Scaling of servers by request rate