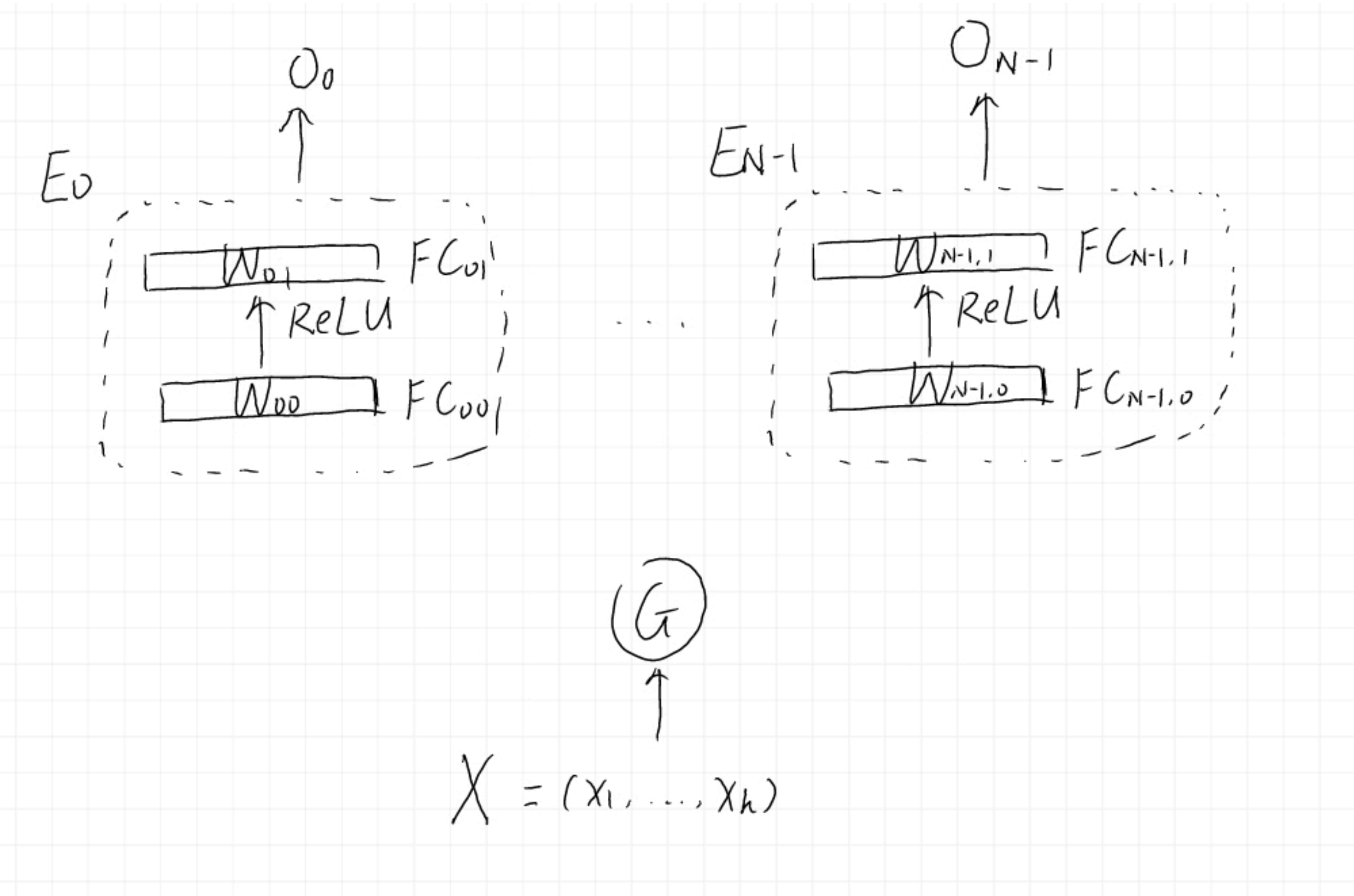
0. Notations:

- N: number of experts, $N > 1$
- d: number of slices per expert, $d > 1$
- h: length of embedding vectors
- m: dimension of the first fully connected layer
- n: dimension of the second fully connected layer
- b: batch size
- B_{comp}: computations per second
- B_{comm}: communications per second
- r: each token is routed to top-r experts ($1 \leq r < N$)
- T_0: time consumed to dispatch a batch of tokens when there is no slicing (it is also the time consumed to combine outputs of experts), note that $T_0 \leq \frac{\frac{b}{d} \cdot hr(d-1)}{B_{comm}} = \frac{bhr(d-1)}{dB_{comm}}$
- δ: time saved by doing expert-slicing per batch
- R: ratio of sliced MoE inference time over unsliced MoE inference time

