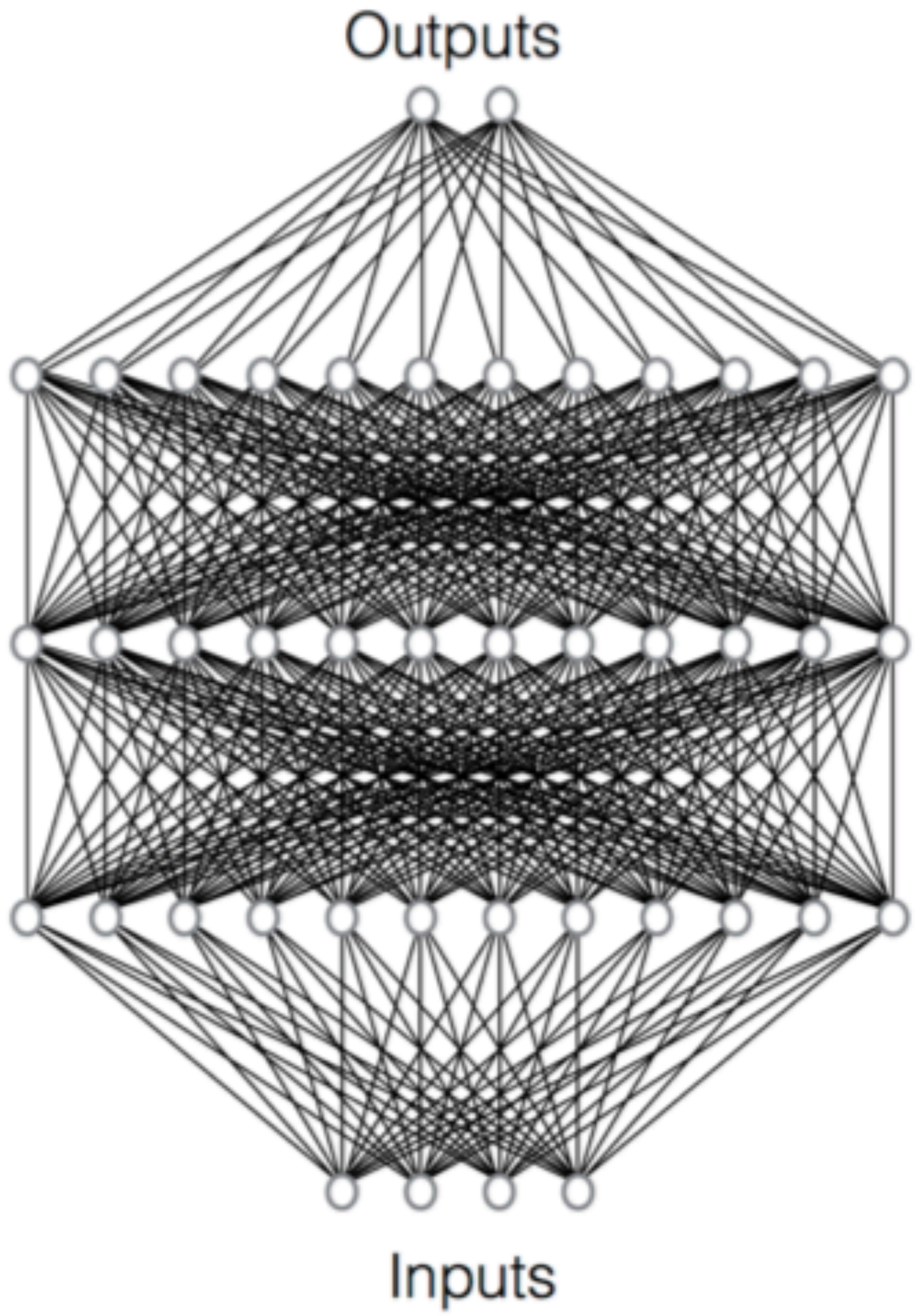
Efficient Distributed Learning via Independent Subnet Training: Results and Trends

Anastasios Kyrillidis
Rice CS

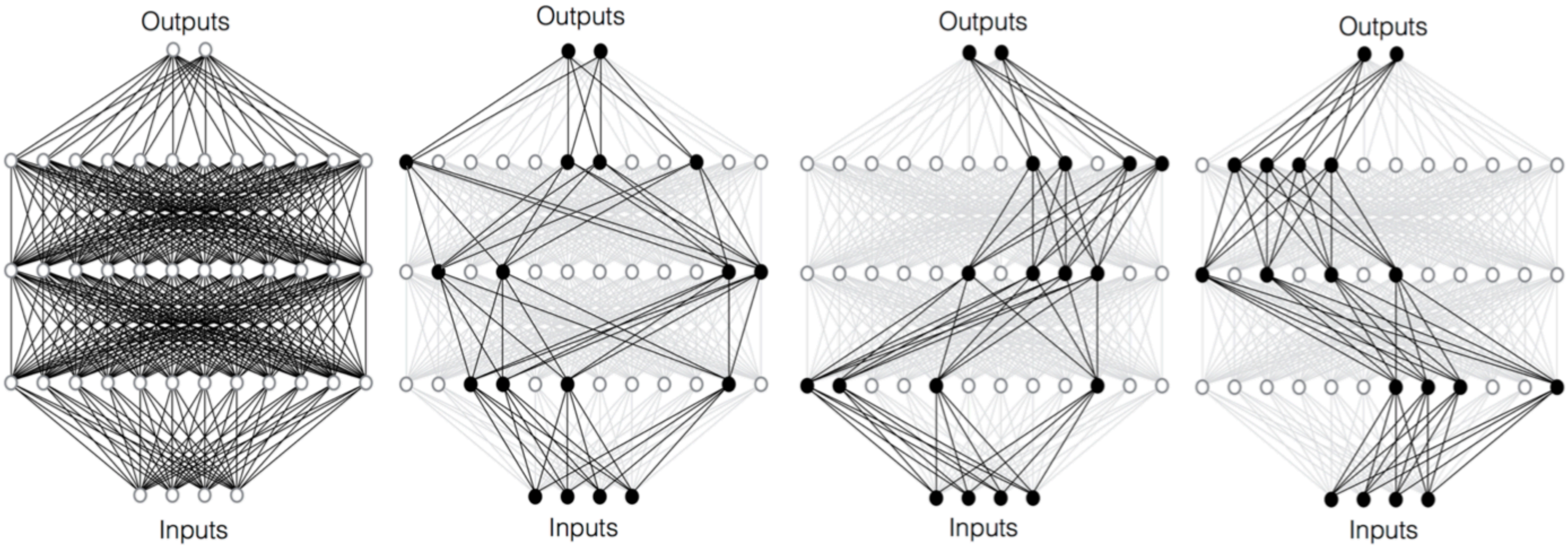
Joint work with: Binhang Yuan, Chen Dun, Cameron Wolfe, Fangshuo Liao,
Qihan Wang, Yuxin Tang, Erdong Hu, Jingkang Yang,
Santiago Segarra, Dimitris Dimitriadis, Chris Jermaine

Independent Subnet Training: NN decomposition



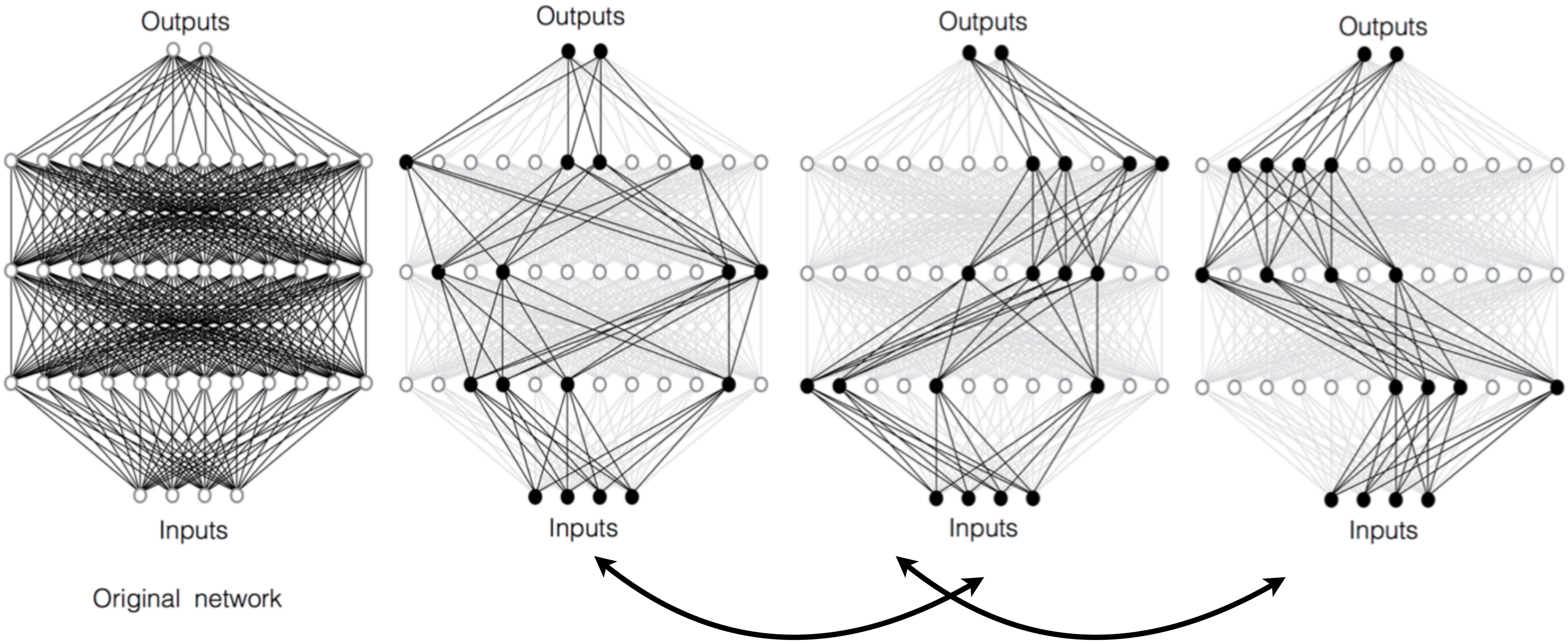
Original network

Independent Subnet Training: NN decomposition



Original network

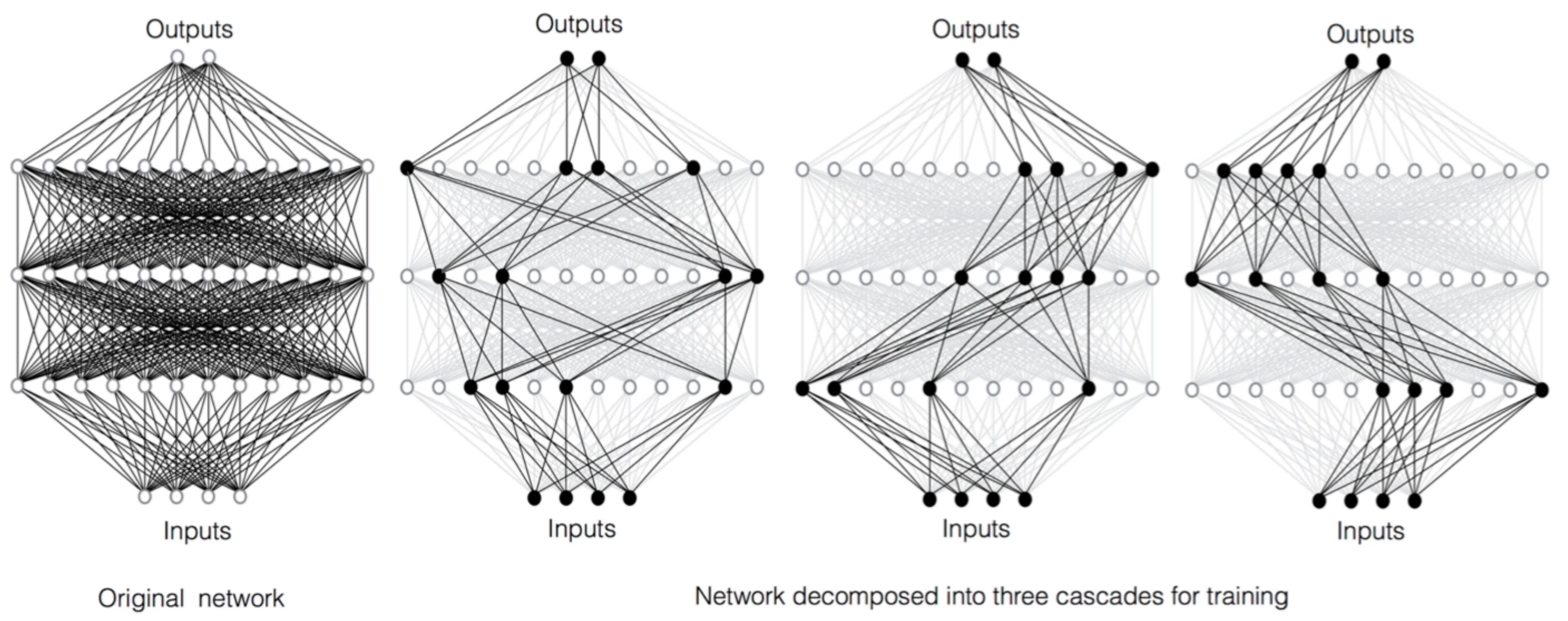
Independent Subnet Training: NN decomposition



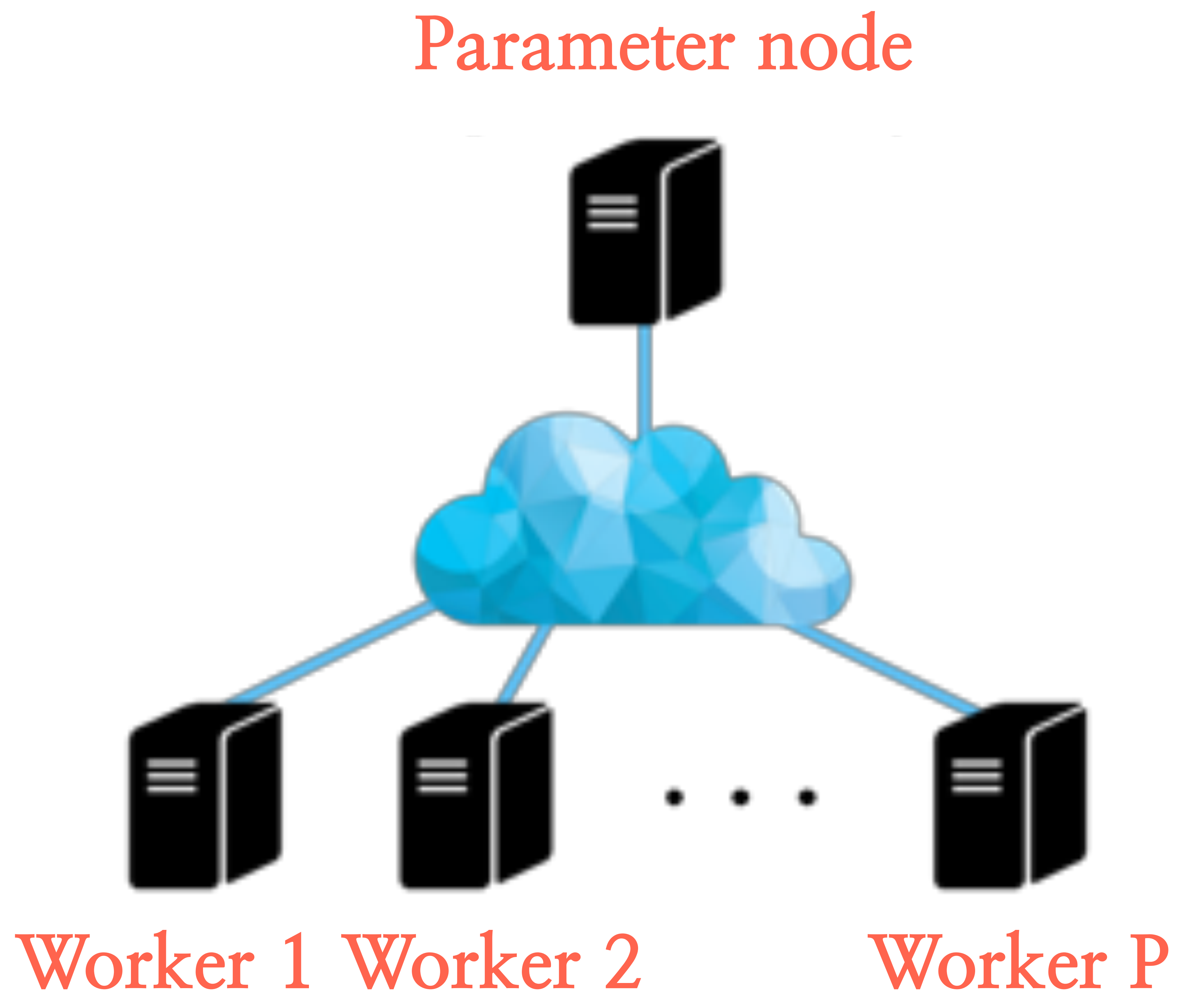
Union of neurons make original network
(Note: union of parameters do not make original network necessarily)

Independent Subnet Training: NN distr. training

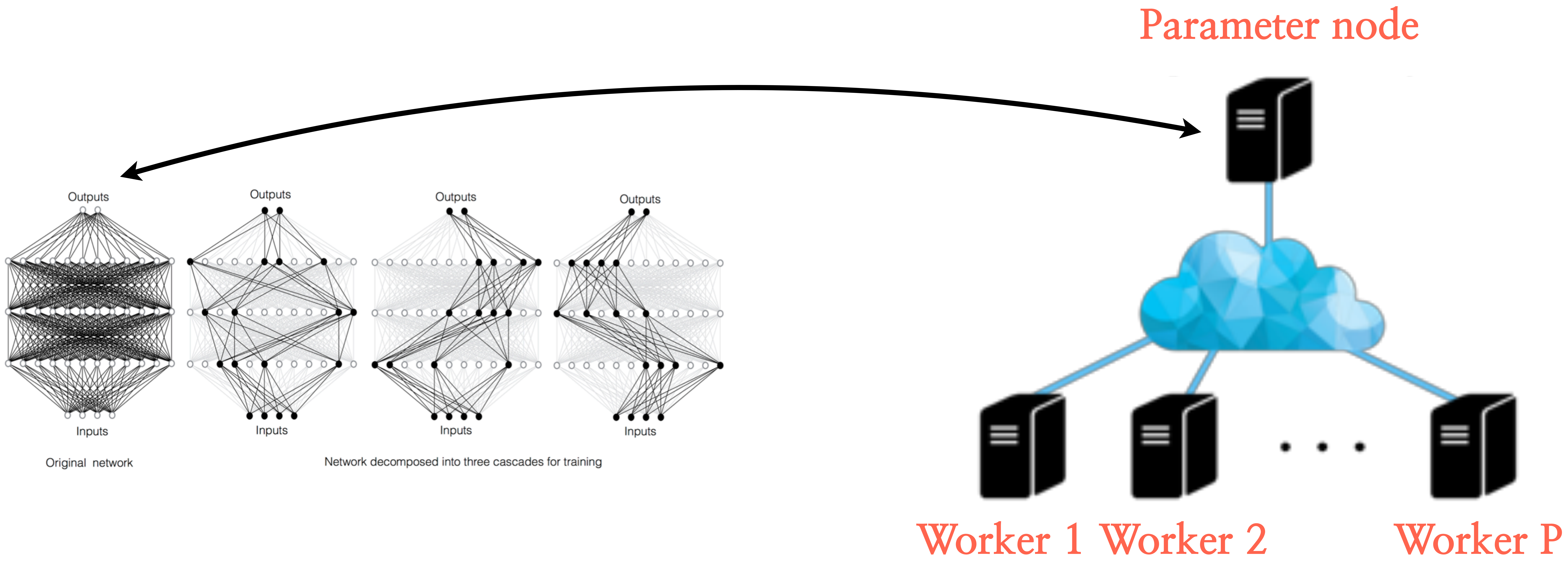
How to decompose a NN:



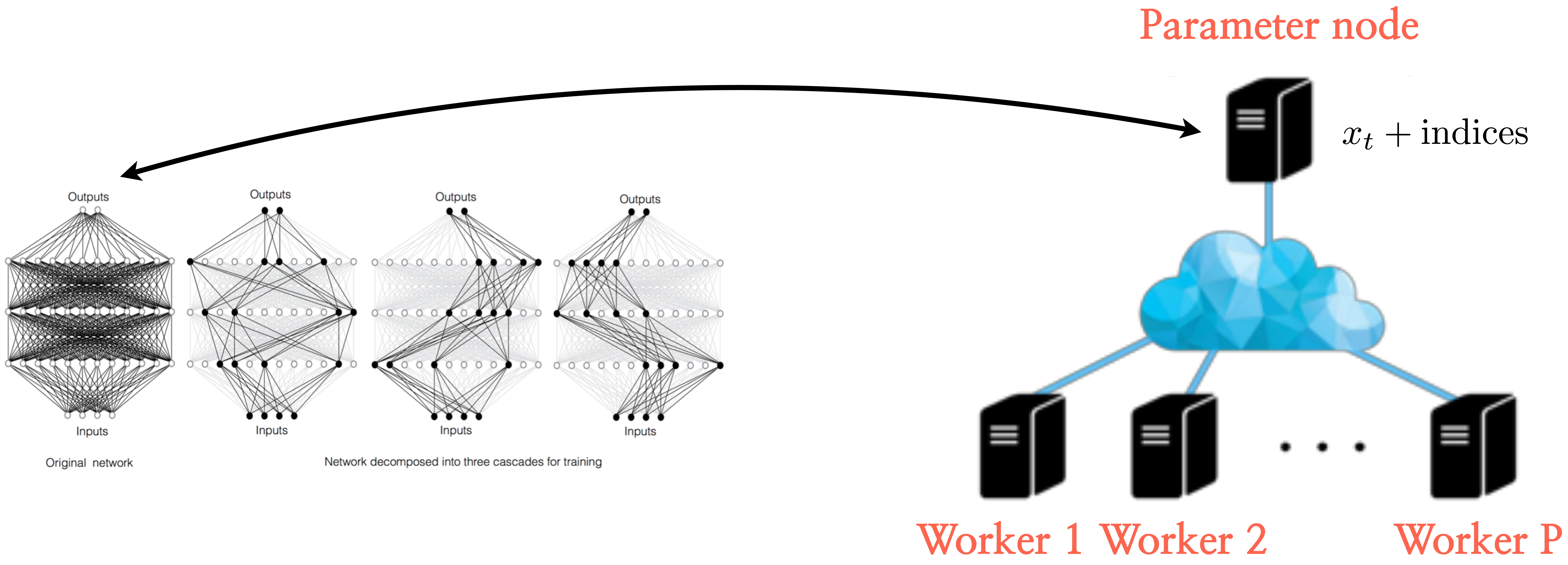
How to train NN in a distributed fashion:



