BAB I

PENDAHULUAN

1.1. Latar Belakang

Dalam tengah bergulirnya era transformasi digital yang berkembang dengan pesat, teknologi cloud computing telah mengukuhkan dirinya sebagai pilar utama dalam mendukung efisiensi penyediaan layanan dan sumber daya komputasi. Paradigma kemudahan akses, skalabilitas, dan pelayanan on-demand yang diusung oleh cloud computing membawa perubahan mendasar dalam pemandangan layanan teknologi informasi. Mulai dari penyimpanan data hingga pemrosesan informasi yang bersifat kompleks, cloud computing menjadi fondasi utama yang memberdayakan berbagai sektor industri dan bisnis(John et al., n.d.).

Namun, seiring dengan pertumbuhan eksponensial adopsi cloud, terutama dalam skala bisnis dan industri, muncul sejumlah tantangan signifikan terkait manajemen beban kerja atau load balancing. Manajemen beban kerja menjadi elemen kunci dalam menjaga distribusi beban kerja di antara server-server cloud agar dilakukan secara merata. Tidak hanya untuk memastikan kinerja sistem yang optimal(Ojha et al., 2020), tetapi juga untuk meminimalkan waktu respons yang dapat mengalami degradasi akibat peningkatan beban kerja yang bersifat fluktuatif.

Peran krusial load balancing dalam ekosistem cloud computing tidak dapat diabaikan, khususnya ketika dihadapkan pada tantangan peningkatan beban kerja yang terjadi dengan cepat dan efisien(Fantacci & Picano, 2020). Dalam dinamika lingkungan cloud yang terus berkembang, load balancing menjadi tulang punggung dalam menjaga distribusi beban kerja di antara server-server. Fungsinya tidak hanya terbatas pada pengaturan giliran pemrosesan data, melainkan juga pada aspek penting dalam mendistribusikan beban kerja dengan seimbang di seluruh infrastruktur. Konsep utama load balancing adalah memastikan bahwa setiap server berkontribusi secara adil dan optimal dalam pemrosesan data, memastikan bahwa sumber daya komputasi dimanfaatkan secara maksimal. Oleh karena itu, load balancing bukan hanya sekadar kebutuhan teknis, melainkan juga

menjadi suatu tuntutan mendesak untuk menjaga stabilitas dan performa optimal dalam lingkungan cloud yang terus berkembang(Elnagar et al., 2022).

Dalam konteks ini, load balancing menjadi elemen penting untuk merespons dinamika permintaan layanan yang terus meningkat. Sejalan dengan pertumbuhan eksponensial pengguna dan permintaan layanan, kemampuan load balancing untuk menyesuaikan diri dengan fluktuasi ini menjadi krusial. Hal ini tidak hanya berdampak pada kinerja sistem secara keseluruhan tetapi juga menentukan pengalaman pengguna akhir. Oleh karena itu, pengembangan dan peningkatan strategi load balancing yang cerdas dan adaptif menjadi semakin mendesak untuk memastikan bahwa infrastruktur cloud dapat memberikan layanan dengan efisien, handal, dan optimal, sesuai dengan tuntutan lingkungan komputasi awan yang terus berkembang pesat.

Algoritma Round Robin, sebagai salah satu strategi utama dalam load balancing, merupakah pendekatan yang meratakan pembagian tugas pemrosesan data di antara sejumlah server secara bergantian(Andhyka & Badri, 2019). Dengan memberikan giliran secara adil kepada setiap server, algoritma ini bertujuan untuk memastikan distribusi beban kerja yang seimbang dan efisien. Fungsinya sebagai instrumen kritis dalam ekosistem cloud computing sangat signifikan, karena mampu mengoptimalkan kinerja sistem dengan meminimalkan waktu respons(Abed & Younis, 2019). Meski demikian, dalam menghadapi dinamika yang terus berkembang dalam lingkungan cloud, tantangan meningkatkan efektivitas algoritma Round Robin menjadi semakin kompleks dan menarik perhatian.

Di tengah kompleksitas tersebut, peningkatan terhadap kemampuan adaptif algoritma Round Robin menjadi aspek yang esensial. Kemampuan untuk secara dinamis menyesuaikan diri dengan fluktuasi permintaan layanan dan perubahan beban kerja dapat menjadi kunci untuk menjaga keseimbangan yang optimal di lingkungan cloud yang dinamis(Prasad, 2022). Oleh karena itu, penelitian lebih lanjut dan pengembangan strategi yang dapat meningkatkan daya tanggap dan kecerdasan adaptif algoritma Round Robin menjadi perlu guna menjawab tuntutan terkini dalam pengelolaan beban kerja di lingkungan komputasi awan yang terus berkembang pesat(Kushwaha et al., 2021).