1. Structure of Moe (according to GShard):

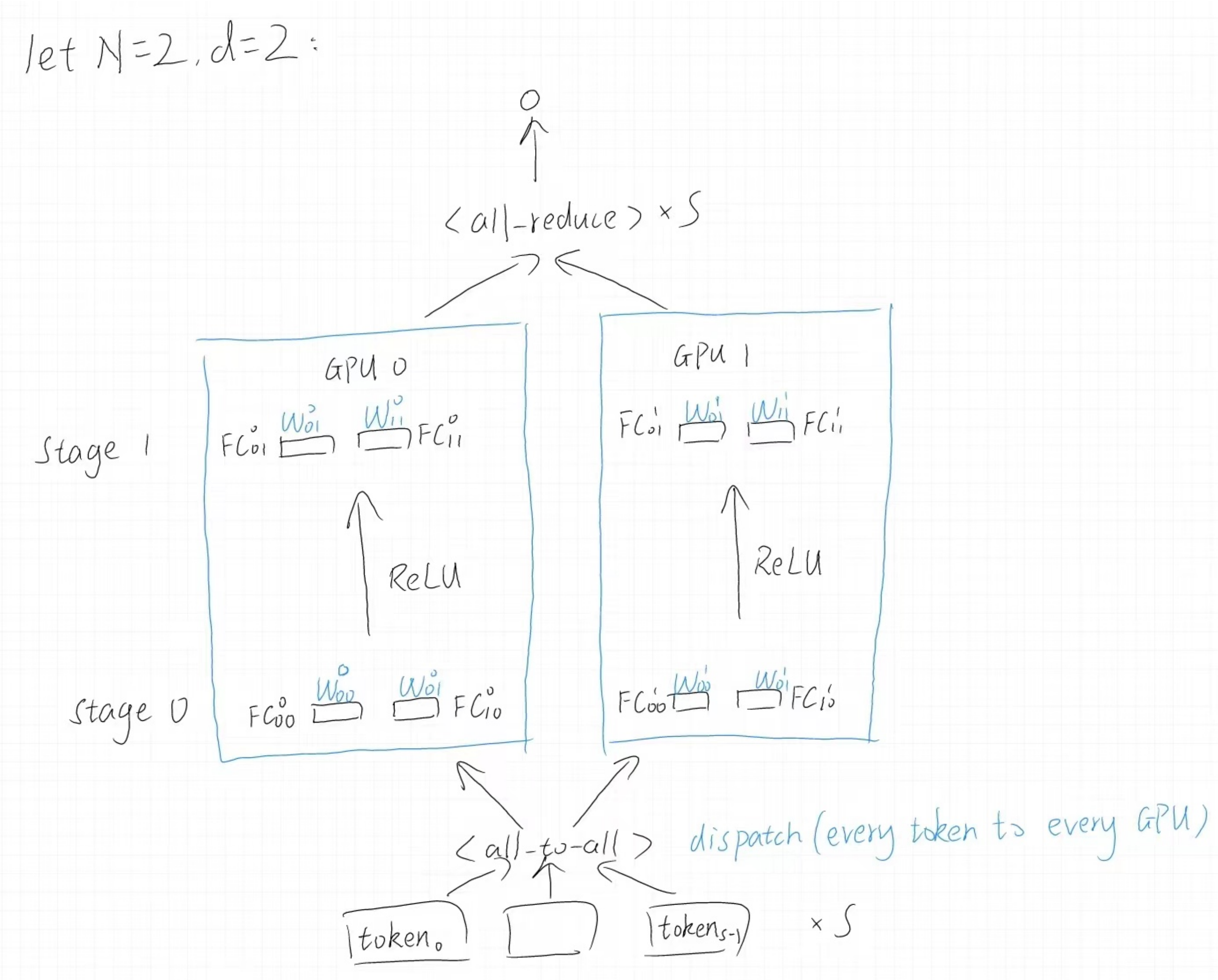


FC_{ij} stands for the j-th fully connected layer of the i-th expert, and W_{ij} denotes the matrix corresponding to FC_{ij}.

