Name of API	Arguments	Semantics of API
`cudaMallocRemote`	`&src, size`	Allocate `size` bytes of memory to `TensorNode` and return ptr to `src`
`cudaFreeRemote`	`&src`	Deallocate remote memory under `TensorNode` DIMMs
`cudaMemcpyAsync`	`&src, &dst, size, direction`	Copy `size` bytes from `src` to `dst`. `direction` can be set as `LocalToRemote` or `RemoteToLocal`

: TensorNode runtime API extensions. The proposed runtime API extensions build upon prior work [13, 37] which enables remote memory (de)allocation under a GPU-side disaggregated memory system, and DMA-invoked data copy operations across local and remote memory. [13, 37] include further details on the system software support required for these runtime CUDA API extensions in a GPU-side disaggregated memory system.