
L-GRECO: LAYERWISE-ADAPTIVE GRADIENT COMPRESSION FOR EFFICIENT AND ACCURATE DEEP LEARNING

Mohammadreza Alimohammadi^{1*} Ilia Markov^{1*} Elias Frantar¹ Dan Alistarh^{1,2}

MLSys 2024

背景：梯度同步正在成为分布式训练瓶颈

同步数据并行的每一步，都要在所有GPU之间交换并聚合完整梯度。

模型更大

GPU 更多

通信开销成为制约训练效率的主要因素之一

问题1：现有压缩通常忽略层间差异

量化 (QSGD)

降低每个梯度值的表示位宽；实现简单，但压缩空间有限。

稀疏化 (TopK)

只传输少量重要元素；压缩率高，但需要误差反馈与参数调节。

低秩近似 (PowerSGD)

利用梯度矩阵的低秩结构；通信量显著下降，但各层最优秩并不相同。

共同局限：通常给所有层设置同一压缩参数

问题2：全局 TopK 会破坏计算—通信重叠

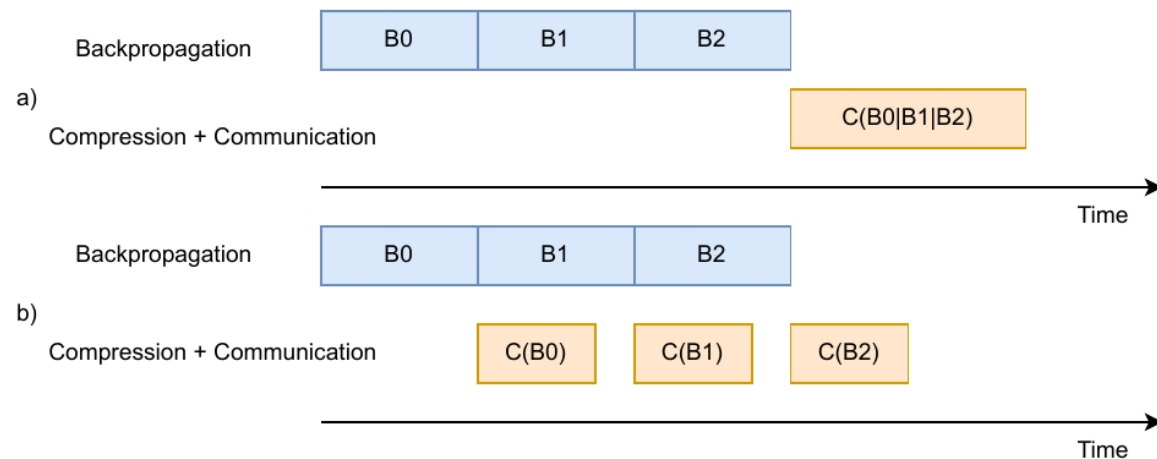
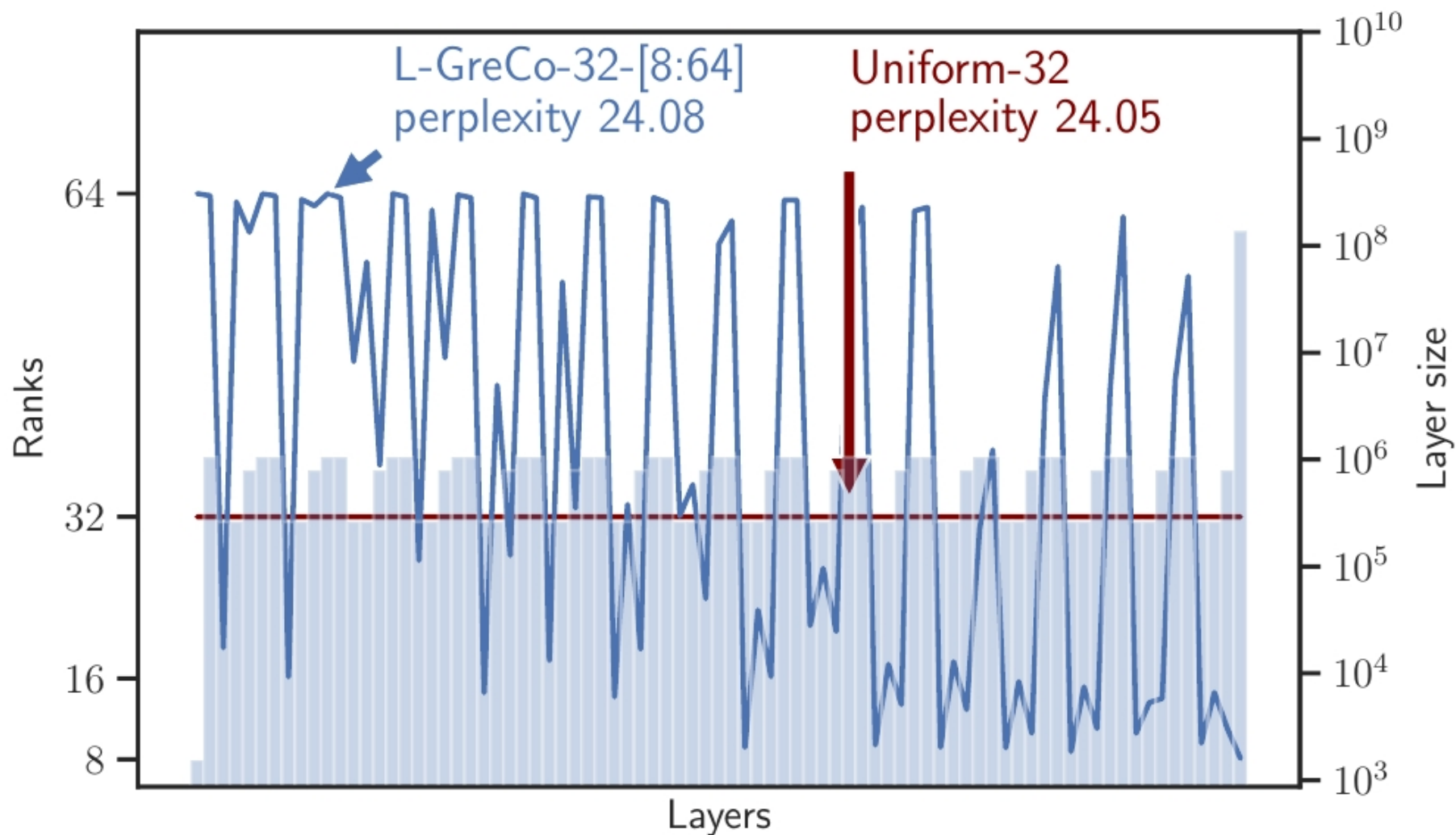


