IntSGD: Adaptive Floatless Compression of Stochastic Gradients

Konstantin Mishchenko, Bokun Wang,
Dmitry Kovalev, Peter Richtárik

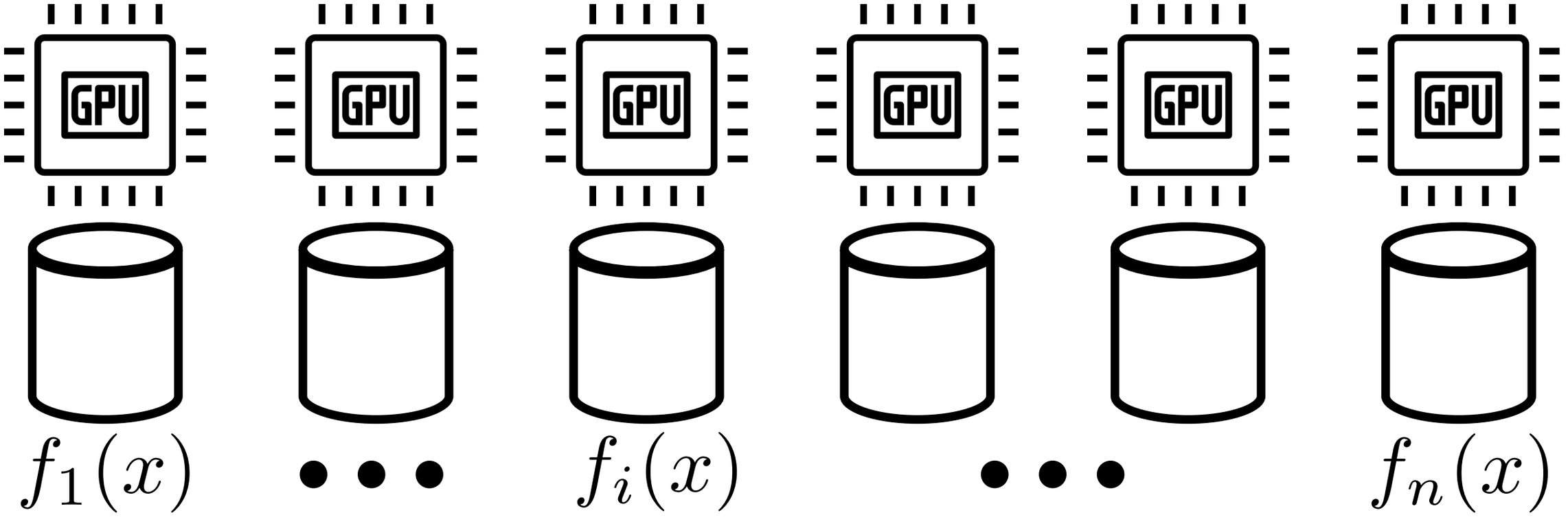
ICLR

Talk plan

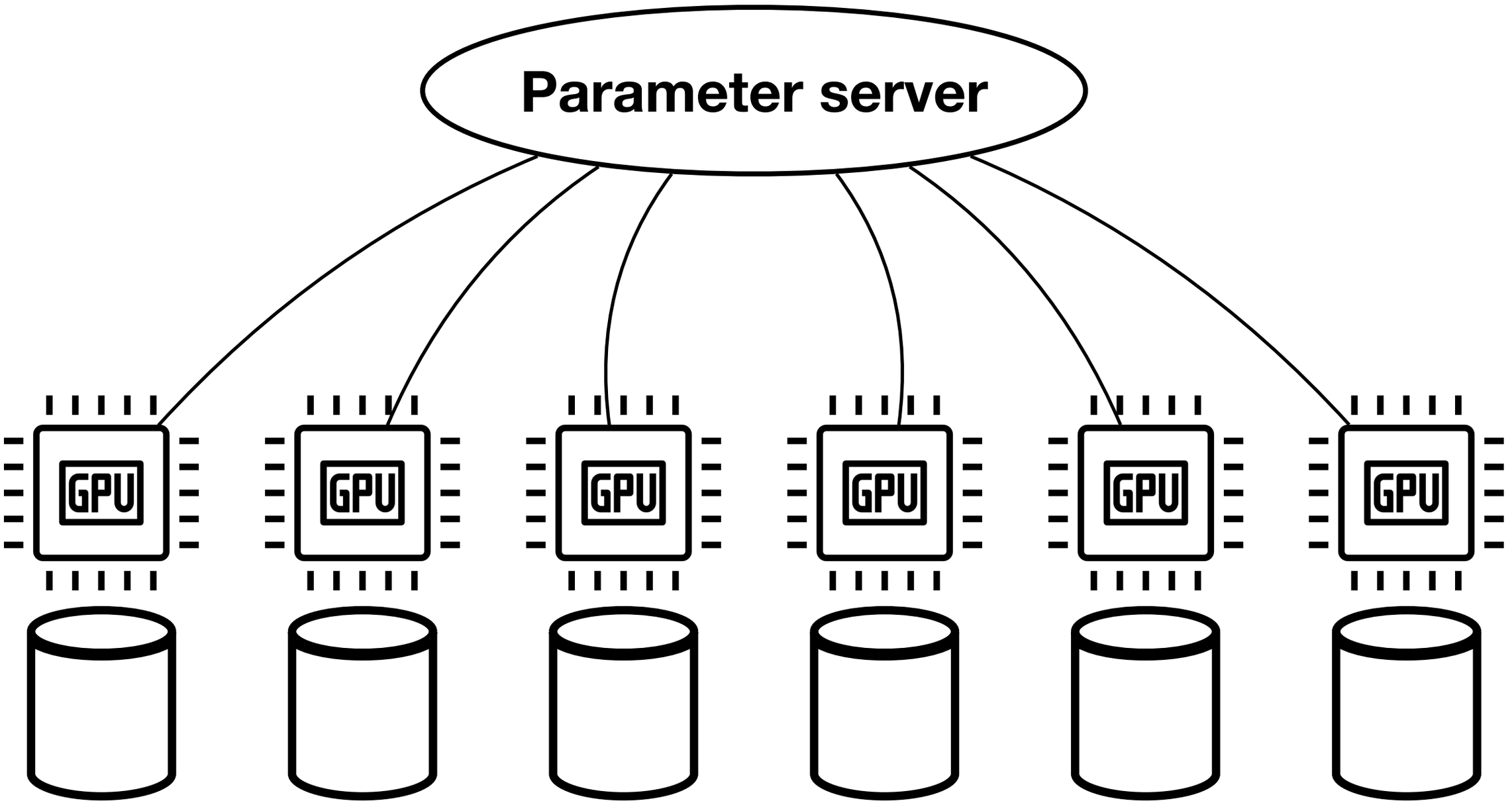
1. Motivation
2. Integer compression
3. Theory
4. Numerical results
5. Summary

Distributed training

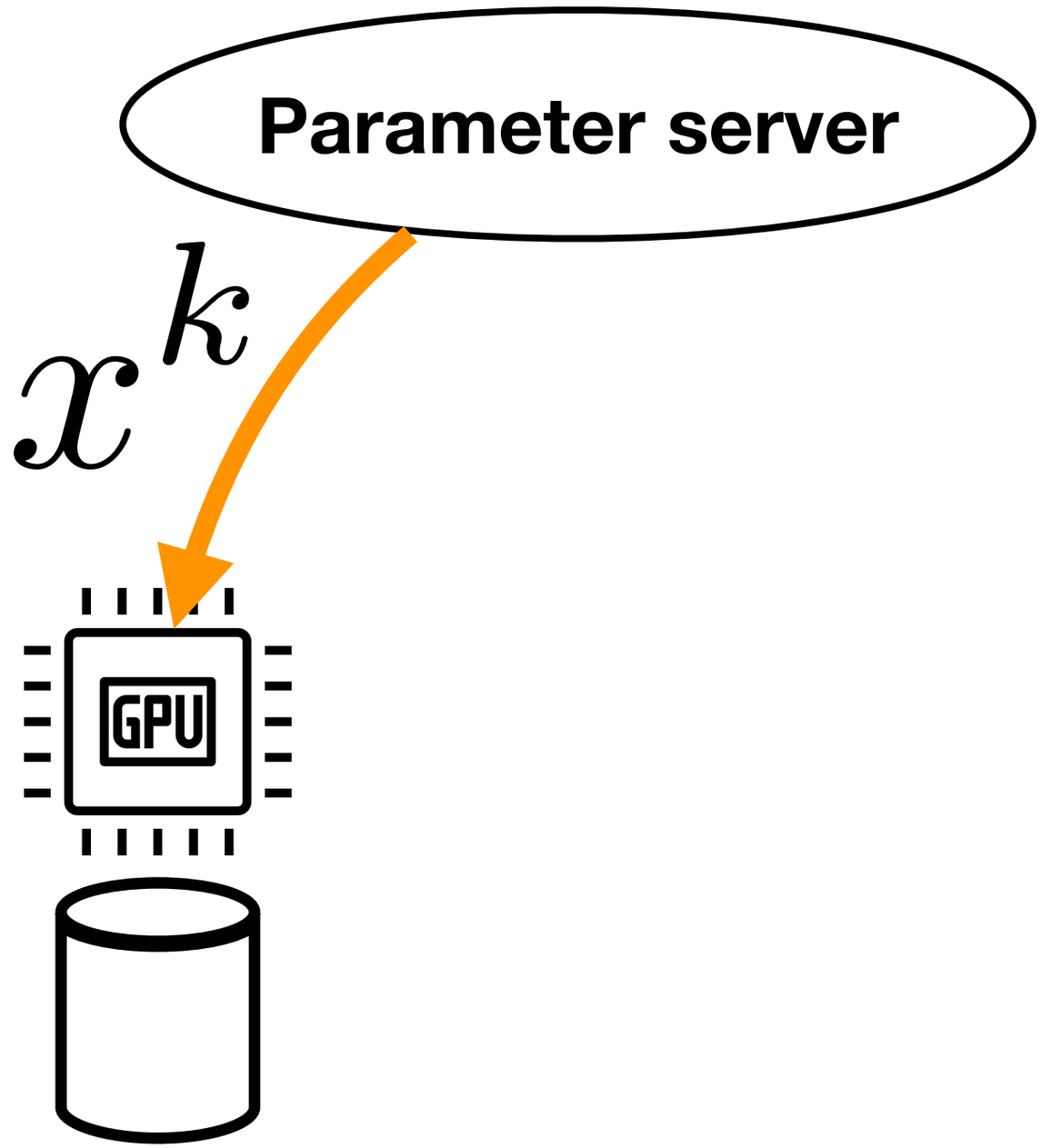
$$\min_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$



