ISSN: 2634 - 8853

Journal of Engineering and Applied Sciences Technology

SCIENTIFIC Research and Community

Review Article

Open Access

Dynamic Capacity Planning in Data Centers for Latency Sensitive Workloads like Immersive Media & AR/VR: A Decision-Engine Approach

Anurag Reddy*, Anil Naik and Sandeep Reddy

Director Capacity Engineering, CloudFlare, USA

ABSTRACT

This paper explores the intricacies of data center planning, specifically focusing on the management of internet traffic and the optimization of capacity for edge servers. In an era where the internet is integral to daily life, effective capacity planning is essential to meet the escalating demands on network infrastructure. The paper introduces factors influencing internet and traffic demands, highlights the concept of capacity planning, and outlines key elements such as traffic analysis, scalability, edge server deployment, and resource monitoring, with a special focus on the implications in futuristic applications like Augmented & Virtual Reality

Two primary methodologies for ensuring required capacity are discussed: the Built to Forecast Methodology and the Decision Engine Methodology. The Built to Forecast approach relies on historical data and usage patterns for proactive infrastructure building, while the Decision Engine approach utilizes intelligent algorithms for dynamic capacity adjustments based on real-time data specific to AR/VR workloads. The latter incorporates push-pull mechanisms and multi-echelon inventory management, addressing the unique challenges posed by the dynamic nature of AR/VR traffic.

The paper delves into challenges and inefficiencies associated with the Built to Forecast Methodology, emphasizing scalability issues, the risk of technological obsolescence, inefficient energy consumption, and operational strain. In response, the Decision Engine Methodology is presented as a more adaptive alternative, incorporating push-pull mechanisms and multi-echelon inventory management for efficient supply chain optimization.

The conclusion underscores the critical decision between a recommendation-based model and a built-to-forecast approach, emphasizing the trade-offs involved. The optimal choice depends on organizational requirements, risk tolerance, and industry characteristics, with the recommendation-based model excelling in scenarios of high demand variability. Ultimately, the paper advocates for a tailored and effective inventory management strategy aligned with the unique needs and circumstances of the organization.

***Corresponding author**

Anurag Reddy, Director Capacity Engineering, CloudFlare, USA.

Received: March 06, 2023; **Accepted:** March 15, 2023; **Published:** March 24, 2023

Keywords: Internet Demand, Customer Ramps, DDoS, Cyber Attacks, Push and Pull, Multi-Echelon Inventory Positioning, Variability in Equipment Demand, Immersive Media

Introduction

Managing internet traffic and ensuring sufficient capacity for edge servers are crucial aspects of providing a seamless online experience for customers. The internet has become an integral part of our daily lives, with users expecting fast and reliable access to various online services. This increasing reliance on the internet has led to significant demands on network infrastructure, and effective capacity planning is essential to meet these demands.

Internet and Traffic Demands

The internet serves as a global network connecting millions of users and devices. With the proliferation of online content, applications, and services, the volume of internet traffic has grown exponentially. Users engage in activities such as streaming videos, accessing cloud-based applications, online gaming, and more, contributing to a diverse set of traffic demands.

Factors Influencing Internet and Traffic Demands

- **Content Richness:** The rise of multimedia content, high-definition videos, and interactive & immersive applications like AR/VR has led to a surge in data consumption.
- **Connected Devices:** The increasing number of internet-connected devices, including smartphones, IoT devices, and smart home appliances, adds to the overall demand on network resources.
- **Remote Work and Collaboration:** The trend toward remote work has amplified the need for reliable internet connectivity, with video conferencing and collaboration tools becoming integral to business operations.
- **E-commerce and Online Services:** The growth of e-commerce platforms and online services has resulted in a higher volume of transactions and data transfer over the internet.

Introduction to Capacity Planning

Capacity planning is a proactive approach to ensuring that the network infrastructure, particularly at the edge locations, can handle current and future traffic demands efficiently. It involves

assessing, forecasting, and allocating resources to prevent bottlenecks and performance issues. Here are key elements of capacity planning:

- **Traffic Analysis:** Understanding the patterns and types of internet traffic helps in predicting peak usage times, identifying popular content, and optimizing network resources accordingly.
- **Scalability:** Designing a network that can scale horizontally or vertically to accommodate increased traffic without compromising performance is essential for handling sudden surges in demand.
- **Edge Server Deployment:** Placing edge servers strategically in various locations helps reduce latency and improve content delivery. Capacity planning involves determining the optimal number and distribution of edge servers.
- **Resource Monitoring:** Continuous monitoring of server performance, bandwidth utilization, and other relevant metrics allows for real-time adjustments and proactive planning for future enhancements.