Figure 10. Scheme of communication in case of (a) Global TopK compression and (b) standard bucket-wise communication. Global TopK has to collect entire model gradients in buckets before it compresses and synchronizes them, whereas standard approach allows to overlap compression and communication with backpropagation.

必须等全部梯度就绪后再压缩和通信

解决方案：不同层需要不同压缩强度



作用位置：L-GreCo 位于训练与通信之间

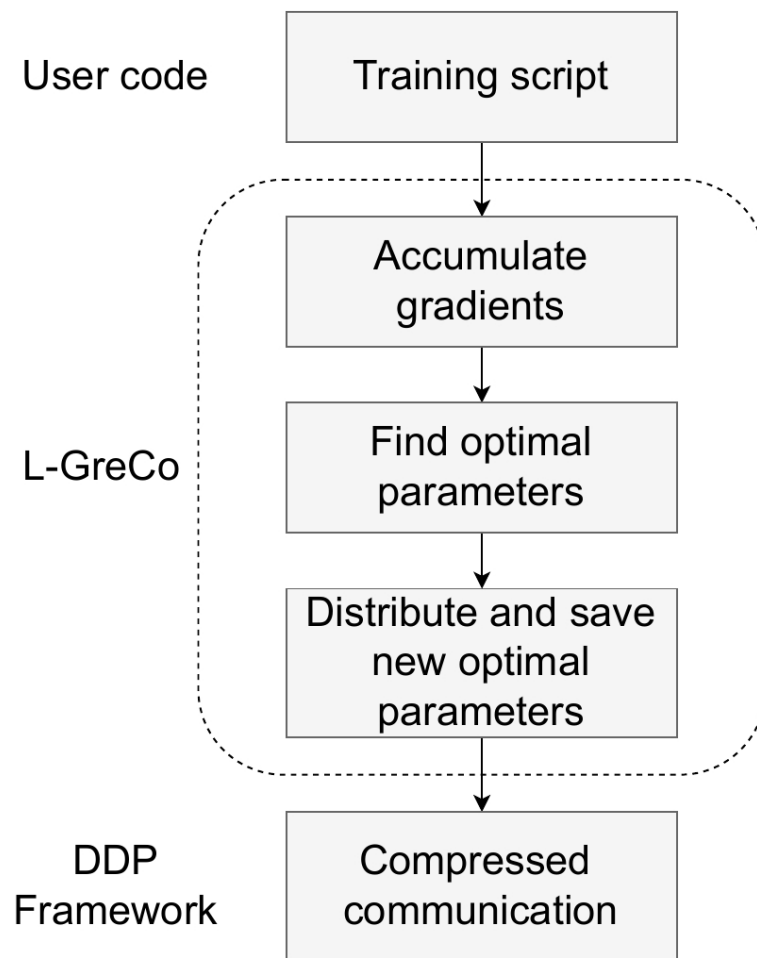


Figure 2. Schema of L-GreCo integration.

数学建模：逐层选择可写成受约束优化

$\mathcal{E}_{\max}$ 由一个已知能够恢复精度的均匀压缩配置给出

$$\text{minimize } \sum_{\ell=1}^L \text{size}(\ell, c^\ell) \quad \text{s.t.} \quad \sum_{\ell=1}^L \text{error}(\ell, c^\ell) \leq \mathcal{E}_{\max}.$$

目标：最小化总传输量

约束：各层 L2 压缩误差之和不超 $\mathcal{E}_{\max}$

决策变量：每一层的压缩参数 c^ℓ

核心算法：动态规划求解逐层配置

```
for c ∈ C do
    DP[1][Errors[1][c]] = Costs[1][c]
    PD[1][Errors[1][c]] = c
```

	误差：1	误差：2
第一层		
第二层		

Errors[1][c] = 本层梯度 - 低秩分解后的精度

DP[层数][误差值] = 累计通信量

PD[层数][误差值] = 选择的rank

Emax目标：是根据rank32算出来的近似全精度的误差

Errors[1][c]:每一层梯度模拟低秩矩阵，算出来的误差

计算的频率：每epoch

补充：误差多为浮点型，分栏离散化近似

核心算法：动态规划求解逐层配置

```
9: for  $c \in C$  do
10:    $DP[1][Errors[1][c]] = Costs[1][c]$ 
11:    $PD[1][Errors[1][c]] = c$ 
12: end for
13: // Dynamic programming algorithm
14: for Layer  $l_i := 2..N$  do
15:   for  $c_i \in C$  do
16:     for  $e_i := Errors[l_i][c_i]..D$  do
17:        $t = DP[l_i - 1][e_i - Errors[l_i][c_i]] + Costs[l_i][c_i]$ 
18:       if  $t < DP[l_i][e_i]$  then
19:          $DP[l_i][e_i] = t$ 
20:          $PD[l_i][e_i] = c_i$ 
21:       end if
22:     end for
23:   end for
24: end for
25:  $err_{min} = argmin(DP[N])$ 
26: // Reconstruction of the optimal parameters
27: for  $l_i = N..1$  do
28:    $result[l_i] = PD[l_i][err_{min}]$ 
29:    $err_{min} = err_{min} - Errors[l_i][result[l_i]]$ 
30: end for
31: return result
```

3 4 5 6 7

ei: 枚举本层每个可能的误差-当前层的rank误差

当前通信代价更小，更新当前误差下的通信代价

最后误差=通信最少的误差

回溯路径获取每个层rank值

实验配置

QSGD • TopK • PowerSGD

ResNet18 / CIFAR-100 ResNet50 / ImageNet

Transformer-XL / WikiText-103 Transformer-LM / WikiText-103

单节点：4 或 8 × RTX3090 多节点：4 台机器，每台 4 × RTX3090，节点间最高 10 Gbps

Table 1. Timing overheads for L-GreCo in relation to the total training time. Numbers in brackets represent error computation.