Algoritma Least Connections merupakan salah satu strategi yang banyak diterapkan dalam mengelola distribusi beban pada sistem jaringan(PUTRA SUJARWO et al., 2023). Konsep dasar dari algoritma ini adalah untuk mengarahkan lalu lintas pengguna ke server yang memiliki jumlah koneksi terendah pada suatu waktu tertentu. Dengan kata lain, server dengan beban koneksi paling minim akan menjadi prioritas dalam menangani permintaan pengguna baru(Oyediran et al., 2024). Dengan menerapkan algoritma Least Connections, tujuan utama adalah mencapai distribusi beban yang seimbang di antara server-server yang tersedia, sehingga mampu mengoptimalkan pemanfaatan sumber daya dan menghindari situasi di mana beberapa server mengalami beban yang berlebihan sementara yang lain mungkin belum termanfaat sepenuhnya(Oduwole et al., 2022). Penerapan algoritma ini tidak hanya bertujuan untuk meningkatkan efisiensi penggunaan sumber daya, tetapi juga untuk meningkatkan ketersediaan dan responsivitas layanan jaringan secara keseluruhan. Dalam konteks ini, algoritma Least Connections menjadi instrumen penting dalam mencapai tujuan manajemen distribusi beban yang optimal di lingkungan jaringan yang dinamis dan seringkali berubah(Oduwole et al., 2022; Wira Harjanti et al., 2022a).

Sementara itu, ada beberapa penyedia layanan cloud terkemuka, seperti Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP), Cloudflare dan beberapa penyedia lainnya yang berperan penting dalam transformasi digital. Setiap penyedia menawarkan beragam layanan termasuk komputasi, penyimpanan, basis data, kecerdasan buatan, dan solusi terkait lainnya, membentuk lanskap teknologi cloud yang sangat kompleks.

Keberadaan penyedia layanan cloud juga menciptakan dampak positif terhadap stabilitas dan kehandalan sistem secara keseluruhan. Infrastruktur yang terdistribusi ini memungkinkan untuk secara dinamis menyesuaikan distribusi beban kerja, bahkan dalam menghadapi fluktuasi permintaan yang signifikan.

Perancangan VPS dan multi-region load balancing menggunakan Google Cloud Platform (GCP) bertujuan untuk menciptakan infrastruktur cloud yang tangguh dan dapat diandalkan. Dengan menggunakan layanan VPS dari GCP, aplikasi dapat dijalankan dengan efisien dan aman di lingkungan virtual yang

terisolasi. Sementara itu, penggunaan multi-region load balancing memungkinkan distribusi lalu lintas yang seimbang di berbagai pusat data Google di seluruh dunia, meningkatkan ketersediaan aplikasi dan responsif terhadap permintaan pengguna secara global(Grevalby et al., 2023). Dengan kombinasi ini, infrastruktur cloud dapat mengatasi tantangan kinerja dan ketersediaan dengan lebih baik, memberikan pengalaman pengguna yang optimal.

Penerapan algoritma Round Robin pada Google Cloud Platform (GCP) umumnya dilakukan dalam konteks distribusi lalu lintas di antara beberapa instance atau server yang identik. Dalam kasus ini, Round Robin akan secara bergantian mengarahkan permintaan pengguna ke setiap instance yang tersedia, memastikan bahwa beban kerja terdistribusi secara merata di antara sumber daya yang ada. GCP menyediakan layanan Load Balancer yang dapat dikonfigurasi untuk menggunakan algoritma Round Robin dalam mengatur distribusi lalu lintas(Gao & Wu, 2022). Melalui pengaturan konfigurasi yang sesuai, Load Balancer akan secara otomatis menangani penyebaran lalu lintas di antara instance-instance yang diinginkan, membantu mengoptimalkan kinerja aplikasi dan meningkatkan ketersediaan layanan. Dengan menerapkan Round Robin pada GCP, pengguna dapat mengoptimalkan penggunaan sumber daya mereka, menghindari overloading pada satu instance tertentu, dan meningkatkan ketersediaan layanan secara keseluruhan(Desmulyati & Perdana Putra, 2019).

Kombinasi antara algoritma load balancing seperti Round Robin dan solusi terkemuka seperti Google Cloud Platform (GCP) menciptakan sinergi yang dapat memberikan kinerja optimal dalam lingkungan cloud computing yang dinamis. Penelitian lebih lanjut terhadap integrasi antara algoritma load balancing dan solusi penyedia cloud seperti GCP menjadi penting untuk terus mengembangkan strategi yang adaptif dan cerdas dalam mengatasi tantangan load balancing di era cloud computing yang terus berkembang. Dengan begitu, sistem cloud dapat memberikan layanan yang andal, efisien, dan responsif, sesuai dengan tuntutan zaman yang terus berubah.

