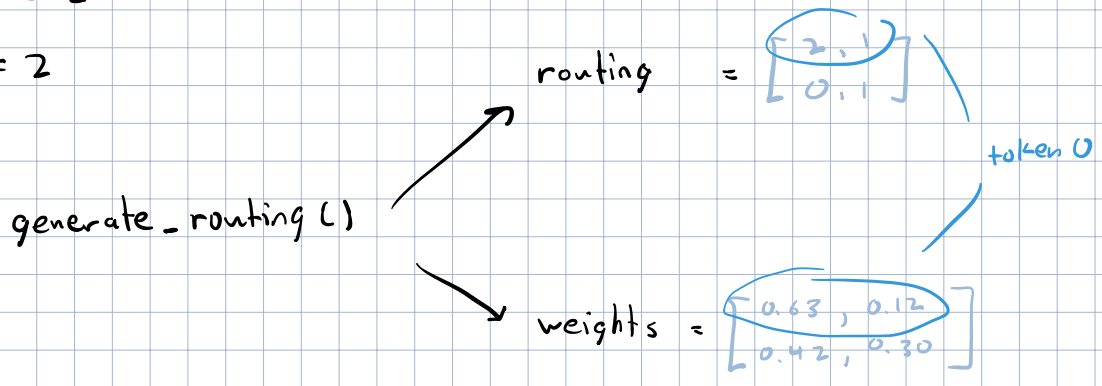
SIMULATION

- HYPERPARAMETERS

- Num_experts = 4
- Seqlen = 2
- Top_K = 2



- Hot_ratio = 0.5
- Hot_weight = 0.8

=> 50% of experts get 80% of routings

weights → npu_identify() → labelled weights

[0.63, 0.12 / 0.42, 0.30]

[0.63, 0.12, 0, 0 / 0.42, 0.30, 1, 1]

src npu id

token id

PERFORMANCE MODEL

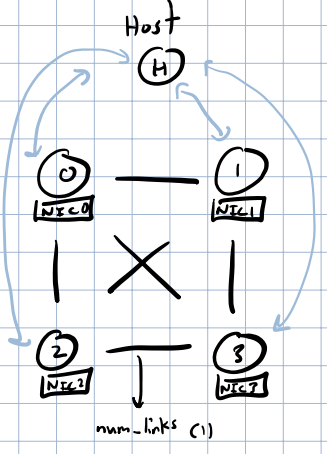
labelled weights + routing → convert_to_bytes() → load dictionary

{0: {0: 0, 1: 2}, 1: {0: 3, 1: 0}}

src: 0, 1 — dest: 0, 1 — load: 0, 2, 3, 0

* Num_links = 1
* Num_nodes = 4

Mock 1st round



* initial cpu delay

loop through load and send packets until one or multiple of these:

1) No more load to send (DONE)

2) No more available links to send remaining load

3) No more bandwidth in desired links (leads to 2))

round_time = (most packets in one link) * packet prep delay

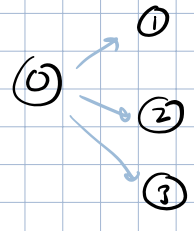
+ (largest bandwidth usage) / intra_bw

+ Base_delay GPU → CPU confirmation

Round robin between dest. nodes for * round_robin_max_packets per dest

= 2

{0: {1: 3, 2: 4, 3: 1}}



#1) 0 → 1 2B

#2) 0 → 2 2B

#3) 0 → 3 1B

#4) 0 → 1 1B

#5) 0 → 2 2B

Assuming a balanced routing,

$$\text{Round\_time} = \left(*\text{initial\_cpu\_delay} + *\text{base\_delay} + \left\lfloor \frac{*\text{intra\_bw}}{*\text{packet\_size}} \right\rfloor \left(*\text{packet\_prep\_delay} + \frac{*\text{packet\_size}}{*\text{intra\_bw}}\right)\right)$$

$$\text{Num\_round} = \frac{* \text{load\_per\_node} \cdot *\text{num\_nodes}}{\left(\frac{*\text{num\_links} \cdot *\text{num\_nodes} \cdot (*\text{num\_nodes} - 1)}{2}\right) \cdot *\text{intra\_bw}}$$

$$\text{Total\_time} = \text{Round\_time} \cdot \text{Num\_rounds}$$

* load = {0: {0:0, 1:10, 2:10},
1 = {0:10, 1:0, 2:10},
2 = {0:10, 1:10, 2:0}}

⇓

load_per_node = 20