In summary, as internet traffic demands continue to evolve, effective capacity planning becomes a linchpin for delivering a responsive and reliable online experience. It involves a combination of analyzing traffic patterns, deploying scalable infrastructure, optimizing edge server placement, and implementing ongoing monitoring to adapt to changing user needs.

Methods

There are two primary methodologies for ensuring that the required capacity is always available:

Built to Forecast Methodology

This method relies on analyzing historical data and usage patterns to make predictions about future capacity needs. Organizations using this approach proactively build and scale their infrastructure based on these forecasts, aiming to stay ahead of anticipated demand. Forecasting demand at edge locations is a critical aspect of efficient resource management. The accuracy of this forecast is paramount, considering the inherent variability in demand patterns. This variability plays a pivotal role in shaping not only the total number of servers needed to meet demand but also influences the strategic decision on how these servers are deployed in batches.

To delve deeper into the forecasting process, it begins by meticulously analyzing demand requirements at each edge location. The goal is to capture the nuances of demand fluctuations, acknowledging that these patterns can be diverse and dynamic. This granular understanding allows for a more precise estimation of the resources needed. The significance of this forecast extends beyond just determining the number of servers. It also plays a crucial role in shaping the deployment strategy. The decision on whether to deploy servers in large batches or smaller increments is directly influenced by the variability identified in the demand forecast. This strategic choice impacts operational efficiency and resource utilization.

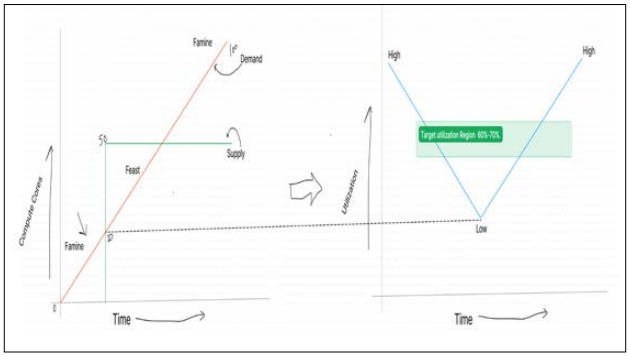
Moving on to the expansion phase, the timeline for deployment is a multifaceted consideration. It involves both fixed and variable lead times, each contributing to the overall time required for expansion. Fixed lead times encapsulate logistical aspects, including the readiness of colocation spaces. This readiness encompasses factors such as the availability of dark fiber and the deployment of network racks. Additionally, a portion of the installation process, involving tasks like unboxing, colocation-specific checks, and provisioning, is part of the fixed lead time.

On the other hand, variable lead time introduces flexibility into the deployment timeline. This variability is associated with the installation process per rack. Acknowledging and accounting for this variability ensures adaptability in the face of unforeseen challenges or changes in the expansion plan.

Forecast Variability and Batch Deployment Process

The forecast not only determines the total server count needed but also assesses the variability in demand over the lead time of the next deployment.

- In edge locations, where demand exhibits significant variability, the organization faces the challenge of deploying capacity over more extended time horizons than strictly necessary.
- The traditional approach of sending one large wave of servers results in inefficient resource utilization. The day of deploying large capacity introduces a peak in utilization, leading to unnecessary overhead and consumption of operational resources such as power and manpower.
- The process of deploying large batches, with extended lead times, creates an environment of feast or famine within the data center either running at high capacity or carrying unnecessary excess.



Challenges and Inefficiencies

This deployment strategy not only results in unnecessary capital expenditure and operational expenses but also poses additional challenges. The prolonged lead times render the data center operationally inflexible, making it difficult to adapt to changes in demand within the predefined lead time. Moreover, the organizations, in their attempts to counteract this inflexibility, often resort to deploying large batches, thereby escalating capital expenditure without optimizing resource utilization.

