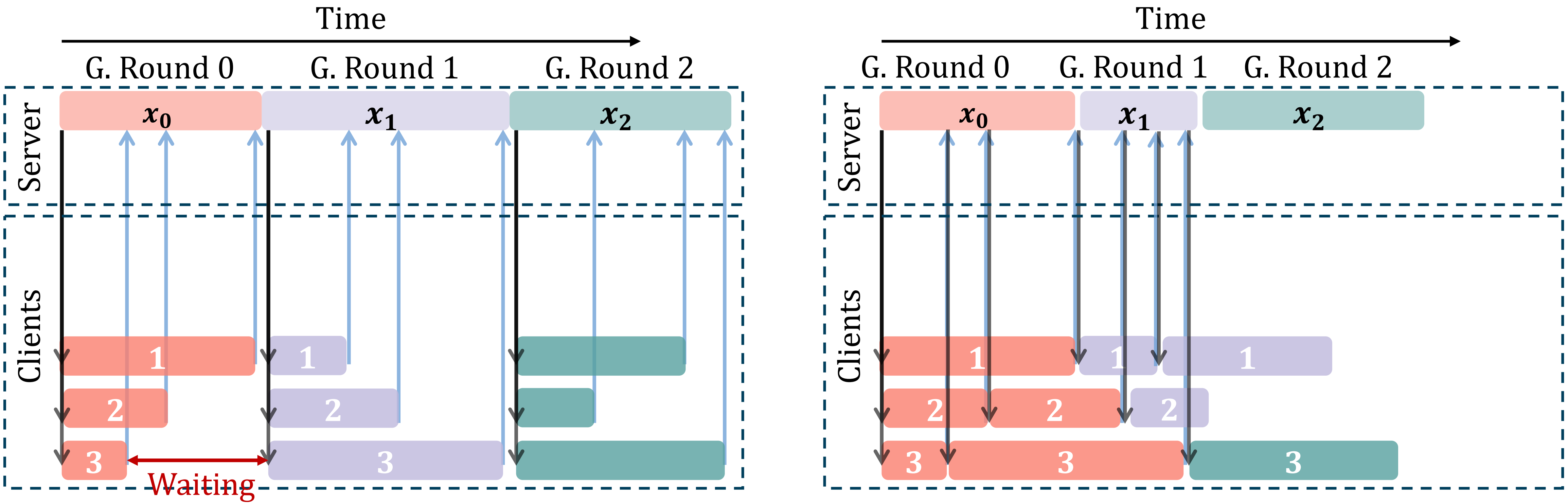
Tackling the Data Heterogeneity in Asynchronous Federated Learning with Cached Update Calibration

Yujia Wang, Yuanpu Cao, Jingcheng Wu, Ruoyu Chen, and Jinghui Chen.

Asynchronous FL: Background

From synchronous FL to asynchronous FL (FedAsync*, FedBuff**): improve the training efficiency



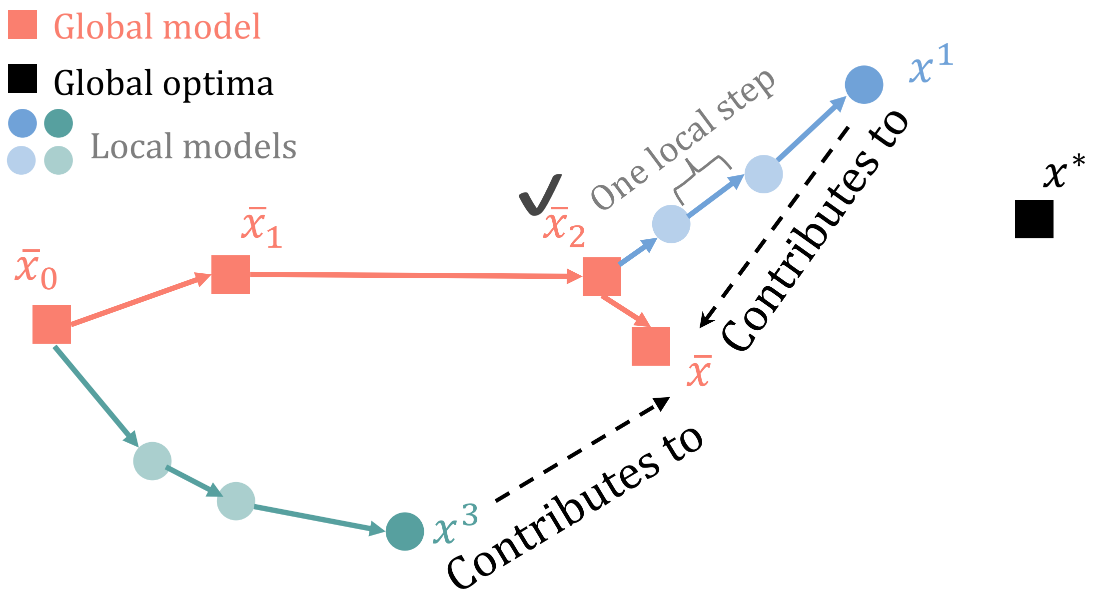
*Xie, Cong, Sanmi Koyejo, and Indranil Gupta. "Asynchronous federated optimization."
**Nguyen, John, et al. "Federated learning with buffered asynchronous aggregation."

