# SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

**An Autonomous Institution**

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING(IoT and Cybersecurity Including BCT)**

COURSE NAME : cloud service management

IV YEAR / VII SEMESTER

Unit II-

Topic : Demand and Capacity matching

- Cloud capacity optimization is the process of improving and adjusting cloud resource utilization and efficiency based on actual and changing needs. This process involves monitoring, measuring, and collecting cloud data to identify trends, anomalies, or opportunities for improvement.
- That does not mean that cloud platforms automatically optimize resource allocation. For most types of cloud services, it's left to the user to determine how many resources cloud workloads will require at any given moment.
- AWS Aurora is one attempt to solve this problem; it automatically allocates resources based on workload need.

**Why the cloud needs capacity management**

- Consider a cloud server that hosts several web applications. Proper capacity management ensures that the server runs on a virtual server instance with enough CPU, memory and storage resources to support the applications, but not so many resources that a significant portion goes unused.

- **Provide insight into long-term IT planning**. For example, capacity management can help determine which workloads to move to the cloud. Workloads with fast-changing capacities are ideal candidates for the cloud, where resource allocations can be easily scaled up and down.

- **Determine which infrastructural and application architectures align with your needs**.
- For instance, if you have a virtual server with routinely fluctuating capacity demands, you might find that serverless functions would be a better way to host that workload.
- Serverless functions allow you to allocate large amounts of resources for short periods in a more cost-effective and easy-to-manage way than is possible with virtual servers. **Arrange the right people and tools**.
- This is a step beyond your team knowing how many resources to allocate to workloads. It's important to find out if you have the organizational resources necessary to assign those resources.
- You'll need staff on hand to perform the necessary provisioning, and those workers should have the requisite skills to work with the tools you use to manage resource allocation.
- **Avoid disruptions to users**. Wrong-sized workloads can create problems for the people who expect a specific application to be ready for them when they need it.
- When your workload capacities are well managed, you minimize your risk of having applications or servers fail.