$o_i = W_{i1}\mathrm{Activation}(\mathrm{W_{i0}x}) \cdot \mathrm{G_i(x)}, \mathrm{G_i(x)} \in [0, 1]$

2. Slicing (N=d=2)

2.1 Structure



Assume for token x is routed to expert E_0.

$$W_{00} = \begin{bmatrix} W^0_{00} \\ W^1_{00} \end{bmatrix}, W_{01} = \begin{bmatrix} W^0_{01} & W^1_{01} \end{bmatrix}$$

$$t = \mathrm{Activation}(\mathrm{W_{00}x}) = \begin{bmatrix} \mathrm{Activation}(\mathrm{W^0_{00}x}) \\ \mathrm{Activation}(\mathrm{W^1_{00}x}) \end{bmatrix}$$

$$o = W_{01}t = W^0_{01}\mathrm{Activation}(\mathrm{W^0_{00}x}) + \mathrm{W^1_{01}}\mathrm{Activation}(\mathrm{W^1_{00}x})$$

2.2 Analysis

By doing expert-slicing, computation and communication cost change as follows:

	Before	After
Number of multiplications per GPU per token	$hm + mn$	$(hm + mn)/2$
Number of additions per GPU per token	$(h-1)m + (m-1)n$	$(h-1)m/2 + (m-1)n/2$
Number of activations per GPU per token	$m + n$	$m/2 + n/2$
Communication cost per batch	$2T_0$	$\frac{bh}{2B_{comm}} + \frac{bh}{2B_{comm}} = \frac{bh}{B_{comm}}$

3 General Situation

	Before	After
Number of multiplications per GPU per token	$(hm + mn) \cdot \frac{N}{d}$	$(hm + mn) \cdot \frac{N}{d^2}$
Number of additions per GPU per token	$[(h-1)m + (m-1)n] \cdot \frac{N}{d}$	$[(h-1)m + (m-1)n] \cdot \frac{N}{d^2}$
Number of activations per GPU per token	$(m+n) \cdot \frac{N}{d}$	$(m+n) \cdot \frac{N}{d^2}$
Communication cost per batch	$2T_0$	$\frac{bh(d-1)}{dB_{comm}} + \frac{bh(d-1)}{dB_{comm}} = \frac{2bh(d-1)}{dB_{comm}}$

$$\delta = 2b(1 - \frac{1}{d})\frac{hm+mn}{B_{comp}} \cdot \frac{N}{d} + 2T_0 - \frac{2bh(d-1)}{dB_{comm}}$$

Since $1 - \frac{1}{d} > 0$ while $T_0 \leq \frac{bh(d-1)}{dB_{comm}}$, there is a trade-off between less computation cost and more communication cost. If δ proves **positive**, then expert-slicing is effective.

Moment Estimation of T_0, δ, R

Denote the number of communications required for token $x_{ij}(0 \leq j \leq \frac{b}{d} - 1)$ on GPU $i(0 \leq i \leq d-1)$ as $G_{ij} \in [\frac{(r-1)N}{d}, \frac{rN}{d}]$, the number of communications between GPU i and other GPUs as M_i.

It is well-known that $\binom{d}{r} = \binom{d-1}{r-1} + \binom{d-1}{r}$.

$$E(G_{ij}) = \frac{N}{d}[(r-1) \cdot \frac{\binom{d-1}{r-1}}{\binom{d}{r}} + r \cdot \frac{\binom{d-1}{r}}{\binom{d}{r}}] = \frac{Nr(d-1)}{d^2}$$

$$E(M_i) = E(\sum_{j=0}^{\frac{b}{d}-1} G_{ij}) = \sum_{j=0}^{\frac{b}{d}-1} E(G_{ij}) = \frac{bNr(d-1)}{d^3}$$

$$E(T_0) = E(\frac{hM_i}{B_{comm}}) = \frac{bNhr(d-1)}{d^3 B_{comm}}, 1 \leq r < N$$

For sparsely-gated MoE network, assuming $m = 4h$, $n = h$, $r = 1$, we have

$$\hat{\delta} = 2bh(1 - \frac{1}{d})[\frac{8h}{B_{comp}} \cdot \frac{N}{d} + \frac{1}{B_{comm}}(\frac{N}{d^2} - 1)]$$

$$\hat{\delta} > 0 \Leftrightarrow h > \frac{1}{8}\frac{B_{comp}}{B_{comm}}(1 - \frac{N}{d^2}) \cdot \frac{d}{N}$$

$$\hat{R} = \frac{\frac{16h^2 \cdot \frac{N}{d^2}}{Bcomp} + \frac{2h(d-1)}{dBcomm}}{\frac{16h^2 \cdot \frac{N}{d}}{Bcomp} + \frac{2Nh(d-1)}{d^3 Bcomm}} \rightarrow \frac{1}{d}(h \rightarrow \infty)$$