Asynchronous FL: Background

What makes Asynchronous FL less efficient?

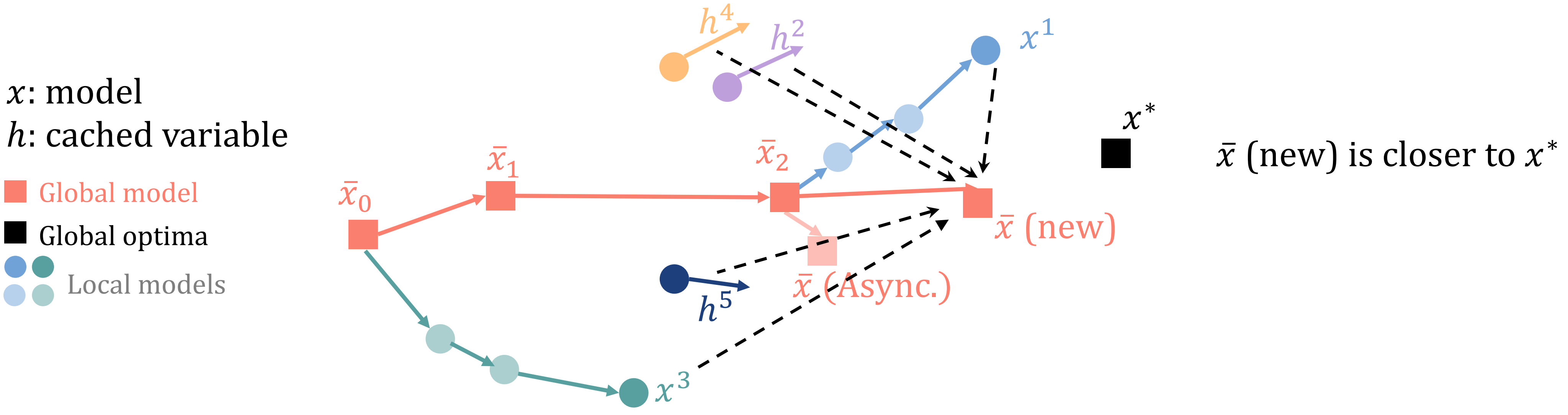
- Client 1 and Client 3 may differ in data distribution
- x^1 is computed from a latest global model $\bar{x}_2$
- x^3 is computed from an outdated model $\bar{x}_0$, but x^3 is update to $\bar{x}_2$

➢ The delay of Client 3 hurts convergence



Cache-Aided Asynchronous FL (CA^{2}FL)

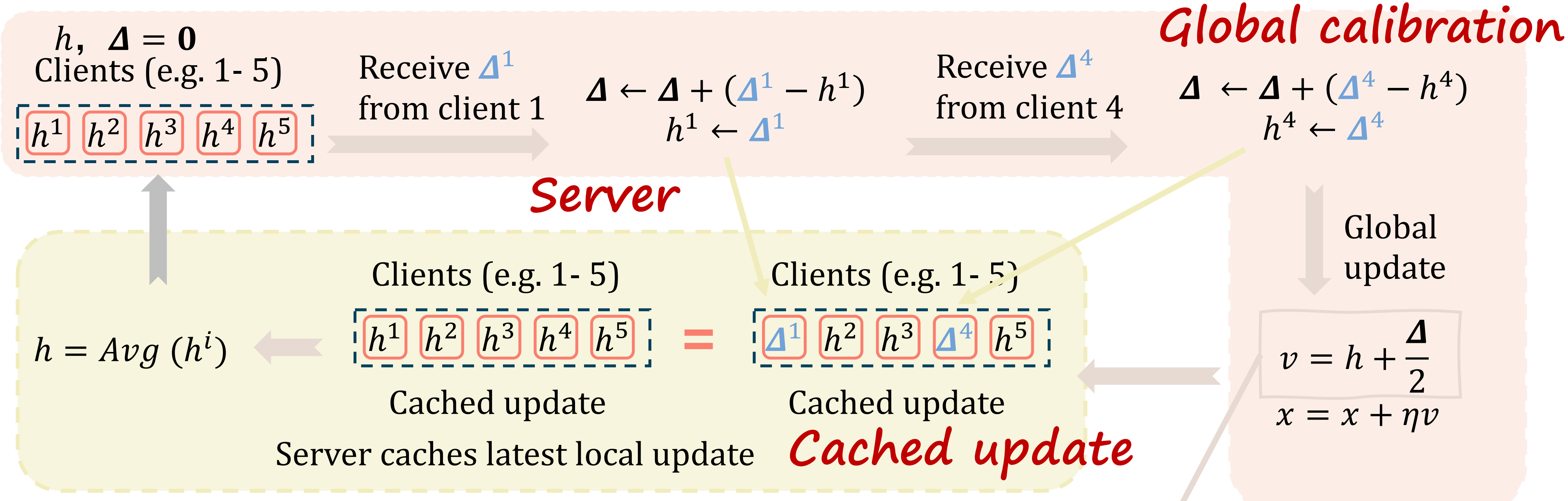
We want some help ("cached variable") from other clients, even they don't participate