Compression Method	Transformer-XL	ResNet50
PowerSGD	0.56% [0.49%]	0.15% [0.14%]
QSGD	0.14% [0.13%]	0.04% [0.03%]
TopK	0.38% [0.35%]	0.33% [0.30%]

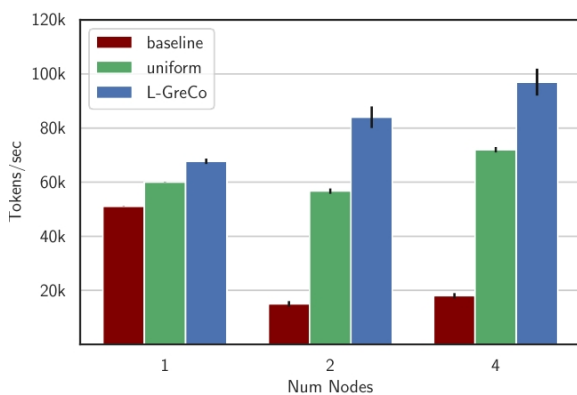
总开销最高 0.56%，其中动态规划只占极小比例。

同等精度下，压缩率最高提高 $5.2\times$

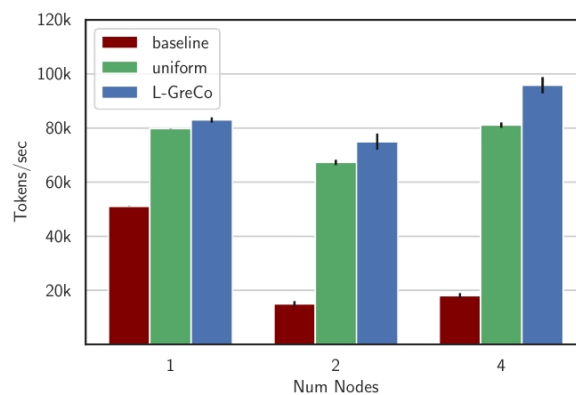
Compression approach	Parameter choice	ResNet18 on CIFAR-100			ResNet50 on ImageNet		
		Default param	Accuracy	Compression ratio	Default param	Accuracy	Compression ratio
Baseline	N/A	-	76.60 ± 0.40	1.0	-	76.88 ± 0.16	1.0
QSGD	uniform	4 bit	76.80 ± 0.40	7.8	4 bit	77.38 ± 0.10	7.7
	L-Greco		76.46 ± 0.21	$8.6 [1.10\times]$		76.77 ± 0.25	$11.0 [1.41\times]$
TopK	uniform	1%	75.73 ± 0.46	48.1	1%	76.85 ± 0.06	45.6
	L-Greco		75.66 ± 0.35	$182.0 [3.78\times]$		77.04 ± 0.27	$122.0 [2.67\times]$
PowerSGD	uniform	rank 4	76.36 ± 0.28	72.2	rank 4	76.50 ± 0.37	66.5
	L-Greco		76.43 ± 0.37	$133.9 [1.85\times]$		76.33 ± 0.27	$96.2 [1.44\times]$

Compression approach	Compression parameters	TransformerXL on WIKITEXT-103			TransformerLM on WIKITEXT-103		
		Default param	Perplexity	Compression ratio	Default param	Perplexity	Compression ratio
Baseline	N/A	-	23.82 ± 0.10	1.0	-	29.34 ± 0.12	1.0
QSGD	uniform	4 bit	23.82 ± 0.1	7.8	4 bit	29.39 ± 0.10	7.8
	L-Greco		24.11 ± 0.09	$9.1 [1.16\times]$		30.03 ± 0.16	$9.9 [1.26\times]$
TopK	uniform	10%	24.13 ± 0.14	4.9	10%	29.29 ± 0.09	4.9
	L-Greco		24.19 ± 0.13	$12.8 [2.61\times]$		29.08 ± 0.20	$25.6 [5.2\times]$
PowerSGD	uniform	rank 32	24.08 ± 0.12	14.0	rank 32	29.98 ± 0.09	15.0
	L-Greco		24.09 ± 0.15	$20.8 [1.48\times]$		30.19 ± 0.09	$26.5 [1.76\times]$