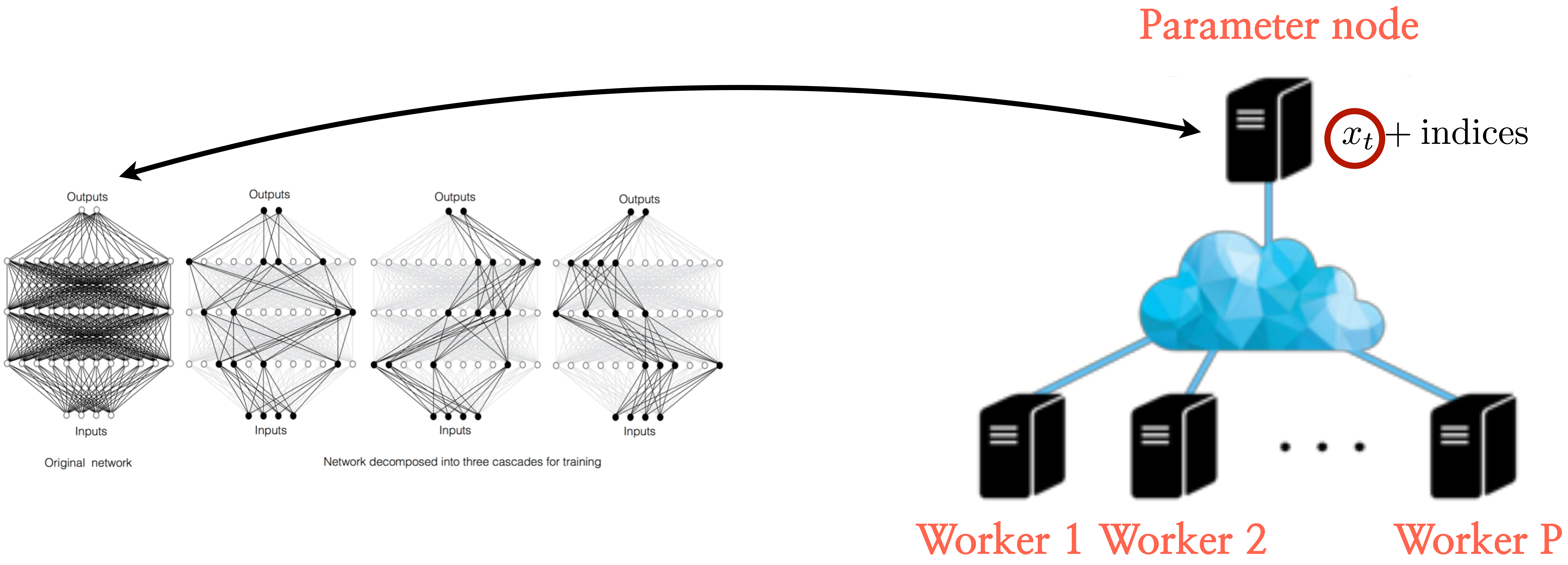
Independent Subnet Training: NN distr. training



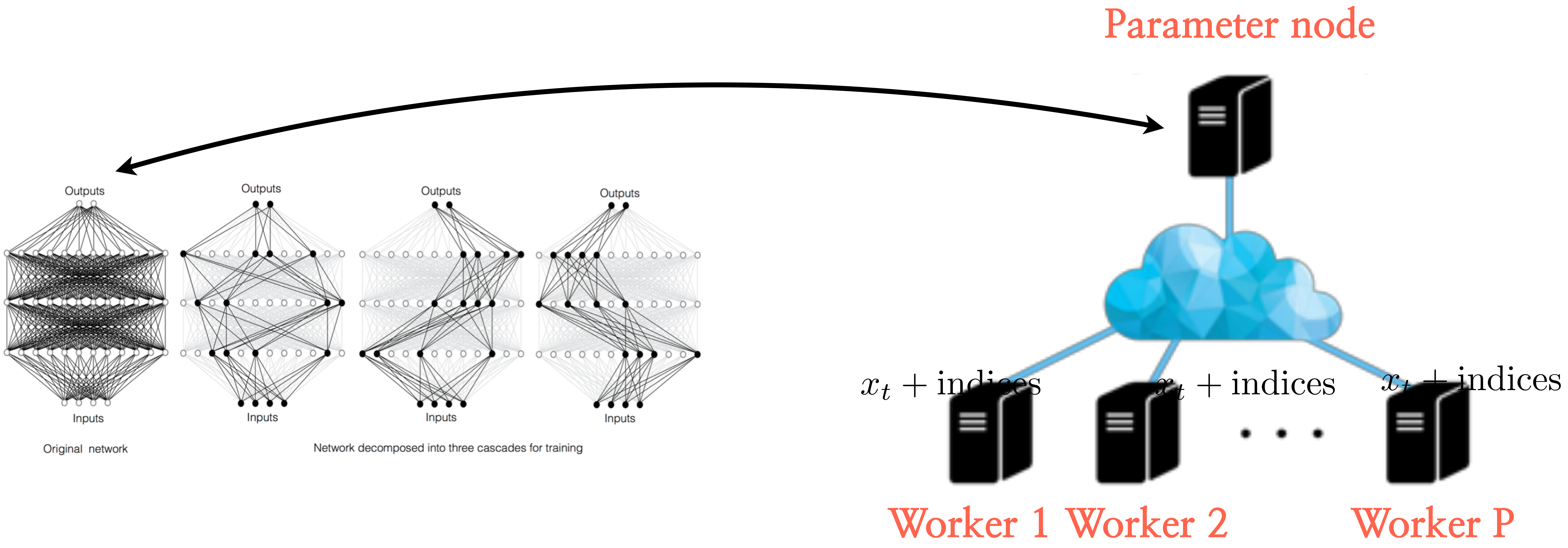
Independent Subnet Training: NN distr. training



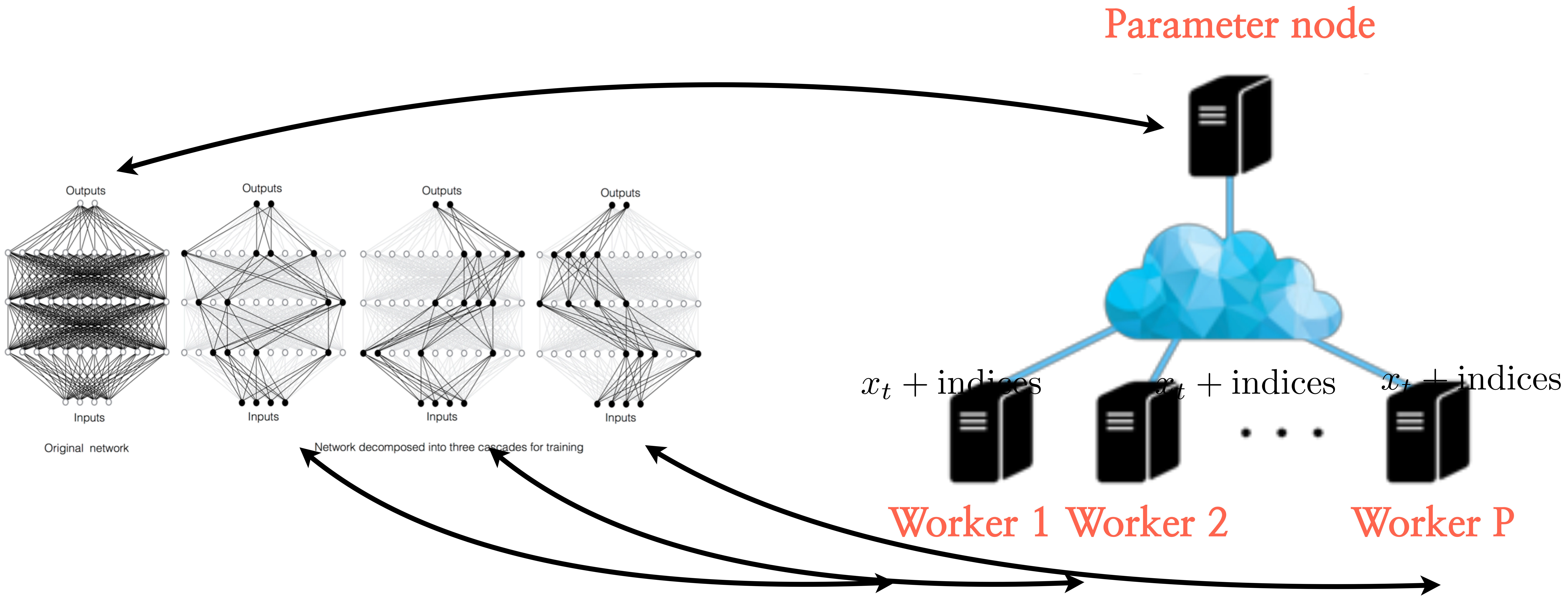
Independent Subnet Training: NN distr. training



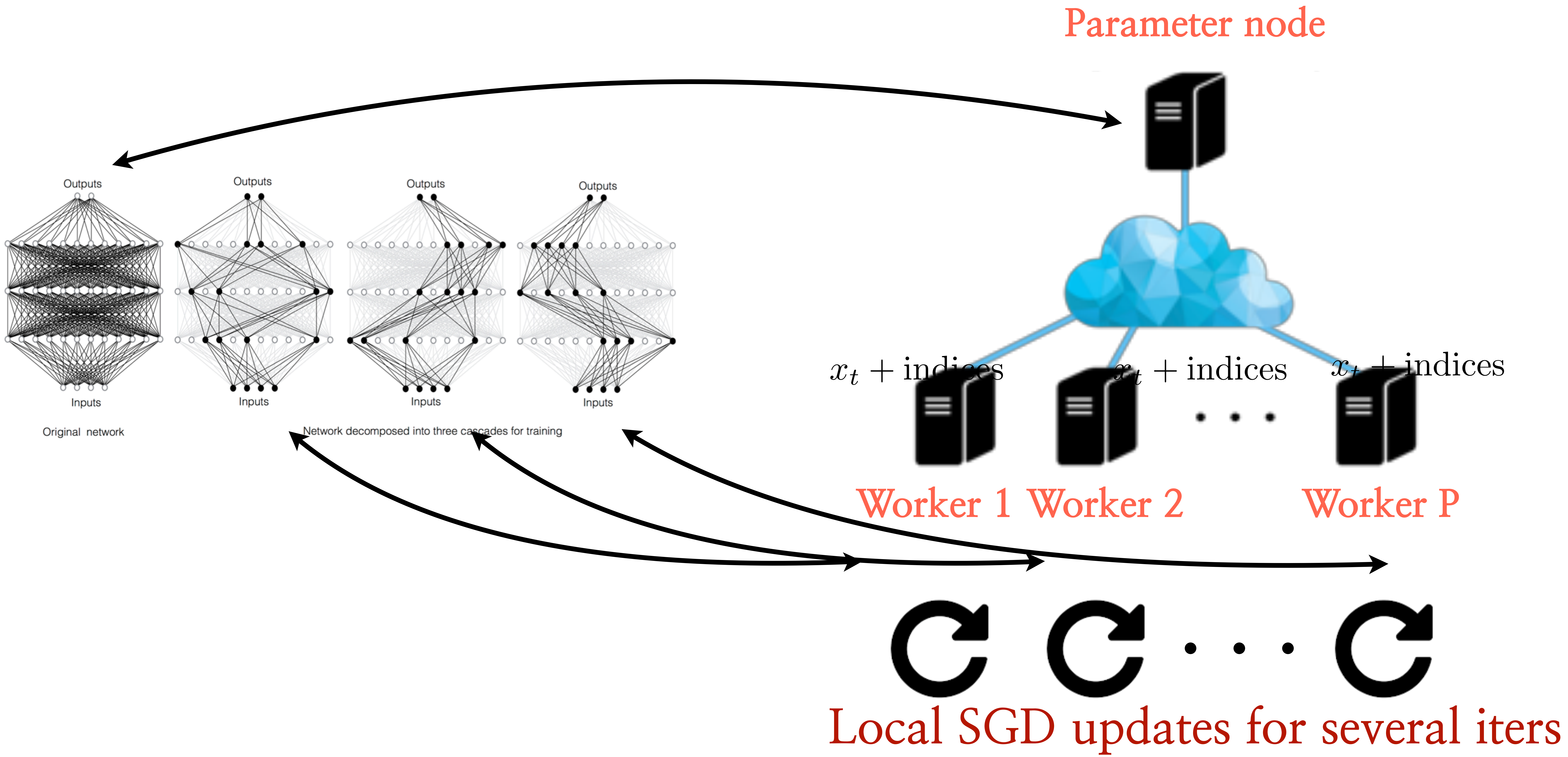
Independent Subnet Training: NN distr. training



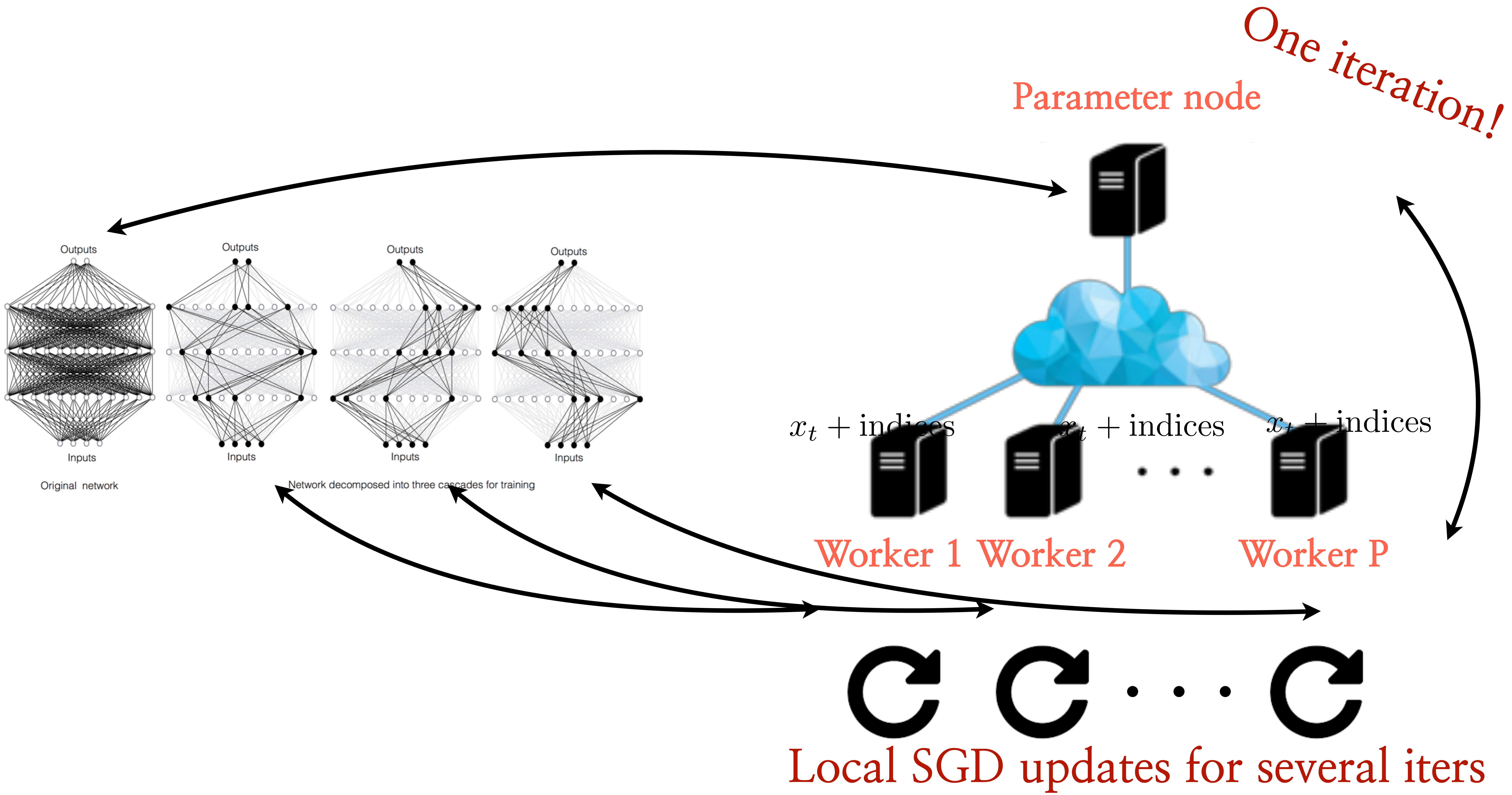
Independent Subnet Training: NN distr. training



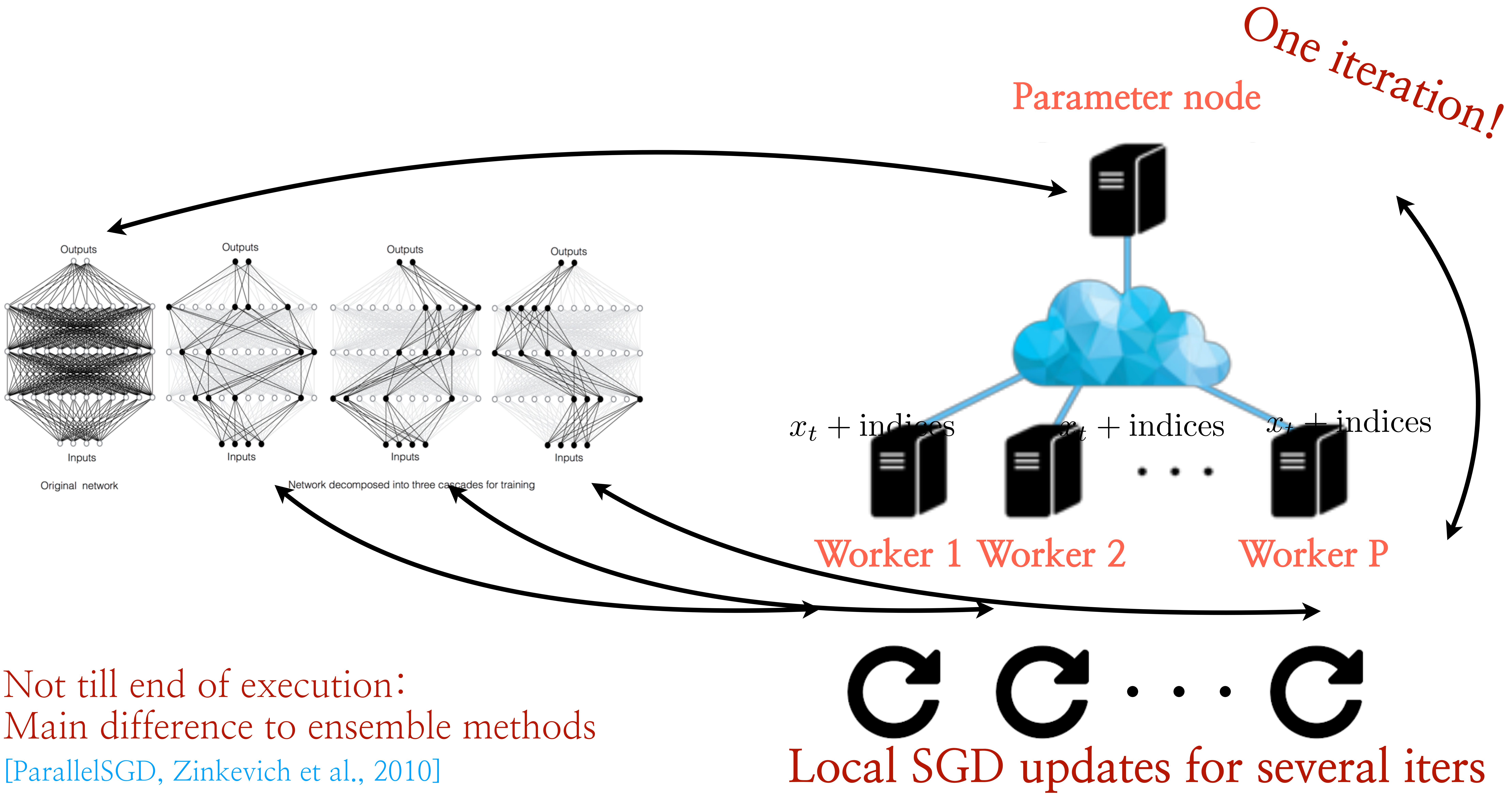
Independent Subnet Training: NN distr. training



Independent Subnet Training: NN distr. training



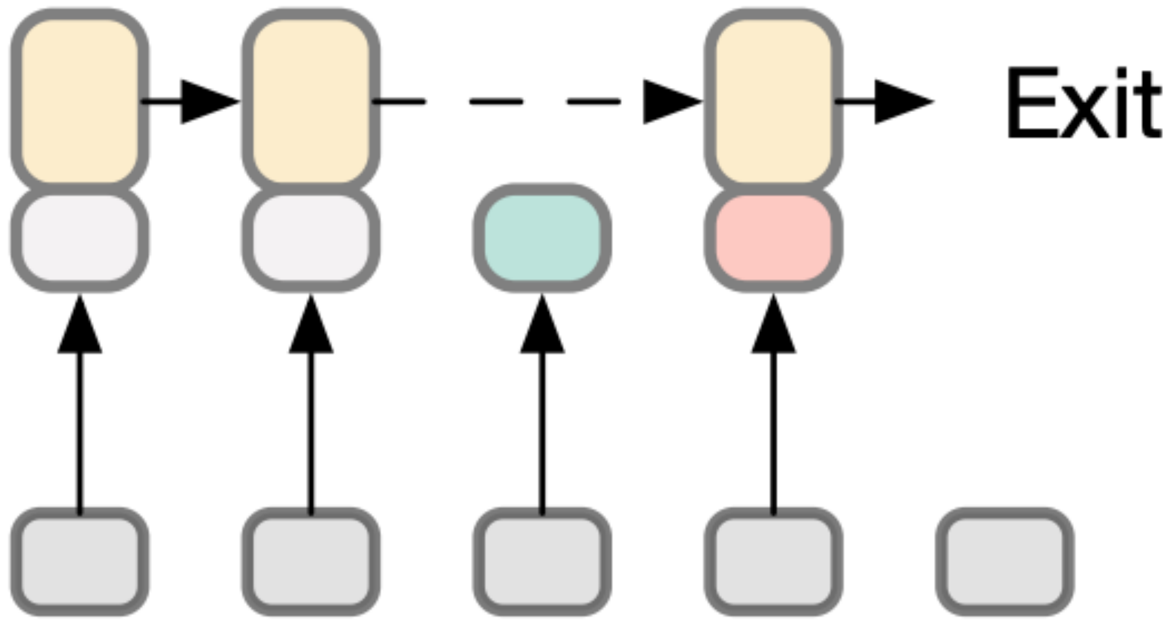
Independent Subnet Training: NN distr. training



Not till end of execution:
Main difference to ensemble methods
[ParallelSGD, Zinkevich et al., 2010]

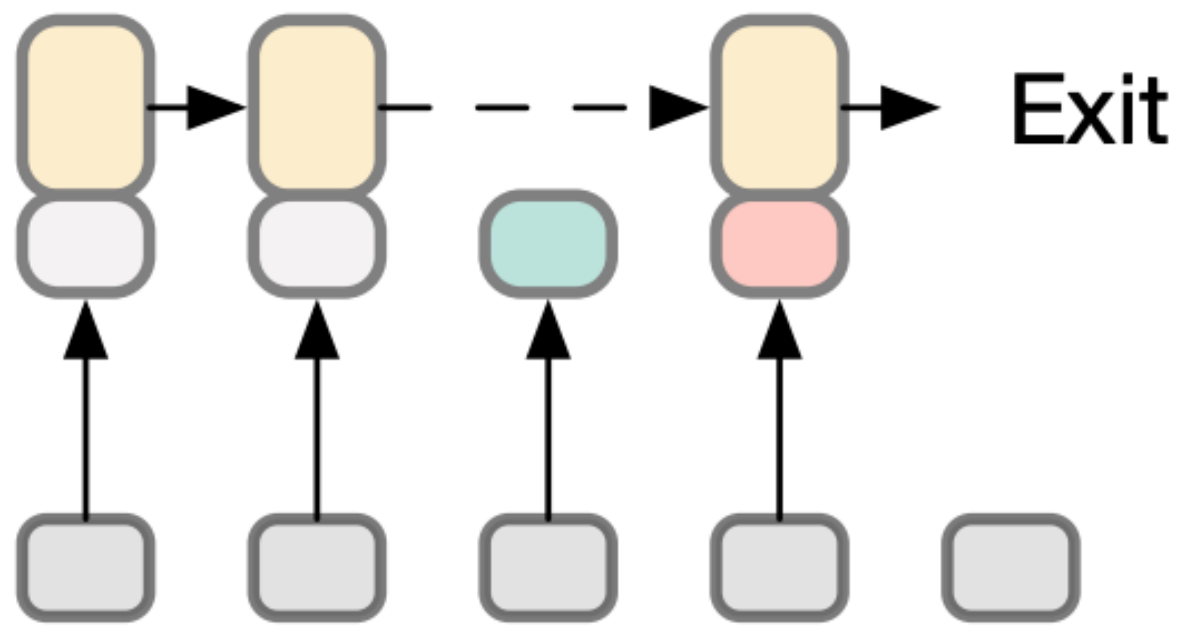
Algorithms: Dynamic/Approximate Training

- Skimming: dynamically allocating computation to different time steps, based on the input tokens
  [Huang et al., 2016; Yu et al., 2017; Campos et al., 2018; Li et al., 2019]

