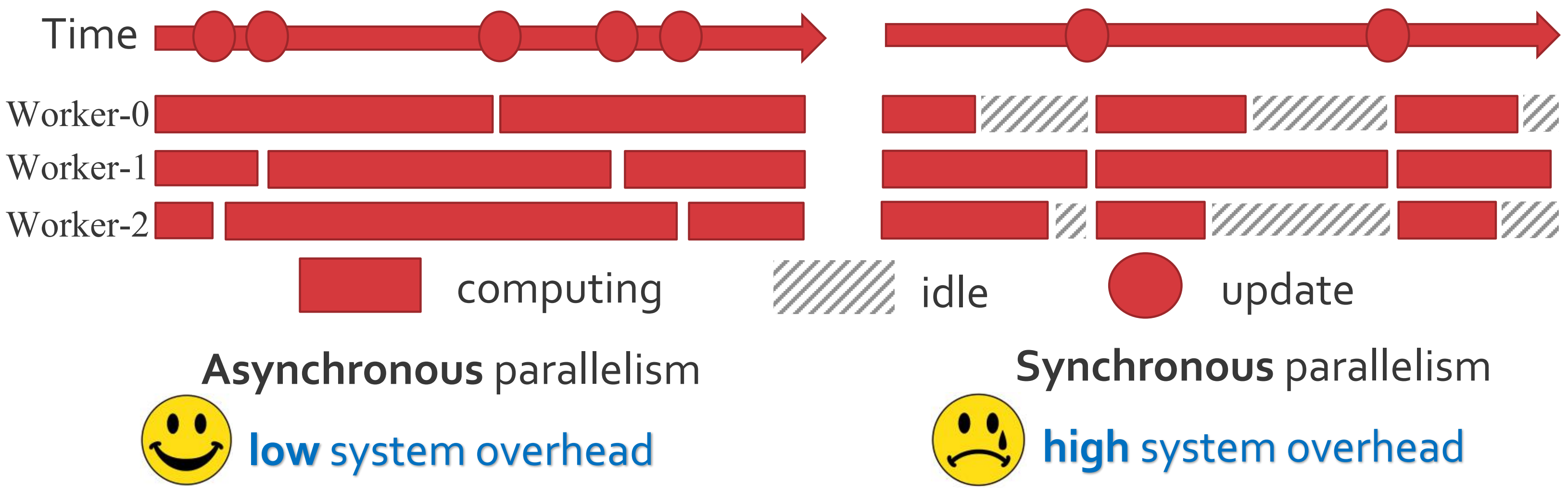
Asynchronous Parallel Stochastic Gradient for **Nonconvex** Optimization

Xiangru Lian, Yijun Huang, Yuncheng Li, and Ji Liu
University of Rochester

Speaker: Xiangru Lian

- **Nonconvex** optimization: Deep Learning, NLP, Recommendation, etc.
- **Asy**nchronous **S**tochastic **G**radient (**AsySG**): popular & powerful in large scale problems.

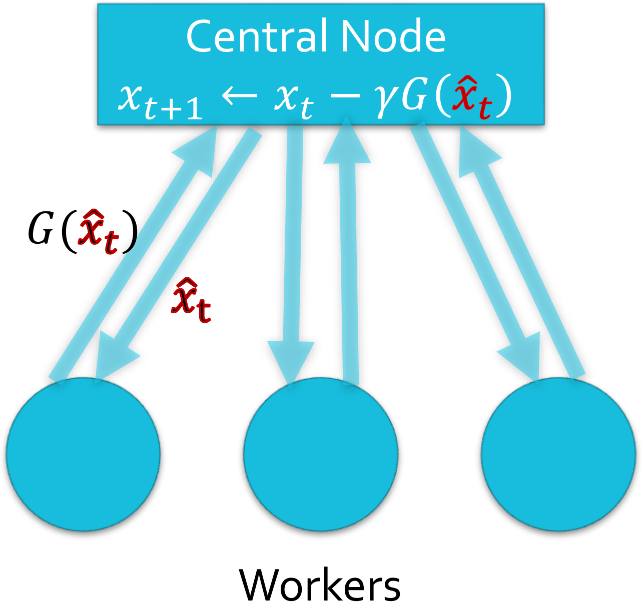


For **AsySG** in **nonconvex** optimization, the theoretical analysis is still limited.

