This paper presents prediction-based dynamic resource allocation algorithms to scale video transcoding service on a given Infrastructure as a Service cloud. The proposed algorithms provide mechanisms for allocation and deallocation of virtual machines (VMs) to a cluster of video transcoding servers in a horizontal fashion. We use a two-step load prediction method, which allows proactive resource allocation with high prediction accuracy under real-time constraints. For cost-efficiency, our work supports transcoding of multiple on-demand video streams concurrently on a single VM, resulting in a reduced number of required VMs. We use video segmentation at group of pictures level, which splits video streams into smaller segments that can be transcoded independently of one another. The approach is demonstrated in a discrete-event simulation and an experimental evaluation involving two different load patterns.