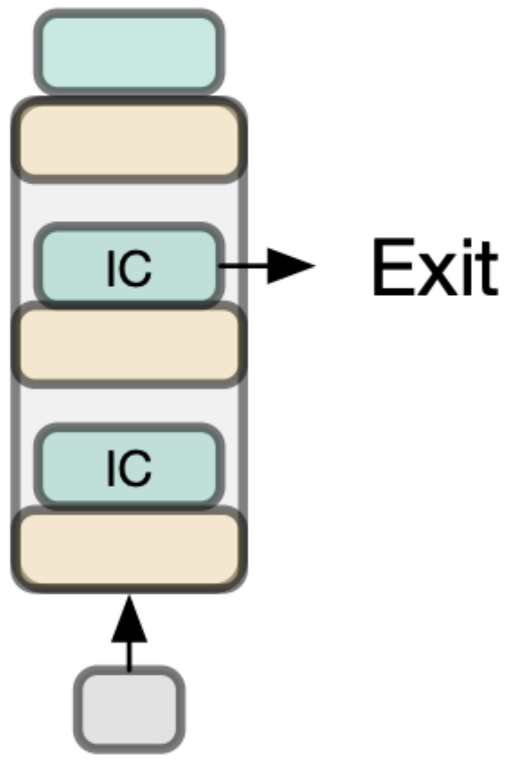


Algorithms: Dynamic/Approximate Training

- Skimming: dynamically allocating computation to different time steps, based on the input tokens
  [Huang et al., 2016; Yu et al., 2017; Campos et al., 2018; Li et al., 2019]

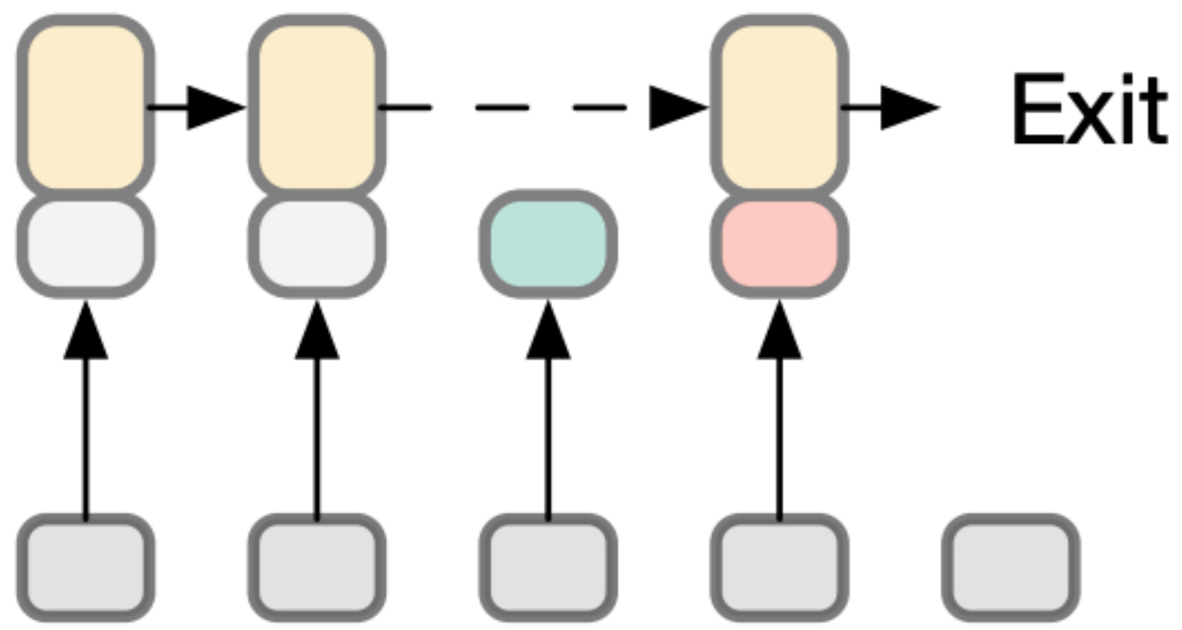


- Early exit/local objectives per layer
  [Kaya et al., 2019; Zhou et al., 2020; Xin et al., 2020, 2021; Sun et al., 2021]

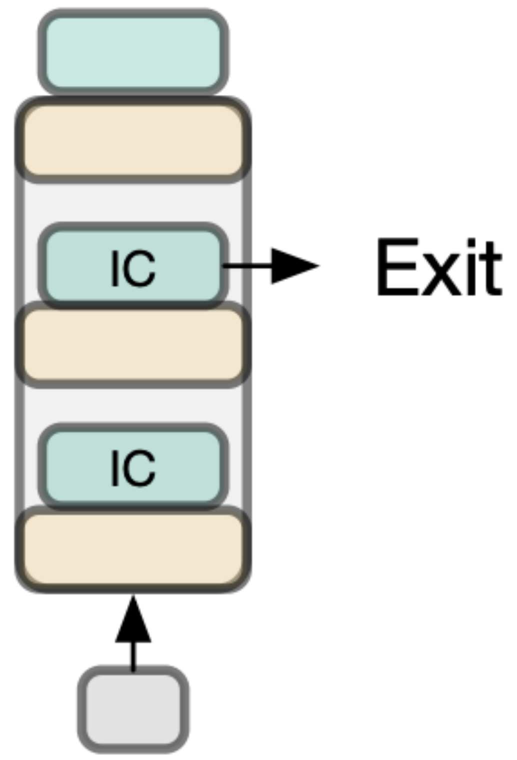


Algorithms: Dynamic/Approximate Training

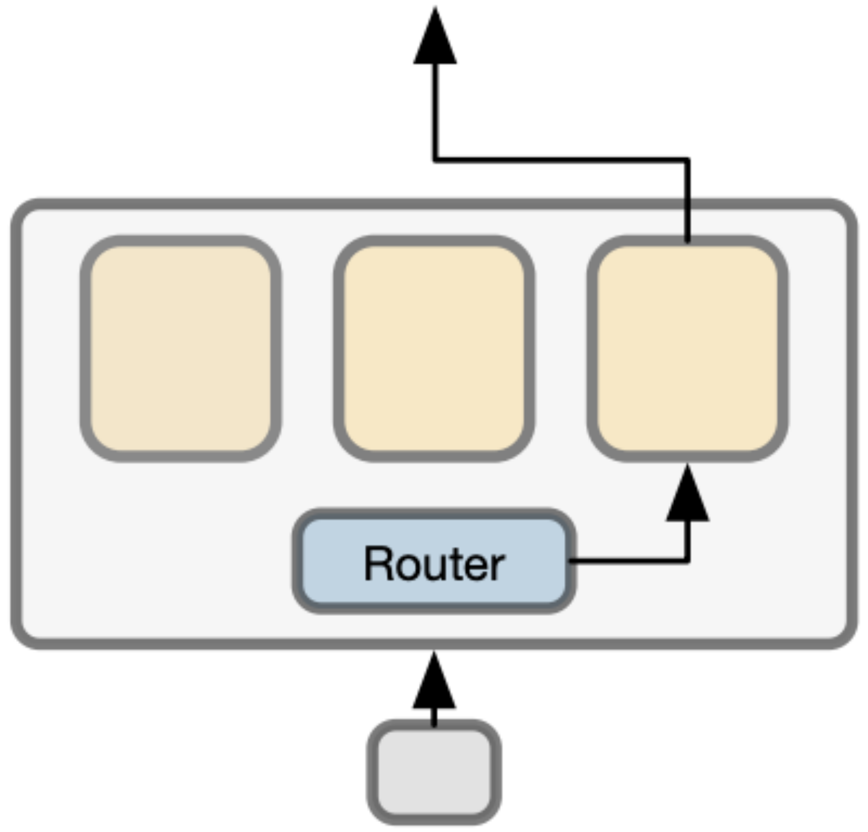
- Skimming: dynamically allocating computation to different time steps, based on the input tokens
  [Huang et al., 2016; Yu et al., 2017; Campos et al., 2018; Li et al., 2019]



- Early exit/local objectives per layer
  [Kaya et al., 2019; Zhou et al., 2020; Xin et al., 2020, 2021; Sun et al., 2021]

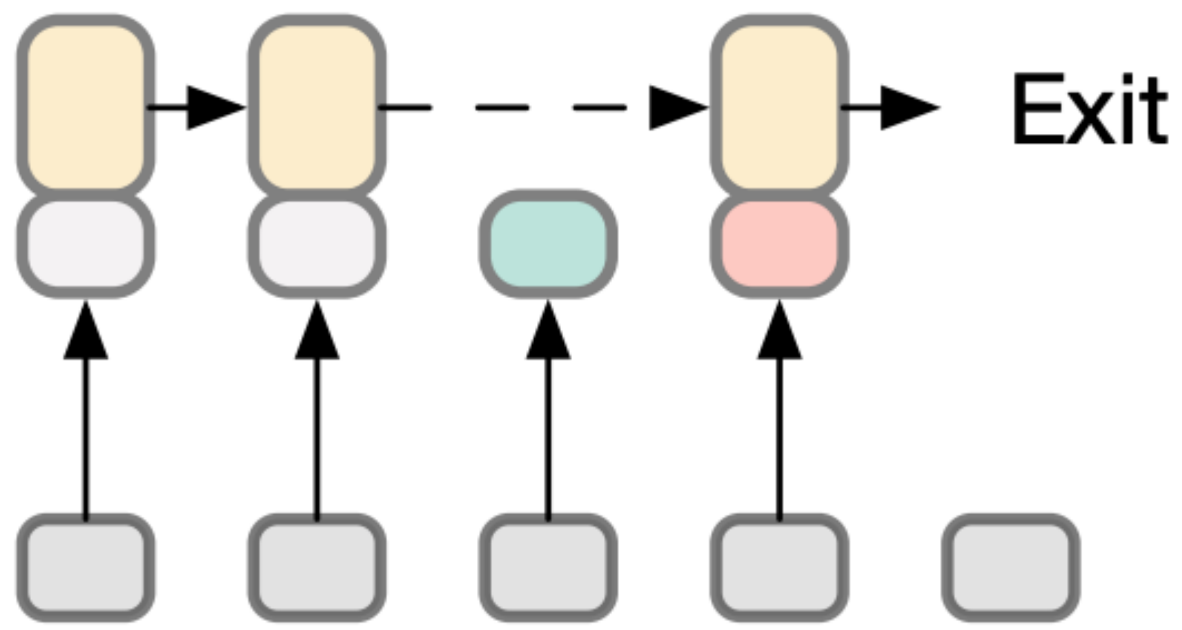


- Sparsely activated MoEs: a layer typically contains multiple sub-networks (i.e., "experts") —> Structured Dropout.
  [Jacobs et al., 1991; Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2021, Lewis et al. 2021]
  - Examples: Sparsely, GShard, Switch, BASE, DTS, Hash, THOR, etc.

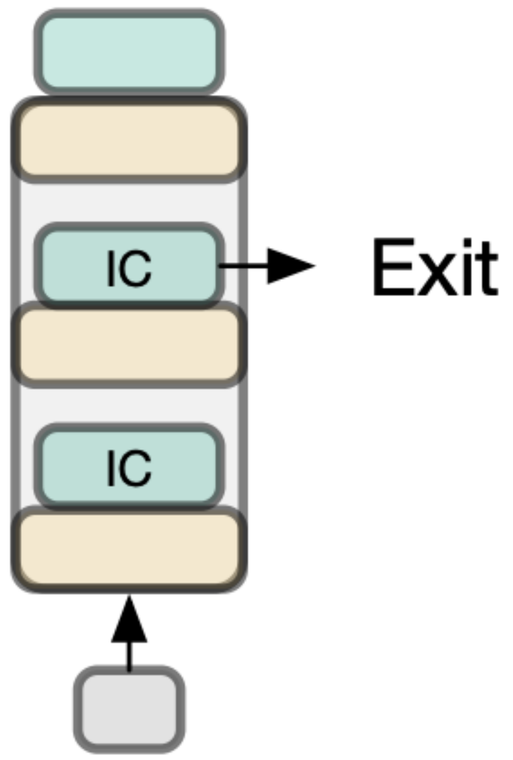


Algorithms: Dynamic/Approximate Training

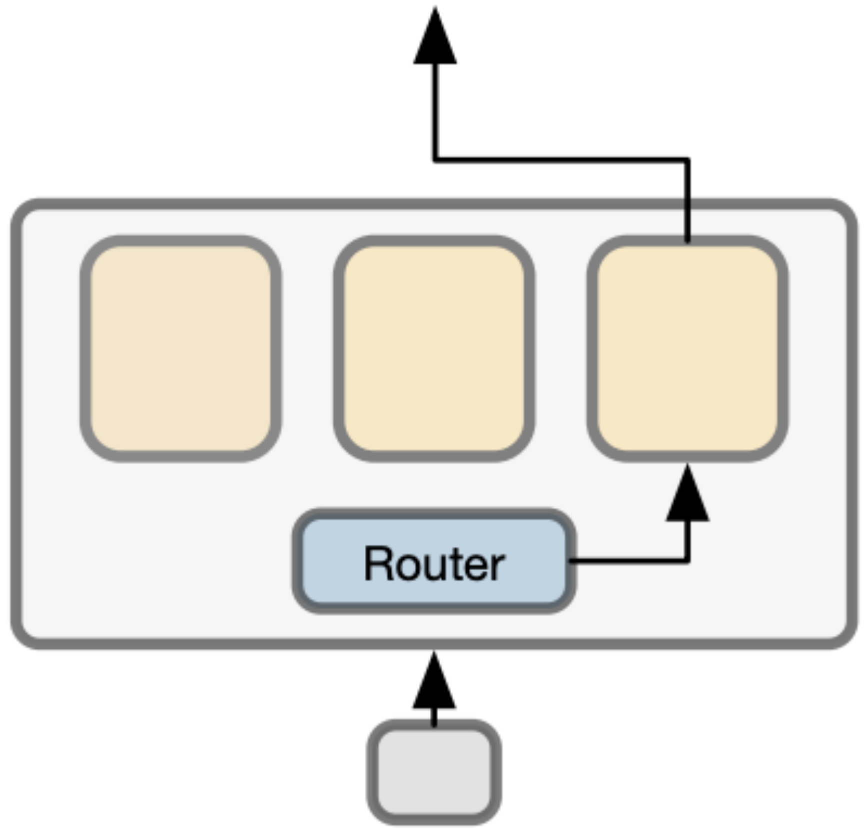
- Skimming: dynamically allocating computation to different time steps, based on the input tokens
  [Huang et al., 2016; Yu et al., 2017; Campos et al., 2018; Li et al., 2019]



- Early exit/local objectives per layer
  [Kaya et al., 2019; Zhou et al., 2020; Xin et al., 2020, 2021; Sun et al., 2021]



- Sparsely activated MoEs: a layer typically contains multiple sub-networks (i.e., “experts”) —> Structured Dropout.
  [Jacobs et al., 1991; Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2021, Lewis et al. 2021]
  - Examples: Sparsely, GShard, Switch, BASE, DTS, Hash, THOR, etc.



- IST: unstructured distributed dropout

Systems: Dynamic/Approximate Distributed Training

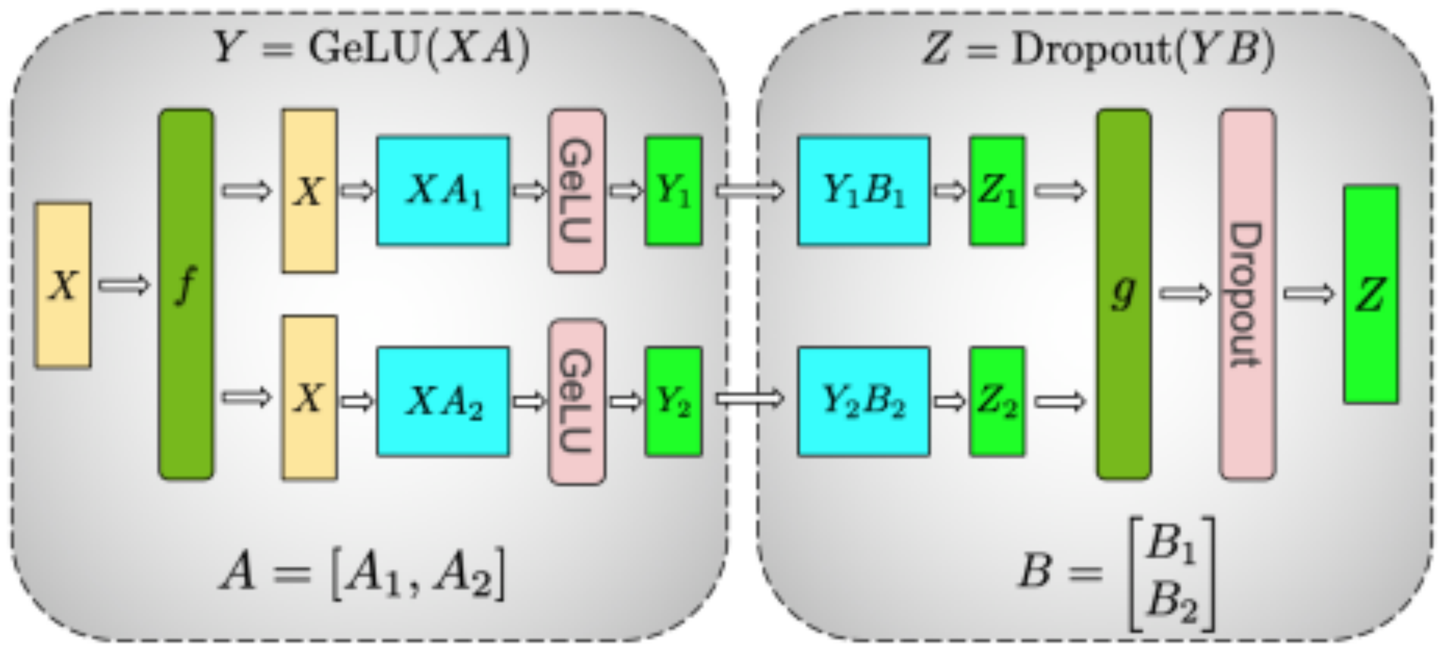
- Classical approaches: data parallel, model parallel
  [Zinkevich et al., 2010; Stich, 2019; Huang et al. 2019]

Systems: Dynamic/Approximate Distributed Training

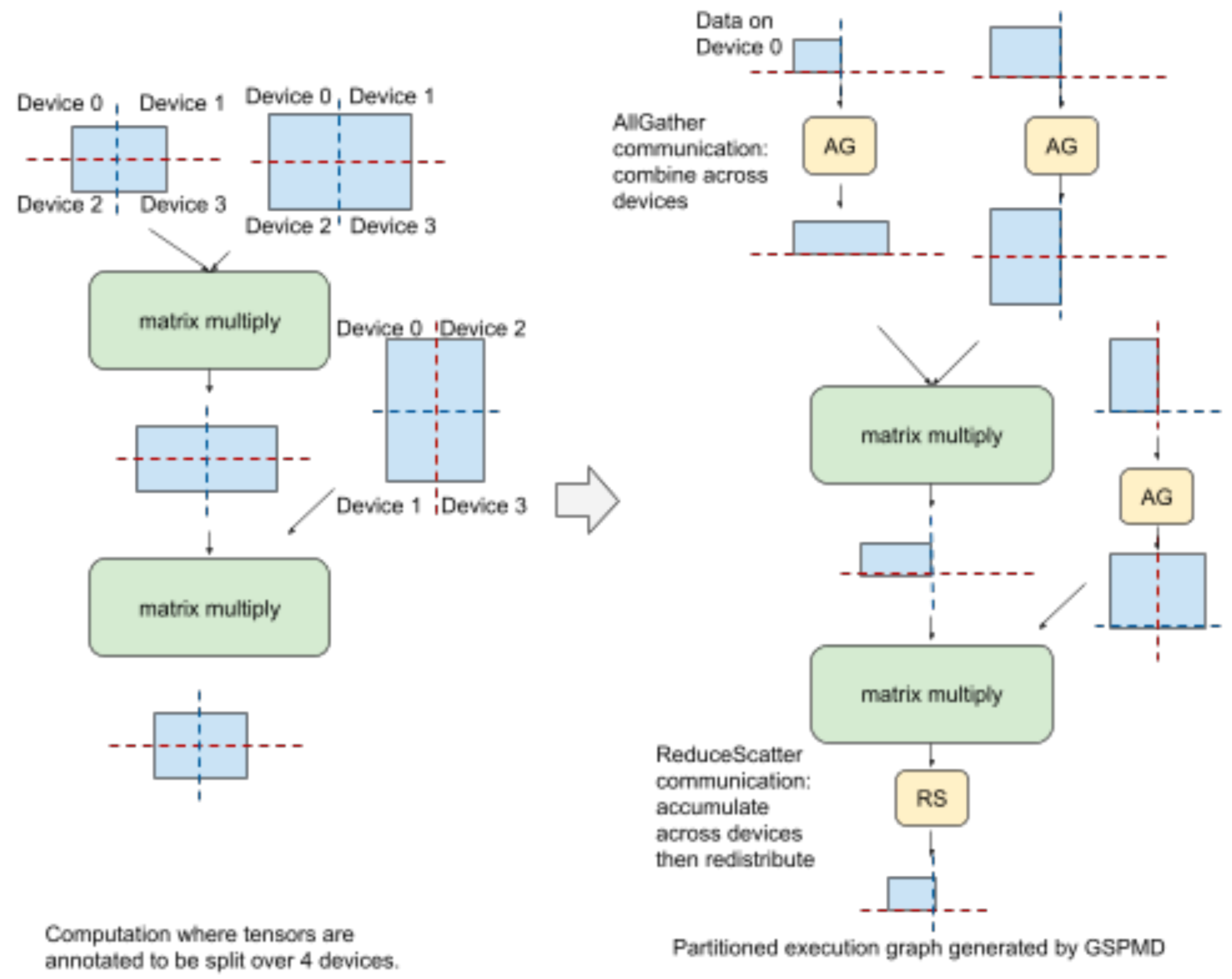
- Classical approaches: data parallel, model parallel
  [Zinkevich et al., 2010; Stich, 2019; Huang et al. 2019]
- More recent approaches: Model tensor parallelism
  3D parallelism
  [Shoeybi et al., 2019; Narayanan et al., 2021; Rae et al., 2021; Rasley et al., 2020]

Systems: Dynamic/Approximate Distributed Training

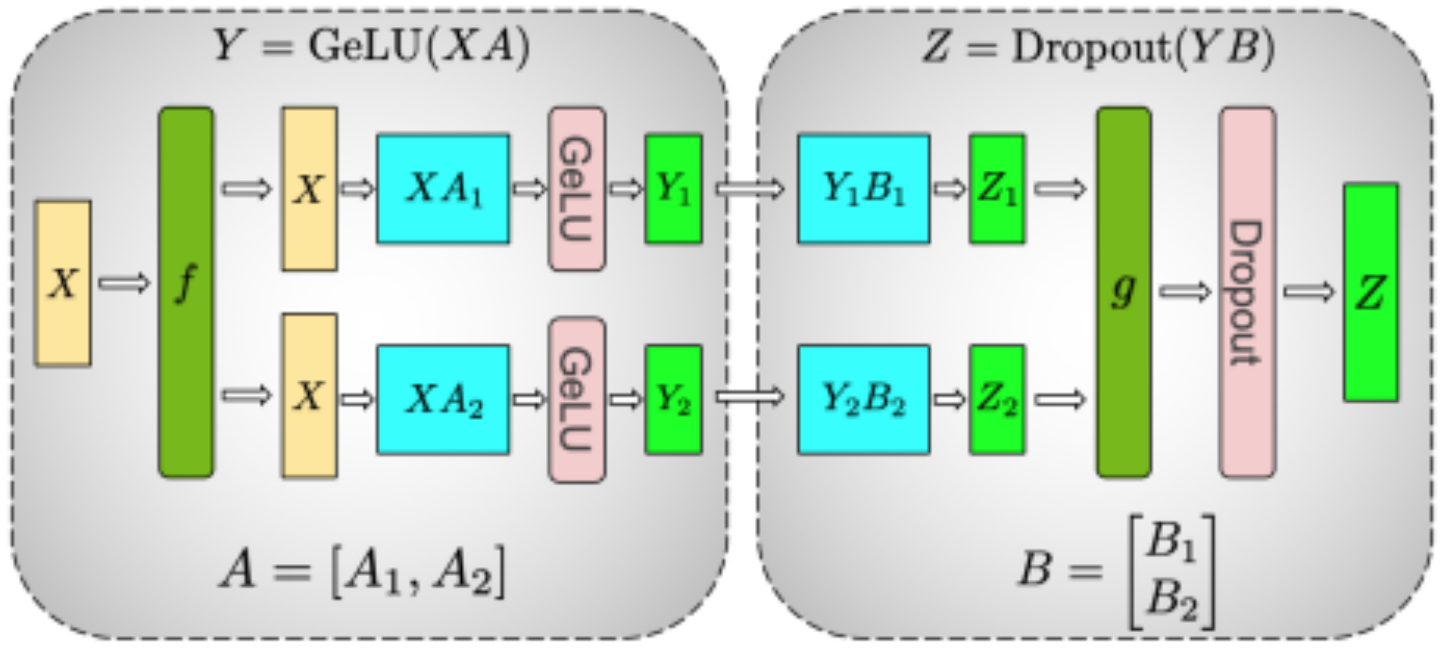
- Classical approaches: data parallel, model parallel
  [Zinkevich et al., 2010; Stich, 2019; Huang et al. 2019]
- More recent approaches: Model tensor parallelism
  3D parallelism
  [Shoeybi et al., 2019; Narayanan et al., 2021; Rae et al., 2021; Rasley et al., 2020]



