מדיניות תזמון

במימוש ה-load balancer שלנו נעשה שימוש במדיניות תזמון מבוססת חיזוי זמן סיום משוער, תוך התחשבות בסוג הבקשה, סוג השרתים והעומס הנוכחי על כל שרת.

לכל בקשה חדשה, מחושב עבור כל שרת $i \in \{1,2,3\}$:

$$\text{finish\_time}_i = \max(\text{current\_time}, \text{server\_free\_time}_i) + \text{request\_cost}$$

כלומר סכום הזמן הנוכחי (אם זו הבקשה הראשונה) או הזמן בו השרת פנוי (עבור שאר הבקשות) עם זמן טיפול הבקשה, כאשר $\text{request\_cost} = \text{request\_time} \times \text{multiplier}$ ו-multiplier מחושב לפי:

Server Type \ Request type	M (Music)	V (Video)	P (Picture)
VIDEO	X2	X1	X1
MUSIC	X1	X3	X2

השרת עם הזמן הקצר ביותר נבחר לביצוע הבקשה ומעודכן הזמן בו השרת יהיה פנוי.

לאחר קבלת תגובה מהשרת, הזמן מתעדכן לזמן אמת (במקום להניח שעיבוד הבקשה לקח בדיוק את הזמן הצפוי).

במקרה של כמה שרתים עם זמני סיום דומים, ייבחר הראשון מביניהם ללא עדיפות כלשהי.