Esha Choukse, Michael B. Sullivan, Mike O'Connor, Mattan Erez, Jeff Pool, David Nellans, and Stephen W. Keckler, "Buddy Compression: Enabling Larger Memory for Deep Learning and HPCWorkloads on GPUs," in *arxiv.org*, 2019

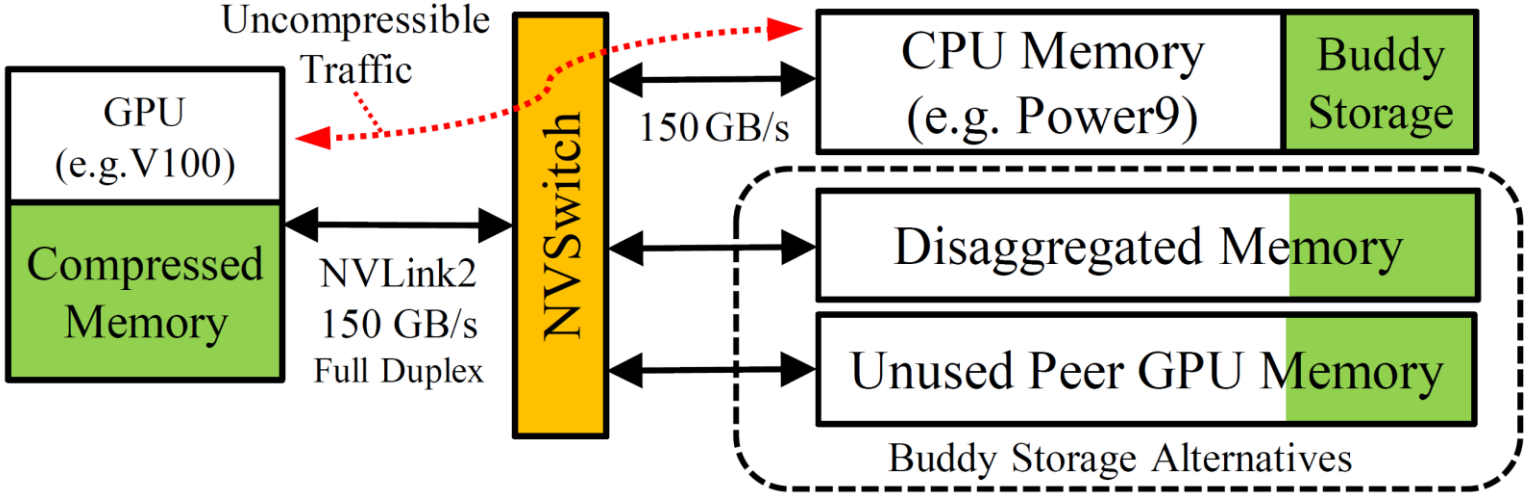

Fig. 2: A target system for Buddy Compression. Any larger NVLink-connected memory is used as buddy storage. Overall organization is like NVIDIA DGX-2 [20].