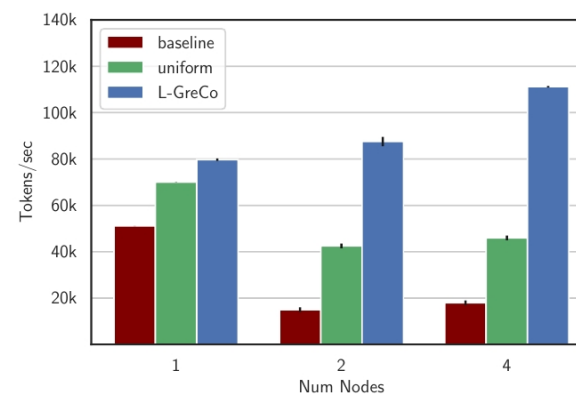
多节点带宽瓶颈下收益更显著



(a) PowerSGD

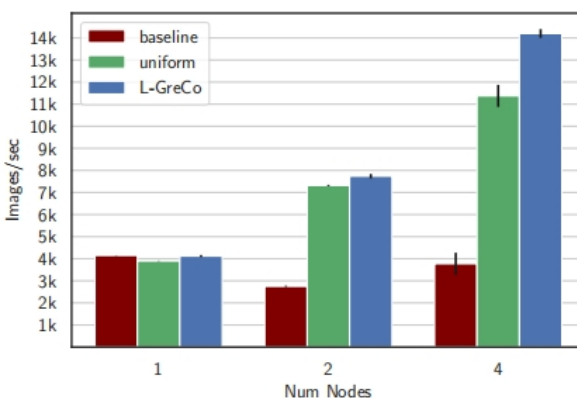


(b) QSGD

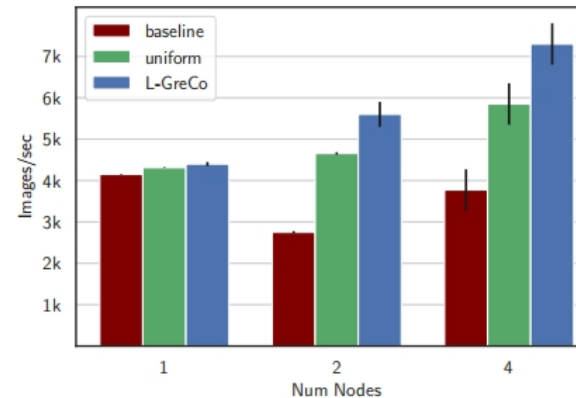


(c) TopK

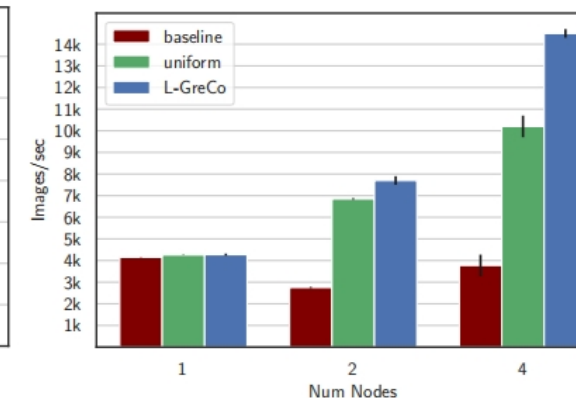
相对均匀压缩: Transformer-XL 最高 2.4×



(a) PowerSGD



(b) QSGD



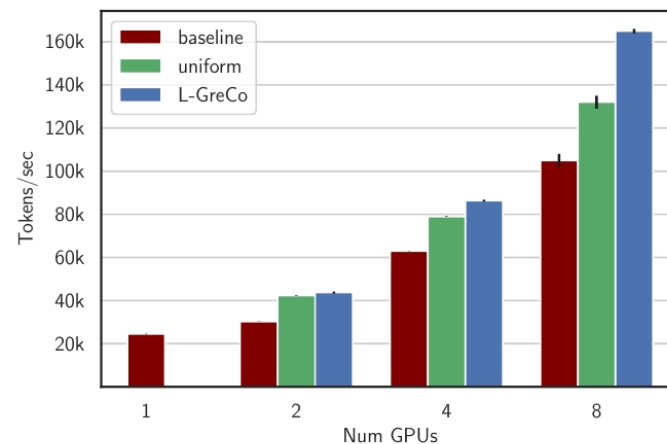