Distributed training

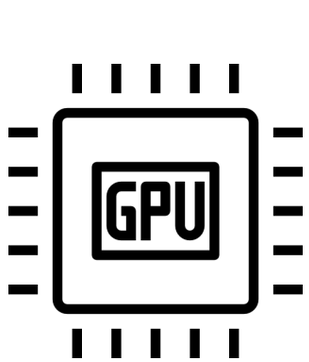


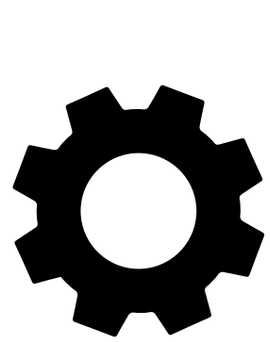
Distributed training



Distributed training

Parameter server

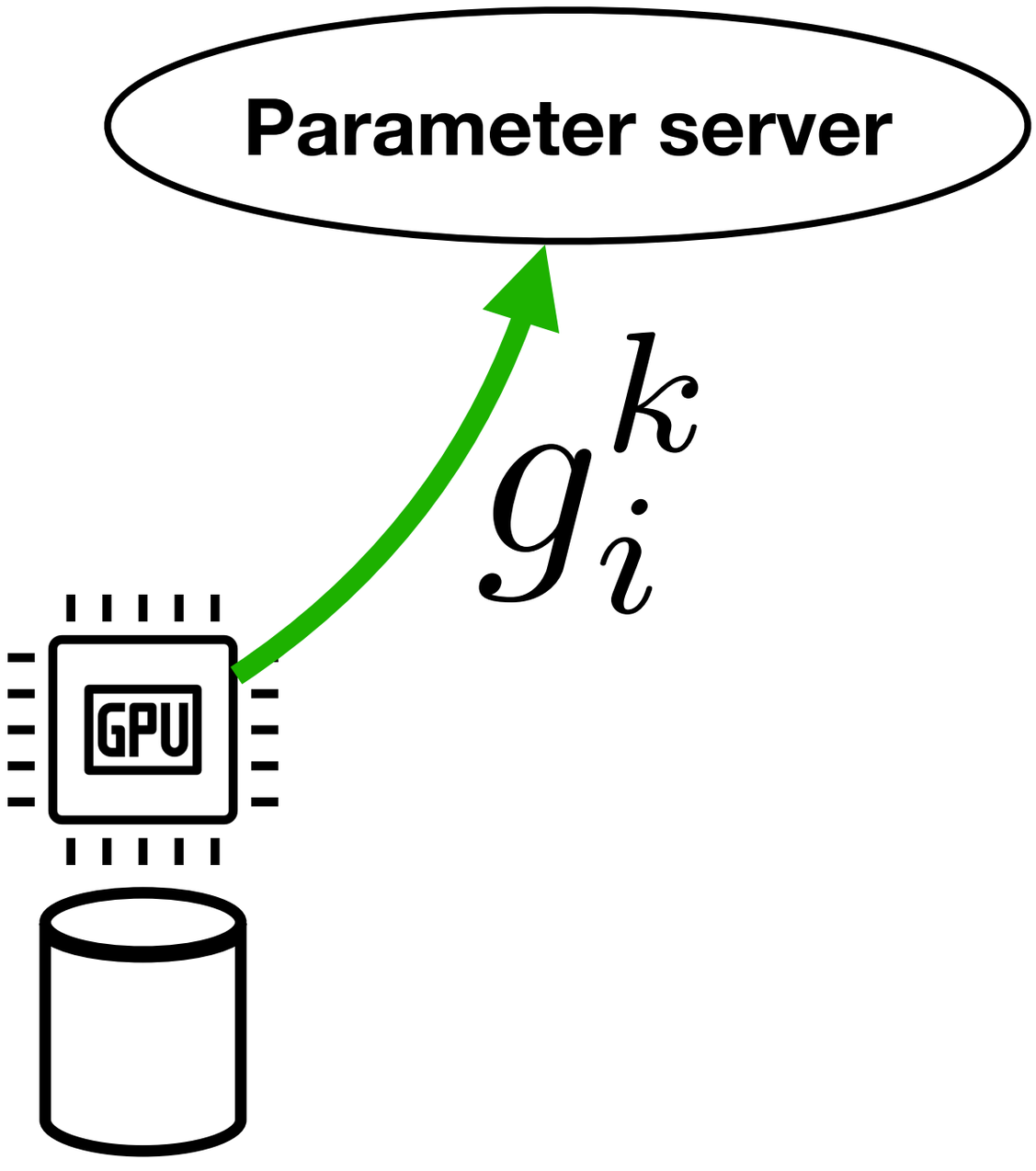




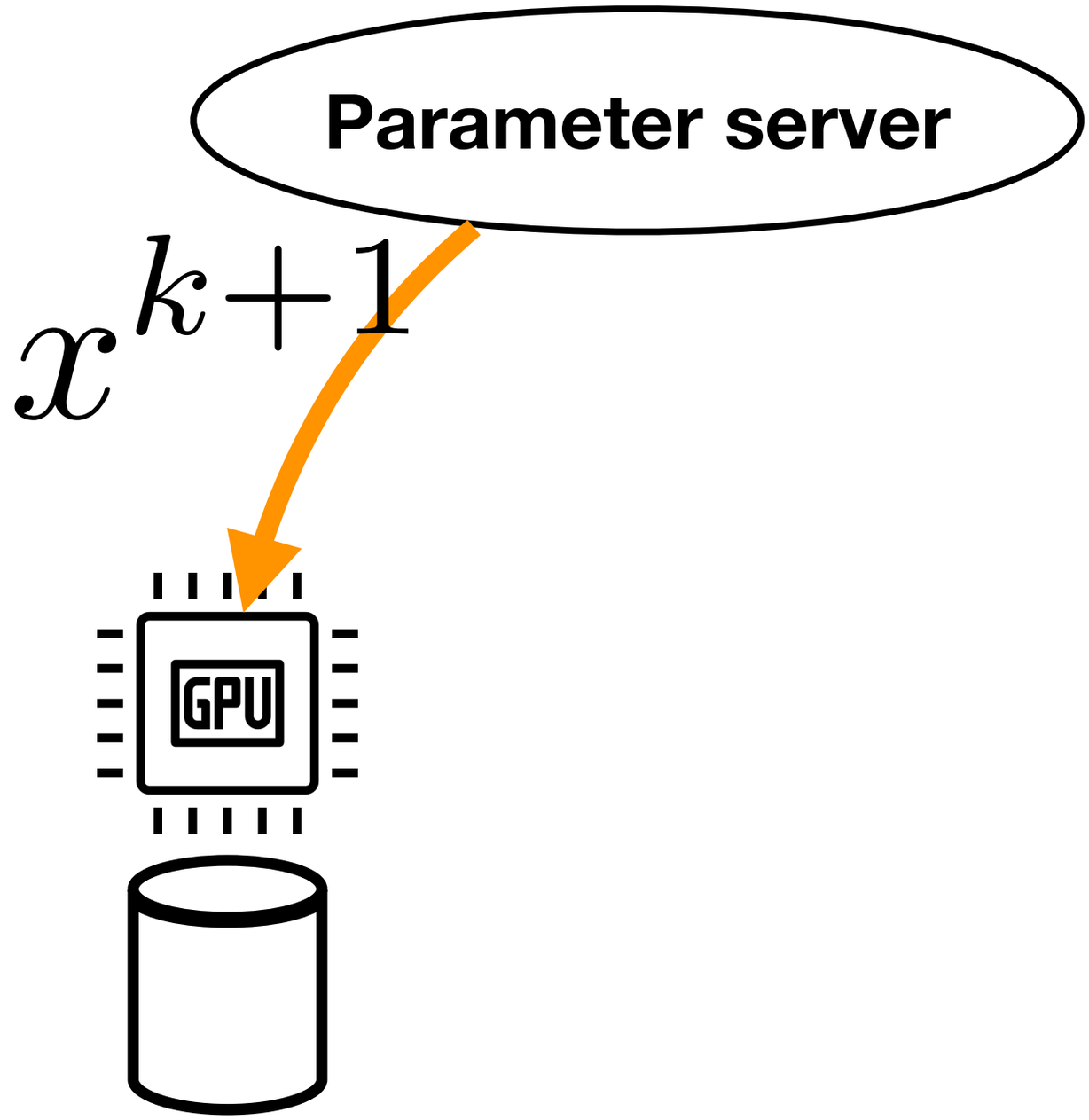
$$g_i^k$$

$$\mathbb{E}[g_i^k] = \nabla f_i(x^k)$$

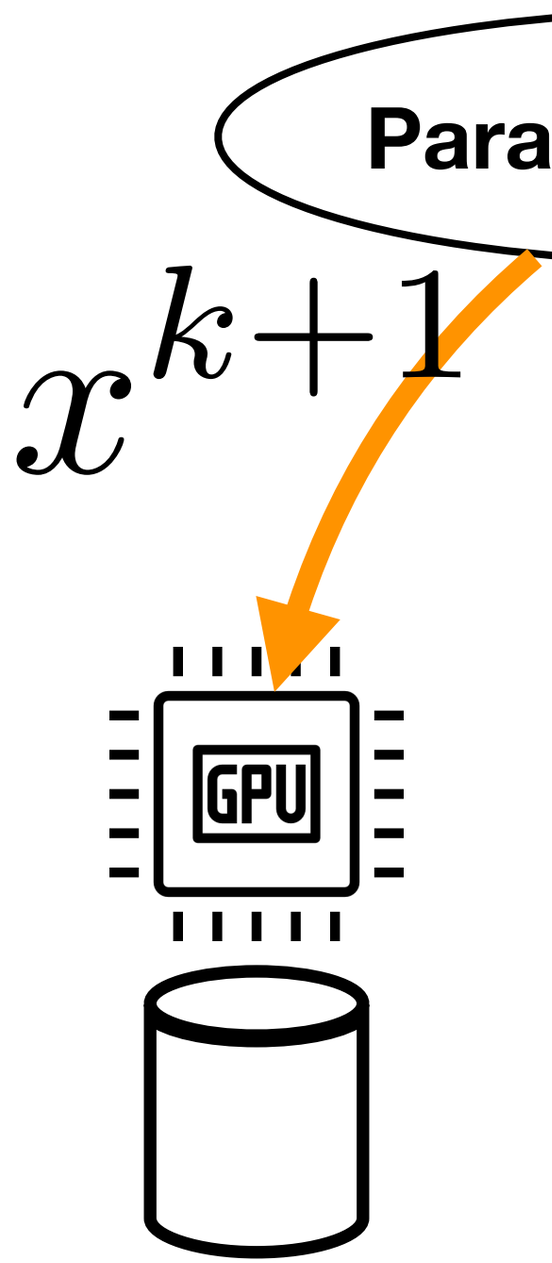
Distributed training



Distributed training

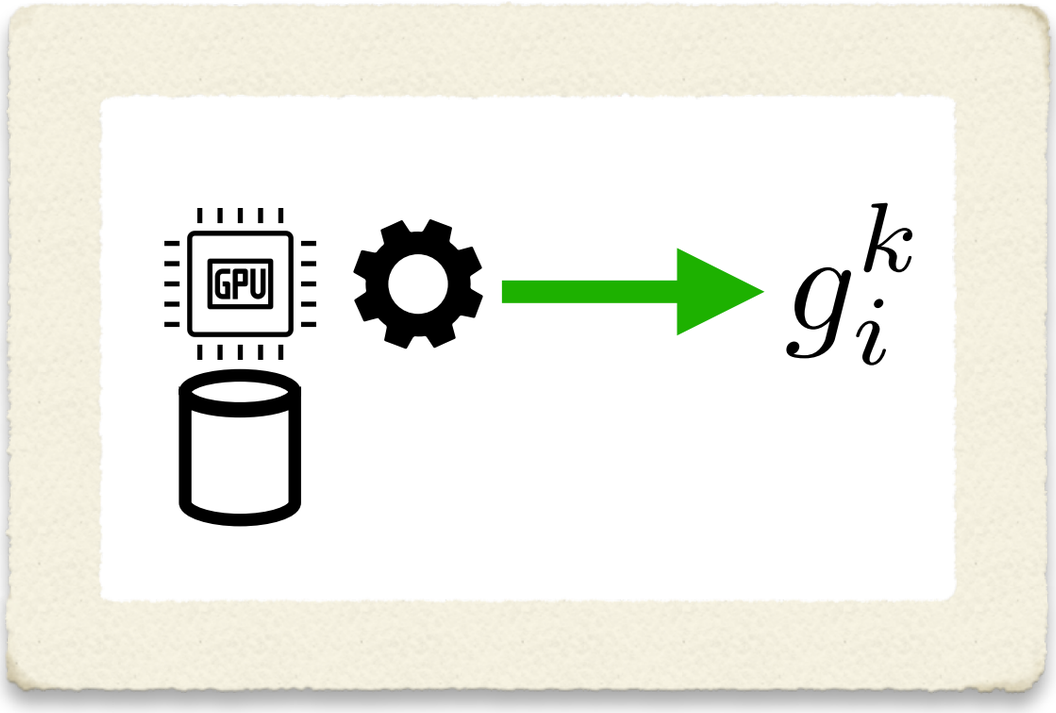


Distributed training



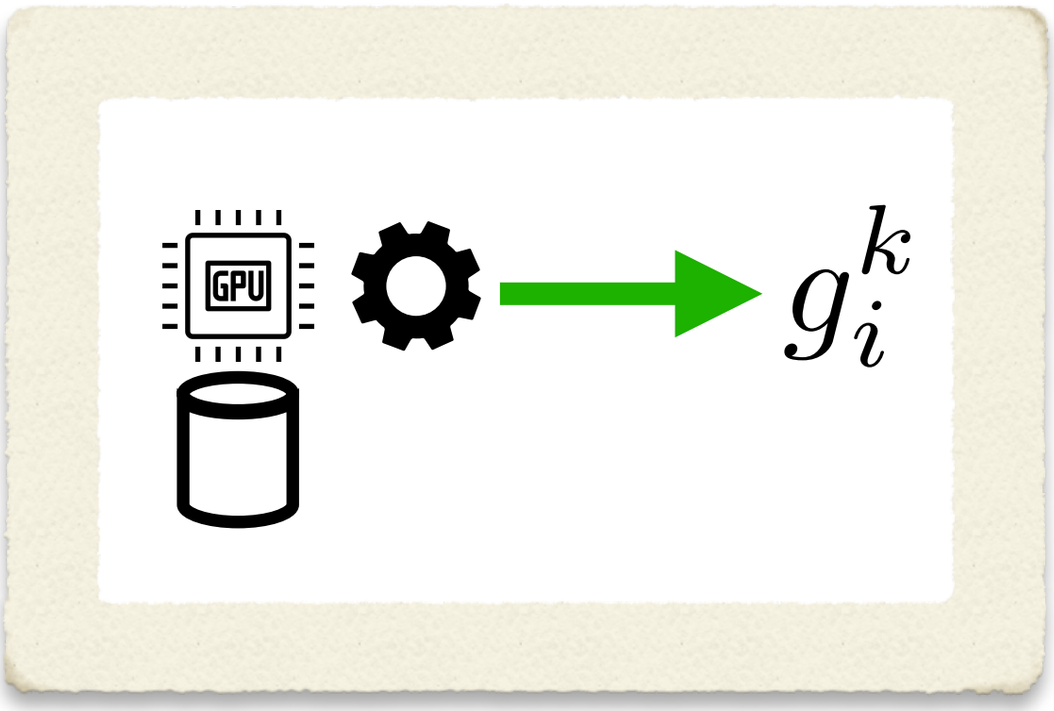
$$x^{k+1} = x^k - \eta_k \frac{1}{n} \sum_{i=1}^{n} g_i^k$$