Our Main Results: Proved 1) Convergence of AsySG, 2) Linear speedup in parallelism.

Asynchronous Stochastic Gradient Algorithm (AsySG)

$\boldsymbol{G}(\cdot)$**:** stochastic gradient. $\boldsymbol{x}$**:** optimization variable.



All workers run concurrently:

1. (**Read**): read $\hat{x}_t$ from the central node.
2. (**Compute**): compute $G(\hat{x}_t)$ using local data.
3. (**Update**): update x in the central node ***without locks:***

$$x_{t+1} \leftarrow x_t - \gamma G(\hat{x}_t).$$

Key challenges in analysis

- $\hat{\boldsymbol{x}}_t \neq x_t$;
- **Different implementations => Different forms of $\hat{\boldsymbol{x}}_t$.**

Example
Cluster Implementation
Multicore Implementation

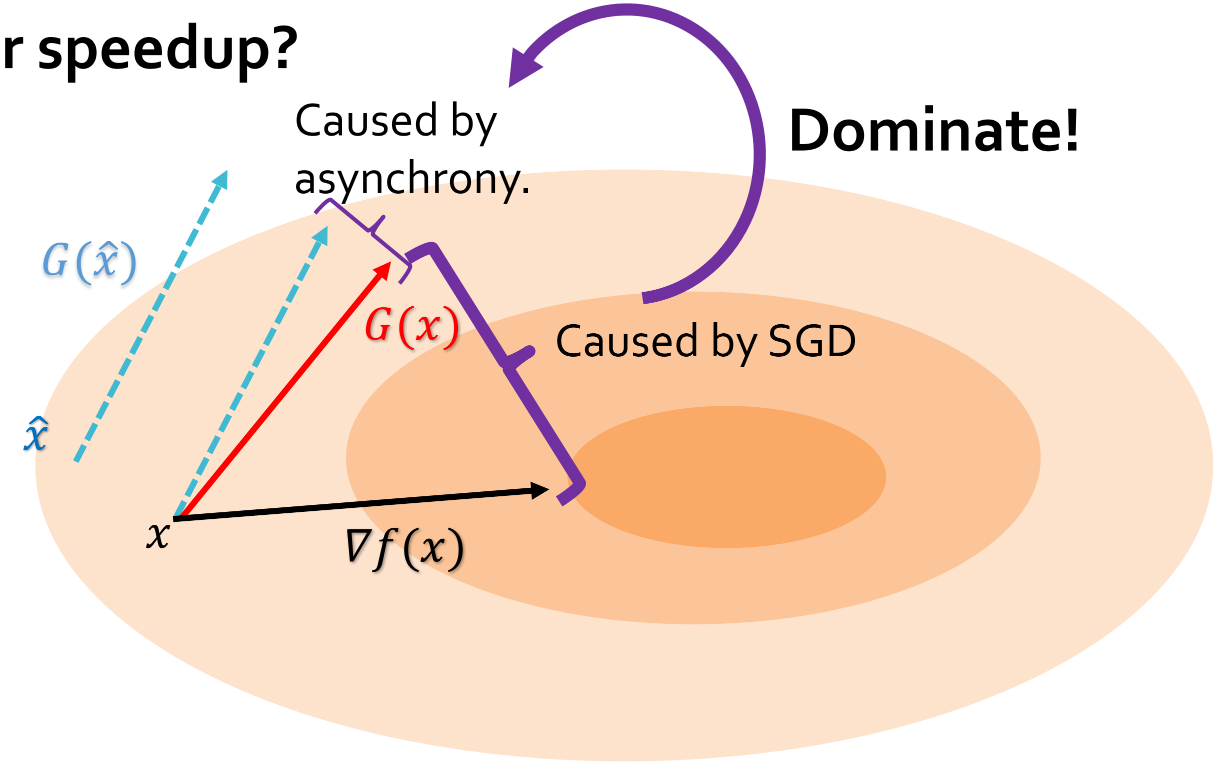
Our Results

K := # of iterations.

- **Q: Does AsySG converge?**
  A: Yes, the rate is consistent with SGD.
- **Q: How much speedup?**
  A: ***Linear speedup*** up to $O(\sqrt{K})$ workers.

Poster: Tonight 7 PM #63 @ 210 C

Why linear speedup?



Thank all the reviewers for their constructive comments!