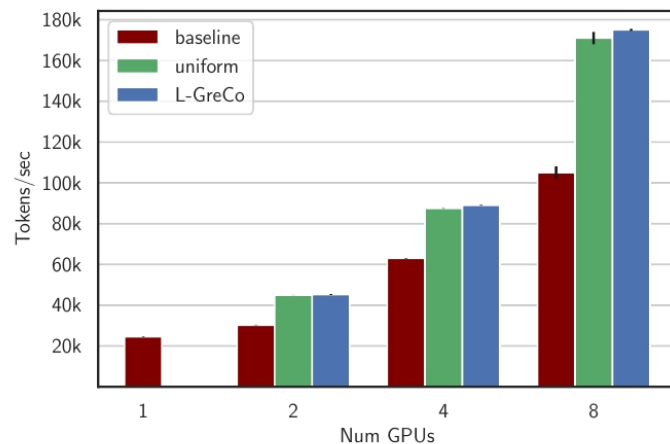
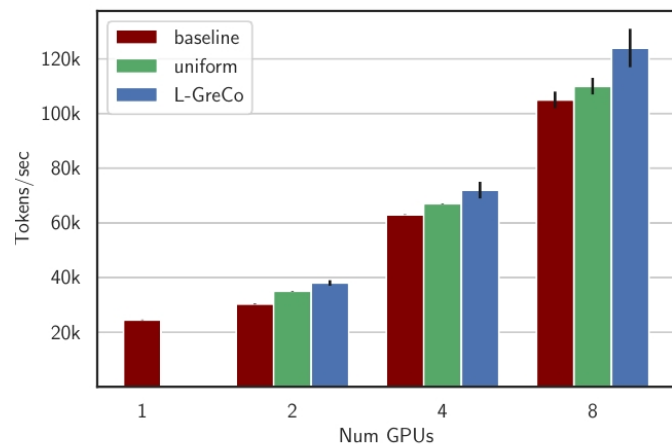
(c) TopK

Figure 5. Throughput for ResNet50/ImageNet. Multi node, each node has 4 RTX3090 GPUs.

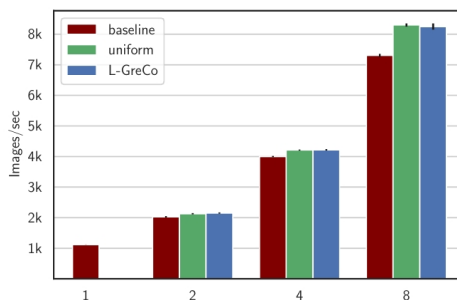
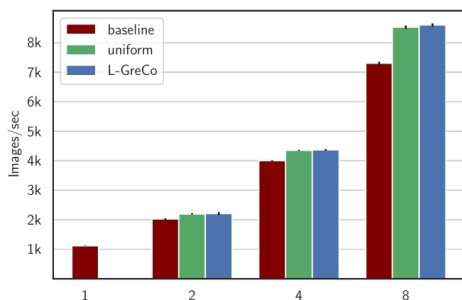
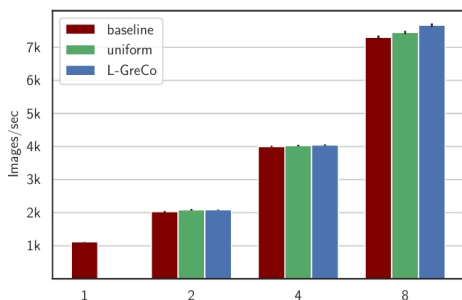
ResNet50 最高 1.45×