Communication bottleneck

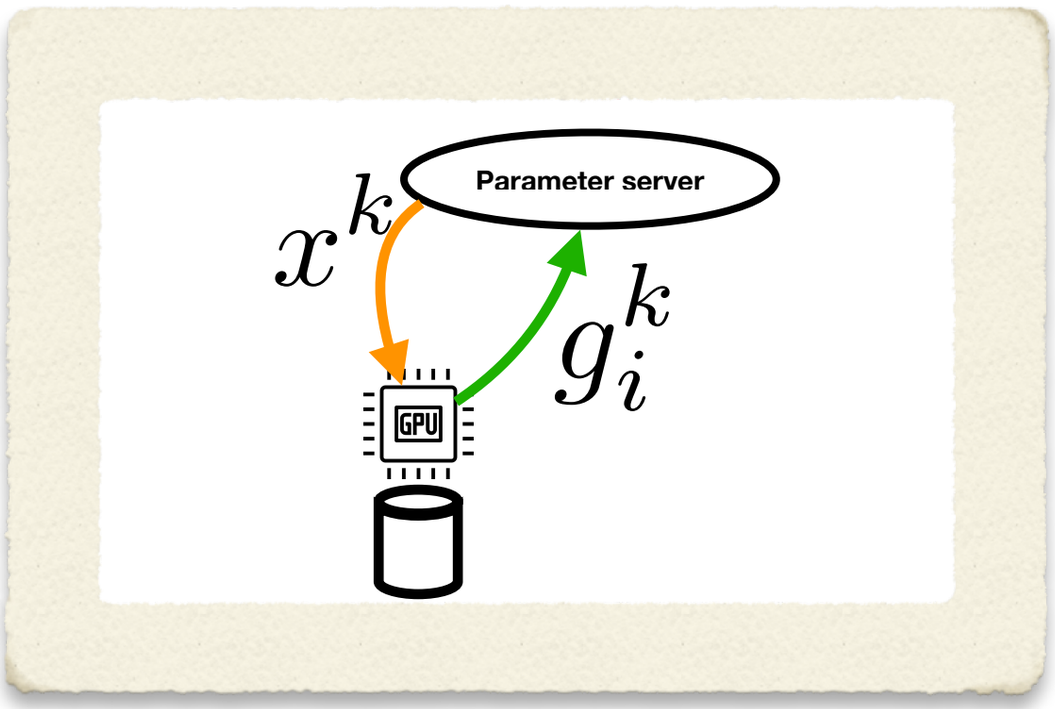


Fast computation

Communication bottleneck

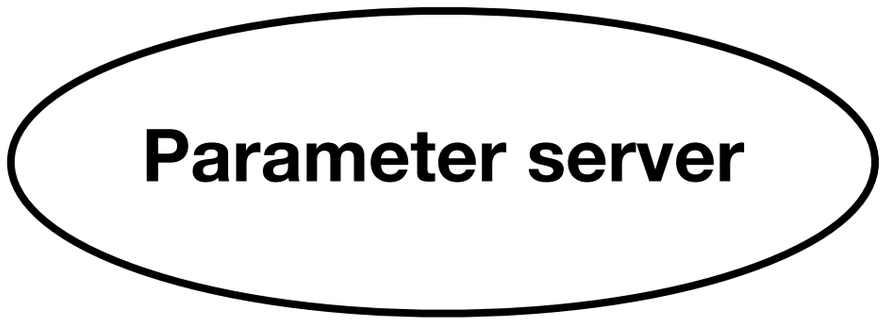


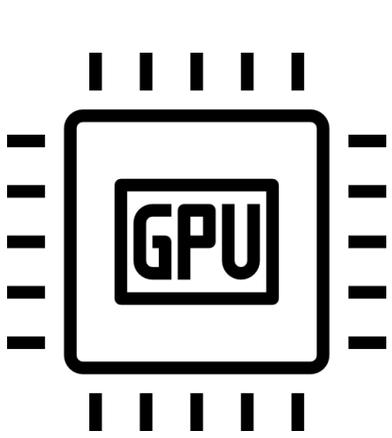
Fast computation



Slow communication

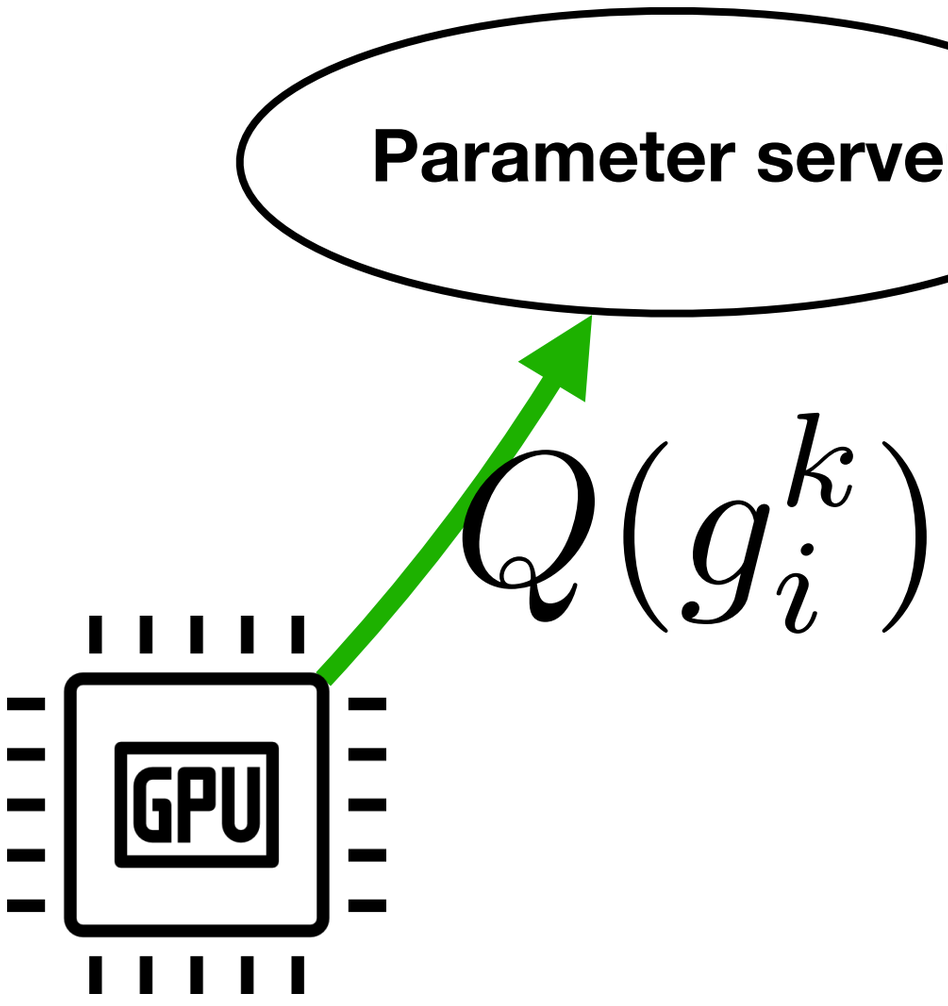
Solution:
Gradient Compression



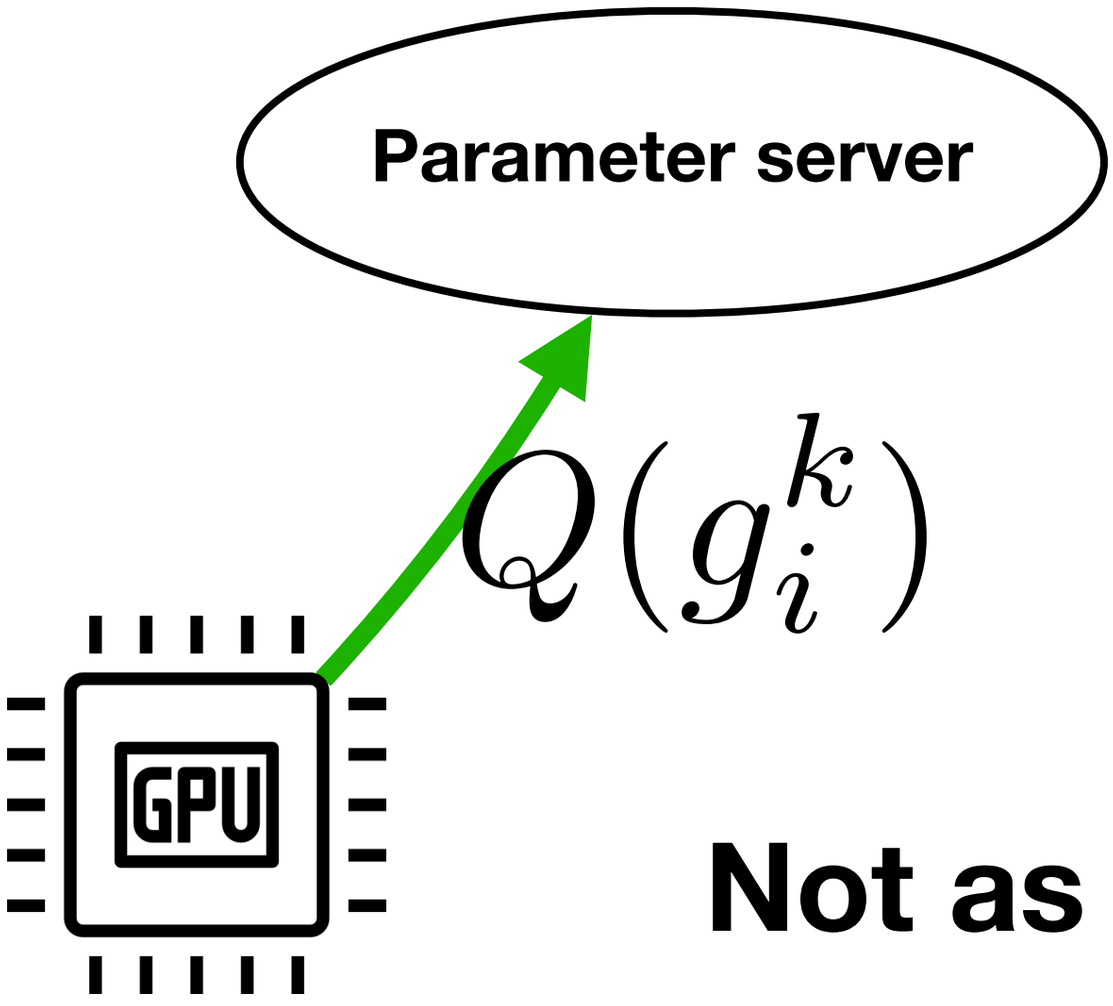


$$g_i^k \longrightarrow Q(g_i^k)$$

Solution:
Gradient Compression



Solution:
Gradient Compression



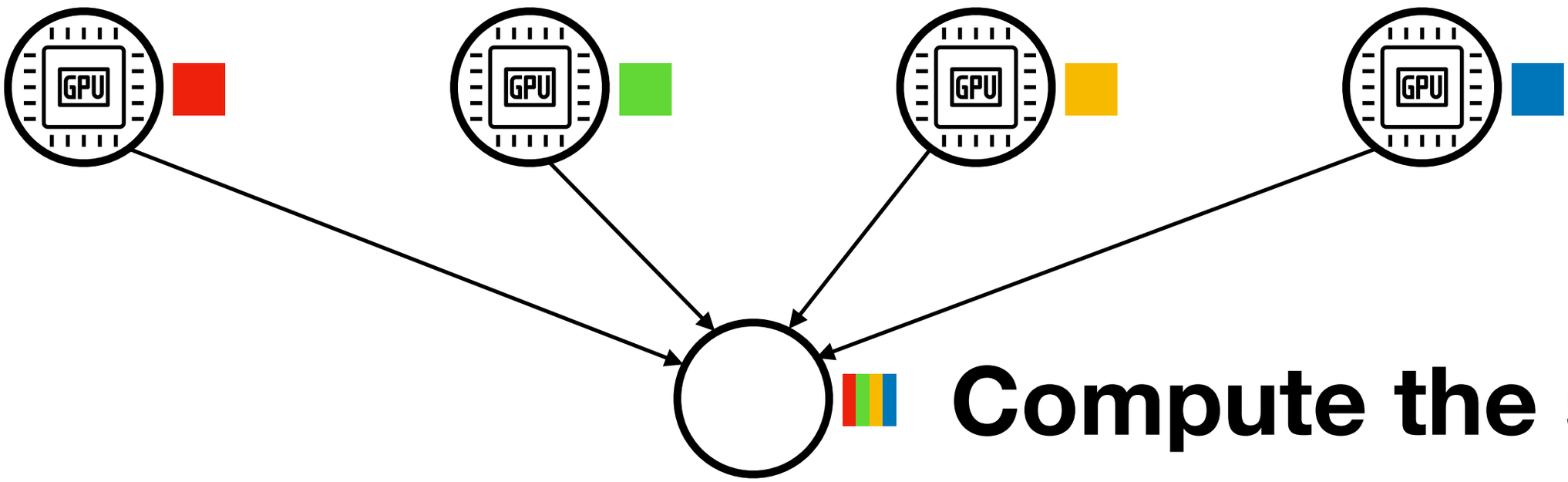
Not as straightforward as may seem

All-reduce vs. gather

All-reduce:

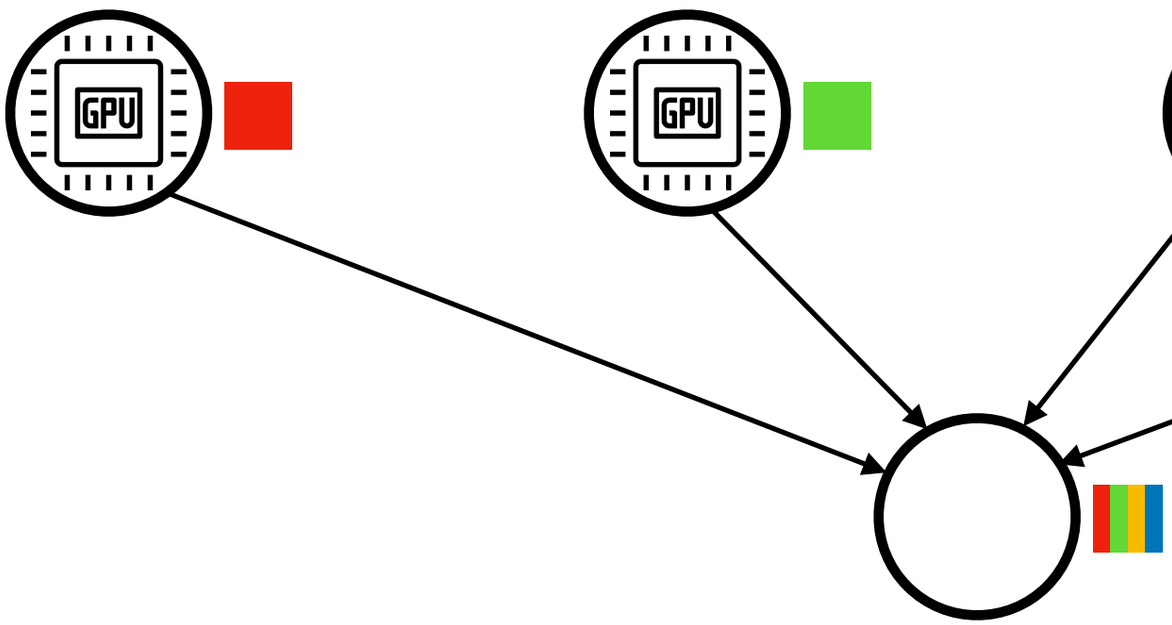
All-reduce vs. gather

All-reduce:



All-reduce vs. gather

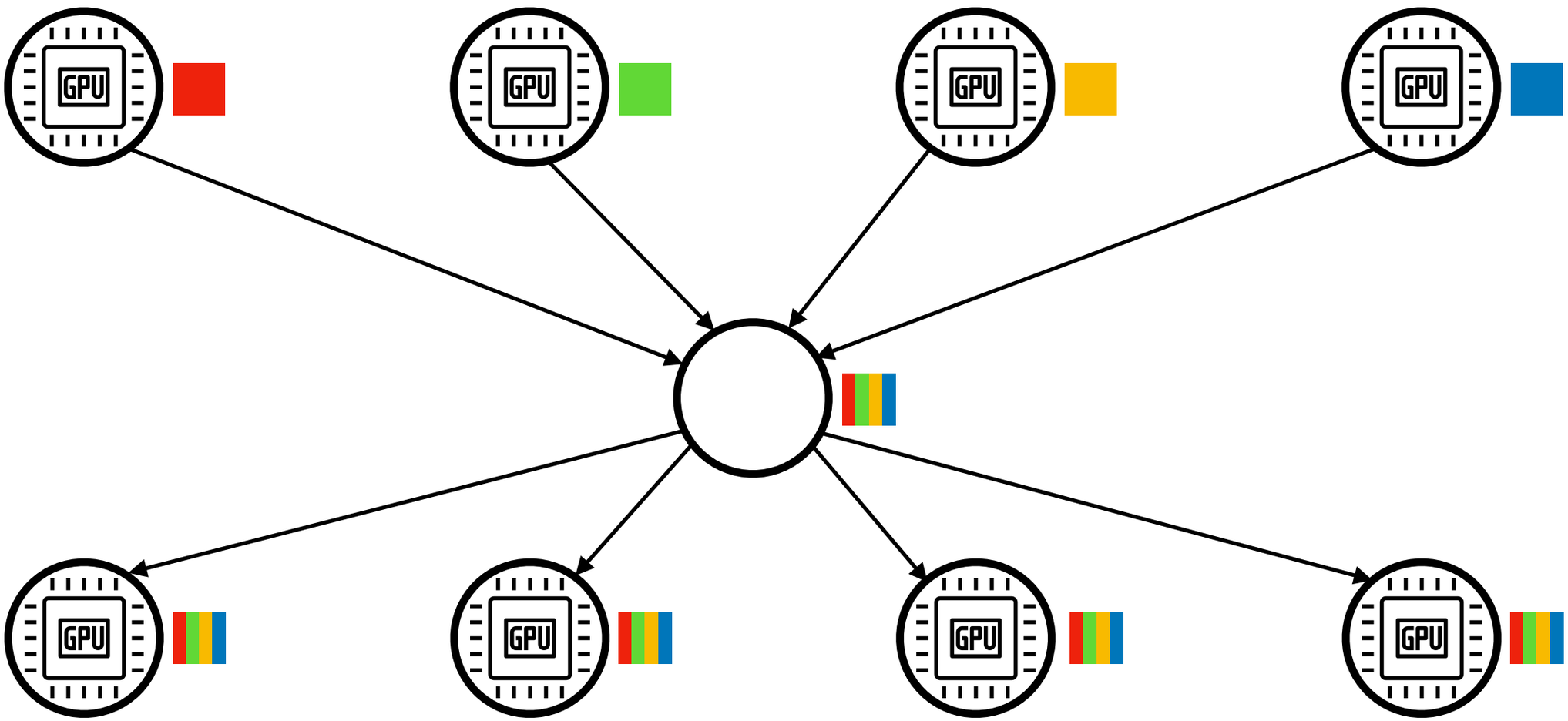
All-reduce:



Compute the sum (hard to compress because of in-place summation)

All-reduce vs. gather

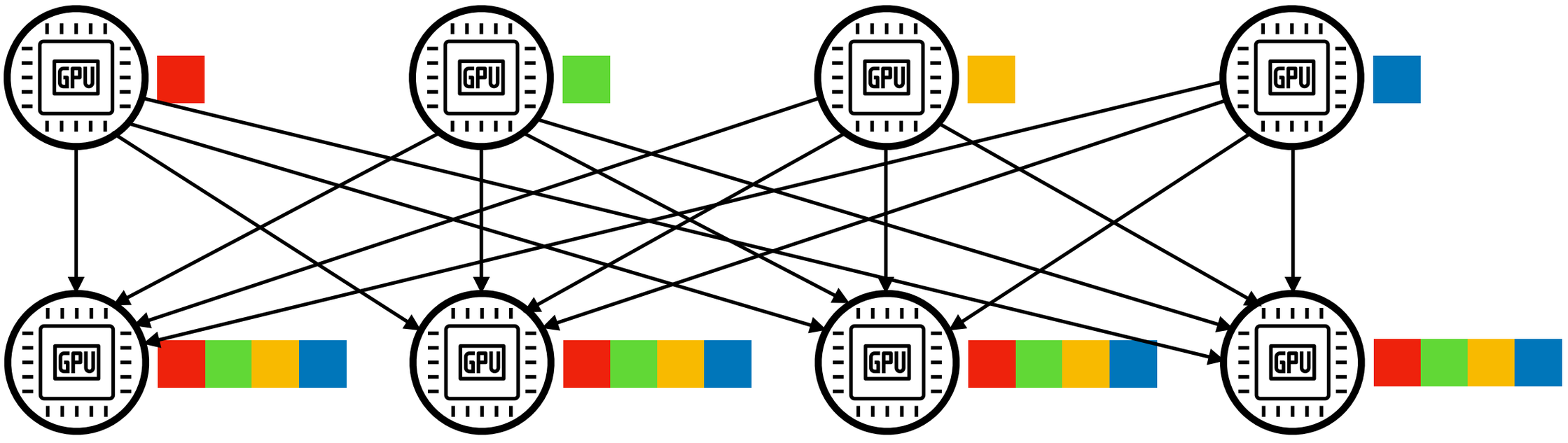
All-reduce:



Broadcast the result

All-reduce vs. gather

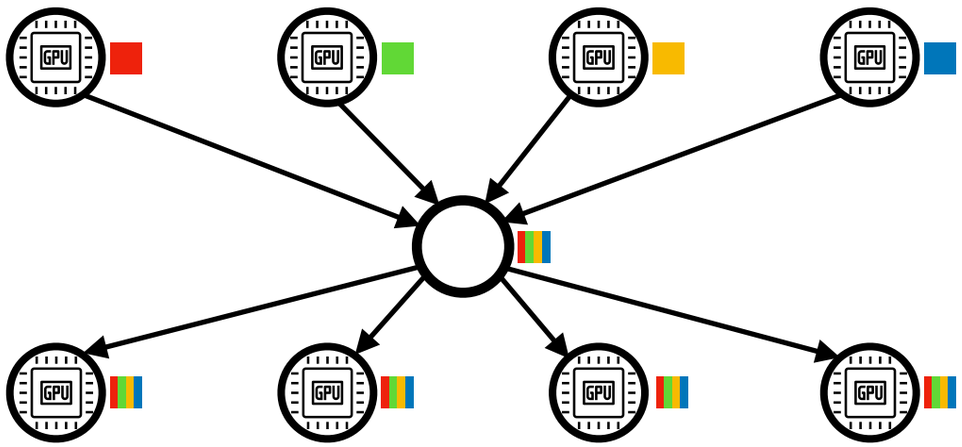
All-gather:



Communicate everything
(potentially compressed)

All-reduce vs. gather

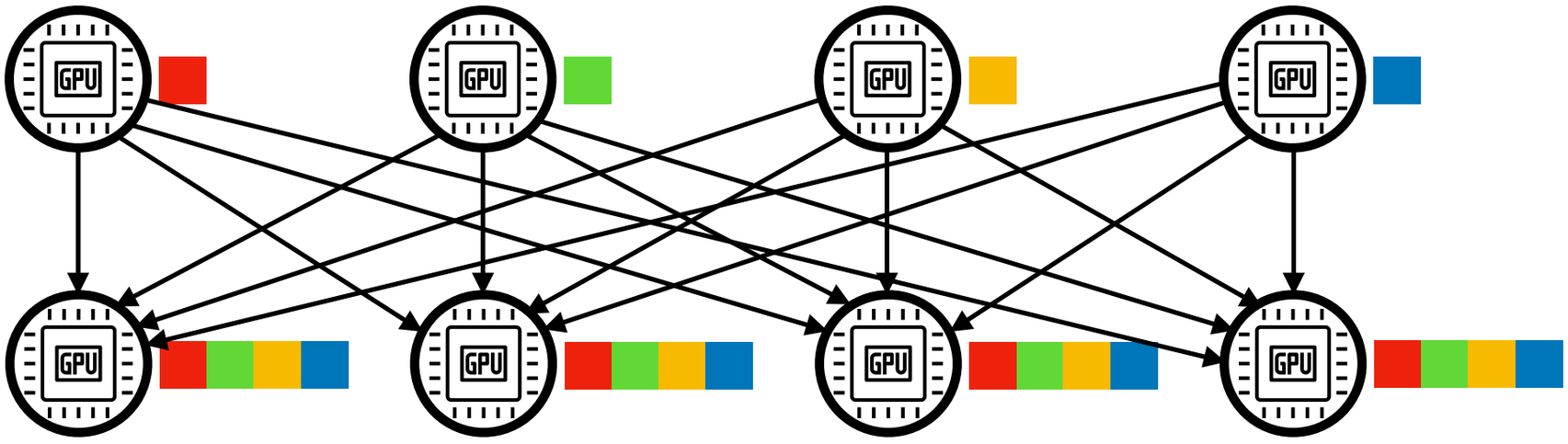
All-reduce:



Communicates

average;

All-gather:

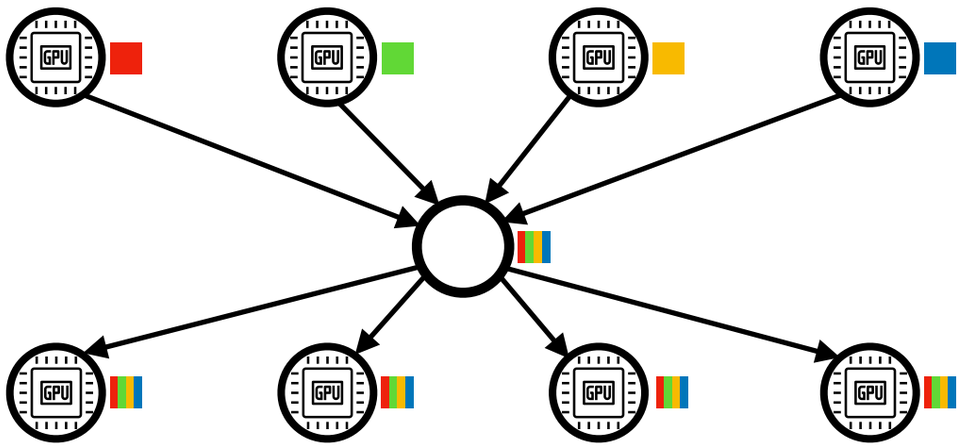


Communicates

everything;

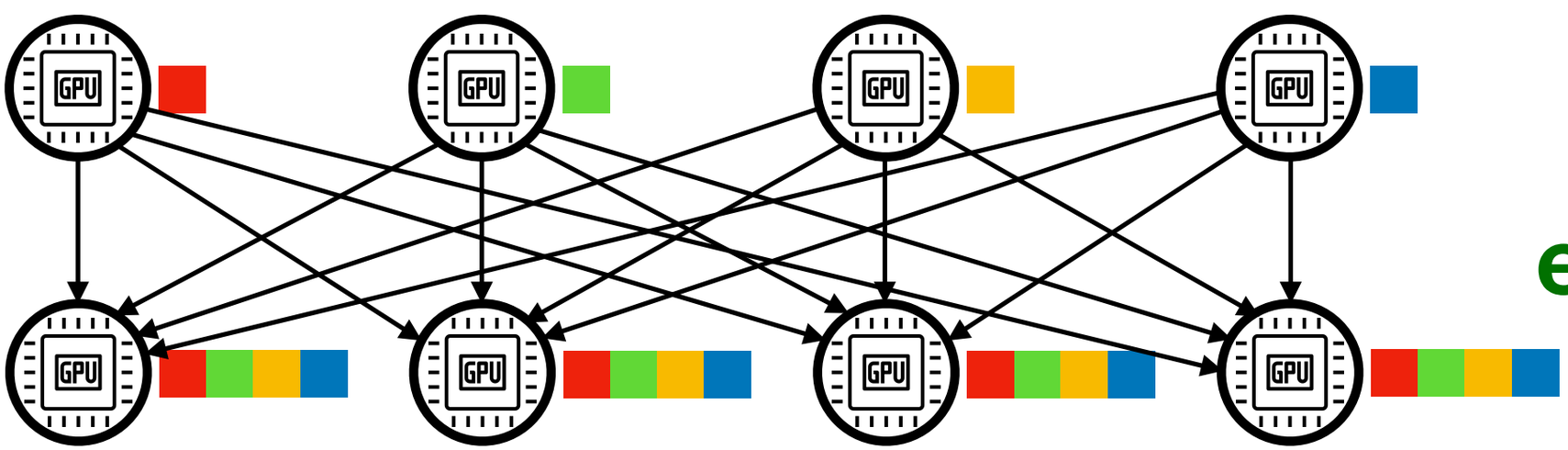
All-reduce vs. gather

All-reduce:



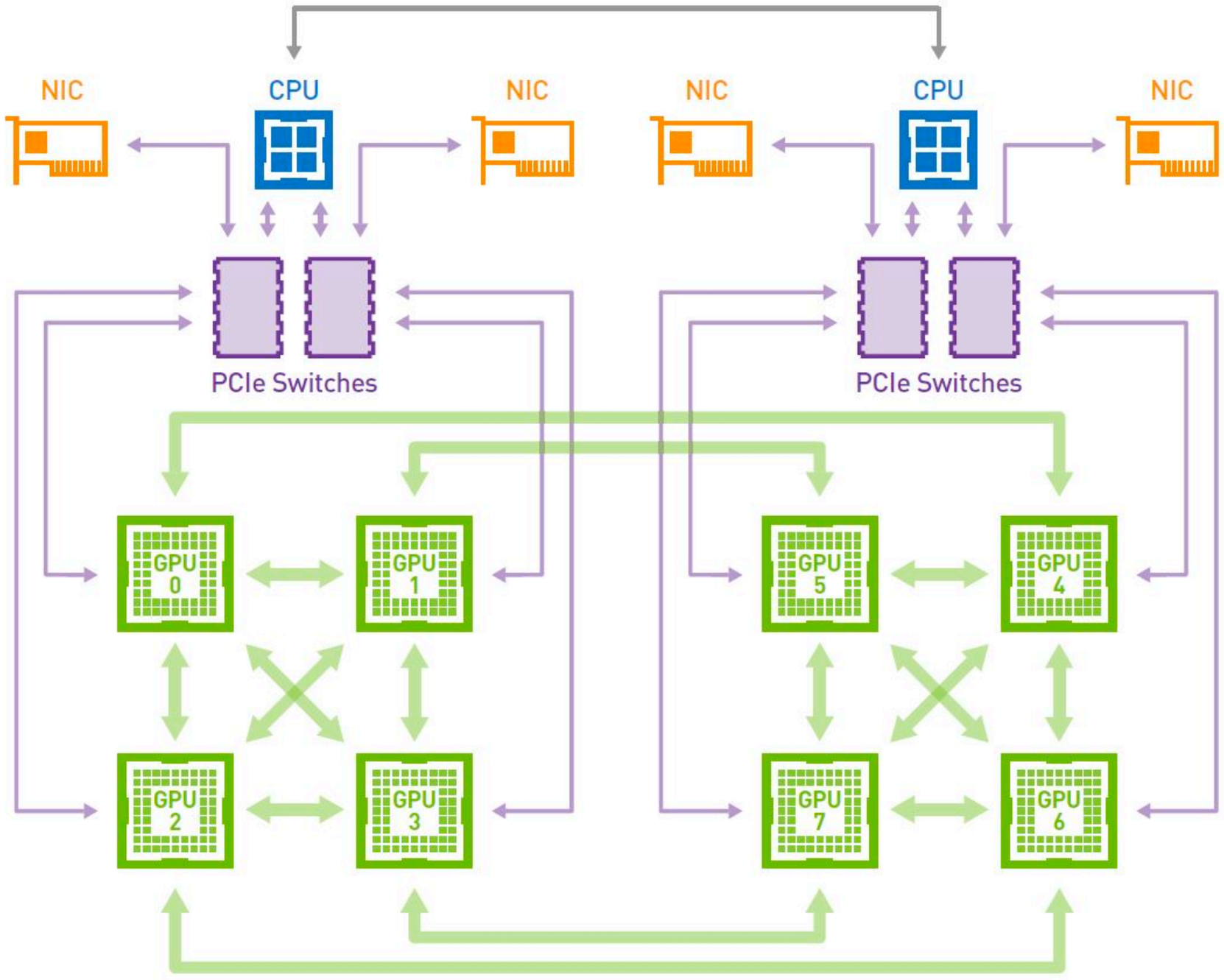
Communicates
average;
hard to compress

All-gather:

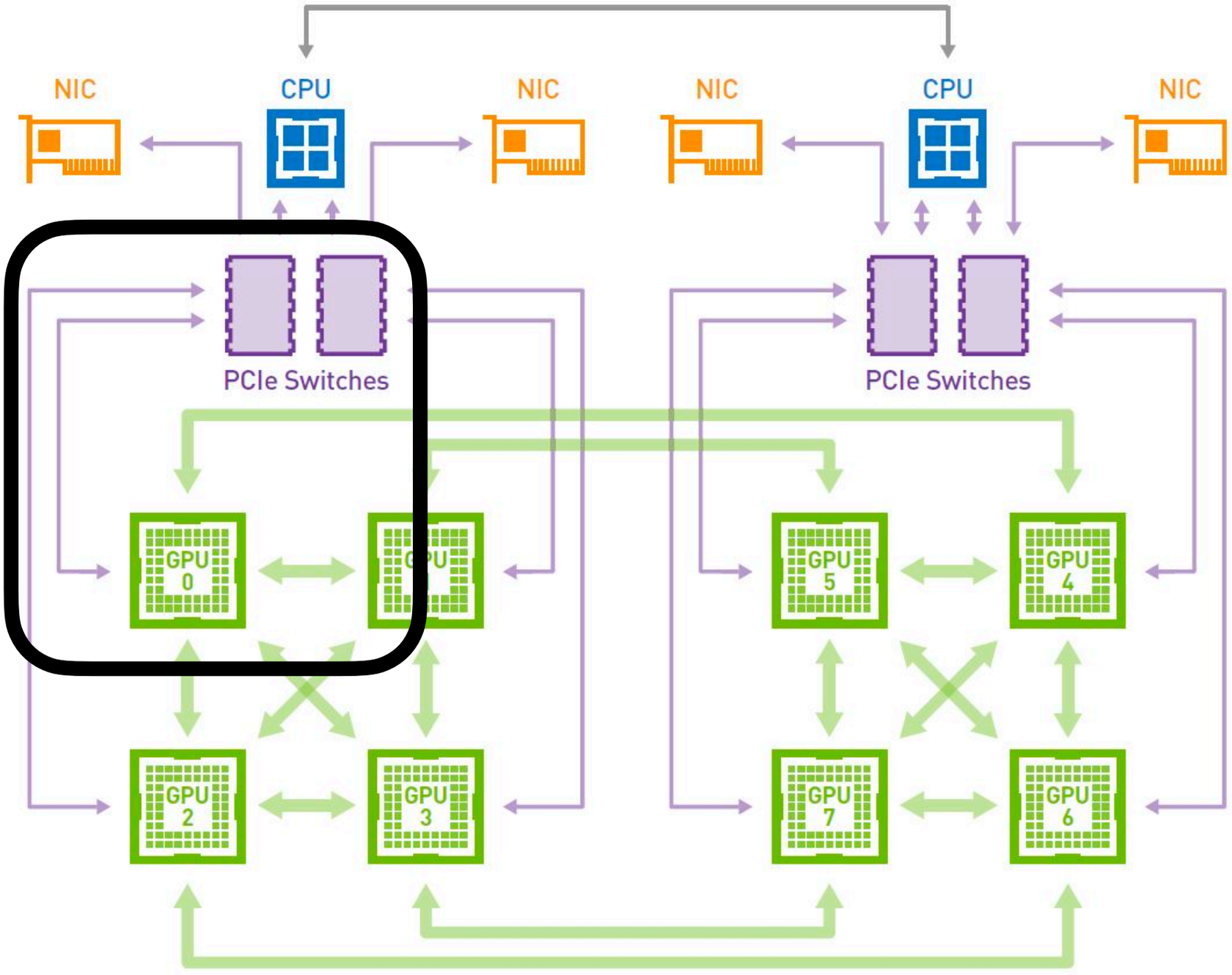


Communicates
everything;
easy to compress

Nvidia's NVSwitch



Nvidia's NVSwitch



Switch

1. Fast inter-communication

Switch

1. Fast inter-communication
2. Limiting compute capabilities

Switch

1. Fast inter-communication
2. Limiting compute capabilities
3. Switch supports only integers

Switch

1. Fast inter-communication
2. Limiting compute capabilities
3. Switch supports only integers

Most compressors can’t operate on switch

Switch

1. Fast inter-communication
2. Limiting compute capabilities
3. Switch supports only integers

Most compressors can’t operate on switch. Those that can do not support all-reduce

IntSGD

$$Q(g_i^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_i^k)$$

IntSGD

$$Q(g_i^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_i^k)$$

Randomized rounding to integer

IntSGD

$$Q(g_i^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_i^k)$$

$$\mathcal{I}nt(t) \stackrel{\text{def}}{=} \begin{cases} [t]+1, & \text{with probability } p_t \stackrel{\text{def}}{=} t-[t], \\ [t], & \text{with probability } 1-p_t, \end{cases}$$

IntSGD

$$Q(g_i^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_i^k)$$

Scaling vector

IntSGD

$$Q(g_i^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_i^k)$$

Example:

$$\begin{pmatrix} 0.089 \\ -0.01 \\ 0.05 \\ 0.023 \end{pmatrix}$$

IntSGD

$$Q(g_i^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_i^k)$$

Example:

$$\begin{pmatrix} 0.089 \\ -0.01 \\ 0.05 \\ 0.023 \end{pmatrix} \xrightarrow{\alpha_k} \begin{pmatrix} 8.9 \\ -1 \\ 5 \\ 2.3 \end{pmatrix}$$

IntSGD

$$Q(g_i^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_i^k)$$

Example:

$$\begin{pmatrix} 0.089 \\ -0.01 \\ 0.05 \\ 0.023 \end{pmatrix} \xrightarrow{\alpha_k} \begin{pmatrix} 8.9 \\ -1 \\ 5 \\ 2.3 \end{pmatrix} \xrightarrow{\mathcal{I}nt} \begin{pmatrix} 9 \\ -1 \\ 5 \\ 2 \end{pmatrix}$$

IntSGD

$$Q(g_i^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_i^k)$$

Example:

$$\begin{pmatrix} 0.089 \\ -0.01 \\ 0.05 \\ 0.023 \end{pmatrix} \xrightarrow{\alpha_k} \begin{pmatrix} 8.9 \\ -1 \\ 5 \\ 2.3 \end{pmatrix} \xrightarrow{\mathcal{I}nt} \begin{pmatrix} 9 \\ -1 \\ 5 \\ 2 \end{pmatrix} \xrightarrow{\frac{1}{\alpha_k}} \begin{pmatrix} 0.09 \\ -0.01 \\ 0.05 \\ 0.02 \end{pmatrix}$$

IntSGD

$$Q(g_i^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_i^k)$$

Properties

$$\mathbb{E}_Q[Q(g_i^k)] = g_i^k$$

IntSGD

$$Q(g_i^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_i^k)$$

Properties

$$\mathbb{E}_Q[Q(g_i^k)] = g_i^k$$

$$\mathbb{E}[\|Q(g_i^k) - g_i^k\|^2] \leq \frac{d}{4\alpha_k^2}$$

IntSGD

$$Q(g_i^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_i^k)$$

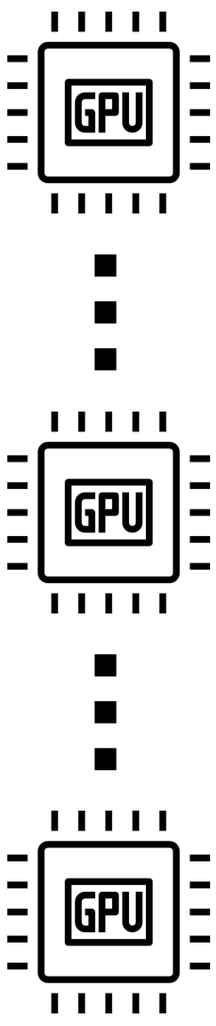
Properties

$$\mathbb{E}_Q[Q(g_i^k)] = g_i^k$$

$$\mathbb{E}[\|Q(g_i^k) - g_i^k\|^2] \leq \frac{d}{4\alpha_k^2}$$

Does not depend on the gradients

IntSGD



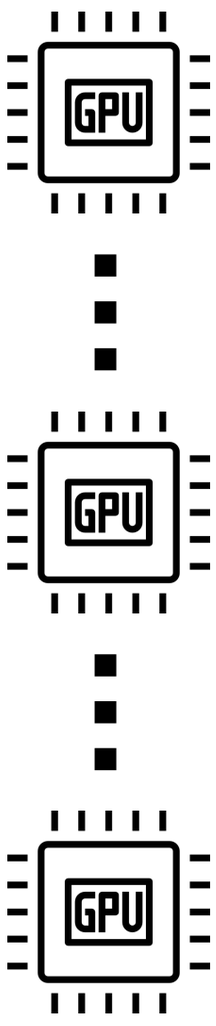
$$Q(g_1^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_1^k)$$

$$Q(g_i^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_i^k)$$

$$Q(g_n^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_n^k)$$

$$\alpha_k = \frac{\sqrt{d}}{\sqrt{2n\|x^k - x^{k-1}\|^2/\eta_k^2 + \varepsilon^2}}$$

IntSGD



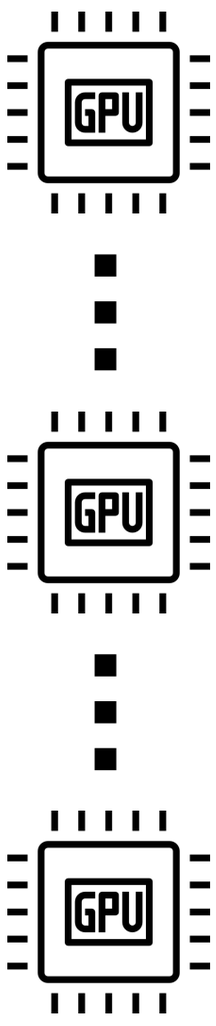
$$Q(g_1^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_1^k)$$

$$Q(g_i^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_i^k)$$

$$Q(g_n^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_n^k)$$

$$\alpha_k = \frac{\sqrt{d}}{\sqrt{2n\|x^k - x^{k-1}\|^2/\eta_k^2 + \varepsilon^2}}$$

IntSGD



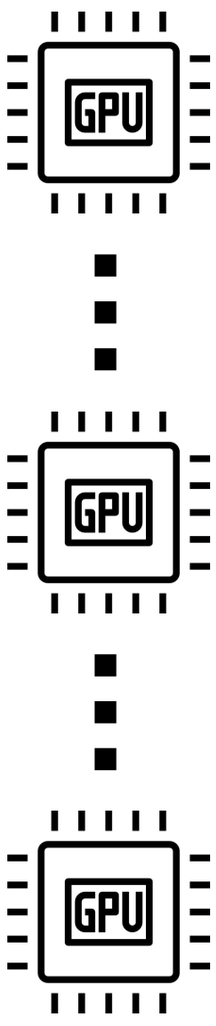
$$Q(g_1^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_1^k)$$

$$Q(g_i^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_i^k)$$

$$Q(g_n^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_n^k)$$

$$\alpha_k = \frac{\sqrt{d}}{\sqrt{2n\|x^k - x^{k-1}\|^2/\eta_k^2 + \varepsilon^2}}$$

IntSGD



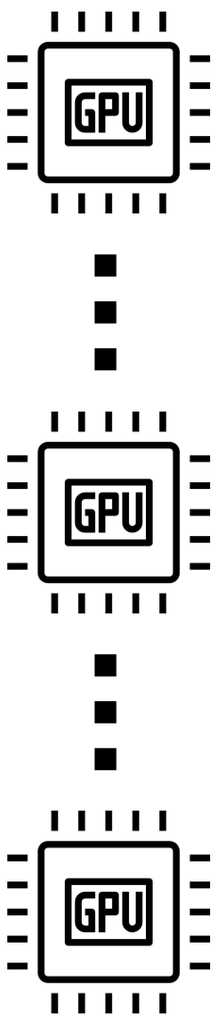
$$Q(g_1^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_1^k)$$

$$Q(g_i^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_i^k)$$

$$Q(g_n^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_n^k)$$

$$\alpha_k = \frac{\sqrt{d}}{\sqrt{2n\|x^k - x^{k-1}\|^2 / \eta_k^2 + \varepsilon^2}}$$

IntSGD



$$Q(g_1^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_1^k)$$

$$Q(g_i^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_i^k)$$

$$Q(g_n^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_n^k)$$

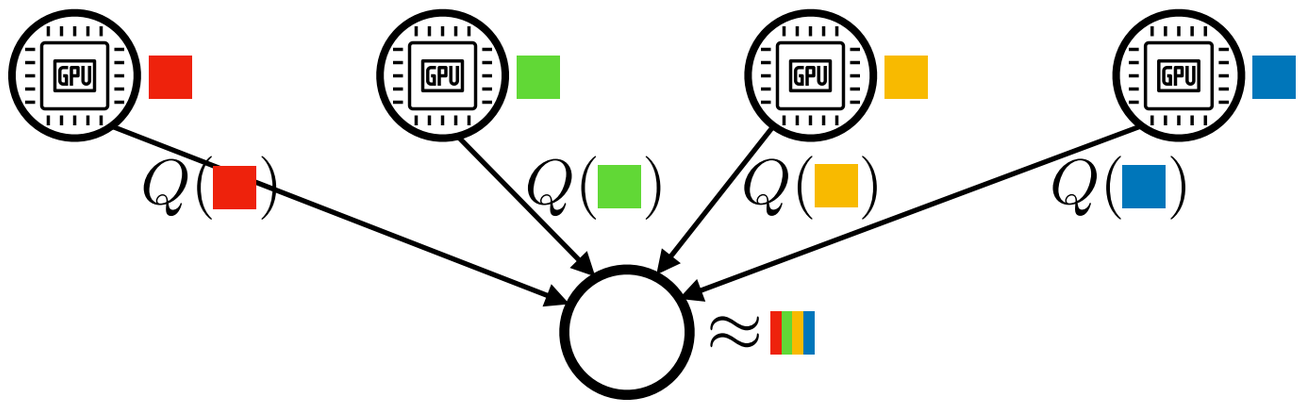
$$\alpha_k = \frac{\sqrt{d}}{\sqrt{2n\|x^k - x^{k-1}\|^2/\eta_k^2 + \varepsilon^2}}$$