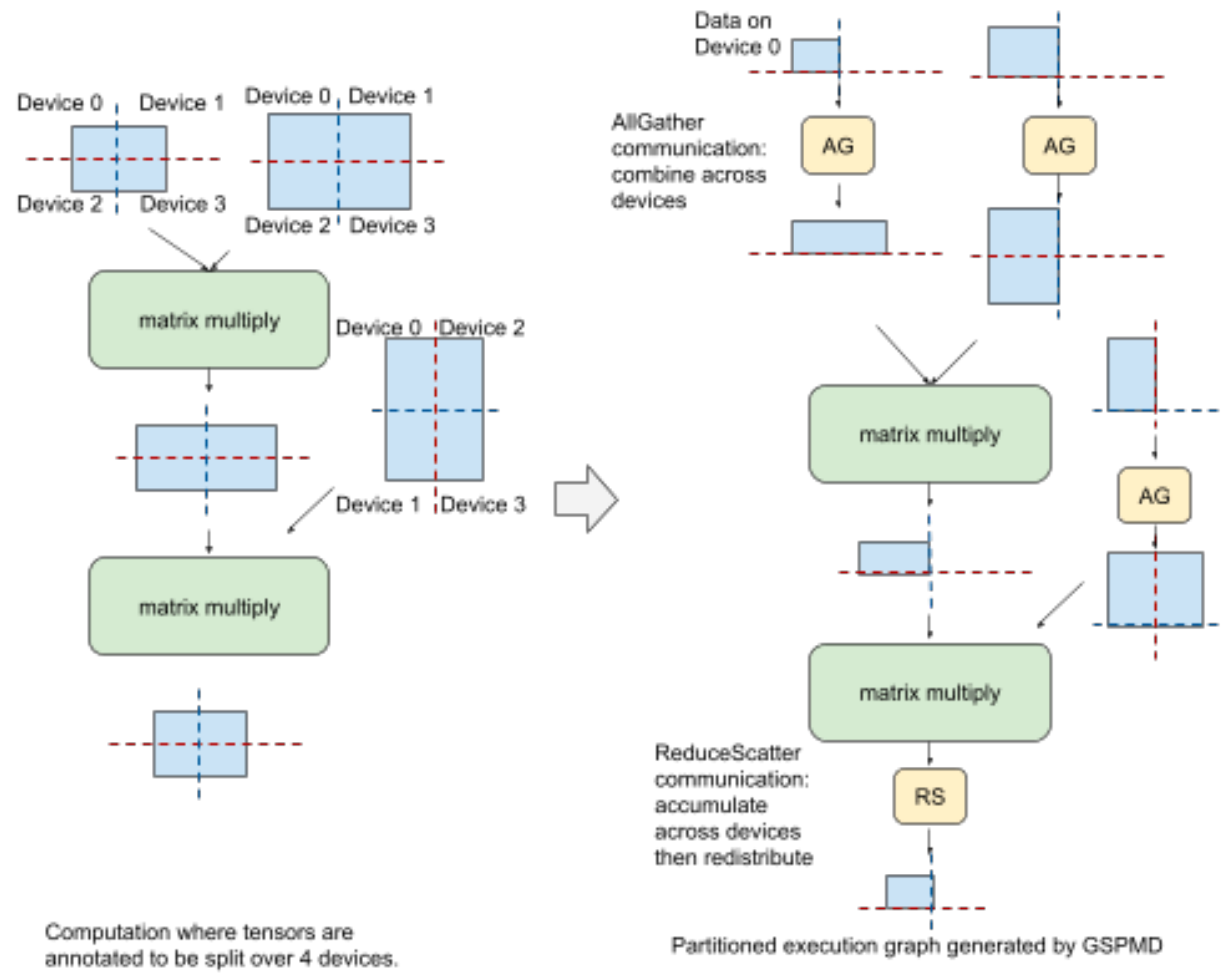
- Some notable implementations:
  - Megatron
  - Amazon SageMaker Model Parallelism
  - Zero-Infinity/DeepSpeed
  - Google's GSPMD



Systems: Dynamic/Approximate Distributed Training



- Classical approaches: data parallel, model parallel
  [Zinkevich et al., 2010; Stich, 2019; Huang et al. 2019]
- More recent approaches: Model tensor parallelism
  3D parallelism
  [Shoeybi et al., 2019; Narayanan et al., 2021; Rae et al., 2021; Rasley et al., 2020]



- Some notable implementations:
  - Megatron
  - Amazon SageMaker Model Parallelism
  - Zero-Infinity/DeepSpeed
  - Google's GSPMD
- IST: End-to-end approximate model tensor/3D parallelism

Published at VLDB 2022

Distributed Learning of Neural Networks using Independent Subnet Training

Binhang Yuan
Rice University
by8@rice.edu

Cameron R. Wolfe
Rice University
crw13@rice.edu

Chen Dun
Rice University
cd46@rice.edu

Yuxin Tang
Rice University
yuxin.tang@rice.edu

Anastasios Kyrillidis
Rice University
anastasios@rice.edu

Chris Jermaine
Rice University
cmj4@rice.edu

https://arxiv.org/pdf/1910.02120.pdf

Published at UAI 2022

ResIST: Layer-Wise Decomposition of ResNets for Distributed Training

Chen Dun * [1] **Cameron R. Wolfe** * [1] **Chris Jermaine** [1] **Anastasios Kyrillidis** [1]

https://arxiv.org/pdf/2107.00961.pdf

ON THE CONVERGENCE OF SHALLOW NEURAL NETWORK TRAINING WITH RANDOMLY MASKED NEURONS

Fangshuo Liao, Anastasios Kyrillidis
Department of Computer Science
Rice University
Houston, TX 77005, USA
{Fangshuo.Liao,anastasios}@rice.edu

https://arxiv.org/pdf/2112.02668.pdf

Published at TMLR

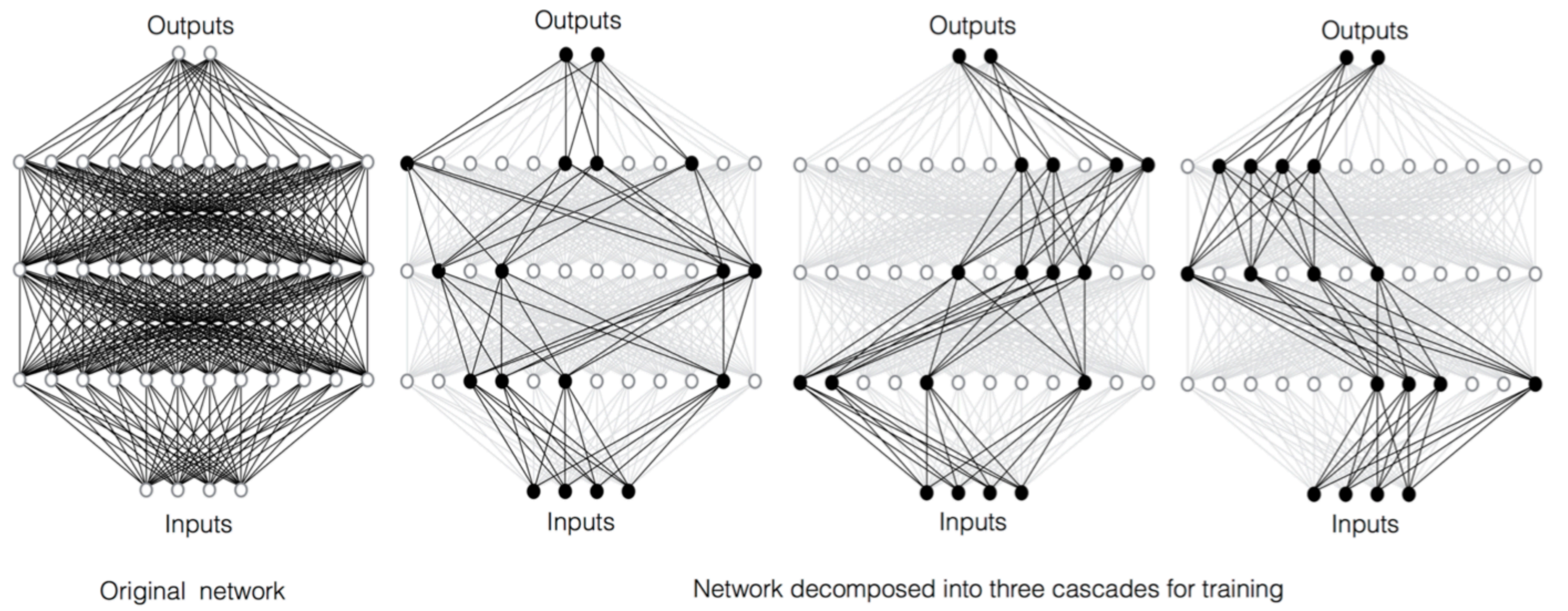
GIST: Distributed Training for Large-Scale Graph Convolutional Networks

Cameron Wolfe * [1] **Jingkang Yang** * [2] **Arindam Chowdhury** [3] **Chen Dun** [1] **Artun Bayer** [3] **Santiago Segarra** [3]
Anastasios Kyrillidis [1]

https://arxiv.org/pdf/2102.10424.pdf

Accepted at Journal of Applied and Computational Topology

Published at VLDB 2022



Published at UAI 2022

ResIST: Layer-Wise Decomposition of ResNets for Distributed Training

Chen Dun [*1] **Cameron R. Wolfe** [*1] **Chris Jermaine** [1] **Anastasios Kyrillidis** [1]

https://arxiv.org/pdf/2107.00961.pdf

On the Convergence of Shallow Neural Network Training with Randomly Masked Neurons

Fangshuo Liao, Anastasios Kyrillidis
Department of Computer Science
Rice University
Houston, TX 77005, USA
{Fangshuo.Liao,anastasios}@rice.edu

https://arxiv.org/pdf/2112.02668.pdf

Published at TMLR

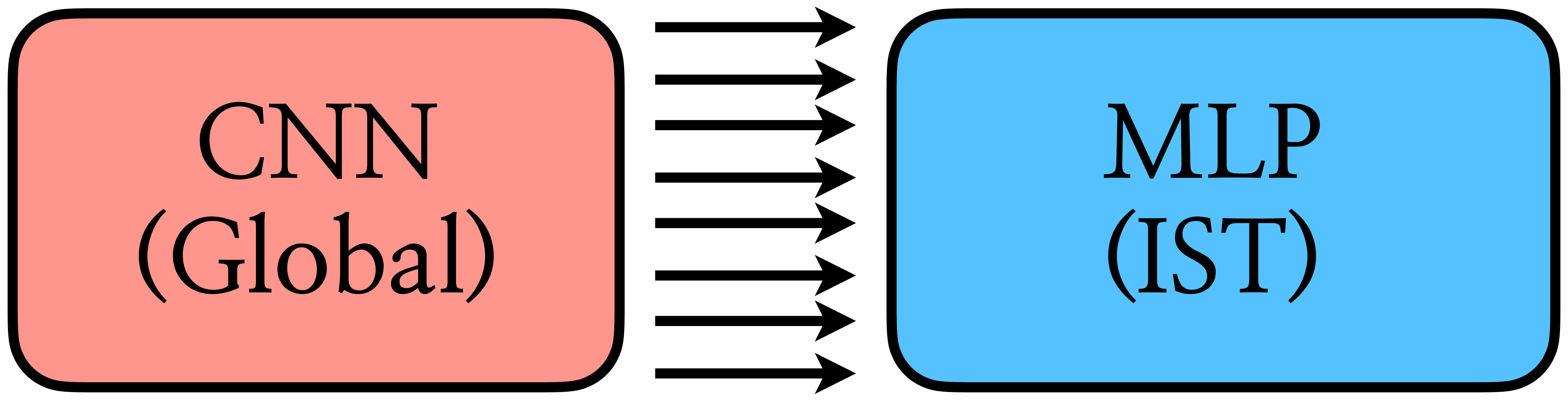
GIST: Distributed Training for Large-Scale Graph Convolutional Networks

Cameron Wolfe [*1] **Jingkang Yang** [*2] **Arindam Chowdhury** [3] **Chen Dun** [1] **Artun Bayer** [3] **Santiago Segarra** [3]
Anastasios Kyrillidis [1]

https://arxiv.org/pdf/2102.10424.pdf

Accepted at Journal of Applied and Computational Topology

Published at VLDB 2022



Published at UAI 2022

ResIST: Layer-Wise Decomposition of ResNets for Distributed Training

Chen Dun [*1] Cameron R. Wolfe [*1] Chris Jermaine [1] Anastasios Kyrillidis [1]

https://arxiv.org/pdf/2107.00961.pdf

ON THE CONVERGENCE OF SHALLOW NEURAL NETWORK TRAINING WITH RANDOMLY MASKED NEURONS

Fangshuo Liao, Anastasios Kyrillidis
Department of Computer Science
Rice University
Houston, TX 77005, USA
{Fangshuo.Liao,anastasios}@rice.edu

https://arxiv.org/pdf/2112.02668.pdf

Published at TMLR

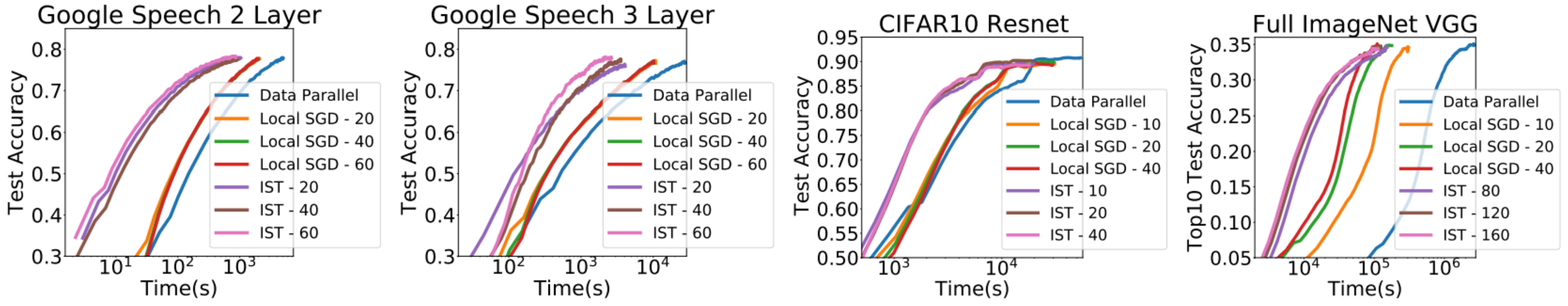
GIST: Distributed Training for Large-Scale Graph Convolutional Networks

Cameron Wolfe [*1] Jingkang Yang [*2] Arindam Chowdhury [3] Chen Dun [1] Artun Bayer [3] Santiago Segarra [3]
Anastasios Kyrillidis [1]

https://arxiv.org/pdf/2102.10424.pdf

Accepted at Journal of Applied and Computational Topology

Published at VLDB 2022



Published at UAI 2022

ResIST: Layer-Wise Decomposition of ResNets for Distributed Training

Chen Dun [*1] **Cameron R. Wolfe** [*1] **Chris Jermaine** [1] **Anastasios Kyrillidis** [1]

https://arxiv.org/pdf/2107.00961.pdf

On the Convergence of Shallow Neural Network Training with Randomly Masked Neurons

Fangshuo Liao, Anastasios Kyrillidis
Department of Computer Science
Rice University
Houston, TX 77005, USA
{Fangshuo.Liao,anastasios}@rice.edu

https://arxiv.org/pdf/2112.02668.pdf

Published at TMLR

GIST: Distributed Training for Large-Scale Graph Convolutional Networks

Cameron Wolfe [*1] **Jingkang Yang** [*2] **Arindam Chowdhury** [3] **Chen Dun** [1] **Artun Bayer** [3] **Santiago Segarra** [3]
Anastasios Kyrillidis [1]

https://arxiv.org/pdf/2102.10424.pdf

Accepted at Journal of Applied and Computational Topology

Published at VLDB 2022

Distributed Learning of Neural Networks using Independent Subnet Training

Binhang Yuan
Rice University
by8@rice.edu

Cameron R. Wolfe
Rice University
crw13@rice.edu

Chen Dun
Rice University
cd46@rice.edu

Yuxin Tang
Rice University
yuxin.tang@rice.edu

Anastasios Kyrillidis
Rice University
anastasios@rice.edu

Chris Jermaine
Rice University
cmj4@rice.edu

https://arxiv.org/pdf/1910.02120.pdf

Published at UAI 2022

ResIST: Layer-Wise Decomposition of ResNets for Distributed Training

Chen Dun [*1] Cameron R. Wolfe [*1] Chris Jermaine [1] Anastasios Kyrillidis [1]

https://arxiv.org/pdf/2107.00961.pdf

ON THE CONVERGENCE OF SHALLOW NEURAL NETWORK TRAINING WITH RANDOMLY MASKED NEURONS

Fangshuo Liao, Anastasios Kyrillidis
Department of Computer Science
Rice University
Houston, TX 77005, USA
{Fangshuo.Liao,anastasios}@rice.edu

https://arxiv.org/pdf/2112.02668.pdf

Published at TMLR

GIST: Distributed Training for Large-Scale Graph Convolutional Networks

Cameron Wolfe [*1] Jingkang Yang [*2] Arindam Chowdhury [3] Chen Dun [1] Artun Bayer [3] Santiago Segarra [3]
Anastasios Kyrillidis [1]

https://arxiv.org/pdf/2102.10424.pdf

Accepted at Journal of Applied and Computational Topology

Published at VLDB 2022

Distributed Learning of Neural Networks using Independent Subnet Training

Binhang Yuan
Rice University
by8@rice.edu

Cameron R. Wolfe
Rice University
crw13@rice.edu

Chen Dun
Rice University
cd46@rice.edu

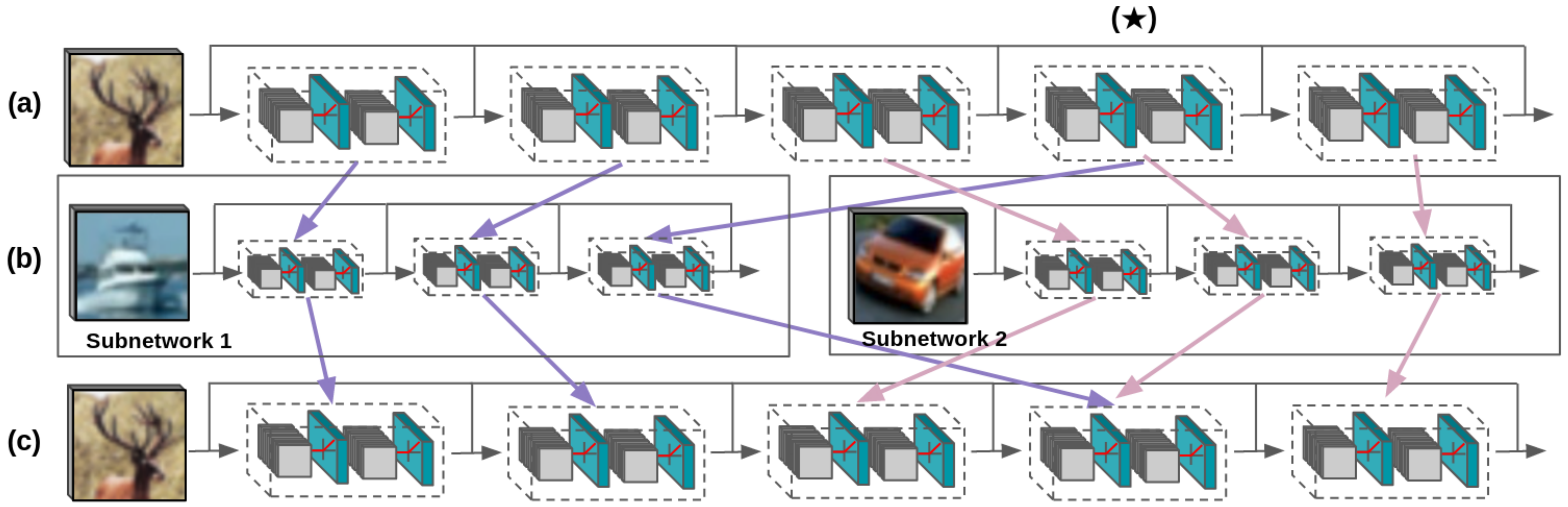
Yuxin Tang
Rice University
yuxin.tang@rice.edu

Anastasios Kyrillidis
Rice University
anastasios@rice.edu

Chris Jermaine
Rice University
cmj4@rice.edu

https://arxiv.org/pdf/1910.02120.pdf

Published at UAI 2022



ON THE CONVERGENCE OF SHALLOW NEURAL NETWORK TRAINING WITH RANDOMLY MASKED NEURONS

Fangshuo Liao, Anastasios Kyrillidis
Department of Computer Science
Rice University
Houston, TX 77005, USA
{Fangshuo.Liao,anastasios}@rice.edu

https://arxiv.org/pdf/2112.02668.pdf

Published at TMLR

GIST: Distributed Training for Large-Scale Graph Convolutional Networks

Cameron Wolfe [*1] **Jingkang Yang** [*2] **Arindam Chowdhury** [3] **Chen Dun** [1] **Artun Bayer** [3] **Santiago Segarra** [3] **Anastasios Kyrillidis** [1]

https://arxiv.org/pdf/2102.10424.pdf

Accepted at Journal of Applied and Computational Topology

Published at VLDB 2022

Distributed Learning of Neural Networks using Independent Subnet Training

Binhang Yuan
Rice University
by8@rice.edu

Cameron R. Wolfe
Rice University
crw13@rice.edu

Chen Dun
Rice University
cd46@rice.edu

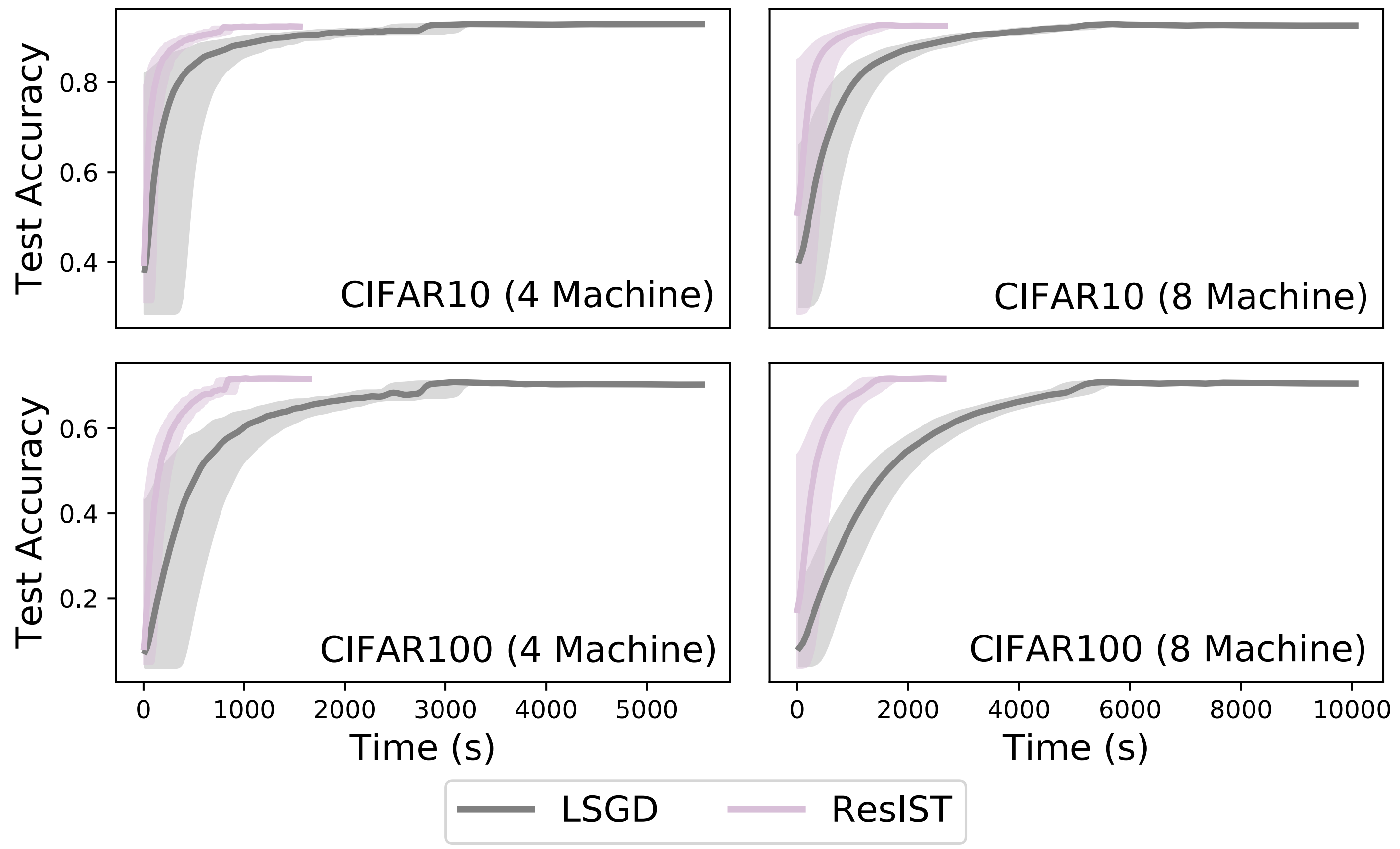
Yuxin Tang
Rice University
yuxin.tang@rice.edu

