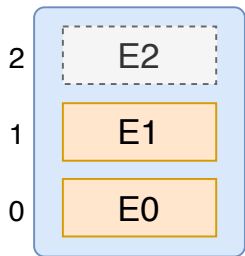


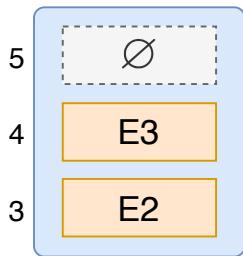
Logical Experts: E0, E1, ..., E7

Physical Experts: 0, 1, ..., 11

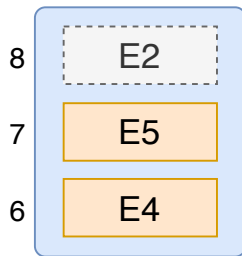
Physical Slots



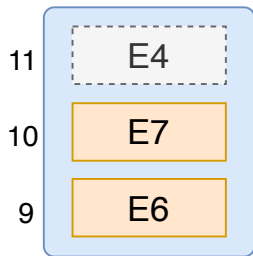
EP Rank 0



EP Rank 1



EP Rank 2



EP Rank 3



Main Experts (Persistent).

**Weight
Grad.
Optim.**

$N_{\text{layers}} \times |\mathcal{E}_r|$
(Layer-independent)



Redundant Slots (Transient).

**Weight
Grad.
Optim.**

$1 \times N_{\text{slot}}$ (Layer-reused)
None