

# დაგეგმვის ალგორითმების მიმოხილვა GPU-ს ფრაგმენტაციის პრევენციისთვის

*ირაკლი კობერიძე<sup>ა</sup>*

ელ-ფოსტა: irakli.koberidze776@ens.tsu.edu.ge

<sup>ა</sup> კომპიუტერულ მეცნიერებათა დეპარტამენტი, ზუსტ და საბუნებისმეტყველო მეცნიერებათა ფაკულტეტი, ივანე ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტი.

## ანოტაცია

თანამედროვე AI ინფრასტრუქტურა კრიტიკული გამოწვევის წინაშე დგას რაც გამოიხატება ძვირადღირებული GPU რესურსების აუთვისებლობით. მაშინ როცა GPU-ს გაზიარება იძლევა ფრაქციული განაწილების საშუალებას, მას შემოაქვს ფრაგმენტაციის პრობლემა, სადაც თავისუფალი რესურსის ნაწილები გამოუყენებელი ხდება აპარატურული უწყვეტობის შეზღუდვის გამო. ეს ტექსტი ანალიზებს დაგეგმარების ჰეურისტიკების ევოლუციას, რომლებიც შექმნილია ამ საკითხის გადასაწყვეტად. ჩვენ მივყვებით გზას პრაგმატული, კლასტერზე ორიენტირებული პოლიტიკებიდან და დახვეწილი მრავალგანზომილებიანი "ბინ-ფეიკინგ" (bin-packing) ალგორითმებიდან მოწინავე თეორიულ მოდელებამდე. ჩვენ ვაჩვენებთ, რომ მიუხედავად იმისა, რომ ეს მიდგომები გვთავაზობენ თანდათანობით გაუმჯობესებებს, მათ მაინც გააჩნიათ შეზღუდვები. ანალიზი კულმინაციას აღწევს "ფრაგმენტაციის გრადიენტული დაშვებით" (FGD), უახლესი ალგორითმით. FGD შემოიტანს ფრაგმენტაციის სტატისტიკურ, დატვირთვაზე ორიენტირებულ საზომს და გეგმავს ამოცანებს მომავალი რესურსების გამოუყენებლობის მინიმიზაციისთვის. ეს მიდგომა პირდაპირ პასუხობს წინა მეთოდების მთავარ შეზღუდვას და გვთავაზობს გადაწყვეტას GPU კლასტერების ეფექტურად გაზიარებულ სერვისებად გარდასაქმნელად, რაც აუცილებელია მდგრადი და მასშტაბირებადი AI პროგრესისთვის.