Anastasios Kyrillidis
Rice University
anastasios@rice.edu

Chris Jermaine
Rice University
cmj4@rice.edu

https://arxiv.org/pdf/1910.02120.pdf

Published at UAI 2022



On the Convergence of Shallow Neural Network Training with Randomly Masked Neurons

Fangshuo Liao, Anastasios Kyrillidis
Department of Computer Science
Rice University
Houston, TX 77005, USA
{Fangshuo.Liao,anastasios}@rice.edu

https://arxiv.org/pdf/2112.02668.pdf

Published at TMLR

GIST: Distributed Training for Large-Scale Graph Convolutional Networks

Cameron Wolfe [*1] **Jingkang Yang** [*2] **Arindam Chowdhury** [3] **Chen Dun** [1] **Artun Bayer** [3] **Santiago Segarra** [3] **Anastasios Kyrillidis** [1]

https://arxiv.org/pdf/2102.10424.pdf

Accepted at Journal of Applied and Computational Topology

Published at VLDB 2022

Distributed Learning of Neural Networks using Independent Subnet Training

Binhang Yuan
Rice University
by8@rice.edu

Cameron R. Wolfe
Rice University
crw13@rice.edu

Chen Dun
Rice University
cd46@rice.edu

Yuxin Tang
Rice University
yuxin.tang@rice.edu

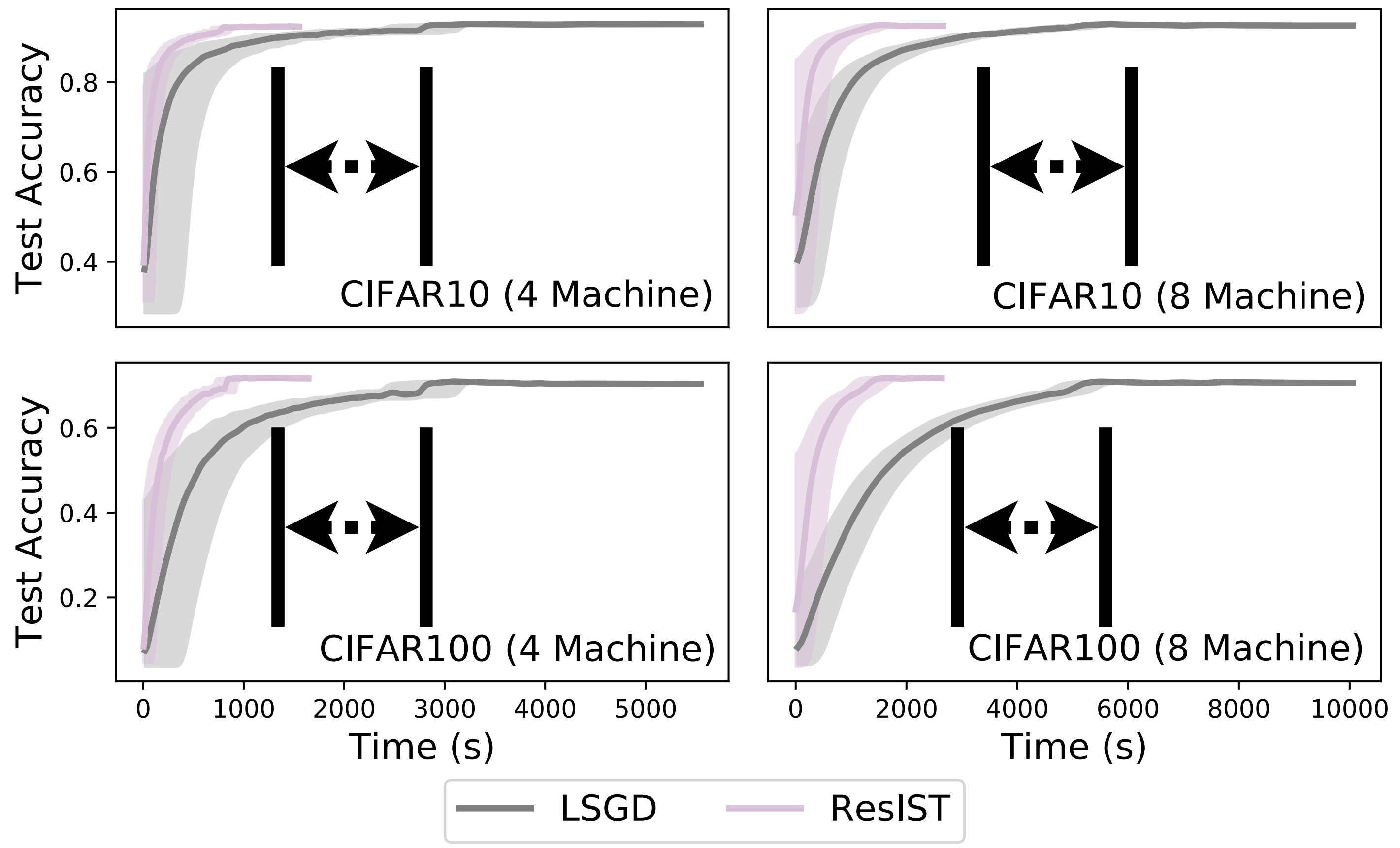
Anastasios Kyrillidis
Rice University
anastasios@rice.edu

Chris Jermaine
Rice University
cmj4@rice.edu

https://arxiv.org/pdf/1910.02120.pdf

Published at UAI 2022

Acceleration (wall clock) while pertaining accuracy!



ON THE CONVERGENCE OF SHALLOW NEURAL NETWORK TRAINING WITH RANDOMLY MASKED NEURONS

Fangshuo Liao, Anastasios Kyrillidis
Department of Computer Science
Rice University
Houston, TX 77005, USA
{Fangshuo.Liao,anastasios}@rice.edu

https://arxiv.org/pdf/2112.02668.pdf

Published at TMLR

GIST: Distributed Training for Large-Scale Graph Convolutional Networks

Cameron Wolfe [*1] Jingkang Yang [*2] Arindam Chowdhury [3] Chen Dun [1] Artun Bayer [3] Santiago Segarra [3] Anastasios Kyrillidis [1]

https://arxiv.org/pdf/2102.10424.pdf

Accepted at Journal of Applied and Computational Topology

Published at VLDB 2022

Distributed Learning of Neural Networks using Independent Subnet Training

Binhang Yuan
Rice University
by8@rice.edu

Cameron R. Wolfe
Rice University
crw13@rice.edu

Chen Dun
Rice University
cd46@rice.edu

Yuxin Tang
Rice University
yuxin.tang@rice.edu

Anastasios Kyrillidis
Rice University
anastasios@rice.edu

Chris Jermaine
Rice University
cmj4@rice.edu

https://arxiv.org/pdf/1910.02120.pdf

Published at UAI 2022

ResIST: Layer-Wise Decomposition of ResNets for Distributed Training

Chen Dun [*1] Cameron R. Wolfe [*1] Chris Jermaine [1] Anastasios Kyrillidis [1]

https://arxiv.org/pdf/2107.00961.pdf

ON THE CONVERGENCE OF SHALLOW NEURAL NETWORK TRAINING WITH RANDOMLY MASKED NEURONS

Fangshuo Liao, Anastasios Kyrillidis
Department of Computer Science
Rice University
Houston, TX 77005, USA
{Fangshuo.Liao,anastasios}@rice.edu

https://arxiv.org/pdf/2112.02668.pdf

Published at TMLR

GIST: Distributed Training for Large-Scale Graph Convolutional Networks

Cameron Wolfe [*1] Jingkang Yang [*2] Arindam Chowdhury [3] Chen Dun [1] Artun Bayer [3] Santiago Segarra [3]
Anastasios Kyrillidis [1]

https://arxiv.org/pdf/2102.10424.pdf

Accepted at Journal of Applied and Computational Topology

Published at VLDB 2022

Distributed Learning of Neural Networks using Independent Subnet Training

Binhang Yuan
Rice University
by8@rice.edu

Cameron R. Wolfe
Rice University
crw13@rice.edu

Chen Dun
Rice University
cd46@rice.edu

Yuxin Tang
Rice University
yuxin.tang@rice.edu

Anastasios Kyrillidis
Rice University
anastasios@rice.edu

Chris Jermaine
Rice University
cmj4@rice.edu

https://arxiv.org/pdf/1910.02120.pdf

Published at UAI 2022

ResIST: Layer-Wise Decomposition of ResNets for Distributed Training

Chen Dun [*1] Cameron R. Wolfe [*1] Chris Jermaine [1] Anastasios Kyrillidis [1]

https://arxiv.org/pdf/2107.00961.pdf

What about theory?

Published at TMLR

GIST: Distributed Training for Large-Scale Graph Convolutional Networks

Cameron Wolfe [*1] Jingkang Yang [*2] Arindam Chowdhury [3] Chen Dun [1] Artun Bayer [3] Santiago Segarra [3] Anastasios Kyrillidis [1]

https://arxiv.org/pdf/2102.10424.pdf

Accepted at Journal of Applied and Computational Topology

Published at VLDB 2022

Distributed Learning of Neural Networks using Independent Subnet Training

Binhang Yuan, Rice University, by8@rice.edu
Cameron R. Wolfe, Rice University, crw13@rice.edu
Chen Dun, Rice University, cd46@rice.edu
Yuxin Tang, Rice University, yuxin.tang@rice.edu
Anastasios Kyrillidis, Rice University, anastasios@rice.edu
Chris Jermaine, Rice University, cmj4@rice.edu

https://arxiv.org/pdf/1910.02120.pdf

Published at UAI 2022

ResIST: Layer-Wise Decomposition of ResNets for Distributed Training

Chen Dun [*1] Cameron R. Wolfe [*1] Chris Jermaine [1] Anastasios Kyrillidis [1]

https://arxiv.org/pdf/2107.00961.pdf

Corollary 3. *Let assumptions (1), (2), and (3) hold. Fix the number of dropout iterations to K, the step size to $\eta = O\left(\lambda_0/n\tau\max\{n,p\}\right)$, and let the number of hidden neurons satisfy $m = \Theta\left(n^5K/\xi\theta\lambda_0^4\delta\right)$. Then the IST algorithm on a two-layer ReLU neural network converges with probability at least $1-\delta$, according to:*

$$\mathbb{E}_{[\mathbf{M}_{k-1}]}\left[\|\mathbf{y}-\mathbf{u}_k\|_2^2\right] \leq \left(1-\frac{1}{4}\eta\theta\tau\lambda_0\right)^k \|\mathbf{y}-\mathbf{u}_0\|_2^2 + O\left(\frac{(1-\xi)^2}{nK} + \frac{\theta-\xi^2}{p} + \left(1-\frac{1}{\tau}\right)\theta^2(1-\xi)\right)$$

Theorem 4. *Let assumptions (1) and (4) hold. Then $\lambda_0 > 0$. Moreover, let $\lambda_{\max}$ denote the maximum eigenvalue of $\mathbf{H}^\infty$. Fix the number of global iterations to K and the number of local iterations to τ. Let the number of hidden neurons be $m = \Omega\left(\frac{n^5\tau^2K\lambda_{\max}}{\lambda_0^6\delta}\right)$, and choose the initialization scale $\kappa = \sqrt{n\lambda_{\max}}\lambda_0^{-1}$. Let $\gamma = (1-p^{-1})^{\frac{1}{3}}$. Then, Algorithm (1) with a constant step-size $\eta = O\left(\frac{\lambda_0}{n^2}\min\left\{\frac{p}{\gamma^2\tau}, 1\right\}\right)$ converges with probability at least $1-\delta$, according to:*

$$\mathbb{E}_{[\mathbf{M}_{k-1}]}\left[\|\mathbf{y}-\mathbf{u}_k\|_2^2\right] \leq \left(\gamma + (1-\gamma)\left(1-\frac{\eta\lambda_0}{2}\right)^\tau\right)^k \|\mathbf{y}-\mathbf{u}_0\|_2^2 + O\left(\frac{\gamma\tau n\kappa^2\lambda_{\max}}{\lambda_0^2}\right).$$

Published at TMLR

GIST: Distributed Training for Large-Scale Graph Convolutional Networks

Cameron Wolfe [*1] Jingkang Yang [*2] Arindam Chowdhury [3] Chen Dun [1] Artun Bayer [3] Santiago Segarra [3] Anastasios Kyrillidis [1]

https://arxiv.org/pdf/2102.10424.pdf

Accepted at Journal of Applied and Computational Topology

Published at VLDB 2022

Distributed Learning of Neural Networks using Independent Subnet Training

Binhang Yuan, Rice University, by8@rice.edu
Cameron R. Wolfe, Rice University, crw13@rice.edu
Chen Dun, Rice University, cd46@rice.edu
Yuxin Tang, Rice University, yuxin.tang@rice.edu
Anastasios Kyrillidis, Rice University, anastasios@rice.edu
Chris Jermaine, Rice University, cmj4@rice.edu

https://arxiv.org/pdf/1910.02120.pdf

Published at UAI 2022

Theorem B.1 (Convergence Rate of Gradient Descent for ResIST). *Assume there are S workers, ℓ local and T global steps. Assume the depth of the whole ResNet is H. Assume for all data indices $i \in [n]$, the data input satisfies $\|\mathbf{x}_i\|_2 = 1$, the data output satisfies $|y_i| = O(1)$, and the number of hidden nodes per layer satisfies $m =$*

$$\Omega\left(\max\left\{\frac{n^4}{\lambda^4_{\min}(\mathbf{K}^{(H)})H^6}, \frac{n^2}{\lambda^2_{\min}(\mathbf{K}^{(H)})H^2}, \frac{n}{\delta}, \frac{n^2\log\left(\frac{Hn}{\delta}\right)}{\lambda^2_{\min}(\mathbf{K}^{(H)})}\right\}\right).$$

Set the step size $\eta = O\left(\frac{\lambda_{\min}(\mathbf{K}^{(H)})H^2}{n^2\ell^2 S}\right)$ in gradient descent in local training iteration, and follow the procedure as in Algorithm 1. Let the squared-norm loss be $L(\theta(t)) := \frac{1}{2}\|\mathbf{y} - f(\theta(t))\|_2^2$, per t global synchronization round, $t = 1, 2, \ldots T$; here, $\mathbf{y}$ corresponds to the data "labels", and $\theta(t)$ and $f(\theta(t))$ represent the parameters and the output of the whole ResNet, respectively, after t-global rounds of ResIST. Here, θ includes weights $\mathbf{W}^{(h)}$ at depth h and the last layer's weights $\mathbf{a}$. Then, with probability at least $1-\delta$ over the random initialization, we have:

$$L(\theta(t)) \leq \left(1 - \frac{\eta\ell\lambda_{\min}(\mathbf{K}^{(H)})}{2}\right)^t \cdot L(\theta(0)).$$

Corollary 3. *Let assumptions (1), (2), and (3) hold. Fix the number of dropout iterations to K, the step size to $\eta = O\left(\lambda_0/n\tau\max\{n,p\}\right)$, and let the number of hidden neurons satisfy $m = \Theta\left(n^5K/\xi\theta\lambda_0^4\delta\right)$. Then the IST algorithm on a two-layer ReLU neural network converges with probability at least $1-\delta$, according to:*

$$\mathbb{E}_{[\mathbf{M}_{k-1}]}\left[\|\mathbf{y}-\mathbf{u}_k\|_2^2\right] \leq \left(1 - \frac{1}{4}\eta\theta\tau\lambda_0\right)^k \|\mathbf{y}-\mathbf{u}_0\|_2^2 + O\left(\frac{(1-\xi)^2}{nK} + \frac{\theta-\xi^2}{p} + \left(1-\frac{1}{\tau}\right)\theta^2(1-\xi)\right)$$

Theorem 4. *Let assumptions (1) and (4) hold. Then $\lambda_0 > 0$. Moreover, let $\lambda_{\max}$ denote the maximum eigenvalue of $\mathbf{H}^\infty$. Fix the number of global iterations to K and the number of local iterations to τ. Let the number of hidden neurons be $m = \Omega\left(\frac{n^5\tau^2K\lambda_{\max}}{\lambda_0^6\delta}\right)$, and choose the initialization scale $\kappa = \sqrt{n\lambda_{\max}}\lambda_0^{-1}$. Let $\gamma = (1-p^{-1})^{\frac{1}{3}}$. Then, Algorithm (1) with a constant step-size $\eta = O\left(\frac{\lambda_0}{n^2}\min\left\{\frac{p}{\gamma^2\tau}, 1\right\}\right)$ converges with probability at least $1-\delta$, according to:*

$$\mathbb{E}_{[\mathbf{M}_{k-1}]}\left[\|\mathbf{y}-\mathbf{u}_k\|_2^2\right] \leq \left(\gamma + (1-\gamma)\left(1-\frac{\eta\lambda_0}{2}\right)^\tau\right)^k \|\mathbf{y}-\mathbf{u}_0\|_2^2 + O\left(\frac{\gamma\tau n\kappa^2\lambda_{\max}}{\lambda_0^2}\right).$$

Published at TMLR

GIST: Distributed Training for Large-Scale Graph Convolutional Networks

Cameron Wolfe [*1] Jingkang Yang [*2] Arindam Chowdhury [3] Chen Dun [1] Artun Bayer [3] Santiago Segarra [3]
Anastasios Kyrillidis [1]

https://arxiv.org/pdf/2102.10424.pdf

Accepted at Journal of Applied and Computational Topology

Published at VLDB 2022

Distributed Learning of Neural Networks using Independent Subnet Training

Binhang Yuan
Rice University
by8@rice.edu

Cameron R. Wolfe
Rice University
crw13@rice.edu

Chen Dun
Rice University
cd46@rice.edu

Yuxin Tang
Rice University
yuxin.tang@rice.edu

Anastasios Kyrillidis
Rice University
anastasios@rice.edu

Chris Jermaine
Rice University
cmj4@rice.edu

https://arxiv.org/pdf/1910.02120.pdf

Published at UAI 2022

ResIST: Layer-Wise Decomposition of ResNets for Distributed Training

Chen Dun [*1] Cameron R. Wolfe [*1] Chris Jermaine [1] Anastasios Kyrillidis [1]

https://arxiv.org/pdf/2107.00961.pdf

ON THE CONVERGENCE OF SHALLOW NEURAL NETWORK TRAINING WITH RANDOMLY MASKED NEURONS

Fangshuo Liao, Anastasios Kyrillidis
Department of Computer Science
Rice University
Houston, TX 77005, USA
{Fangshuo.Liao,anastasios}@rice.edu

https://arxiv.org/pdf/2112.02668.pdf

Published at TMLR

GIST: Distributed Training for Large-Scale Graph Convolutional Networks

Cameron Wolfe [*1] Jingkang Yang [*2] Arindam Chowdhury [3] Chen Dun [1] Artun Bayer [3] Santiago Segarra [3]
Anastasios Kyrillidis [1]

https://arxiv.org/pdf/2102.10424.pdf

Accepted at Journal of Applied and Computational Topology

Published at VLDB 2022

Distributed Learning of Neural Networks using Independent Subnet Training

Binhang Yuan
Rice University
by8@rice.edu

Cameron R. Wolfe
Rice University
crw13@rice.edu

Chen Dun
Rice University
cd46@rice.edu

Yuxin Tang
Rice University
yuxin.tang@rice.edu

Anastasios Kyrillidis
Rice University
anastasios@rice.edu

Chris Jermaine
Rice University
cmj4@rice.edu

https://arxiv.org/pdf/1910.02120.pdf

Published at UAI 2022

ResIST: Layer-Wise Decomposition of ResNets for Distributed Training

Chen Dun [*1] Cameron R. Wolfe [*1] Chris Jermaine [1] Anastasios Kyrillidis [1]

https://arxiv.org/pdf/2107.00961.pdf

ON THE CONVERGENCE OF SHALLOW NEURAL NETWORK TRAINING WITH RANDOMLY MASKED NEURONS

Fangshuo Liao, Anastasios Kyrillidis
Department of Computer Science
Rice University
Houston, TX 77005, USA
{Fangshuo.Liao,anastasios}@rice.edu

https://arxiv.org/pdf/2112.02668.pdf

Published at TMLR

What about more architectures?

