ABSTRAK

SYAHRUL RAMADHAN, Implementasi Mekanisme *Queue* dan *Worker* untuk Optimasi Utilisasi GPU dalam Pemrosesan Layanan AI pada *Backend*. (Dibimbing oleh Muhyiddin AM Hayat, S.Kom., M.T dan Rizki Yusliana Bakti S.Kom., M.T.).

Perkembangan layanan *Artificial Intelligence* (AI) berbasis *Large Language Model* (LLM) meningkatkan kebutuhan akan sistem *backend* yang mampu menangani proses inferensi secara efisien. Proses inferensi yang dijalankan pada *Graphics Processing Unit* (GPU) bersifat komputasi intensif sehingga berpotensi menimbulkan *bottleneck* dan pemanfaatan GPU yang kurang optimal. Penelitian ini bertujuan mengimplementasikan mekanisme *queue* dan *worker* pada *backend* layanan AI berbasis GPU, menganalisis pengaruh variasi jumlah *worker* terhadap performa sistem dan utilisasi GPU, serta menentukan konfigurasi *worker* yang paling optimal. Sistem dikembangkan menggunakan bahasa pemrograman Go dengan *layered architecture*, Redis sebagai *message queue*, *Asynq* sebagai pengelola antrean dan *worker*, serta vLLM sebagai layanan inferensi AI berbasis GPU. Pengujian dilakukan menggunakan K6 dengan beban 50 *virtual users* selama 60 detik pada konfigurasi 1, 2, 4, 8, dan 16 *worker*. Parameter yang dianalisis meliputi *response time*, *throughput*, *error rate*, utilisasi GPU, dan GPU *idle time*. Hasil penelitian menunjukkan bahwa peningkatan jumlah *worker* mampu meningkatkan *throughput* dari 26,60 menjadi 142,91 *request*/s, menurunkan *response time* dari 1904 menjadi 343,54 ms, serta meningkatkan utilisasi GPU dari 80,21% menjadi 100%. Seluruh konfigurasi menghasilkan *error rate* sebesar 0%, sedangkan GPU *idle time* menurun dari 18,3% menjadi 0%. Berdasarkan seluruh parameter pengujian, konfigurasi 8 *worker* memberikan keseimbangan terbaik antara kecepatan respons, kemampuan pemrosesan, dan efisiensi pemanfaatan GPU sehingga menjadi konfigurasi yang paling optimal pada lingkungan pengujian penelitian ini.

Kata Kunci: *Artificial Intelligence*, *Backend*, *Queue*, *Worker Pool*, Redis, *Asynq*, vLLM, Utilisasi GPU, *Response Time*, *Throughput*.