Although x^3 is computed from a very outdated model, the cached update direction h^2, h^4, h^5 can help calibrate the update direction

Cache-Aided Asynchronous FL (CA^{2}FL)

How to appropriately use cache variables?



$\boldsymbol{\Delta}$: global model difference, Δ^i: local update from client i,

Asynchronous: $v = \frac{\Delta_1 + \Delta_4}{2}$ $\quad\Longleftrightarrow\quad$ $\boxed{v = \frac{h_1 + h_2 + h_3 + h_4 + h_5}{5} + \frac{\boldsymbol{\Delta_1} + \boldsymbol{\Delta_4}}{2} - \frac{h_1 + h_4}{2}}$

Investigate the Convergence of Async. FL (FedBuff)

The convergence rate for FedBuff is

τ_{max}: maximum delay
τ_{avg}: average delay

$$O\left(\frac{\sqrt{K}}{\sqrt{TM}}\sigma_g^2 + \frac{1}{\sqrt{TKM}} + \frac{K\tau_{max}\tau_{avg}\sigma_g^2 + \tau_{max}\sigma^2}{T}\right),$$

T: total global rounds, K: #n of local updates, N: #n total clients, M: #n participated clients

The convergence degradation brought by the asynchronous delay τ is amplified by the high data heterogeneity (large σ_g^2)

Investigate the Convergence of CA^{2}FL

ζ_{max}: maximum difference between cached step and current step

The convergence rate for CA^{2}FL is

$$O\left(\frac{1}{\sqrt{TKM}} + \frac{(\tau_{max}+\zeta_{max})\sigma^2}{T}\right),$$

T: total global rounds, K: #n of local updates, M: #n participated clients

Comparing with the convergence of FedBuff

$$O\left(\frac{\sqrt{K}}{\sqrt{TM}}\sigma_g^2 + \frac{1}{\sqrt{TKM}} + \frac{K\tau_{max}\tau_{avg}\sigma_g^2+\tau_{max}\sigma^2}{T}\right)$$

Merged with smaller order terms

Eliminate this term

Experiments

Experiments on image classification and language understanding

Method	Dir(0.3)		Dir(0.1)	
	CNN Acc. & std	ResNet-18 Acc. & std	CNN Acc. & std	ResNet-18 Acc. & std
FedAsync	62.29 ± 0.16	79.8 ± 2.28	-	40.58 ± 2.92
FedBuff	60.74 ± 1.18	78.53 ± 3.31	53.96 ± 0.10	63.03 ± 3.17
CA^{2}FL	**64.40** ± 0.32	**83.79** ± 0.34	**57.62** ± 0.42	**68.37** ± 1.97

Method	MRPC Acc. & std.	SST-2 Acc. & std.	RTE Acc. & std.	CoLA Acc. & std.
FedAsync	**82.86** ± 0.42	87.32 ± 3.76	62.09 ± 0.76	54.53 ± 1.52
FedBuff	78.68 ± 0.41	86.06 ± 3.86	60.07 ± 1.09	55.57 ± 0.94
CA^{2}FL	79.26 ± 0.12	**90.76** ± 1.02	**65.63** ± 0.35	**56.10** ± 0.25

Experiments

	Acc.	FedAsync	FedBuff	CA^{2}FL	FedAvg
CIFAR-10	80%	268.80	291.53	**214.16**	388.64
CIFAR-100	55%	333.47	295.49	**233.49**	476.78
MRPC	80%	2549.54	403.95	**87.39**	97.71
SST-2	90%	2853.5	2079.35	648.71	**572.01**
RTE	63%	815.94	420.83	**79.61**	95.17
CoLA	55%	217.23	144.64	34.75	**0.79**

Matthew's correlation for CoLA

The proposed CA^{2}FL shows advantage in training efficiency

Thank You