Accepted at Journal of Applied and Computational Topology

Published at VLDB 2022

Distributed Learning of Neural Networks using Independent Subnet Training

Binhang Yuan
Rice University
by8@rice.edu

Cameron R. Wolfe
Rice University
crw13@rice.edu

Chen Dun
Rice University
cd46@rice.edu

Yuxin Tang
Rice University
yuxin.tang@rice.edu

Anastasios Kyrillidis
Rice University
anastasios@rice.edu

Chris Jermaine
Rice University
cmj4@rice.edu

https://arxiv.org/pdf/1910.02120.pdf

Published at UAI 2022

ResIST: Layer-Wise Decomposition of ResNets for Distributed Training

Chen Dun [*1] Cameron R. Wolfe [*1] Chris Jermaine [1] Anastasios Kyrillidis [1]

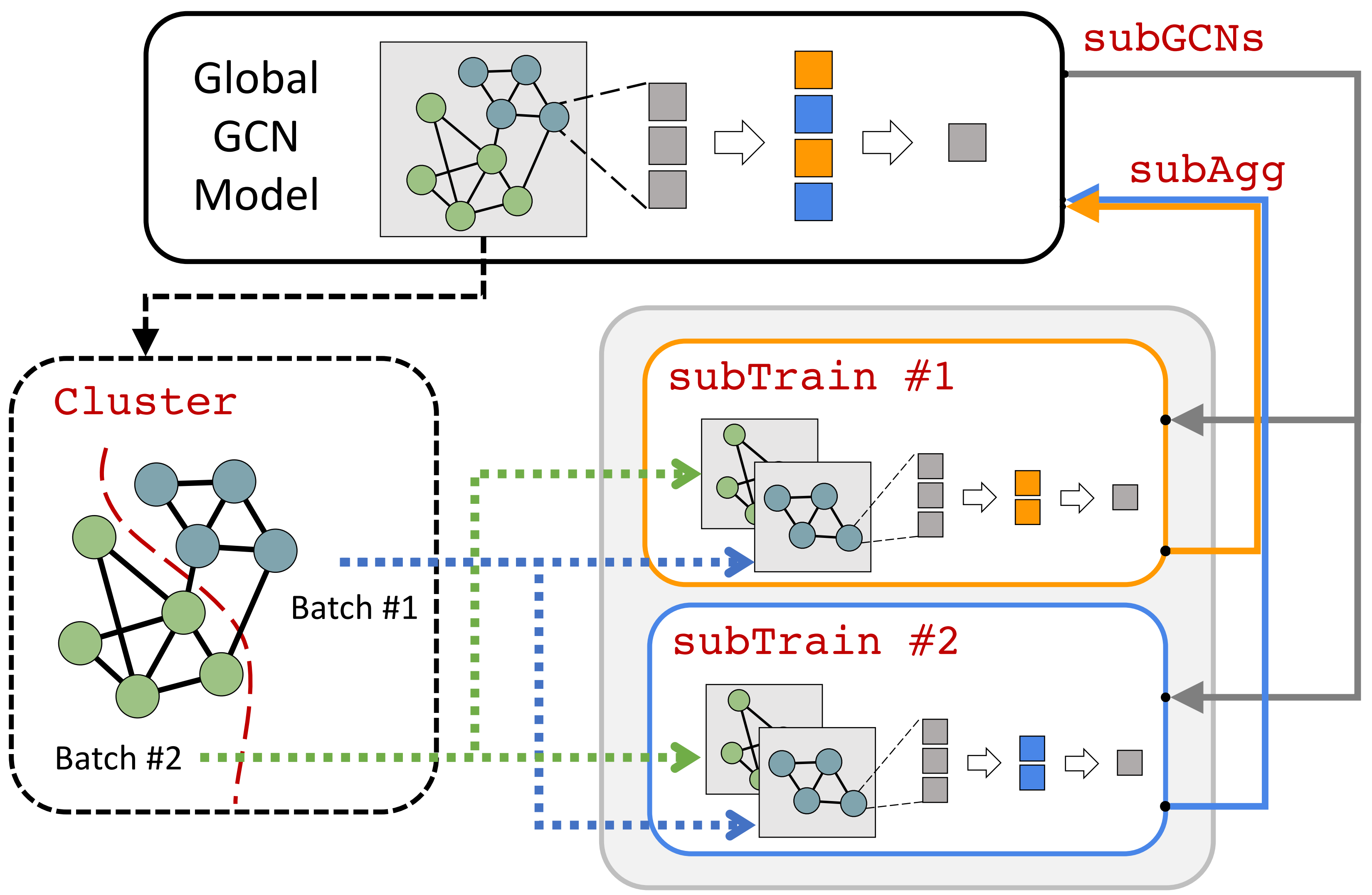
https://arxiv.org/pdf/2107.00961.pdf

On the Convergence of Shallow Neural Network Training with Randomly Masked Neurons

Fangshuo Liao, Anastasios Kyrillidis
Department of Computer Science
Rice University
Houston, TX 77005, USA
{Fangshuo.Liao,anastasios}@rice.edu

https://arxiv.org/pdf/2112.02668.pdf

Published at TMLR



Published at VLDB 2022

Distributed Learning of Neural Networks using Independent Subnet Training

Binhang Yuan, Rice University, by8@rice.edu
Cameron R. Wolfe, Rice University, crw13@rice.edu
Chen Dun, Rice University, cd46@rice.edu
Yuxin Tang, Rice University, yuxin.tang@rice.edu
Anastasios Kyrillidis, Rice University, anastasios@rice.edu
Chris Jermaine, Rice University, cmj4@rice.edu

https://arxiv.org/pdf/1910.02120.pdf

Published at UAI 2022

ResIST: Layer-Wise Decomposition of ResNets for Distributed Training

Chen Dun [*1] Cameron R. Wolfe [*1] Chris Jermaine [1] Anastasios Kyrillidis [1]

https://arxiv.org/pdf/2107.00961.pdf

On the Convergence of Shallow Neural Network Training with Randomly Masked Neurons

Fangshuo Liao, Anastasios Kyrillidis
Department of Computer Science
Rice University
Houston, TX 77005, USA
{Fangshuo.Liao,anastasios}@rice.edu

https://arxiv.org/pdf/2112.02668.pdf

Published at TMLR

L	m	F1 Score (Time in hours)				
		$d_i = 400$	$d_i = 4\ 096$	$d_i = 8\ 192$	$d_i = 16\ 384$	$d_i = 32\ 768$
2	-	89.38 (1.81)	90.58 (5.17)	OOM	OOM	OOM
	2	87.48 (1.25)	90.09 (1.70)	90.87 (2.76)	90.94 (9.31)	90.91 (32.31)
	4	84.82 (1.11)	88.79 (1.13)	89.76 (1.49)	90.10 (2.24)	90.17 (5.16)
	8	82.56 (1.13)	87.16 (1.11)	88.31 (1.20)	88.89 (1.39)	89.46 (1.76)
3	-	89.73 (2.32)	90.99 (9.52)	OOM	OOM	OOM
	2	87.79 (1.56)	90.40 (2.12)	90.91 (4.87)	91.05 (17.7)	OOM
	4	85.30 (1.37)	88.51 (1.42)	89.75 (2.07)	90.15 (3.44)	OOM
	8	82.84 (1.37)	86.12 (1.34)	88.38 (1.37)	88.67 (1.88)	88.66 (2.56)
4	-	89.77 (3.00)	91.02 (14.20)	OOM	OOM	OOM
	2	87.75 (1.79)	90.36 (2.77)	91.08 (6.92)	91.09 (26.44)	OOM
	4	85.32 (1.58)	88.50 (1.65)	89.76 (2.36)	90.05 (4.93)	OOM
	8	83.45 (1.56)	86.60 (1.55)	88.13 (1.61)	88.44 (2.30)	OOM

Table 3: Performance of GraphSAGE models of different widths trained with GIST on Amazon2M. $m =$ "-" refers to the baseline and "OOM" marks experiments that cause out-of-memory errors. *GIST enables training of higher-performing, ultra-wide models.*

Published at VLDB 2022

Distributed Learning of Neural Networks using Independent Subnet Training

Binhang Yuan
Rice University
by8@rice.edu

Cameron R. Wolfe
Rice University
crw13@rice.edu

Chen Dun
Rice University
cd46@rice.edu

Yuxin Tang
Rice University
yuxin.tang@rice.edu

Anastasios Kyrillidis
Rice University
anastasios@rice.edu

Chris Jermaine
Rice University
cmj4@rice.edu

https://arxiv.org/pdf/1910.02120.pdf

Published at UAI 2022

ResIST: Layer-Wise Decomposition of ResNets for Distributed Training

Chen Dun [*1] Cameron R. Wolfe [*1] Chris Jermaine [1] Anastasios Kyrillidis [1]

https://arxiv.org/pdf/2107.00961.pdf

On the Convergence of Shallow Neural Network Training with Randomly Masked Neurons

Fangshuo Liao, Anastasios Kyrillidis
Department of Computer Science
Rice University
Houston, TX 77005, USA
{Fangshuo.Liao,anastasios}@rice.edu

https://arxiv.org/pdf/2112.02668.pdf

Published at TMLR

Theorem 2. *Suppose assumptions 2-4, and property 2 hold. Moreover, suppose in each global iteration the masks are generated from a categorical distribution with uniform mean* $1/m$. *Fix the number of global iterations to* T *and local iterations to* ζ. *If the number of hidden neurons satisfies* $d_1 = \Omega\left(\frac{n^3\zeta^2T^2}{\delta^2\gamma(1-\gamma)^2\lambda_0^4}\left(n+\frac{d}{m^2}\|\bar{\mathbf{A}}^2\|_{1,1}\right)\right)$, *then procedure (3) with constant step size* $\eta = O\left(\frac{\lambda_0}{n^2\|\bar{\mathbf{A}}^2\|_{1,1}}\right)$ *converges according to*

$$\mathbb{E}_{[\mathcal{M}_{t-1}],\mathbf{\Theta}_0,\mathbf{a}}\left[\|\mathbf{y}-\hat{\mathbf{y}}(t)\|_2^2\right] \leq \left(\gamma+(1-\gamma)\left(1-\frac{\eta\lambda_0}{2}\right)^{\zeta}\right)^t \mathbb{E}_{\mathbf{\Theta}_0,\mathbf{a}}\left[\|\mathbf{y}-\hat{\mathbf{y}}(0)\|_2^2\right] + O\left(\frac{(m-1)^2\zeta\|\bar{\mathbf{A}}^2\|_{1,1}nd}{\gamma m^2 d_1}\right)$$

with probability at least $1-\delta$.

Accepted at Journal of Applied and Computational Topology

Landscape of IST–related papers

- Idea of splitting the model into smaller ones (as part of MM) – Google

[Caldas, Konecny, McMahan, Talwalkar, 2018]

Remark: IST is different since each worker receives a different model to aggregate

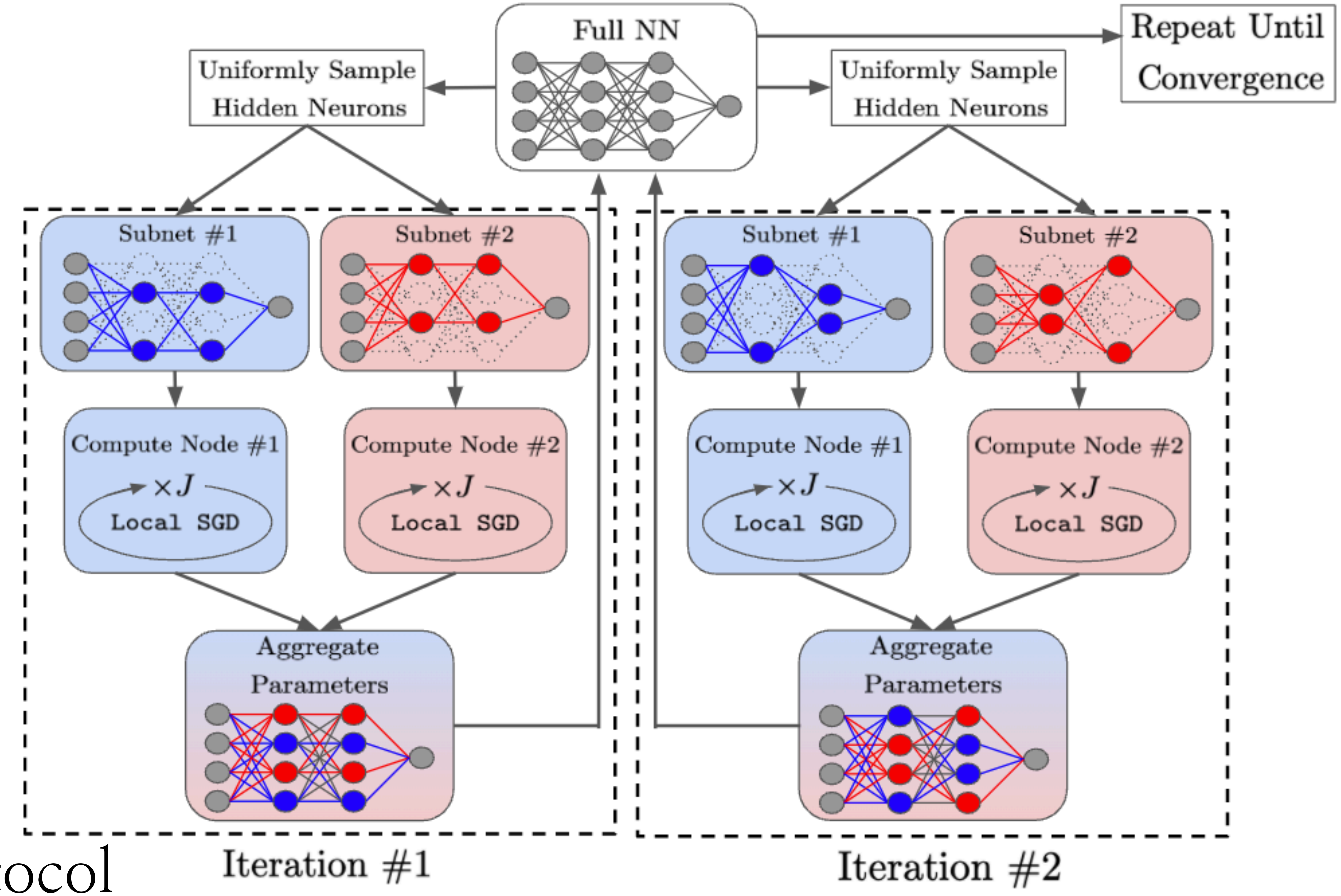
- One concurrent work from a FL perspective (to reduce comm/comp)

[Jiang, Wang, Valls, Ko, Lee, Leung, Tassiulas, 2019]

- Several works after IST:
  - Helios [Xu, Yu, Xiong and Chen, 2019]
  - HeteroFL [Diao, Ding, and Tarokh, 2020]
  - FjORD [Horvath, Laskaridis, Almeida, Leontiadis, Venieris and Lane, 2021]
  - PVT by Google [Yang, Guliani, Beaufays and Motta, 2021]
  - Masked NNs [Mohtashami, Jaggi, Stich, 2021]
  - General theory [Zhou, Lan, Venkataramani and Ding, 2022]
  - Federated Dropout [Wen, Jean, and Huang, 2022]
  - Federated Pruning (Google) [Lin et al., 2022]

[More information at https://akyrillidis.github.io/ist/]

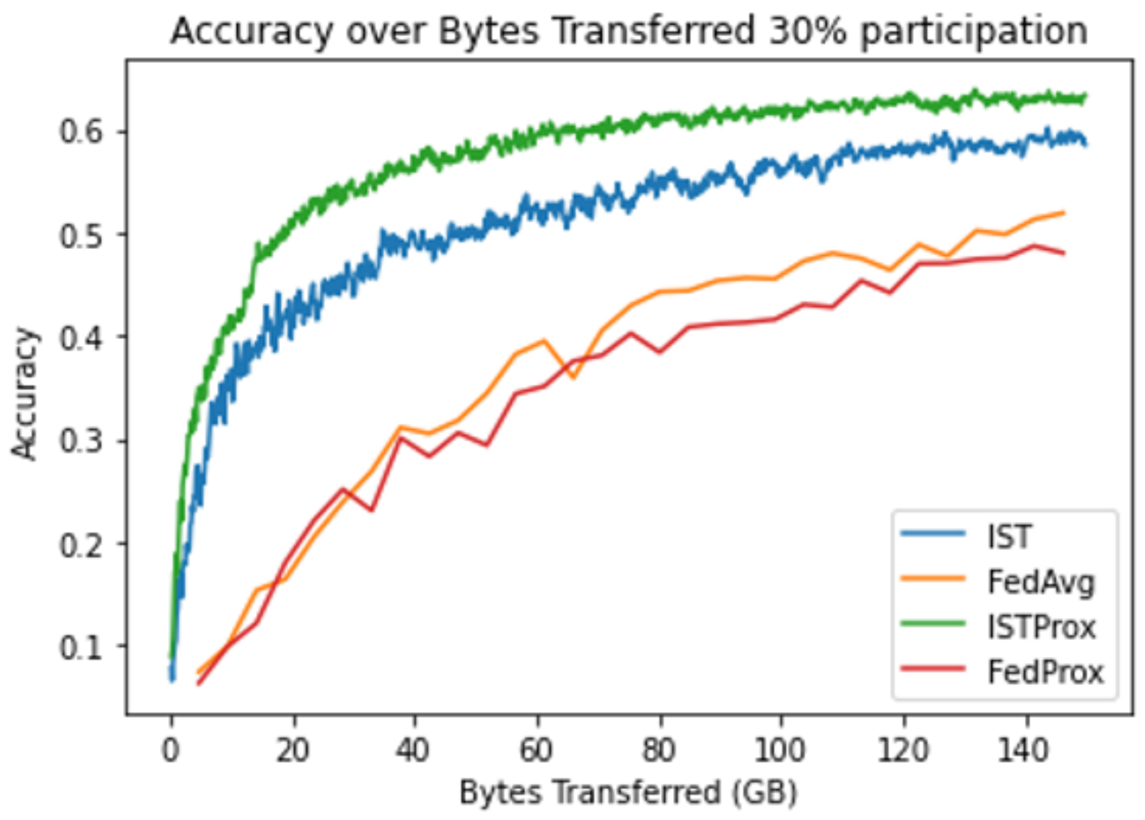
Take home message

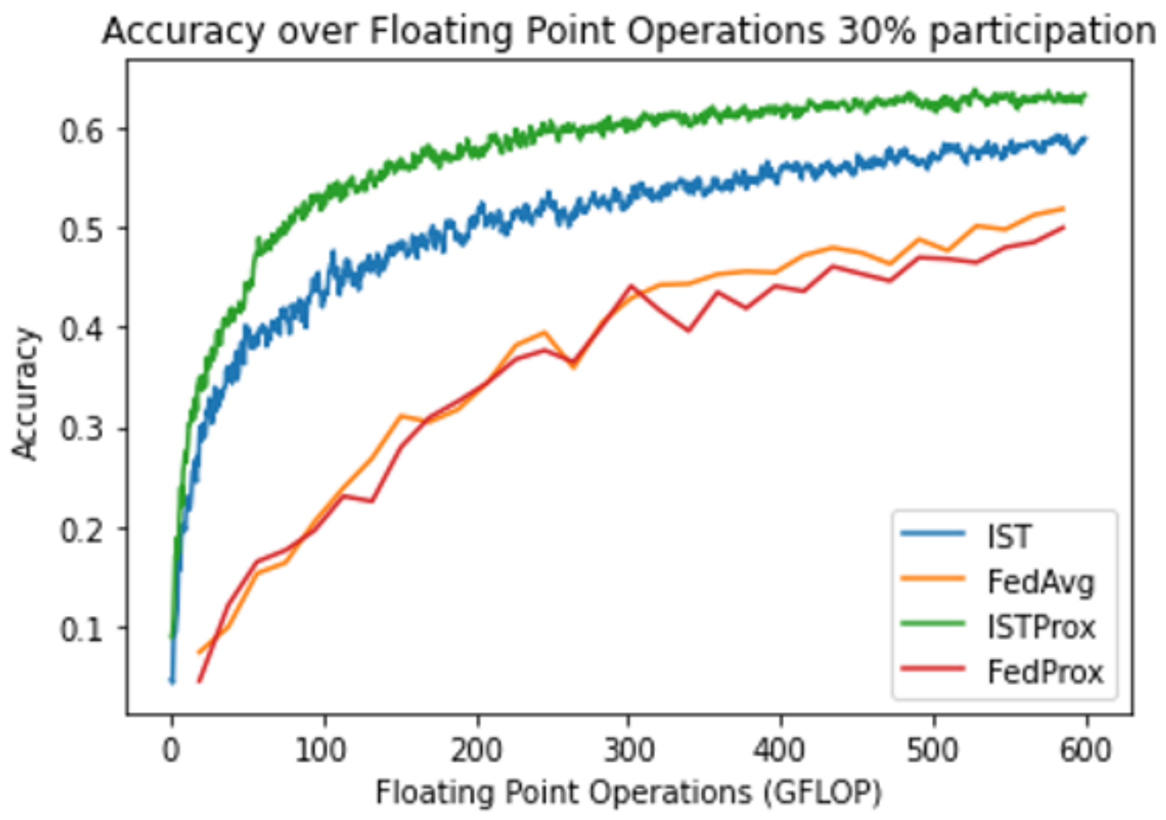


- A.. different distributed protocol
- Potential impact on communication, compute requirements
- We need a clearer view of large-scale models with hundreds of workers
- Unifies well existing models with theory

Where to go from here?

