1 Appendix

Algorithm 1 Layer Redistribution with Migration Coordinator

Input: Current layer assignment L_{curr}, Adjusted layer assignment L_{adj}
Output: Efficient migration with minimal inference disruption
Initialization Phase:
    Pre-allocate memory for potential reassigned layers
    Preload weights for all candidate layers
On Scheduler Trigger of Redistribution:
    $M \leftarrow$ IdentifyMigratedLayers(L_{curr}, L_{adj})
for each layer l in M **do**
    if l is to be sent **then**
        Wait for computation of l to complete on source stage
        Asynchronously send KV cache of l to target stage
    else if l is to be received **then**
        Asynchronously receive KV cache
    end if
end for
During Inference Execution:
for each layer l assigned to target stage **do**
    if $l \in M$ **then**
        Wait until KV cache for l is received
    end if
    Execute forward computation for l
end for