In addition to the above, this deployment approach introduces the following challenges:

- **Scalability Issues:** Rapid changes in demand may lead to difficulties in scaling up or down effectively, causing underutilization or overutilization of resources.
- **Risk of Technological Obsolescence:** Extended lead times increase the risk of deploying technology that may become obsolete before it is even put into full use, leading to wasted investments.
- **Inefficient Energy Consumption:** Large, periodic deployments contribute to inefficient energy consumption, as the data center operates at suboptimal levels during non-peak times.
- **Operational Strain:** Managing large-scale deployments places a strain on operational teams, affecting efficiency and potentially leading to errors in the deployment process.

Addressing these challenges necessitates a fundamental reconsideration of the deployment strategy, urging organizations to explore more dynamic and responsive approaches that can efficiently meet demand while minimizing unnecessary costs and operational constraints.

Decision Engine Methodology

In this approach, intelligent decision engines or algorithms are employed to continuously assess real-time data and adjust capacity dynamically. This methodology emphasizes adaptability, using automation and machine learning to make on-the-fly decisions based on the current demands of internet traffic.

The primary objective of this process is to consistently operate within the target utilization region, ensuring optimal performance for customers and maintaining high capital efficiency. Instead of relying solely on forecast accuracy, the strategy involves strategically positioning the supply chain inventory for nimble execution. The key nodes in a distributed data center supply chain include:

Methodology

The core concept revolves around a comprehensive analysis of the entire supply chain for data center expansion, incorporating multi-echelon inventory management for strategic positioning. This strategic deployment enables the swift deployment of servers when utilization reaches the target level. The overall flow of the data center supply chain involves servers sourced from Manufacturers/ ODM undergoing a meticulous journey through various stages before deployment at Data Center Locations.

For customized or specialized server configurations with uncertain demand, a pull mechanism is adopted. Servers are produced based on actual orders and real-time demand signals, allowing flexibility and responsiveness to dynamic customer requirements.

- **Push-Pull Mechanism:** Within this supply chain, a dynamic interplay of push, pull, and multi-echelon mechanisms is employed to optimize efficiency and responsiveness
- **Push Mechanism for Standardization:** Standard server configurations and components are produced based on forecasts and anticipated demand, optimizing manufacturing processes for items with relatively stable demand.
- **Multi-Echelon Inventory Management:** The multi-echelon approach strategically distributes inventory across different levels within the supply chain, ensuring efficient inventory management and minimizing risks associated with uncertainties in demand and lead times.

This combination of push, pull, and multi-echelon elements ensures a delicate balance in the data center supply chain. It enables the optimization of inventory levels, minimizes lead times, and facilitates the prompt deployment of servers to meet varying demand all while maintaining capital efficiency. The result is an agile and adaptive supply chain aligned with the goal of operating within the target utilization region and delivering optimum performance to customers.

The general flow of any data center supply chain involves servers sourced from Manufacturers/ODM undergoing a meticulous journey through various stages before deployment at Data Center Locations. This process encompasses two critical types of inventory Hardware and Datacenter Space Power

Inventory Positioning Strategy

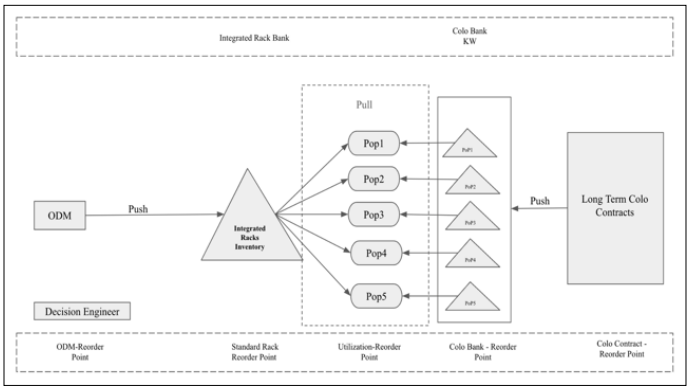
The standard rack inventory is strategically located at the rack integration facility, acting as a buffer with lower variability compared to individual data center demands. This inventory is consistently maintained above a minimum threshold. Operating with a lead time to final deployment measured in weeks, the inventory is flexibly pulled based on requirements to synchronize peak utilization with the target, ensuring optimal resource utilization.

Additionally, servers at the Original Design Manufacturer (ODM) are managed through a reservation system for server assembly capacity and server component inventory. This alignment is based on long-term forecasts reflecting global demand with reduced variability. A pull mechanism triggers server assembly when the inventory at the rack integration facility falls below the specified threshold. Despite the several weeks of lead time involved, the ongoing support from the inventory at the rack integration site enables timely replenishment.