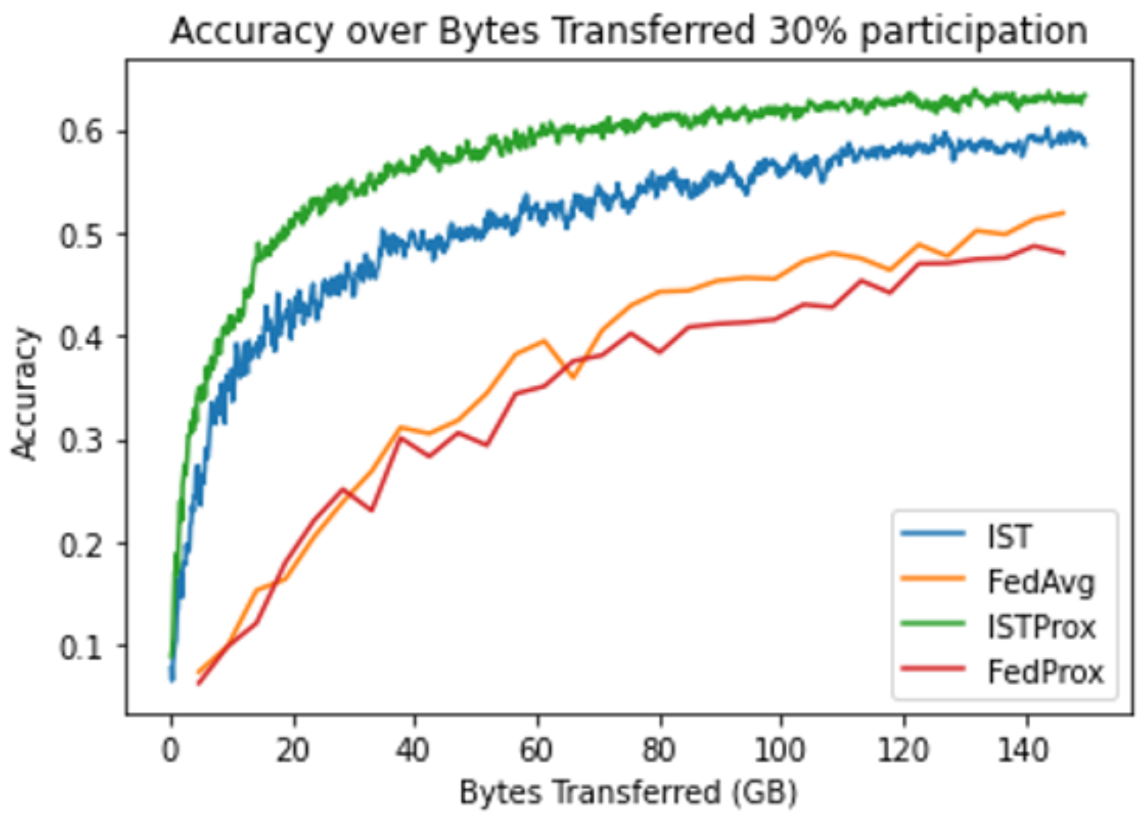
- IST + FL

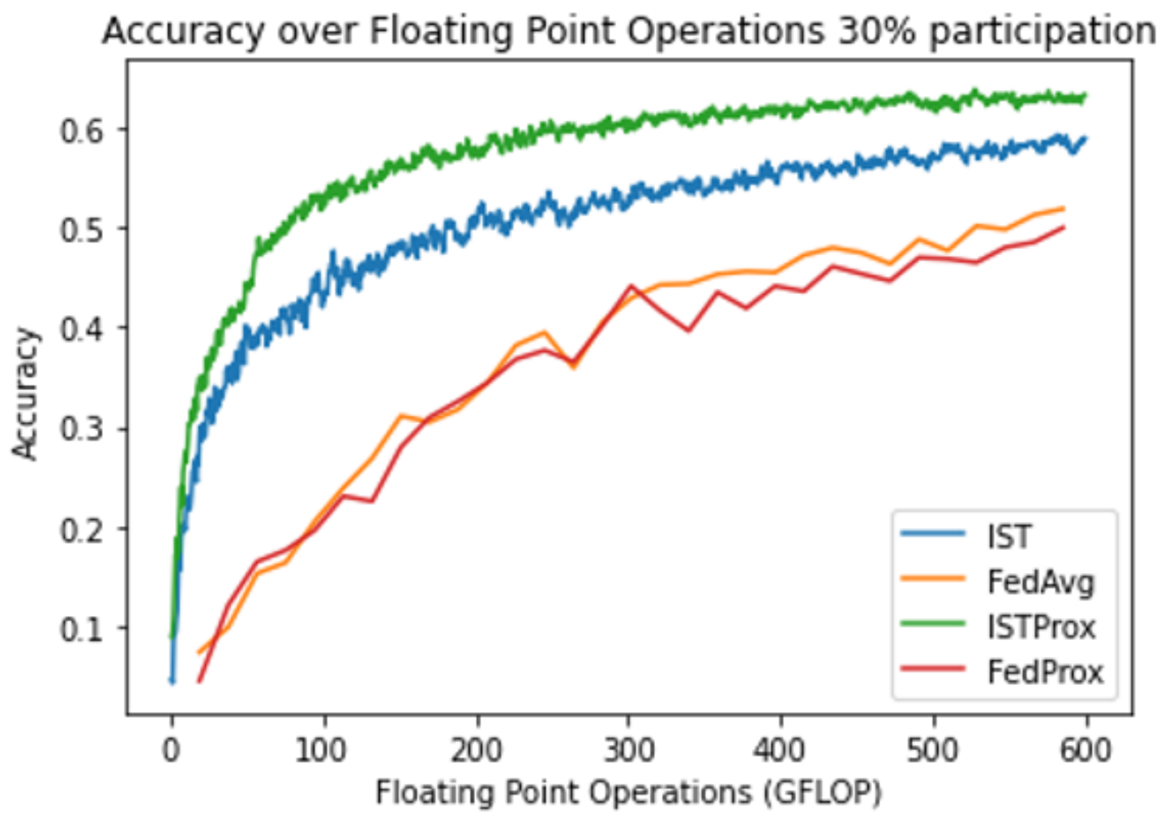




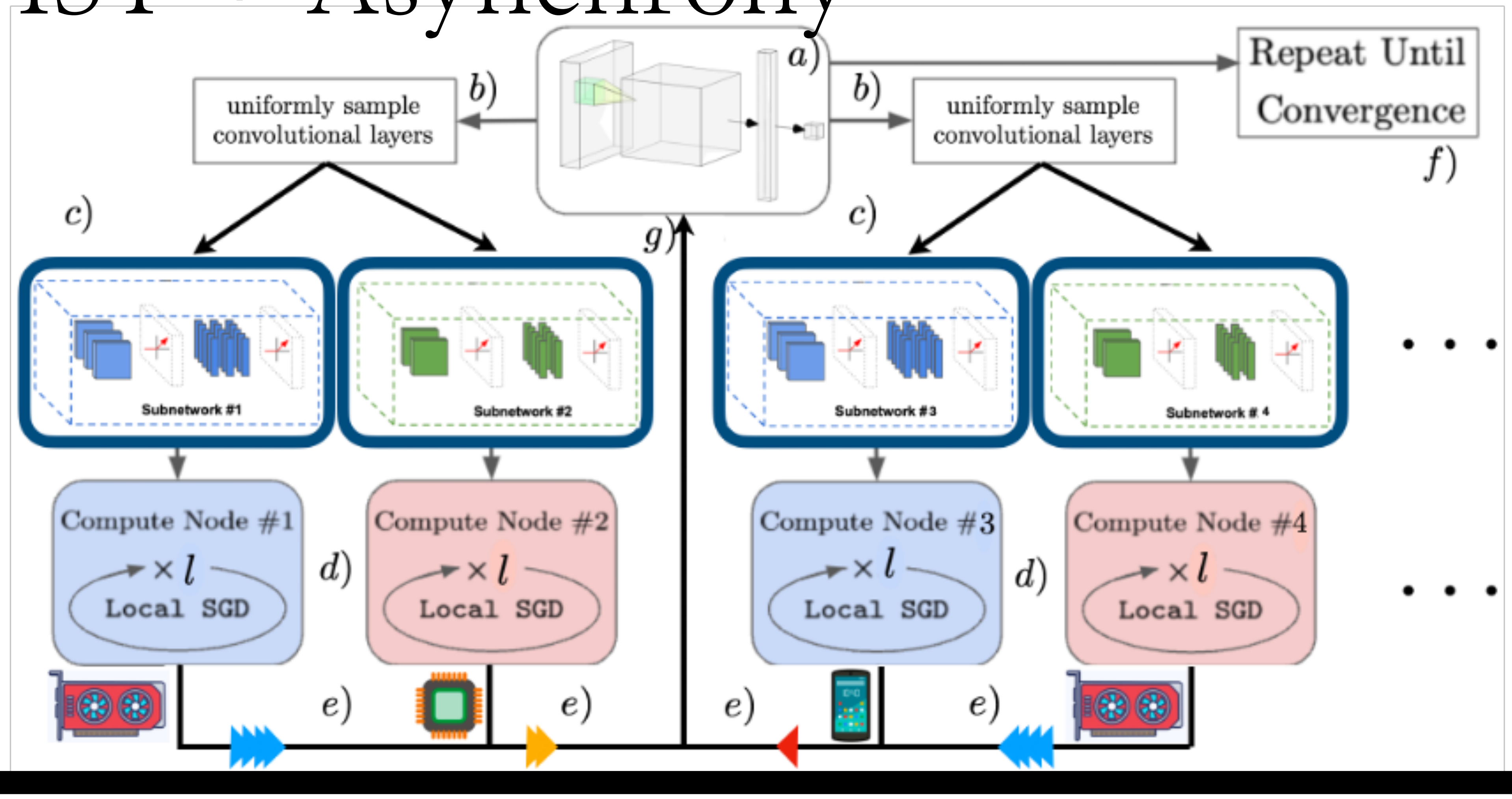
Where to go from here?

- IST + FL



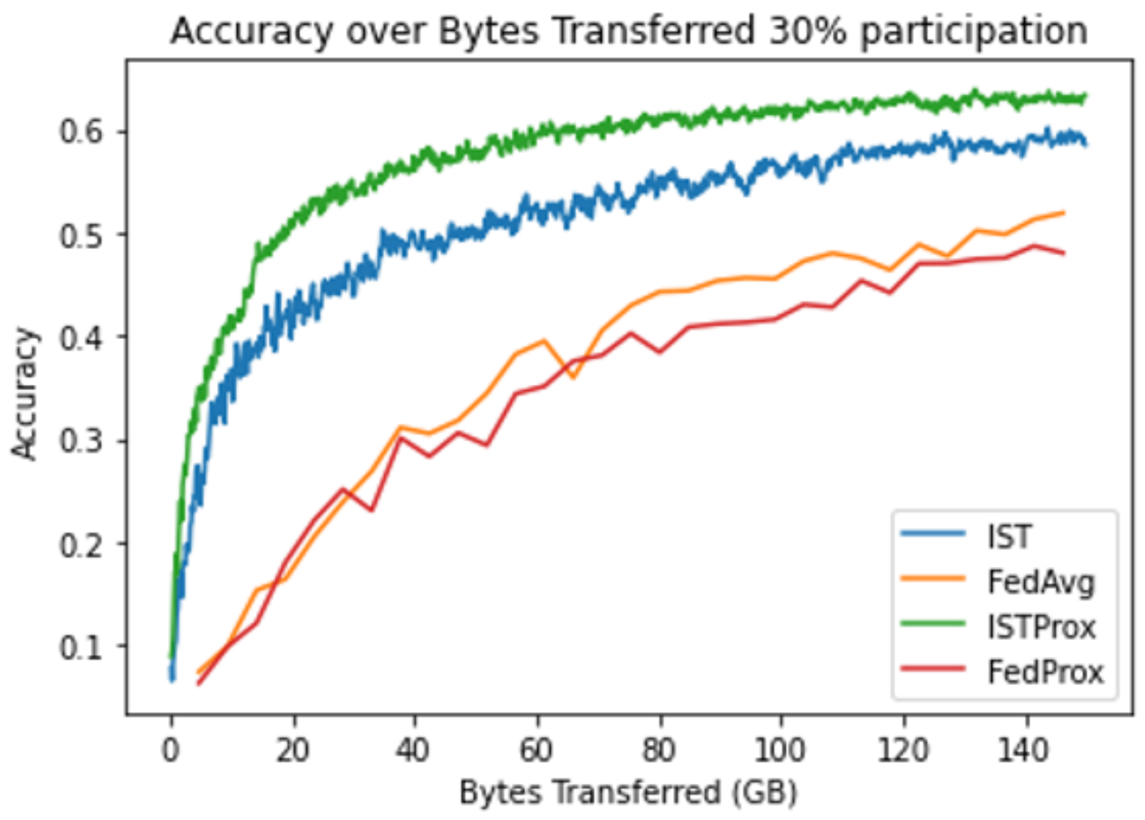


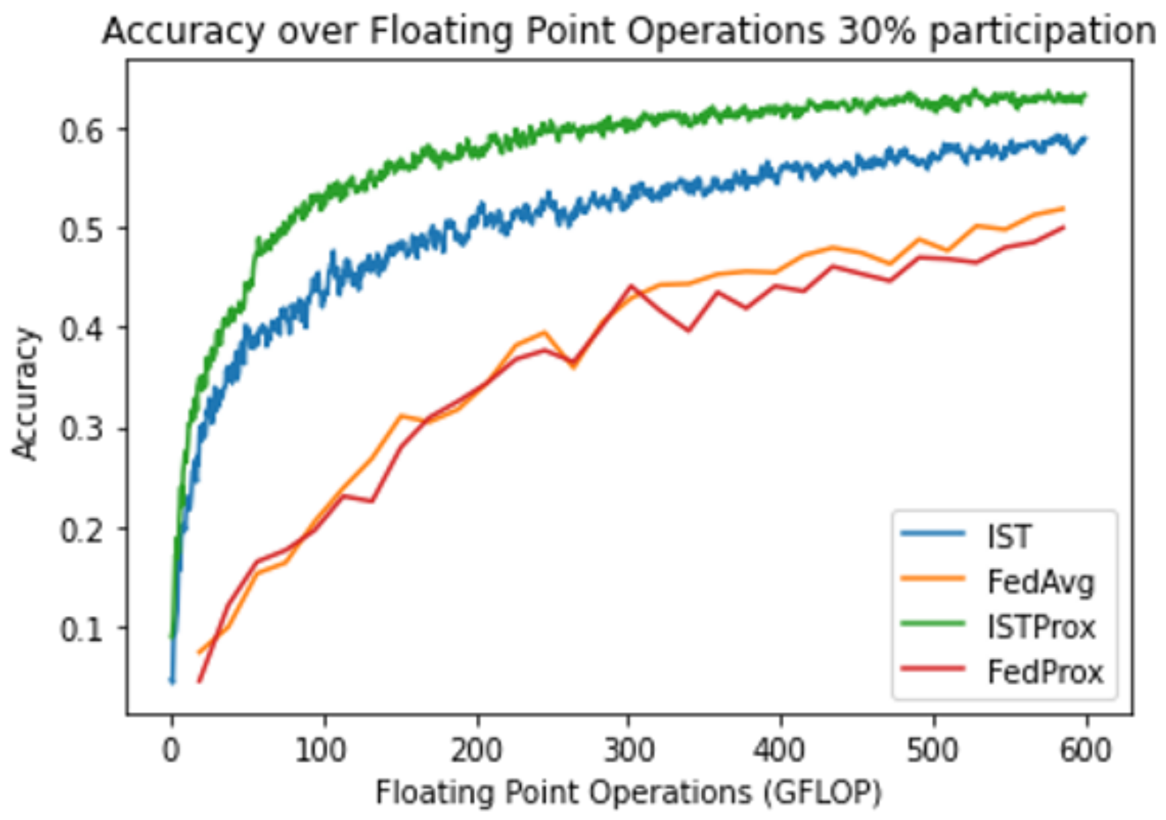
- IST + Asynchrony



Where to go from here?

- IST + FL



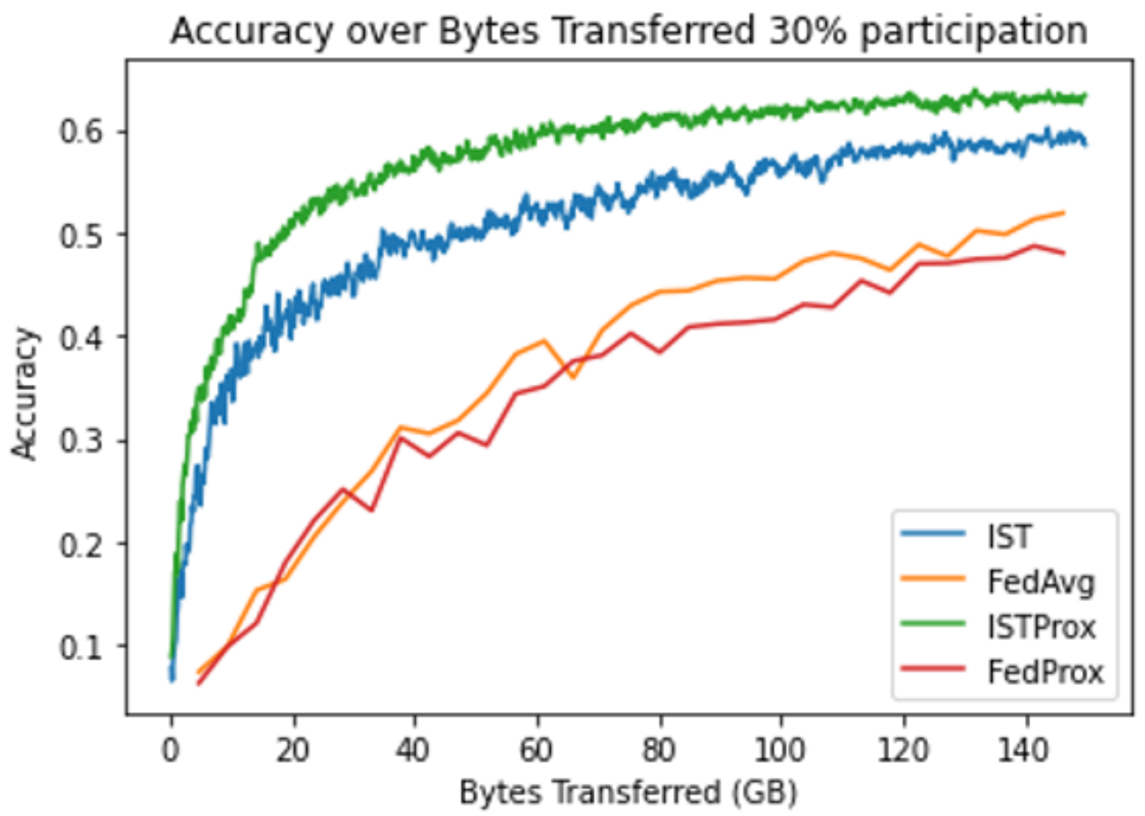


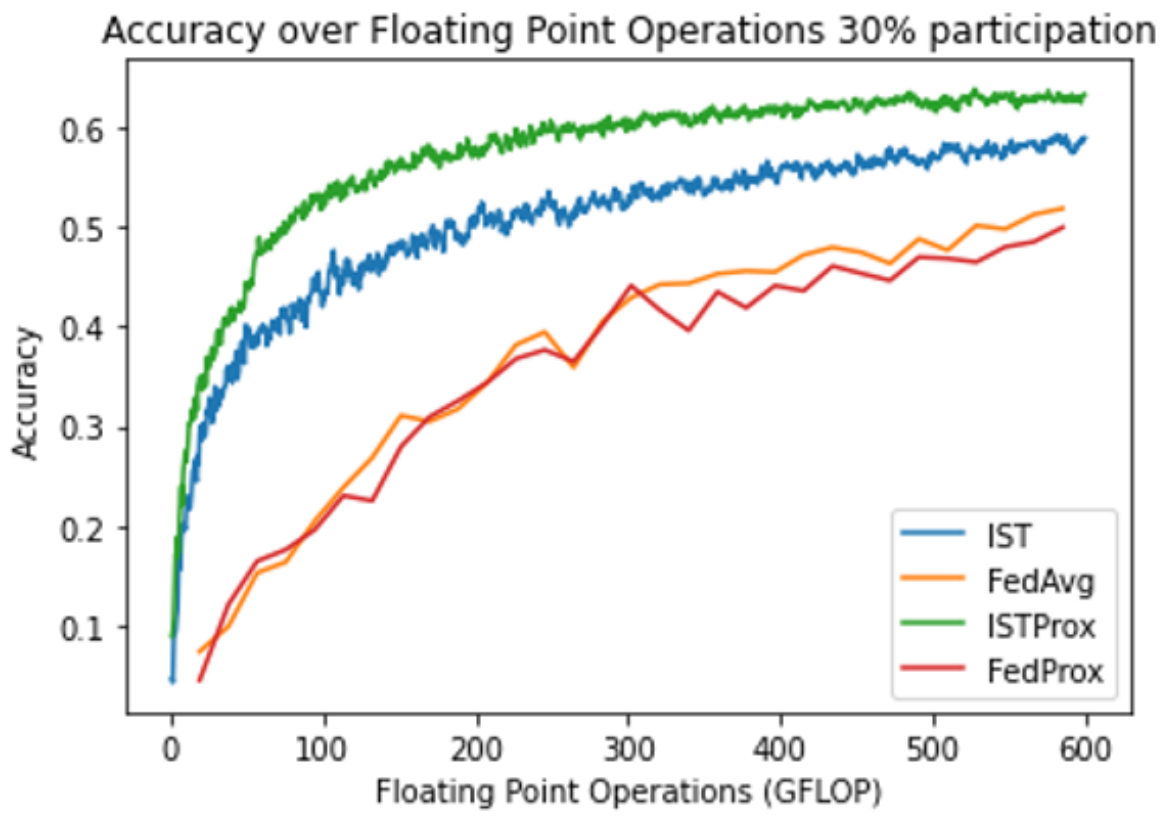
- IST + Asynchrony

$$
\begin{aligned}
\mathbb{E}_{\mathbf{M}_t}\left[\|\mathbf{u}_{t+1}-\mathbf{y}\|_2^2\right] \leq \left(1-\frac{\theta\eta\lambda_0}{4}\right)^t \|\mathbf{u}_0-\mathbf{y}\|_2^2 \\
+ O\left(\frac{\theta\eta\lambda_0^3\xi^2\kappa^2E^2}{n^2} + \frac{\xi^2(1-\xi)^2\theta\eta n^3\kappa^2 d}{m\lambda_0} + \frac{\eta^2\theta^2 n\kappa^2\lambda_0\xi^4E^2}{m^4} + \frac{\xi^2(1-\xi)^2\theta^2\eta^2n^2\kappa^2 d}{m^3\lambda_0}\right. \\
\left. + \frac{\xi^2(1-\xi)^2\theta^2\eta^2\kappa^2\lambda_0E^2}{m^3} + \frac{\xi^2(1-\xi)^2\theta^2\eta^2n^2\kappa^2 d}{m^2\lambda_0} + \frac{n\kappa^2\left(\theta-\xi^2\right)}{S}\right)
\end{aligned}
$$

Where to go from here?

- IST + FL

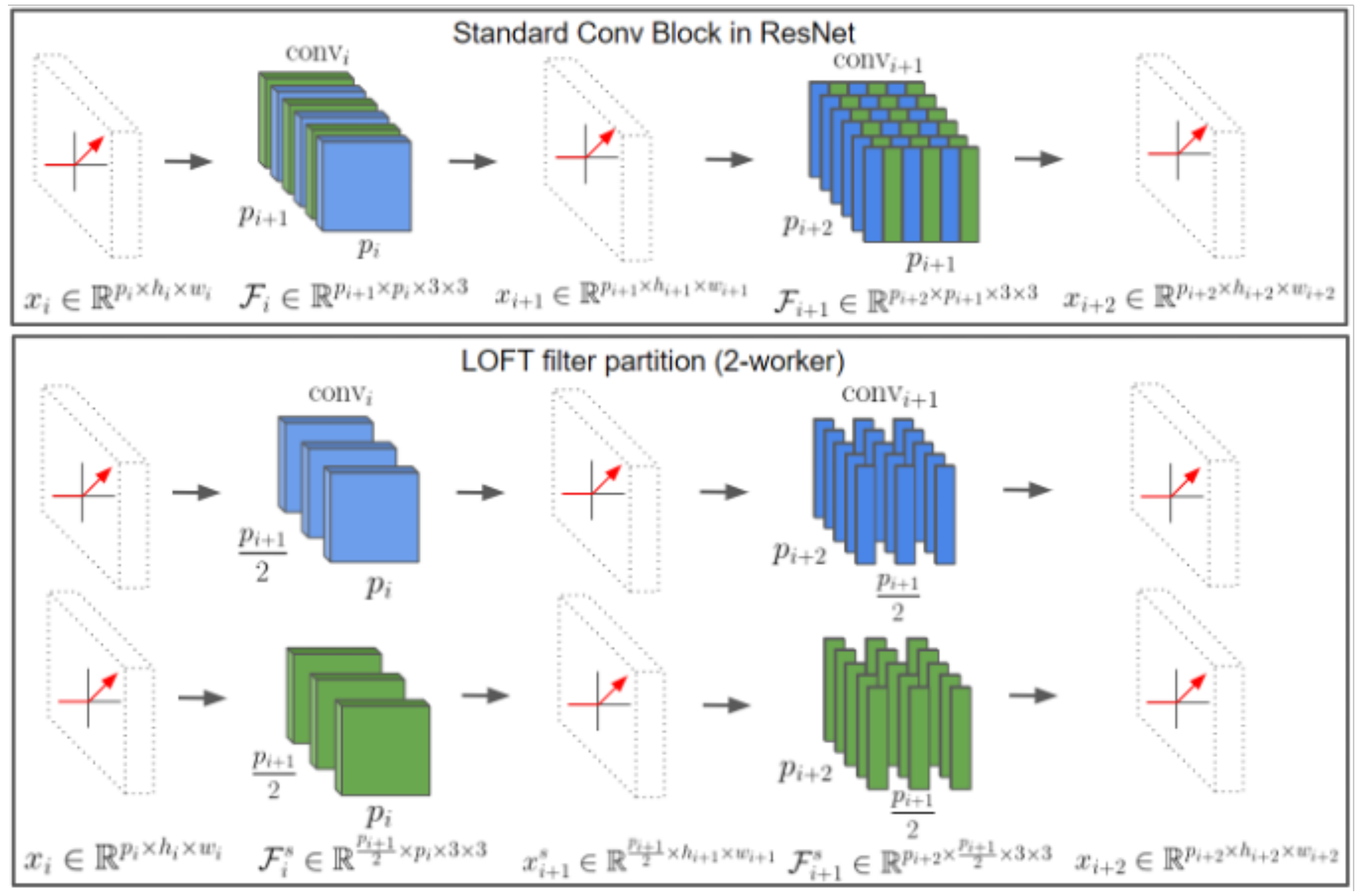




- IST + Asynchrony

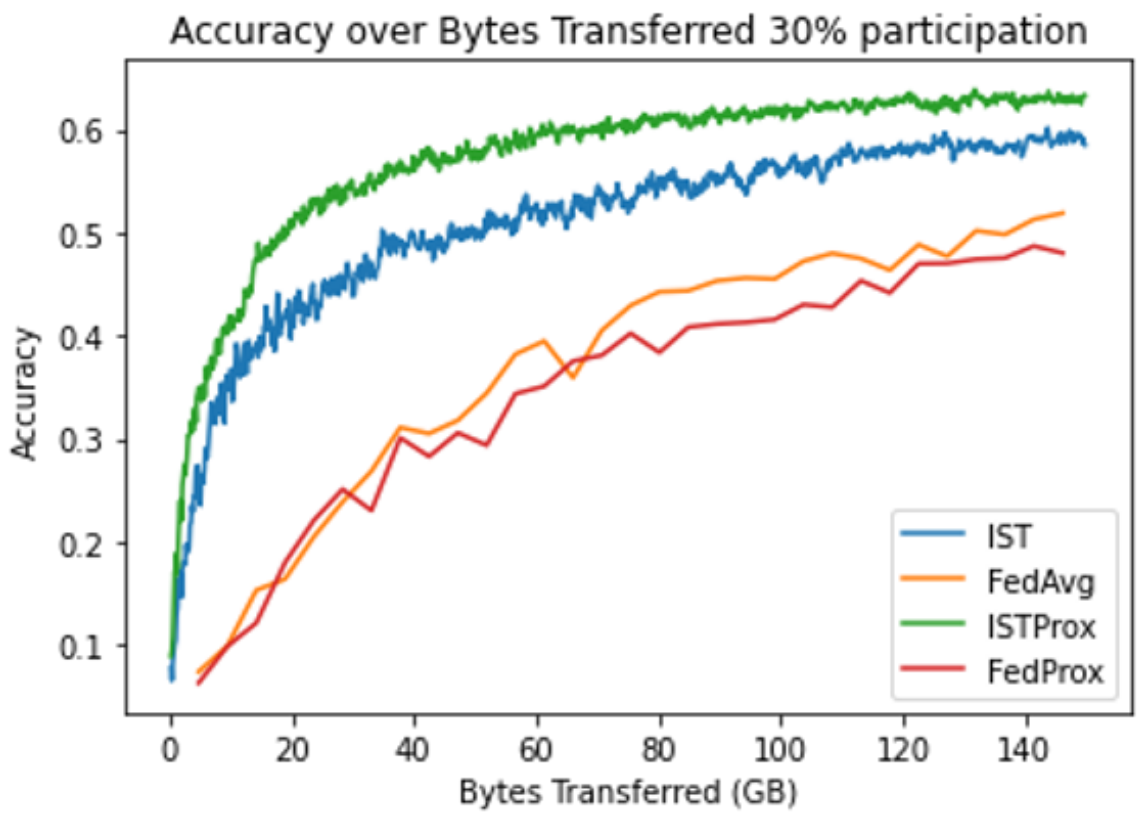
$$\mathbb{E}_{\mathbf{M}_t}\left[\|\mathbf{u}_{t+1}-\mathbf{y}\|_2^2\right] \leq \left(1-\frac{\theta\eta\lambda_0}{4}\right)^t \|\mathbf{u}_0-\mathbf{y}\|_2^2 + O\left(\frac{\theta\eta\lambda_0^3\xi^2\kappa^2E^2}{n^2} + \frac{\xi^2(1-\xi)^2\theta\eta n^3\kappa^2 d}{m\lambda_0} + \frac{\eta^2\theta^2 n\kappa^2\lambda_0\xi^4E^2}{m^4} + \frac{\xi^2(1-\xi)^2\theta^2\eta^2n^2\kappa^2d}{m^3\lambda_0} + \frac{\xi^2(1-\xi)^2\theta^2\eta^2\kappa^2\lambda_0E^2}{m^3} + \frac{\xi^2(1-\xi)^2\theta^2\eta^2n^2\kappa^2d}{m^2\lambda_0} + \frac{n\kappa^2\left(\theta-\xi^2\right)}{S}\right)$$

