Another way to think about the different regions of operation is to consider the how collector and base currents are related. In the graph below, we see that at cutoff, Ib and Ic are equal and are zero. In the linear region, $Ic = \beta Ib$. In the saturated region Ib and Ic are not zero but further increases in Ib barely increase Ic at all.



When we use a BJT as a saturated switch, we usually supply base current well in excess of what is needed to keep the transistor in saturation.