Colocation Inventory Management Strategy

Efficient lead time management in colocation involves breaking down lead time components, including Site Building (spanning multiple quarters), Cage Build with Network Only (measured in months), and Rack Deployment (taking only a few days). The strategic placement of inventory plays a pivotal role in optimizing this process.

Site Building, as the lengthiest lead time component, aligns with long-term forecasts. This alignment allows for strategic planning of the ramp-up, offering flexibility in response to unforeseen changes in demand. This strategic approach becomes particularly crucial during contract negotiations. Cage Build with Network only represents a cost-effective alternative (approximately 20% of the cost of deploying the same cage with all servers). Careful planning of buffer strategies is essential to maintain a specific kilowatt (KW) of unused cage space. This strategic buffer ensures just-in-time deployment, ultimately enhancing capital efficiency.



Through meticulous planning of colocation inventory and strategic positioning, the execution timeline is streamlined to a matter of days. The integration of rack deployment within a pre-capable cage transforms into a low-effort and time-efficient process. This approach guarantees the optimal utilization of resources while simultaneously minimizing operational complexities.

Conclusion

In conclusion, the strategic decision between adopting a recommendation-based model and adhering to a built-to-forecast approach is a critical choice that necessitates a nuanced evaluation of multiple factors. While the built-to-forecast process presents

advantages in long-term planning and budget optimization, it is accompanied by limitations in flexibility, capital efficiency, and adaptability to dynamic shifts in demand. Conversely, the recommendation based model demonstrates agility, responsiveness, and adaptiveness, closely aligning with real-time demand needs and facilitating optimized resource allocation. We introduce the pivotal factors influencing AR/VR-related internet traffic demands and outline the key elements essential for capacity planning in this dynamic landscape, including traffic analysis, scalability considerations, edge server deployment strategies, and real-time resource monitoring tailored to AR/VR workloads.

The optimal selection between these two approaches hinges on specific organizational & application requirements, risk tolerance, and the characteristics of the industry landscape. The recommendation-based model excels in scenarios marked by high demand variability, where quick adaptability is paramount. It not only enables dynamic responses to changes but also prevents unnecessary capital expenditure, fostering operational efficiency. Ultimately, the choice must be tailored to the unique needs and circumstances of the organization, balancing the benefits and trade-offs inherent in each approach to achieve a harmonized and effective inventory management strategy [1-8].

References

1. Noormohammadpour Max, Raghavendra Cauligi (2018) Datacenter Traffic Control: Understanding Techniques and Trade-offs. IEEE Communications Surveys & Tutorials 20: 1492-1525.
2. Gmach D, Rolia J, Cherkasova L, Kemper A (2007) Capacity Management and Demand Prediction for Next Generation Data Centers," IEEE International Conference on Web Services (ICWS 2007), Salt Lake City, UT, USA 43-50.
3. Gu L, Zhou Y, Zhang Z (2020) Multi-echelon Inventory Optimization of Spare Parts considering Cross-region Transshipment and Changing Demand. 2020 Global Reliability and Prognostics and Health Management (PHM-Shanghai), Shanghai, China 1-6.
4. Kwan-Po Wong, Cho-Li Wang (1999) Push-Pull Messaging: a high-performance communication mechanism for commodity SMP clusters. Proceedings of the 1999 International Conference on Parallel Processing, Aizu-Wakamatsu, Japan 12-19.
5. Chen X, Liu W, Chen J, Zhou J (2020) An Edge Server Placement Algorithm in Edge Computing Environment. 2020 12th International Conference on Advanced Infocomm Technology (ICAIT), Macao, China 85-89.
6. Douligeris C, Mitrokotsa A (2003) DDoS attacks and defense mechanisms: a classification. Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology (IEEE Cat. No.03EX795), Darmstadt, Germany 190-193.
7. Marcos H, Gernowo R, Rosyida I, Nurhayati OD (2022) Intelligent Traffic Management System using Internet of Things: A Systematic Literature Review. 2022 IEEE 8th International Conference on Computing, Engineering and Design (ICCED), Sukabumi, Indonesia 1-6.
8. Schroter A (1998) Introduction to the network infrastructure warehouse. NOMS 98 1998 IEEE Network Operations and Management Symposium, New Orleans, LA, USA 1: 210-219.