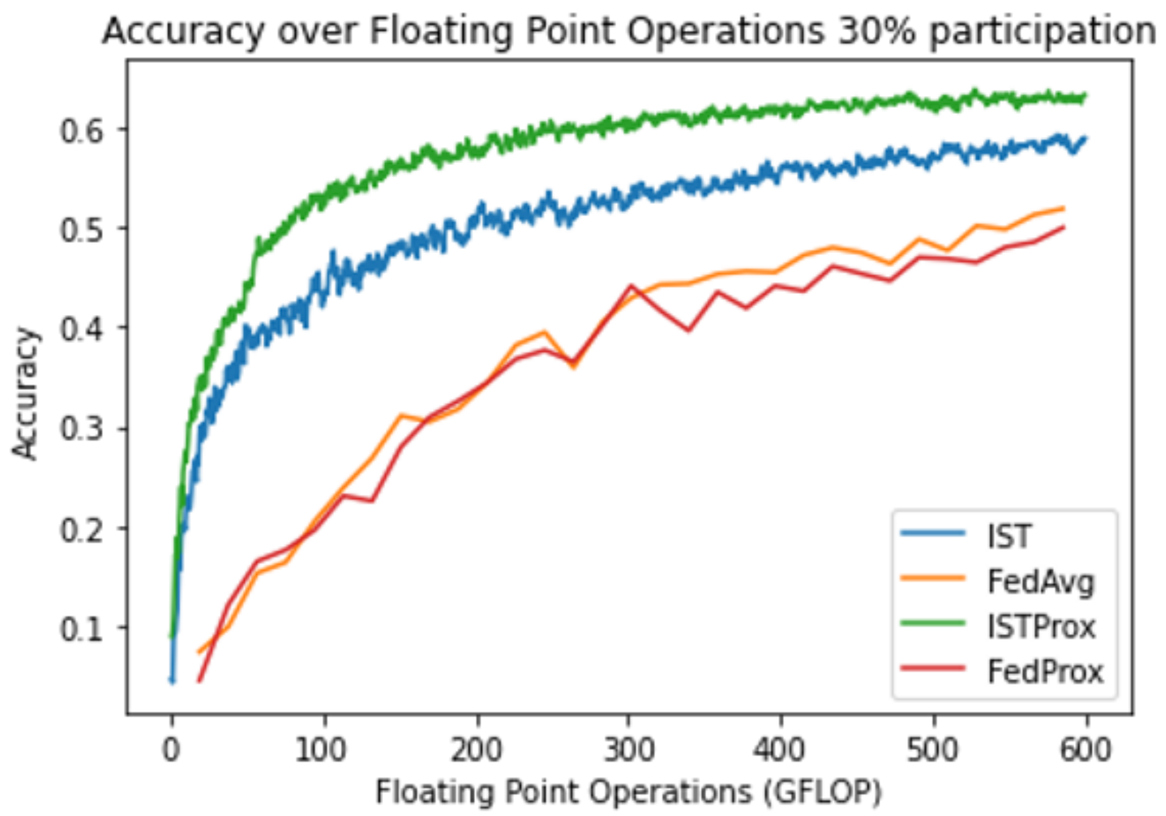
- IST + LTH



Where to go from here?

- IST + FL





- IST + Asynchrony

$$\mathbb{E}_{\mathbf{M}_t}\left[\|\mathbf{u}_{t+1}-\mathbf{y}\|_2^2\right] \le \left(1-\frac{\theta\eta\lambda_0}{4}\right)^t \|\mathbf{u}_0-\mathbf{y}\|_2^2 + O\left(\frac{\theta\eta\lambda_0^3\xi^2\kappa^2E^2}{n^2} + \frac{\xi^2(1-\xi)^2\theta\eta n^3\kappa^2 d}{m\lambda_0} + \frac{\eta^2\theta^2 n\kappa^2\lambda_0\xi^4E^2}{m^4} + \frac{\xi^2(1-\xi)^2\theta^2\eta^2n^2\kappa^2 d}{m^3\lambda_0} + \frac{\xi^2(1-\xi)^2\theta^2\eta^2\kappa^2\lambda_0E^2}{m^3} + \frac{\xi^2(1-\xi)^2\theta^2\eta^2n^2\kappa^2 d}{m^2\lambda_0} + \frac{n\kappa^2\left(\theta-\xi^2\right)}{S}\right)$$

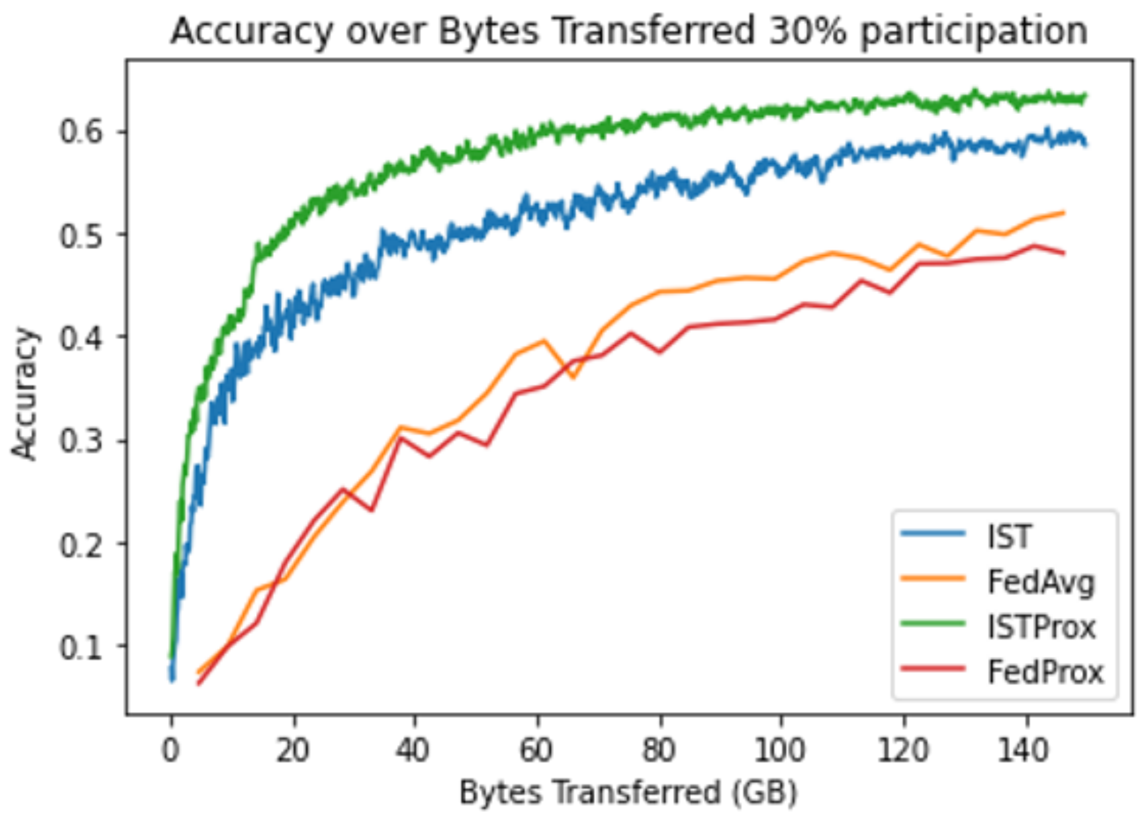
- IST + LTH

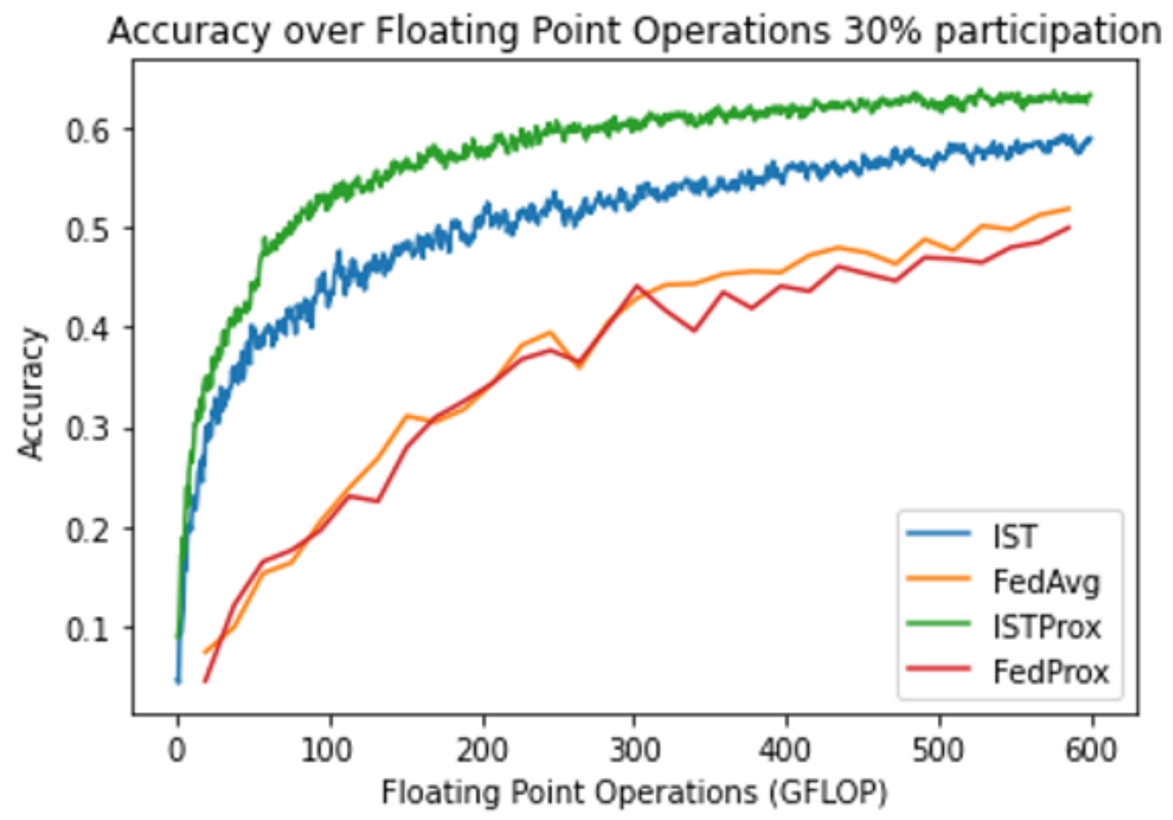
Theorem 3. *Let $f(\cdot,\cdot)$ be a one-hidden-layer CNN with the second layer weight fixed. Assume the number of hidden neurons satisfies $m = \Omega\left(\frac{n^4K^2}{\lambda_0^4\delta^2}\max\{n,d\}\right)$ and the step size satisfies $\eta = O\left(\frac{\lambda_0}{n^2}\right)$: Let Assumptions 1 and 2 be satisfied. Then, with probability at least $1-O(\delta)$ we have:*

$$\mathbb{E}_{[\mathbf{M}_T]}\left[\left\|\mathbf{W}_T-\hat{\mathbf{W}}_T\right\|_F^2\right] + \eta\sum_{t=0}^{T-1}\mathbb{E}_{[\mathbf{M}_T]}\left[\left\|f\left(\mathbf{X},\mathbf{W}_t\right)-f\left(\mathbf{X},\hat{\mathbf{W}}_t\right)\right\|_2^2\right] \le O\left(\frac{n^2\sqrt{d}}{\lambda_0^2\kappa m^{\frac{1}{4}}\sqrt{\delta}} + \frac{2\eta^2T\theta^2(1-\xi)\lambda_0}{S}\right).$$

Where to go from here?

- IST + FL





- IST + Asynchrony

$$
\begin{aligned}
\mathbb{E}_{\mathbf{M}_t}\left[\|\mathbf{u}_{t+1}-\mathbf{y}\|_2^2\right] \leq \left(1-\frac{\theta\eta\lambda_0}{4}\right)^t \|\mathbf{u}_0-\mathbf{y}\|_2^2 \\
+ O\left(\frac{\theta\eta\lambda_0^3\xi^2\kappa^2E^2}{n^2} + \frac{\xi^2(1-\xi)^2\theta\eta n^3\kappa^2 d}{m\lambda_0} + \frac{\eta^2\theta^2 n\kappa^2\lambda_0\xi^4E^2}{m^4} + \frac{\xi^2(1-\xi)^2\theta^2\eta^2 n^2\kappa^2 d}{m^3\lambda_0}\right. \\
\left. + \frac{\xi^2(1-\xi)^2\theta^2\eta^2\kappa^2\lambda_0E^2}{m^3} + \frac{\xi^2(1-\xi)^2\theta^2\eta^2 n^2\kappa^2 d}{m^2\lambda_0} + \frac{n\kappa^2\left(\theta-\xi^2\right)}{S}\right)
\end{aligned}
$$

- IST + LTH

Theorem 3. *Let $f(\cdot,\cdot)$ be a one-hidden-layer CNN with the second layer weight fixed. Assume the number of hidden neurons satisfies $m = \Omega\left(\frac{n^4K^2}{\lambda_0^4\delta^2}\max\{n,d\}\right)$ and the step size satisfies $\eta = O\left(\frac{\lambda_0}{n^2}\right)$: Let Assumptions 1 and 2 be satisfied. Then, with probability at least $1-O(\delta)$ we have:*

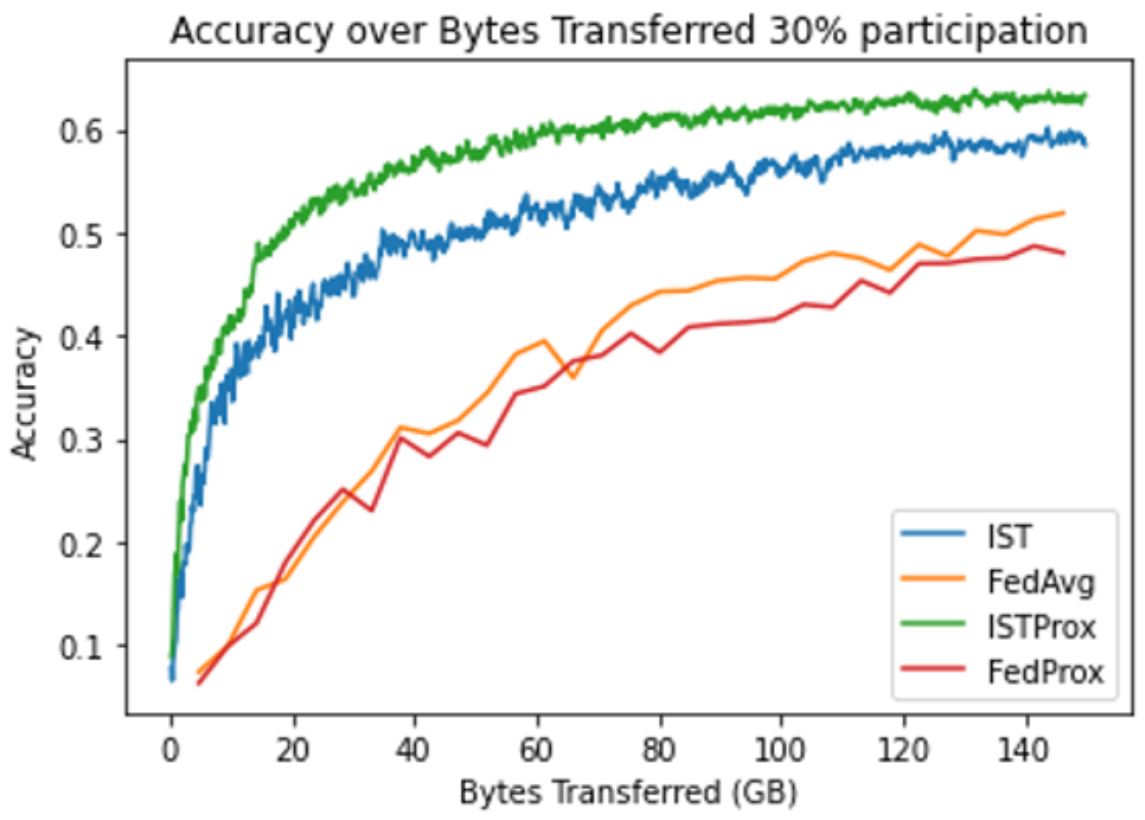
$$
\mathbb{E}_{[\mathbf{M}_T]}\left[\left\|\mathbf{W}_T-\hat{\mathbf{W}}_T\right\|_F^2\right] + \eta\sum_{t=0}^{T-1}\mathbb{E}_{[\mathbf{M}_T]}\left[\left\|f(\mathbf{X},\mathbf{W}_t) - f\left(\mathbf{X},\hat{\mathbf{W}}_t\right)\right\|_2^2\right] \leq O\left(\frac{n^2\sqrt{d}}{\lambda_0^2\kappa m^{\frac{1}{4}}\sqrt{\delta}} + \frac{2\eta^2T\theta^2(1-\xi)\lambda_0}{S}\right).
$$

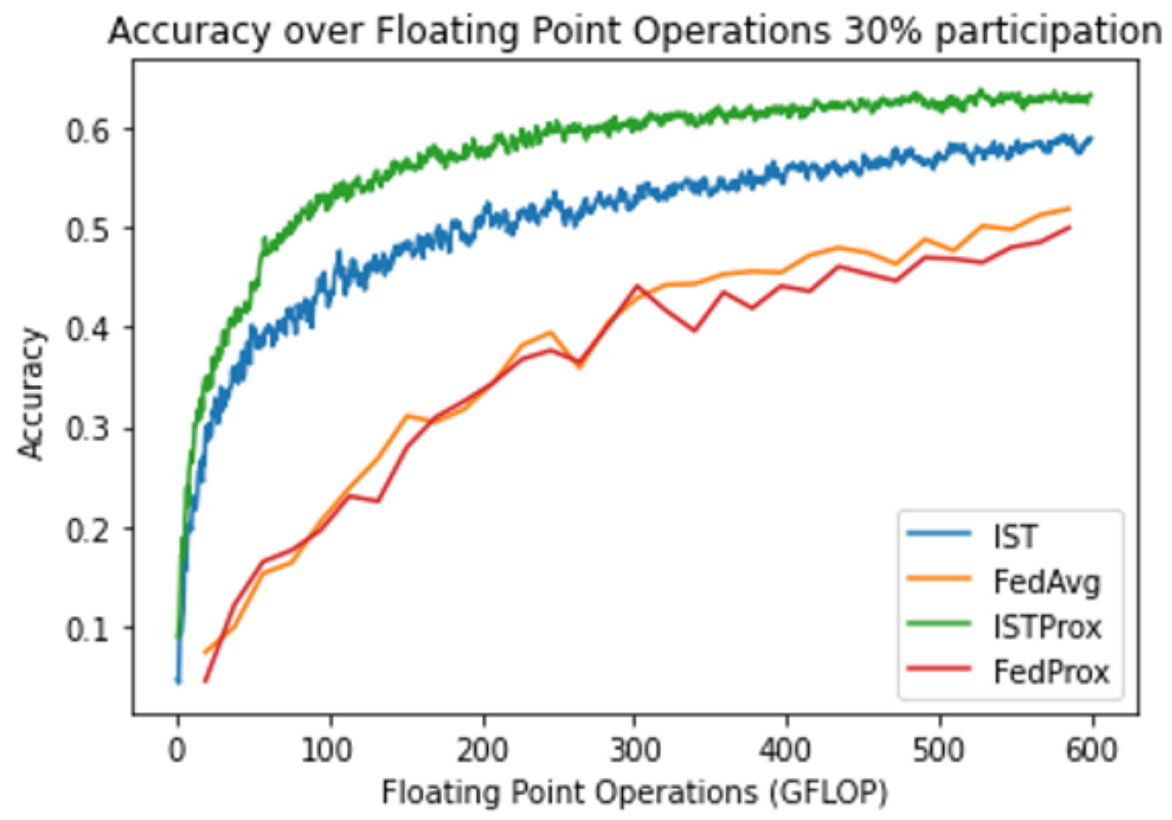
- IST + modern NNs (Transformers)

(Ongoing)

Where to go from here?

- IST + FL





- IST + Asynchrony

$$\mathbb{E}_{\mathbf{M}_t}\left[\|\mathbf{u}_{t+1}-\mathbf{y}\|_2^2\right] \leq \left(1-\frac{\theta\eta\lambda_0}{4}\right)^t \|\mathbf{u}_0-\mathbf{y}\|_2^2 + O\left(\frac{\theta\eta\lambda_0^3\xi^2\kappa^2E^2}{n^2} + \frac{\xi^2(1-\xi)^2\theta\eta n^3\kappa^2 d}{m\lambda_0} + \frac{\eta^2\theta^2 n\kappa^2\lambda_0\xi^4E^2}{m^4} + \frac{\xi^2(1-\xi)^2\theta^2\eta^2n^2\kappa^2 d}{m^3\lambda_0} + \frac{\xi^2(1-\xi)^2\theta^2\eta^2\kappa^2\lambda_0E^2}{m^3} + \frac{\xi^2(1-\xi)^2\theta^2\eta^2n^2\kappa^2 d}{m^2\lambda_0} + \frac{n\kappa^2\left(\theta-\xi^2\right)}{S}\right)$$

- IST + LTH

Theorem 3. *Let $f(\cdot,\cdot)$ be a one-hidden-layer CNN with the second layer weight fixed. Assume the number of hidden neurons satisfies $m = \Omega\left(\frac{n^4K^2}{\lambda_0^4\delta^2}\max\{n,d\}\right)$ and the step size satisfies $\eta = O\left(\frac{\lambda_0}{n^2}\right)$: Let Assumptions 1 and 2 be satisfied. Then, with probability at least $1-O(\delta)$ we have:*

$$\mathbb{E}_{[\mathbf{M}_T]}\left[\left\|\mathbf{W}_T-\hat{\mathbf{W}}_T\right\|_F^2\right] + \eta\sum_{t=0}^{T-1}\mathbb{E}_{[\mathbf{M}_T]}\left[\left\|f(\mathbf{X},\mathbf{W}_t)-f\left(\mathbf{X},\hat{\mathbf{W}}_t\right)\right\|_2^2\right] \leq O\left(\frac{n^2\sqrt{d}}{\lambda_0^2\kappa m^{\frac{1}{4}}\sqrt{\delta}} + \frac{2\eta^2T\theta^2(1-\xi)\lambda_0}{S}\right).$$

- IST + modern NNs (Transformers)

(Ongoing)

Thank you!