הסבר על מדיניות התזמון

הload balancer שמיממשנו משתמש במדיניות תזמון המבוססת על עומס מינימלי משוקלל (weighted least) כדי להפיץ בקשות נכנסות מלקוחות בין השרתים. מדיניות זו מתאימה את עצמה באופן דינמי בהתבסס על העומס הנוכחי והצפוי על כל שרת, וכך מבטיחה הפצה שווה של הבקשות.

אופן פעולת ה load balancer:

1. **מעקב אחר עומס:**
   - הload balancer מחזיק במילון (server_loads) שעוקב אחר העומס הנוכחי על כל שרת.
   - הload balancer מעדכן את העומס הנוכחי לכל שרת כשמגיעה בקשה, בהתחשב בזמן שחלף מאז העדכון האחרון.
2. **טיפול בבקשות:**
   - כאשר מתקבלת בקשה חדשה מלקוח, הload balancer בוחר את השרת עם העומס הצפוי המינימלי לטיפול בבקשה על ידי חישוב פשוט עם המשקלים שניתנו לנו בתרגיל.
   - הבקשה מועברת לשרת הנבחר, ונפתח thread שמטפל בבקשה הזו.
3. **חישוב עומס משוקלל:**
   - לכל סוג בקשה יש משקל המשקף את ההשפעה שלה על העומס של השרת לפי הנחיות התרגיל.
   - העומס הצפוי לכל שרת מחושב על ידי התחשבות בעומס הנוכחי והוספת העומס המשוער מהבקשה הנכנסת, בהתבסס על סוגה ומשך הזמן שלה.

מדיניות תזמון זו מבטיחה שהעומס מחולק באופן שווה בין כל השרתים על ידי התחשבות בעומס הנוכחי והצפוי, מה שעוזר באופטימיזציה של הביצועים ומניעת היווצרות צוואר בקבוק באף שרת בודד.

הערה - רעיון נוסף שלא מומש:

לשמור איזשהו תור של הבקשות על ה loadbalancer, ולבצע בחירה אופטימלית בכל פעם שמתפנה שרת (על ידי סוג של בקטרקינג או ע"י בחירת הבקשה הכי כבדה שמתאימה לשרת שהתפנה), לא מומש מכיוון שלמטרות התרגיל זה יכל לעבוד טוב, אך לדעתינו לא יהיה scalable ולא בדיוק תואם עם עקרונות ה load balancing.