IntSGD

$$Q(g_1^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_1^k)$$

$$Q(g_i^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_i^k)$$

$$Q(g_n^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_n^k)$$



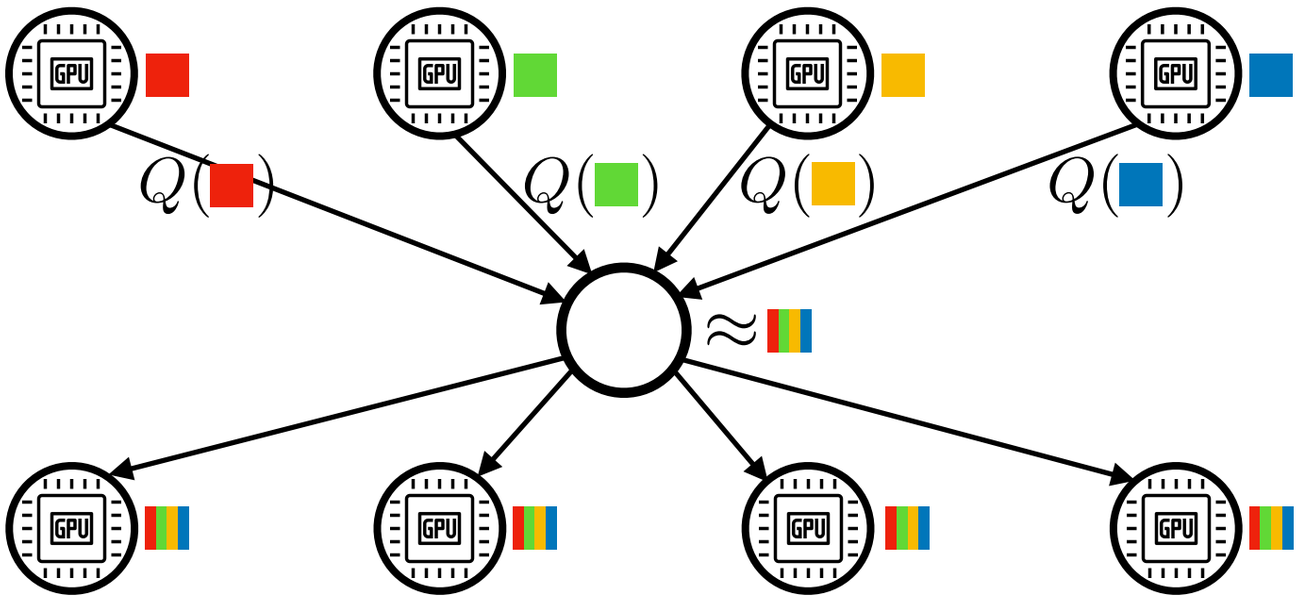
$$\alpha_k = \frac{\sqrt{d}}{\sqrt{2n\|x^k - x^{k-1}\|^2/\eta_k^2 + \varepsilon^2}}$$

IntSGD

$$Q(g_1^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_1^k)$$

$$Q(g_i^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_i^k)$$

$$Q(g_n^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_n^k)$$



$$\alpha_k = \frac{\sqrt{d}}{\sqrt{2n\|x^k - x^{k-1}\|^2/\eta_k^2 + \varepsilon^2}}$$

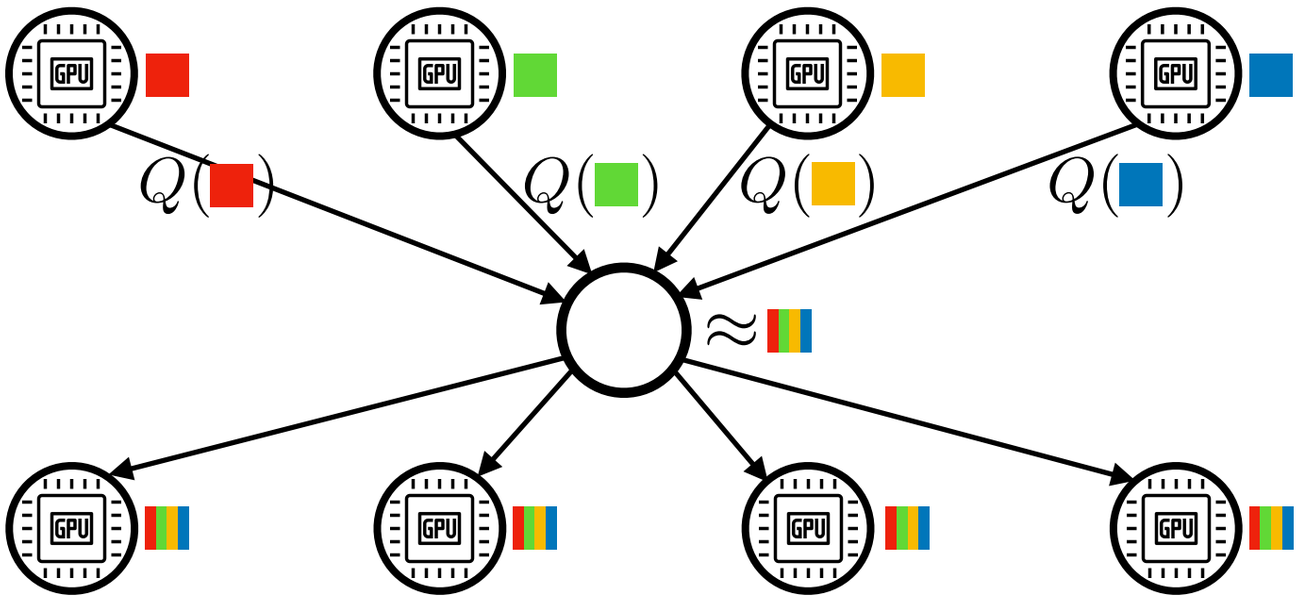
IntSGD

$$Q(g_1^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_1^k)$$

$$Q(g_i^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_i^k)$$

$$Q(g_n^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_n^k)$$

$$\alpha_k = \frac{\sqrt{d}}{\sqrt{2n\|x^k - x^{k-1}\|^2/\eta_k^2 + \varepsilon^2}}$$



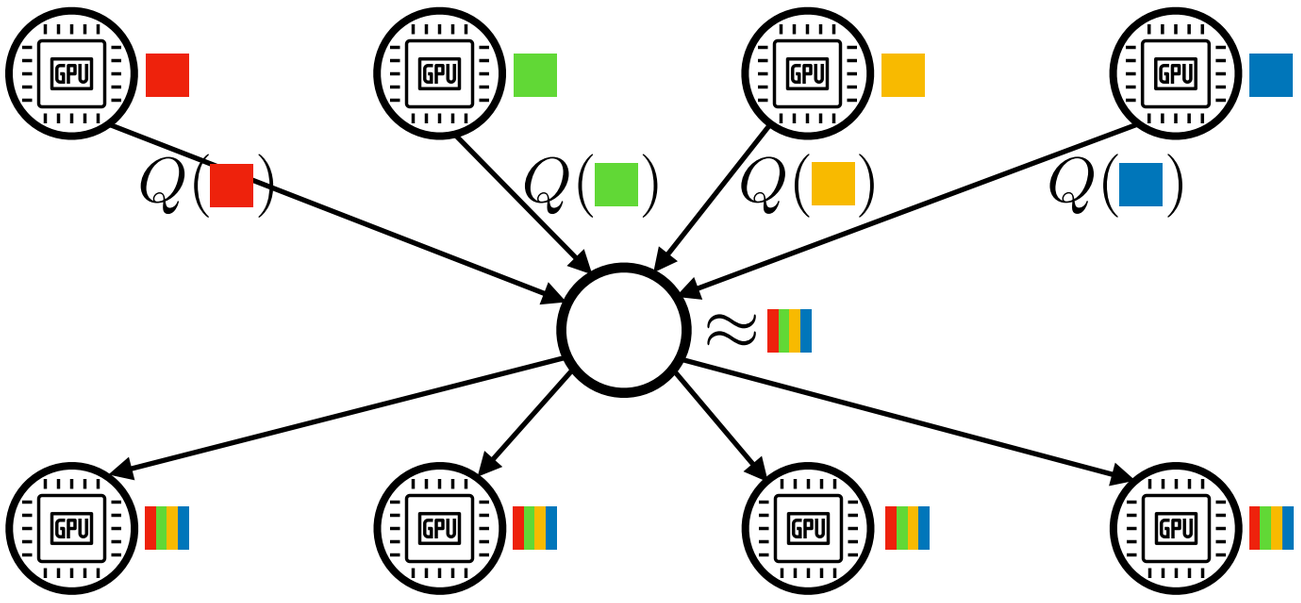
1. float32->int8

IntSGD

$$Q(g_1^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_1^k)$$

$$Q(g_i^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_i^k)$$

$$Q(g_n^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_n^k)$$



$$\alpha_k = \frac{\sqrt{d}}{\sqrt{2n\|x^k - x^{k-1}\|^2/\eta_k^2 + \varepsilon^2}}$$

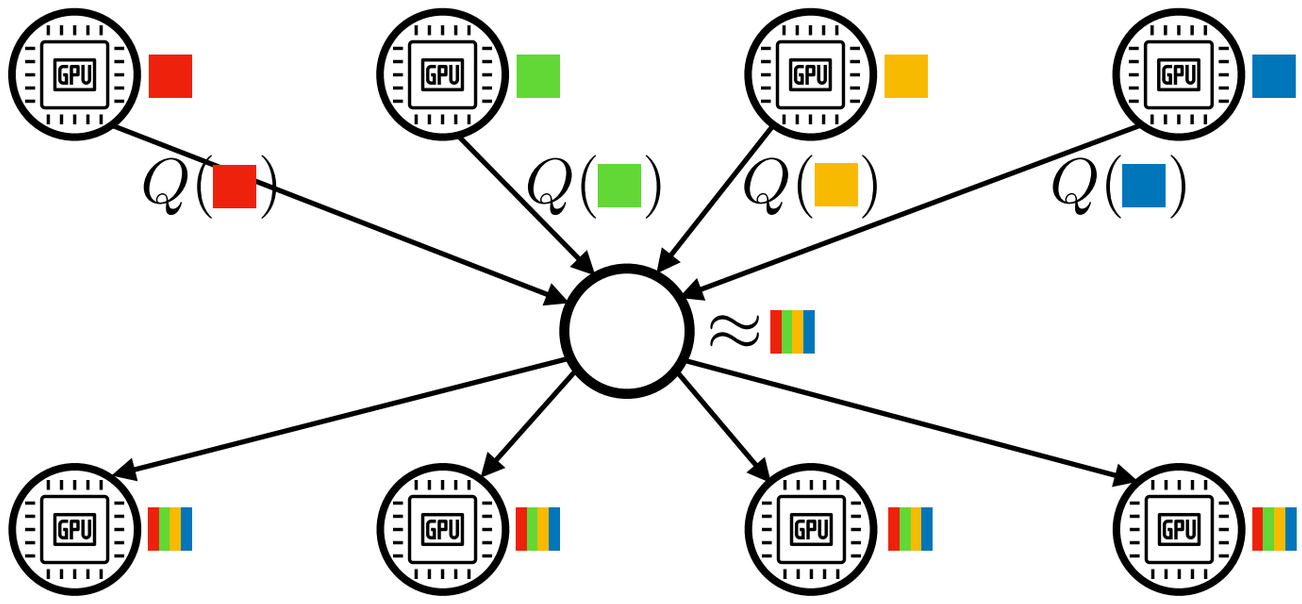
1. **float32->int8**
2. **Supports all-reduce**

IntSGD

$$Q(g_1^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_1^k)$$

$$Q(g_i^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_i^k)$$

$$Q(g_n^k) = \frac{1}{\alpha_k}\mathcal{I}nt(\alpha_k \circ g_n^k)$$



$$\alpha_k = \frac{\sqrt{d}}{\sqrt{2n\|x^k - x^{k-1}\|^2/\eta_k^2 + \varepsilon^2}}$$

1. **float32->int8**
2. **Supports all-reduce**
3. **Can run on switch**

Theory

Theorem.

Smooth nonconvex problems:

$$\mathbb{E}\left[\|\nabla f(\hat{x}^k)\|^2\right] = \mathcal{O}\left(\frac{\sigma+\varepsilon}{\sqrt{kn}} + \frac{f(x^0)-f^{\inf}}{k}\right)$$

Almost the same rate as SGD

Theory

Theorem.

Smooth nonconvex problems:

$$\mathbb{E}\left[\|\nabla f(\hat{x}^k)\|^2\right] = \mathcal{O}\left(\frac{\sigma+\varepsilon}{\sqrt{kn}} + \frac{f(x^0)-f^{\inf}}{k}\right)$$

Smooth convex problems:

$$\mathbb{E}\left[f(\hat{x}^k) - f(x^*)\right] = \mathcal{O}\left(\frac{\sigma_*+\varepsilon}{\sqrt{kn}} + \frac{\|x^0-x^*\|}{k}\right)$$

Theory

Theorem.

Smooth nonconvex problems:

$$\mathbb{E}\left[\|\nabla f(\hat{x}^k)\|^2\right] = \mathcal{O}\left(\frac{\sigma+\varepsilon}{\sqrt{kn}} + \frac{f(x^0)-f^{\inf}}{k}\right)$$

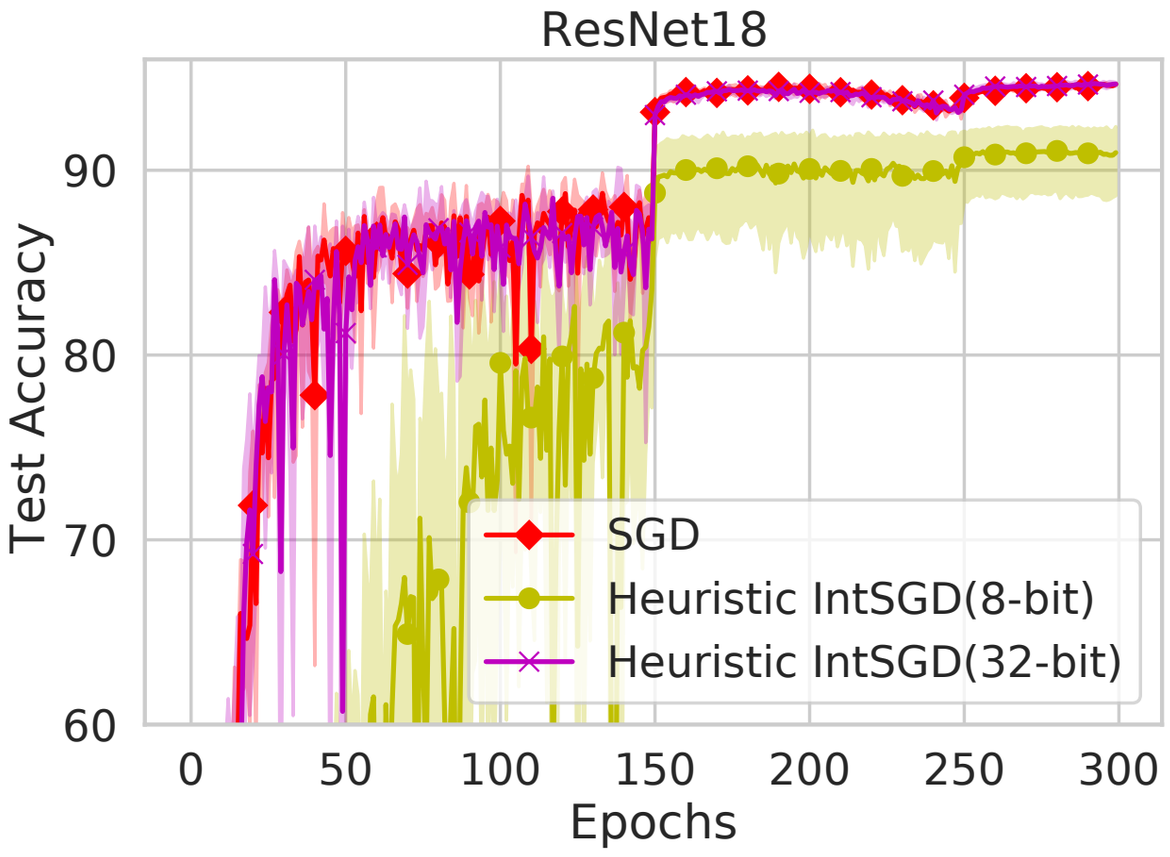
Smooth convex problems:

$$\mathbb{E}\left[f(\hat{x}^k) - f(x^*)\right] = \mathcal{O}\left(\frac{\sigma_*+\varepsilon}{\sqrt{kn}} + \frac{\|x^0-x^*\|}{k}\right)$$