单节点收益取决于模型异质性

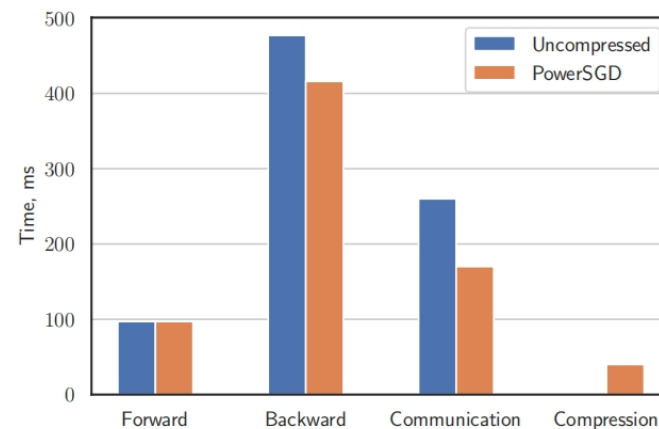
Transformer-XL / WikiText-103



ResNet50 / ImageNet



性能剖析



Transformer-XL 的 PowerSGD / TopK 端到端最高加速约 25%；
ResNet50 已接近理想通信配置，提升空间更小。

与 Accordion 结合可进一步增益

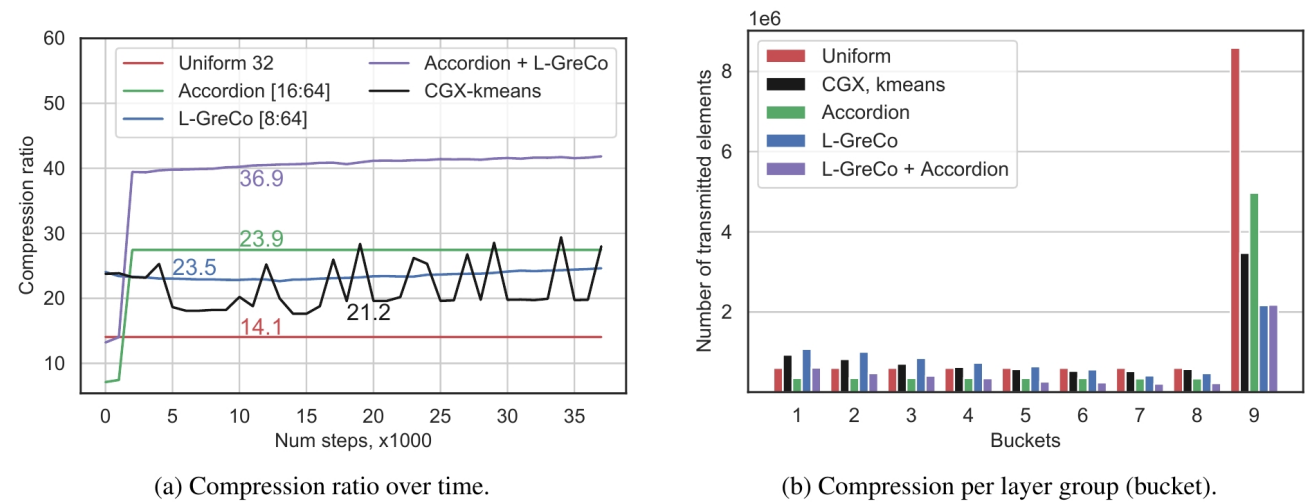


Figure 9. Adaptive compression using L-GreCo versus other methods, for PowerSGD compression on Transformer-XL. The left plot shows the dynamics of the compression ratio during training, marking the *average compression ratio*. The right plot presents the transmitted number of elements per bucket averaged over time. Buckets are in communication order.

Table 4. Comparison of L-GreCo with other adaptive algorithms on Transformer-XL using PowerSGD.

Adaptive algorithm	Param. Range	Ratio	Single node Tokens/s	Multi-node Tokens/s
Uniform	32	14.1	110k	72k
L-GreCo	8 - 64	23.5	144k	150k
Accordion	16, 64	23.9	114k	107k
CGX, kmeans	8 - 64	21.6	124k	112k
L-GreCo + Accordion	8 - 128	36.9	138k	176k

L-GreCo

23.5× 压缩率

144k / 150k tokens/s

L-GreCo + Accordion

36.9× 压缩率 · 176k tokens/s

Accordion 决定训练阶段的压缩强弱；L-GreCo 决定同一阶段内的层间分配，两者近似正交。