Besides, they do not receive

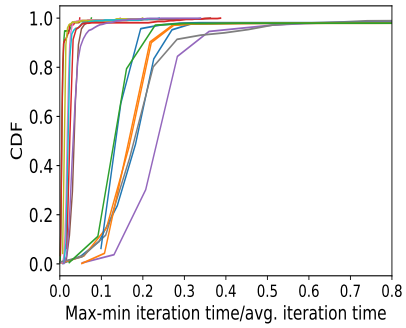
their required resources due to resource unavailability or uncertainty.



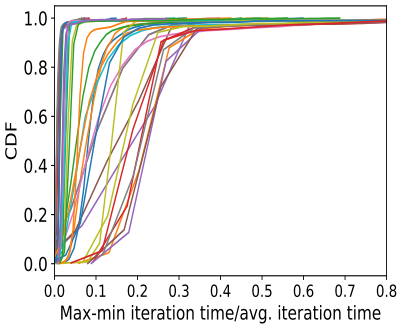
(a) DL jobs with 8 workers in PS architecture.



(b) DL jobs with 4 workers in PS architecture.



(c) DL jobs with 8 workers in all-reduce architecture.



(d) DL of jobs with 4 workers in all-reduce architecture.

Figure 8: Ratio of time difference and average iteration time for different jobs.

Figure 8 shows the ratio of the time difference over the average iteration time for both the PS and all-reduce architectures. Figure 8(a) and Figure 8(b) show the ratio of the time difference over the average iteration time for DL jobs in the PS architecture with 8 workers and 4 workers, respectively. For the jobs with 8 workers, we can observe that the time difference is more than 30% of the average iteration time for more than 80% iterations, which is significantly higher compared to the average iteration time. This indicates the prevalence of straggler occurrence. For almost all the jobs with 4 workers, we can see that the time difference is more than 15% of the average iteration time for more than 80% iterations.

Figure 8(c) and 8(d) show the ratio of the time difference over the average iteration time for DL jobs in the all-reduce architecture with 8 workers and 4 workers, respectively. For the jobs with 8 workers, we can observe that the time difference is more than 10% of average iteration time for more than 80% iterations. For the jobs with 4 workers, we can see that the time difference is more than 7% of average iteration time for more than 80% iterations. The results show the existence of stragglers.

2.2 Existence of PS Stragglers

In a PS architecture with multiple PSs, a PS may become straggler if it is slow than other PSs [2] . We ran 50 jobs that use PS architecture in a computing platform at our institute to test the existence of PS stragglers. First, let's present some concepts. As shown in the following, we compute the PS speed by dividing the parameter size of a PS by its PS time. A PS's PS time is the sum of the average gradient and parameter transfer times (between the PS and all workers) and its PS computing time. The PS speed variation is the difference between maximal and minimal PS speed divided by the minimal PS speed.

$$\begin{aligned}\text{PS time} = &\text{ Average gradient transfer time} \\ &+ \text{Average parameter transfer time} \\ &+ \text{PS computing time}\end{aligned} \tag{1}$$

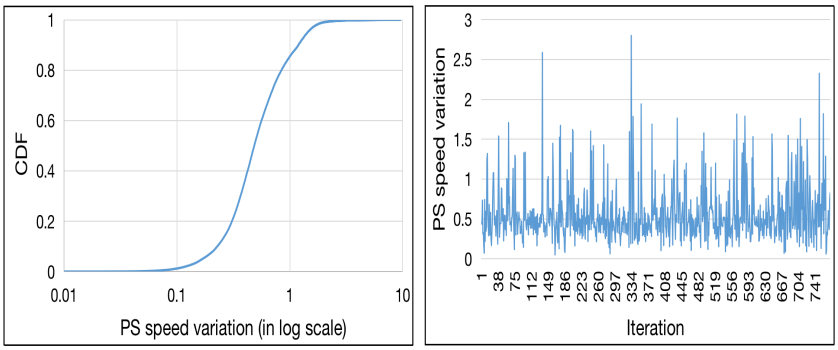
$$\text{PS speed} = \frac{\text{Parameter size of a PS}}{\text{PS time}} \tag{2}$$

$$\text{PS speed variation} = \frac{\text{Max PS speed} - \text{Min PS speed}}{\text{Min PS speed}} \tag{3}$$

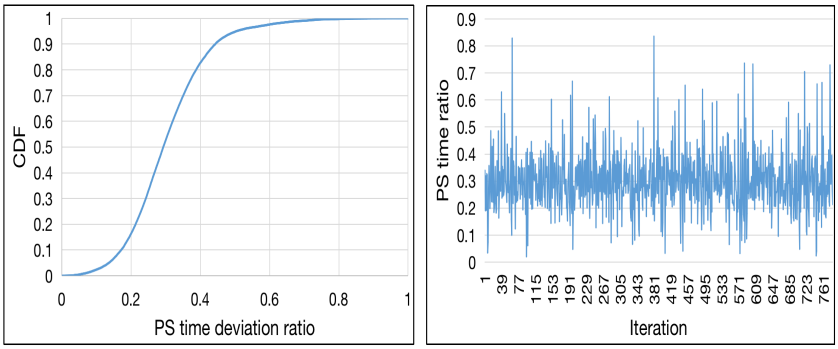
We also compute PS time deviation ratio by dividing the difference between maximal and minimal PS time by iteration time.

$$\text{PS Time Deviation Ratio} = \frac{\text{Max PS time} - \text{Min PS time}}{\text{Iteration time}} \tag{4}$$

Figure 9(b) and Figure 9(a) show the CDF of the PS speed variation [2] and this value over each iteration, respectively. Figure 9(d) and Figure 9(c) show the CDF of the PS time deviation ratio and this value over each iteration, respectively. It can be observed that both PS speed and PS time deviation ratio vary considerably during the entire training. PS speed variation ranges between 0.1 and 2.5 and PS time deviation ratio fluctuates between 0.1 and 0.8. The results show that there exist PS stragglers from the beginning to the end during ML training.



(a) CDF of PS speed variation. (b) PS speed variation.



(c) CDF of PS time deviation ratio. (d) PS time deviation ratio.

Figure 9: Existence of PS stragglers.

References

[1] Microsoft trace. `https://github.com/msr-fiddle/philly-traces`.

[2] Yangrui Chen, Yanghua Peng, Yixin Bao, Chuan Wu, Yibo Zhu, and Chuanxiong Guo. Elastic parameter server load distribution in deep learning clusters. In *Proc. of SoCC*, 2020.

[3] Qizhen Weng, Wencong Xiao, Yinghao Yu, Wei Wang, Cheng Wang, Jian He, Yong Li, Liping Zhang, Wei Lin, and Yu Ding. MLaaS in the wild: Workload analysis and scheduling in large-scale heterogeneous GPU clusters. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*, 2022.

[4] Wencong Xiao, Shiru Ren, Yong Li, Yang Zhang, Pengyang Hou, Zhi Li, Yihui Feng, Wei Lin, and Yangqing Jia. Antman: Dynamic scaling on gpu clusters for deep learning. In *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20)*, pages 533–548, 2020.