Non-smooth convex problems:

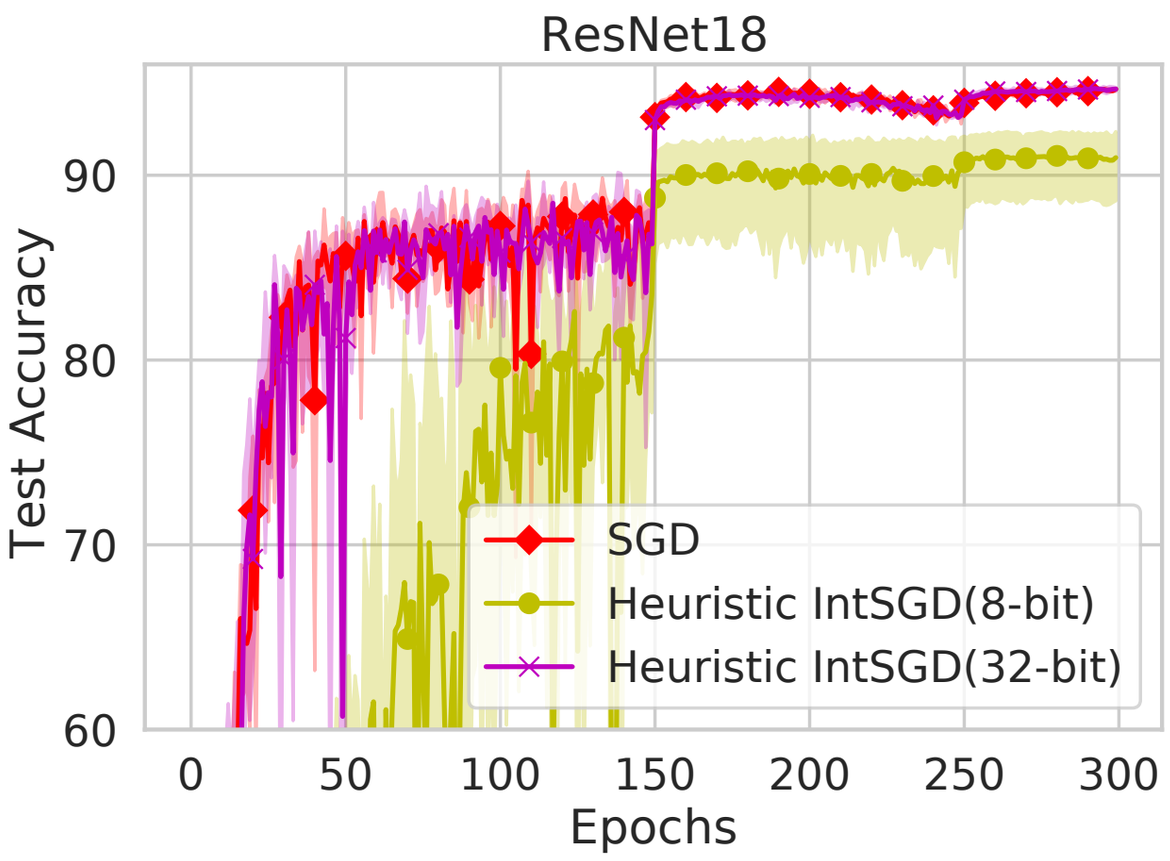
$$\mathbb{E}\left[f(\hat{x}^k) - f(x^*)\right] = \mathcal{O}\left(\frac{\sigma+\varepsilon}{\sqrt{kn}} + \frac{G}{\sqrt{k}}\right)$$