1.2. Identifikasi Masalah

Berdasarkan konteks latar belakang yang telah diuraikan, rumusan masalah dapat dirumuskan seperti:

1. Terjadinya pengarahan semua permintaan pengguna ke satu server tanpa adanya load balancing menyebabkan server rentan terhadap overload, yang dapat mengakibatkan respons aplikasi yang lambat dan downtime.
2. Pendistribusian lalu lintas antara server yang tersedia tidak merata, yang berpotensi menurunkan efisiensi dan kinerja sistem secara keseluruhan.
3. Terjadinya kegagalan terhadap responsivitas layanan seiring meningkatnya beban lalu lintas.

1.3. Tujuan Penelitian

Berdasarkan dari identifikasi masalah yang ada, pada penelitian ini memiliki tujuan diantaranya:

1. Menganalisis efektivitas algoritma Round Robin dan Least Connections dalam mendistribusikan beban kerja.
2. Menganalisis seberapa merata keduanya membagi lalu lintas di antara server-server yang tersedia.
3. Mengevaluasi bagaimana keduanya mengatasi peningkatan beban lalu lintas dan sejauh mana dapat mempertahankan responsivitas layanan.

1.4. Manfaat

Manfaat yang diharapkan dari penelitian ini mencakup:

1. Dapat mengetahui tingkat keefektivan algoritma Round Robin dan Least Connections dalam mendistribisikan beban kerja.
2. Mengetahui keseimbangan beban lalu lintas di antara server-server yang tersedia.
3. Mengetahui responsivitas layanan seiring pertambahan beban kerja.

1.5. Batasan Masalah

Batasan masalah penelitian ini mencakup fokus pada analisis kinerja load balancing, sebagai berikut:

1. Membatasi fokusnya pada lingkup infrastruktur Google Cloud Platform (GCP) sebagai lingkungan tempat diuji coba dan dianalisisnya kinerja algoritma Round Robin dan Least Connection dalam manajemen distribusi beban.
2. Distribusi beban untuk layanan web dan aplikasi yang berjalan di atas infrastruktur cloud, dengan penekanan pada protokol HTTP.
3. Evaluasi kinerja algoritma akan difokuskan pada efektivitas distribusi beban, keseimbangan beban lalu lintas, dan responsivitas layanan

1.6. Sistematika Penulisan

Sistematika pembahasan dibuat agar mempermudah pembaca dalam memahami isi penelitian. Secara garis besar sistematika pembahasan dalam penelitian ini adalah:

BAB 1 Pendahuluan

Bab ini memperkenalkan latar belakang penelitian, menggambarkan urgensi penelitian dalam konteks transformasi digital dan perkembangan cloud computing. Di dalamnya, terdapat pembahasan mengenai konsep load balancing, peranannya dalam sistem cloud computing, serta dilema yang muncul seiring dengan pertumbuhan pengguna dan permintaan layanan.

BAB 2 Tinjauan Pustaka

Pada bab ini, penelitian akan merinci landasan teoritis yang mendukung pemahaman konsep load balancing, algoritma Round Robin dan Least Connections. Disajikan pula tinjauan terhadap penelitian-penelitian terdahulu yang relevan, memberikan konteks bagi penelitian ini dan mendukung pembentukan kerangka konseptual.

BAB 3 Metode Penelitian

Bab ini membahas metodologi yang digunakan dalam penelitian, termasuk rancangan penelitian, populasi dan sampel, teknik pengumpulan data, serta prosedur analisis data. Pemaparan ini bertujuan untuk memberikan gambaran yang jelas tentang langkah-langkah yang diambil untuk mencapai tujuan penelitian.

BAB 4 Hasil dan Pembahasan

Bab ini berfokus pada presentasi hasil-hasil penelitian yang telah diperoleh melalui analisis data. Dalam bentuk tabel, grafik, atau narasi, bab ini akan menyajikan temuan-temuan dan data-data yang relevan dengan tujuan penelitian.

BAB 5 Kesimpulan dan Saran

Bab ini merupakan tahap akhir yang menggambarkan kesimpulan dari seluruh penelitian serta memberikan saran-saran yang dapat diimplementasikan untuk pengembangan selanjutnya. Berikut adalah ringkasan dari kesimpulan dan saran yang diambil dari hasil penelitian