Experiments

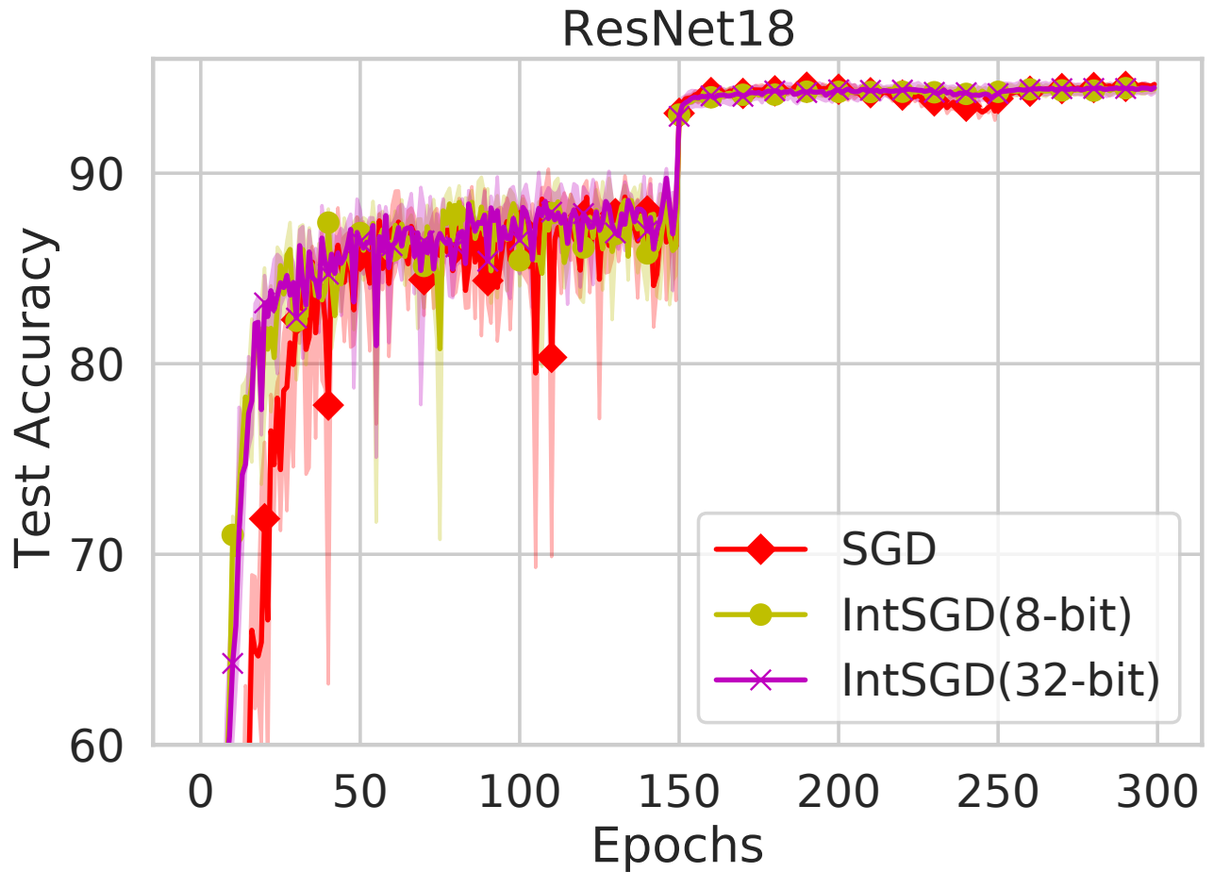


Heuristic int compression

Experiments

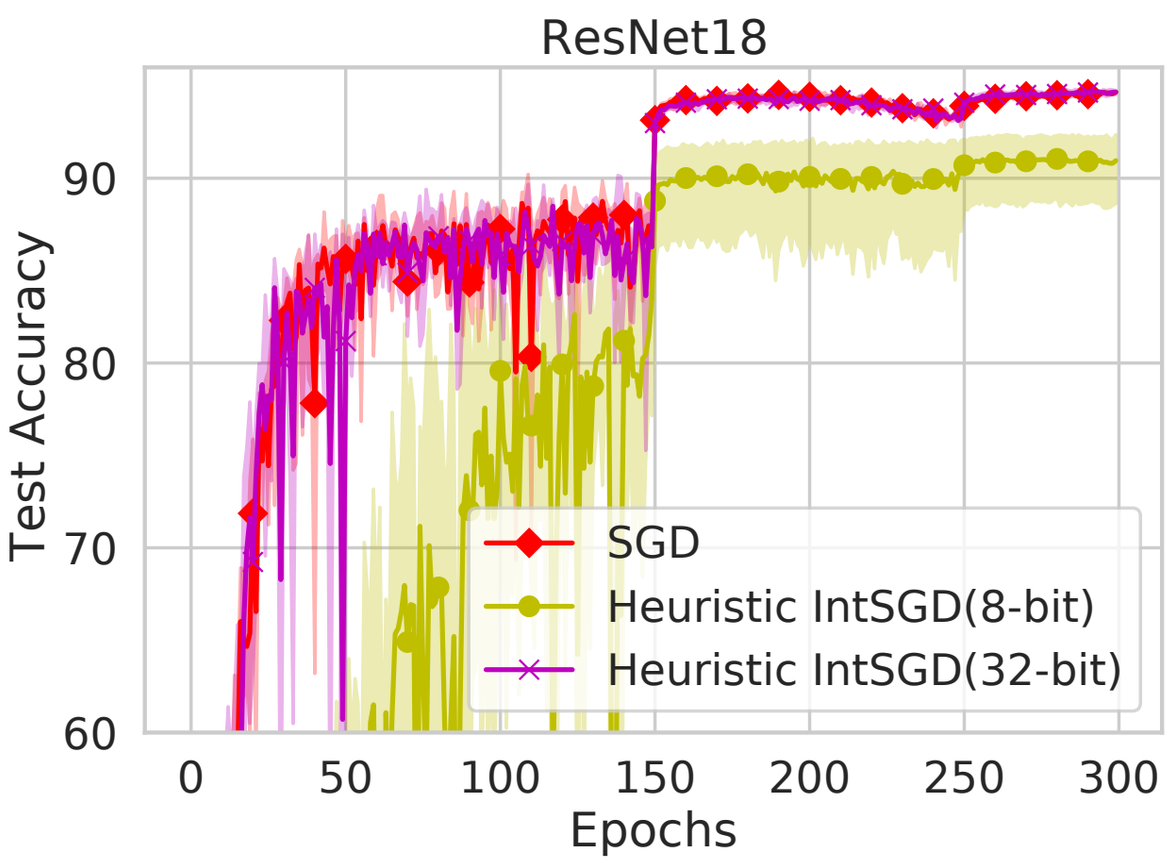


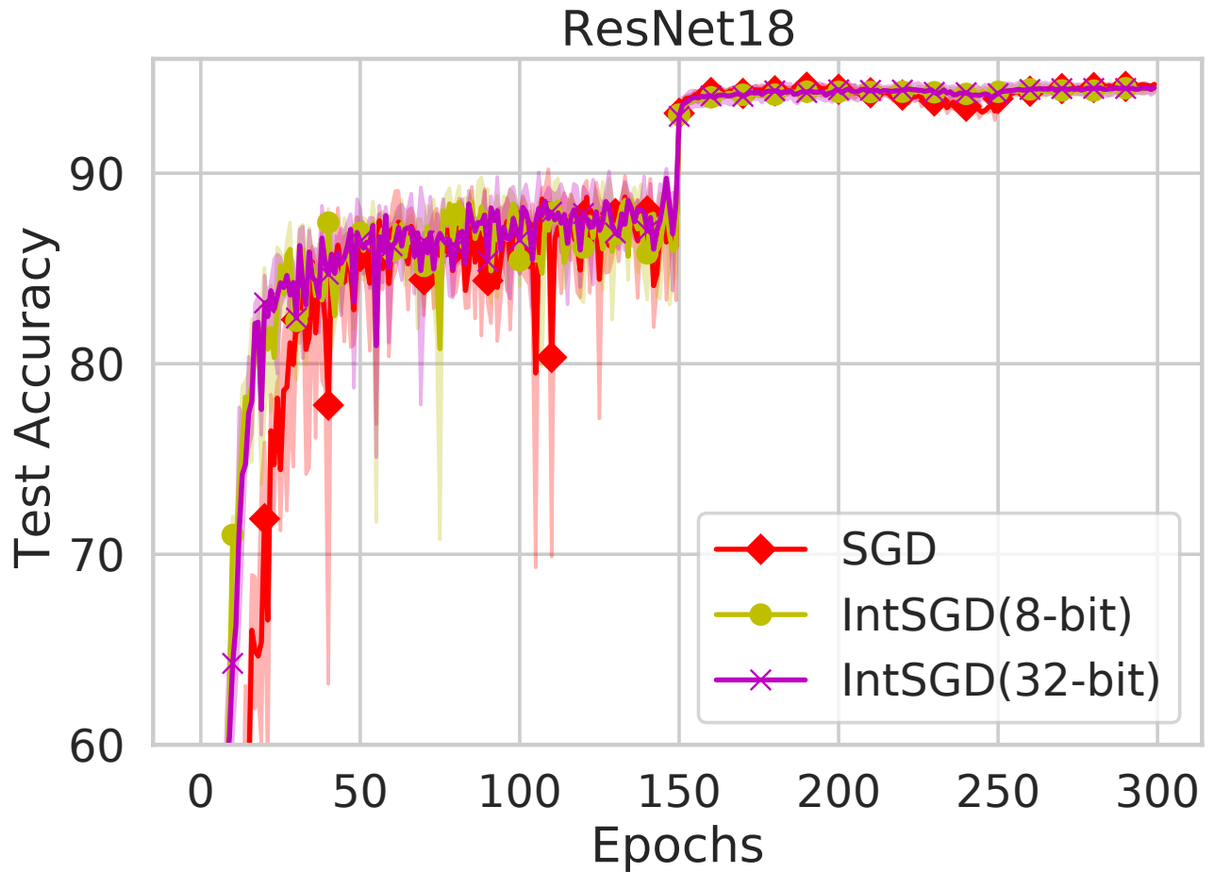
Heuristic int compression



Our compression

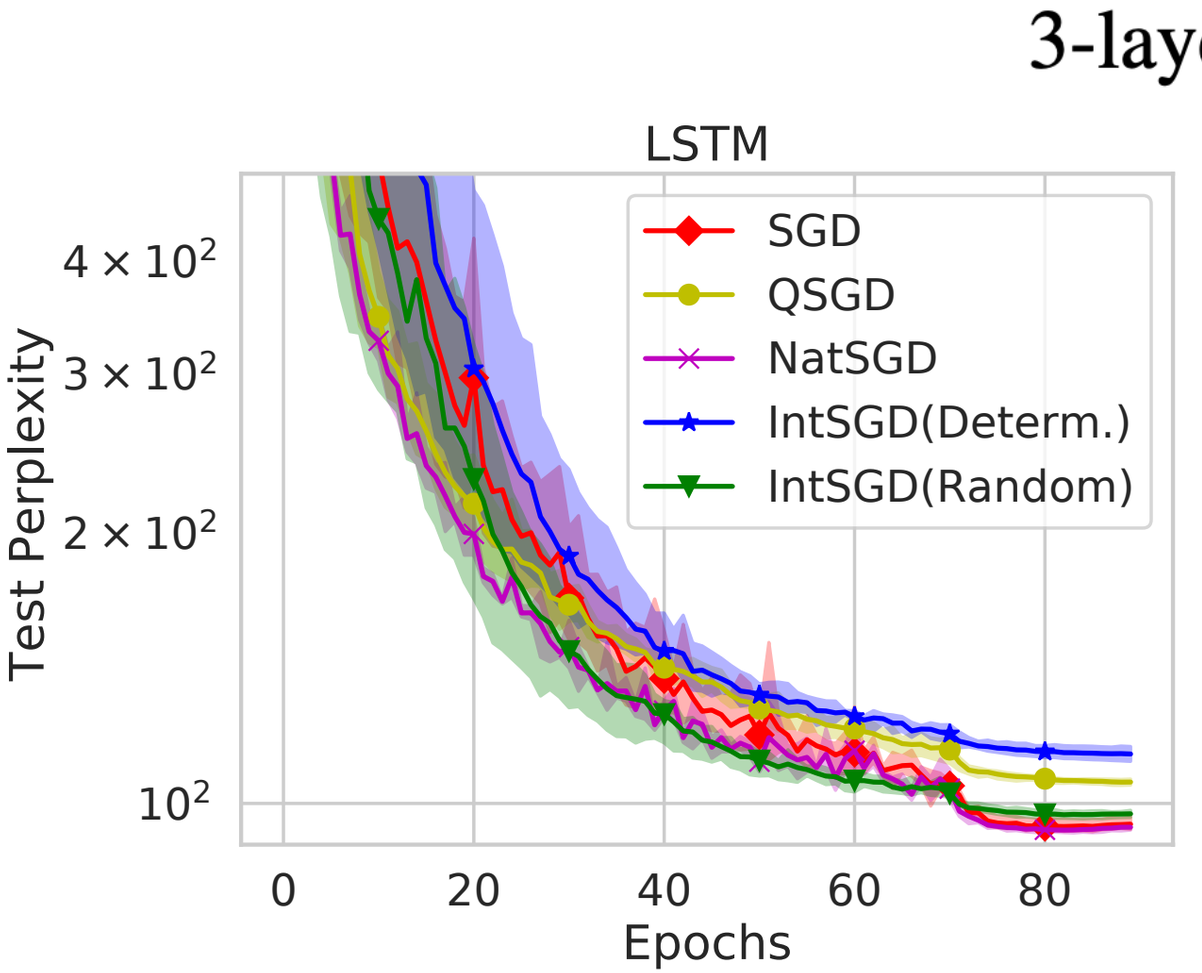
Experiments





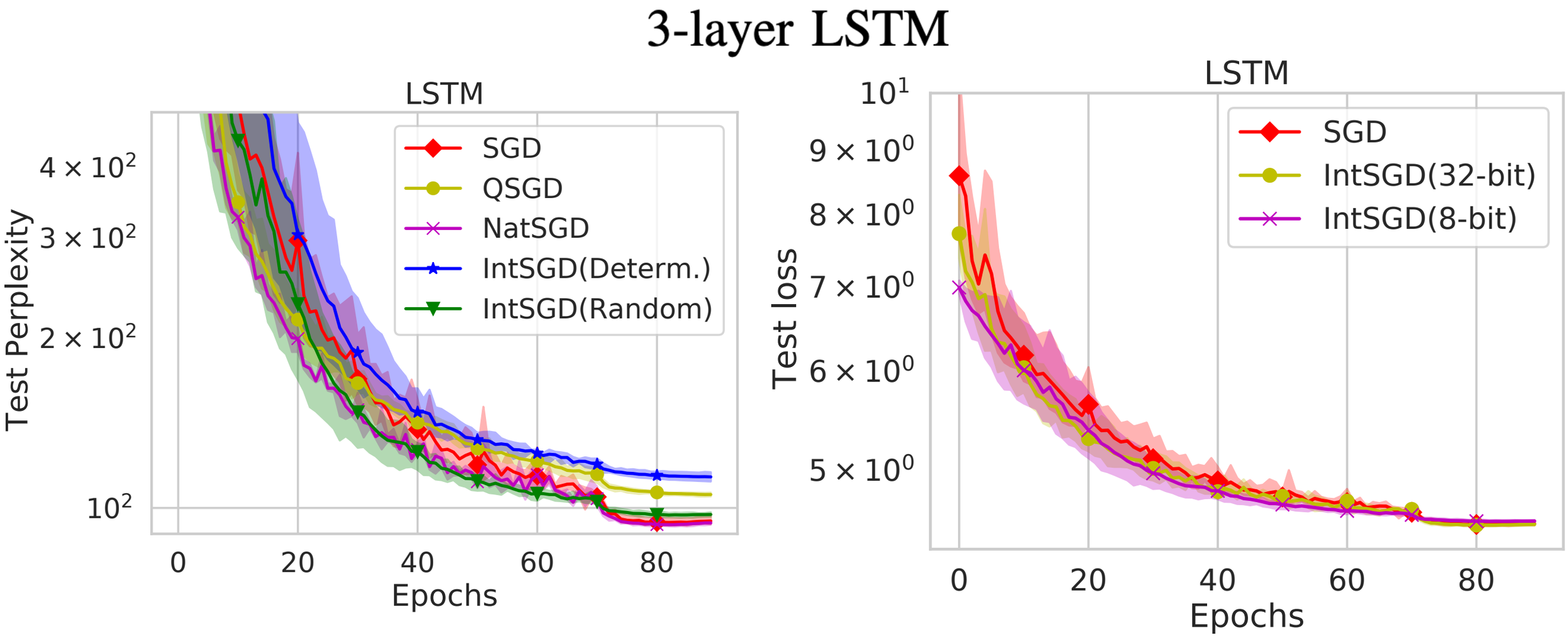
Much more stable
than naive approach

Experiments



Randomization helps

Experiments



8 bits is as good as 32 bits

Experiments

Table : Test accuracy and time breakdown in one iteration (on average) of training ResNet18 on the CIFAR-10 dataset with 16 workers. All numbers of time are in millisecond (ms). In each column, the best one is highlighted in black and the second-best one is highlighted in gray.

Experiments

Table : Test accuracy and time breakdown in one iteration (on average) of training ResNet18 on the CIFAR-10 dataset with 16 workers. All numbers of time are in millisecond (ms). In each column, the best one is highlighted in black and the second-best one is highlighted in gray.

Algorithm	Test Accuracy (%)	Computation Overhead	Communication	Total Time
SGD (All-gather)	**94.65 ± 0.08**	-	261.29 ± 0.98	338.76 ± 0.76
QSGD	93.69 ± 0.03	129.25 ± 1.58	138.16 ± 1.29	320.49 ± 2.11
NatSGD	94.57 ± 0.13	36.01± 1.30	106.27 ± 1.43	197.18 ± 0.25
SGD (All-reduce)	**94.67 ± 0.17**	-	18.48 ± 0.09	74.32 ± 0.06
PowerSGD (EF)	94.33 ± 0.15	7.07 ± 0.03	**5.03 ± 0.07**	67.08 ± 0.06
IntSGD (Determ.)	94.43 ± 0.12	**2.51 ± 0.04**	6.92± 0.07	**64.95 ± 0.15**
IntSGD (Random)	94.55 ± 0.13	**3.20 ± 0.02**	**6.21 ± 0.13**	**65.22 ± 0.08**

Experiments

Table : Test accuracy and time breakdown in one iteration (on average) of training ResNet18 on the CIFAR-10 dataset with 16 workers. All numbers of time are in millisecond (ms). In each column, the best one is highlighted in black and the second-best one is highlighted in gray.

Algorithm	Test Accuracy (%)	Computation Overhead	Communication	Total Time
SGD (All-gather)	**94.65 ± 0.08**	-	261.29 ± 0.98	338.76 ± 0.76
QSGD	93.69 ± 0.03	129.25 ± 1.58	138.16 ± 1.29	320.49 ± 2.11
NatSGD	94.57 ± 0.13	36.01± 1.30	106.27 ± 1.43	197.18 ± 0.25
				32 ± 0.06
				08 ± 0.06
IntSGD (Determ.)	94.43 ± 0.12	**2.51 ± 0.04**	6.92± 0.07	**64.95 ± 0.15**
IntSGD (Random)	94.55 ± 0.13	**3.20 ± 0.02**	**6.21 ± 0.13**	**65.22 ± 0.08**

All-gather is extremely slow

Experiments

Table : Test accuracy and time breakdown in one iteration (on average) of training ResNet18 on the CIFAR-10 dataset with 16 workers. All numbers of time are in millisecond (ms). In each column, the best one is highlighted in black and the second-best one is highlighted in gray.

Algorithm	Test Accuracy (%)	Computation Overhead	Communication	Total Time
SGD (All-gather)	**94.65 ± 0.08**	-	261.29 ± 0.98	338.76 ± 0.76
QSGD	93.69 ± 0.03	129.25 ± 1.58	138.16 ± 1.29	320.49 ± 2.11
NatSGD	94.57 ± 0.13	36.01± 1.30	106.27 ± 1.43	197.18 ± 0.25
				74.32 ± 0.06
				67.08 ± 0.06
IntSGD (Determ.)	94.43 ± 0.12	**2.51 ± 0.04**	6.92± 0.07	**64.95 ± 0.15**
IntSGD (Random)	94.55 ± 0.13	**3.20 ± 0.02**	**6.21 ± 0.13**	**65.22 ± 0.08**

Cluster is not very stable

Experiments

3-layer LSTM Wiki-text2 dataset with 16 workers

Algorithm	Test Loss	Computation Overhead	Communication	Total Time
SGD (All-gather)	**4.52 ± 0.01**	-	733.07 ± 1.04	796.23 ± 1.03
QSGD	4.63 ± 0.01	43.67 ± 0.11	307.63 ± 1.16	399.10 ± 1.25
NatSGD	**4.52 ± 0.01**	64.63 ± 0.12	309.87 ± 1.32	422.49 ± 2.15
SGD (All-reduce)	**4.54 ± 0.03**	-	22.33 ± 0.02	70.46 ± 0.05
PowerSGD (EF)	**4.52 ± 0.01**	**4.22 ± 0.01**	**2.10 ± 0.01**	**54.89 ± 0.02**
IntSGD (Determ.)	4.70 ± 0.02	**3.04 ± 0.01**	**6.94 ± 0.05**	**57.93 ± 0.03**
IntSGD (Random)	**4.54 ± 0.01**	4.76 ± 0.01	7.14 ± 0.04	59.99 ± 0.01

Experiments

3-layer LSTM Wiki-text2 dataset with 16 workers

Algorithm	Test Loss	Computation Overhead	Communication	Total Time
SGD (A				96.23 ± 1.03
Q				99.10 ± 1.25
Na				22.49 ± 2.15
SGD (All-reduce)	4.54 ± 0.03	-	22.33 ± 0.02	70.46 ± 0.05
PowerSGD (EF)	**4.52 ± 0.01**	4.22 ± 0.01	**2.10 ± 0.01**	**54.89 ± 0.02**
IntSGD (Determ.)	4.70 ± 0.02	**3.04 ± 0.01**	**6.94 ± 0.05**	**57.93 ± 0.03**
IntSGD (Random)	4.54 ± 0.01	4.76 ± 0.01	7.14 ± 0.04	59.99 ± 0.01

PowerSGD is faster but requires tuning

Key features

Fast in practice

Algorithm	Supports all-reduce
IntSGD	✓
Heuristic IntSGD	✓
PowerSGD (theoretical)	✓
PowerSGD (practical)	✓
NatSGD	✗
QSGD	✗
SignSGD	✗

Key features

Supports special hardware

Algorithm	Supports all-reduce	Supports switch
IntSGD	✓	✓
Heuristic IntSGD	✓	✓
PowerSGD (theoretical)	✓	✗
PowerSGD (practical)	✓	✗
NatSGD	✗	✓
QSGD	✗	✗
SignSGD	✗	✗

Key features

Won't break for no reason

Algorithm	Supports all-reduce	Supports switch	Provably works
IntSGD	✓	✓	✓
Heuristic IntSGD	✓	✓	✗
PowerSGD (theoretical)	✓	✗	✓
PowerSGD (practical)	✓	✗	✗
NatSGD	✗	✓	✓
QSGD	✗	✗	✓
SignSGD	✗	✗	✓

Key features

Almost no overhead

Algorithm	Supports all-reduce	Supports switch	Provably works	Fast compression
IntSGD	✓	✓	✓	✓
Heuristic IntSGD	✓	✓	✗	✓
PowerSGD (theoretical)	✓	✗	✓	✗(1)
PowerSGD (practical)	✓	✗	✗	✓(1)
NatSGD	✗	✓	✓	✗
QSGD	✗	✗	✓	✓
SignSGD	✗	✗	✓	✓

Key features

Uses no extra memory

Algorithm	Supports all-reduce	Supports switch	Provably works	Fast compression	Works without error-feedback
IntSGD	✓	✓	✓	✓	✓
Heuristic IntSGD	✓	✓	✗	✓	✓
PowerSGD (theoretical)	✓	✗	✓	✗(1)	✗
PowerSGD (practical)	✓	✗	✗	✓(1)	✗
NatSGD	✗	✓	✓	✗	✓
QSGD	✗	✗	✓	✓	✓
SignSGD	✗	✗	✓	✓	✗

Key features

Requires no extra hyperparameters

Algorithm	Supports all-reduce	Supports switch	Provably works	Fast compression	Works without error-feedback	Adaptive
IntSGD	✓	✓	✓	✓	✓	✓
Heuristic IntSGD	✓	✓	✗	✓	✓	✗
PowerSGD (theoretical)	✓	✗	✓	✗(1)	✗	✗(2)
PowerSGD (practical)	✓	✗	✗	✓(1)	✗	✗(2)
NatSGD	✗	✓	✓	✗	✓	N/A
QSGD	✗	✗	✓	✓	✓	N/A
SignSGD	✗	✗	✓	✓	✗	N/A