D. G. Velev, Assoc. prof., Ph.D.
Department of Information Technologies and Communications
University of National and World Economy, Sofia, Bulgaria
P. V. Zlateva, Assoc. prof., Ph.D.
Institute of Control and Systems Research Bulgarian Academy of Sciences, Sofia, Bulgaria

PRINCIPLES OF CLOUD COMPUTING CAPACITY PLANNING

The term «Cloud» is used an abstraction of the Internet infrastructure. The cloud infrastructure provides for a computer infrastructure platform as a service. The cloud computing providers offer online common business applications accessed by Web browsers, and the software and data are stored in servers.

One of the most advertised advantages of the cloud-computing paradigm is the reduction of hardware deployment and installation times. Cloud storage and compute instances must be viewed as another type of resource.

The promise of cloud computing is that it can increase capacity «on-demand» easily and not necessarily automatically. Since every compute instance is a service, many cloud providers put the control of those instances in the hands of their customers. The decision when to launch new instances and their number can be crucially important regarding the rising traffic and computing capabilities and in those circumstances all could end with inefficiencies in cloud capacity.

The current article attempts to propose guidelines for efficient planning of cloud capacity.

Key words: *Capacity planning, cloud computing, resources*

Introduction. The **National Institute of Standards and Technology (NIST)** defines Cloud Computing as based on five key characteristics (on-demand self-service, broad network access, resource pooling, rapid elasticity and measured service), three delivery models (SaaS — Software as a Service, PaaS — Platform as a Service and IaaS — Infrastructure as a Service) plus four deployment models (Public/Consumer Private, Hybrid and Community) and describes Cloud computing as a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [8]. The cloud model promotes availability and is composed of five essential characteristics, three delivery models, four deployment models in two types external and internal.

Capacity planning is long-term decision that establishes a firms' overall level of resources. It extends over time horizon long enough to

obtain resources. Capacity decisions affect the production lead time, customer responsiveness, operating cost and company ability to compete. Inadequate capacity planning can lead to the loss of the customer and business. Excess capacity can drain the company's resources and prevent investments into more lucrative ventures. The question of when capacity should be increased and by how much are the critical decisions.

Capacity Planning. Capacity is defined as the maximum amount of work that an organization is capable of completing in a given period of time with the following calculation, Capacity = (number of machines or workers) × (number of shifts) × (utilization) × (efficiency). A discrepancy between the capacity of an organization and the demands of its customers results in inefficiency, either in under-utilized resources or unfulfilled customers. The broad classes of capacity planning are lead strategy, lag strategy, and match strategy [2, 3, 10].

- Lead strategy is adding capacity in anticipation of an increase in demand. Lead strategy is an aggressive strategy with the goal of luring customers away from the company's competitors. The possible disadvantage to this strategy is that it often results in excess inventory, which is costly and often wasteful.
- Lag strategy refers to adding capacity only after the organization is running at full capacity or beyond due to increase in demand. This is a more conservative strategy. It decreases the risk of waste, but it may result in the loss of possible customers.
- Match strategy is adding capacity in small amounts in response to changing demand in the market. This is a more moderate strategy.

Capacity planning is the process of determining the production capacity needed by an organization to meet changing demands for its products. In the context of capacity planning, «capacity» is the maximum amount of work that an organization is capable of completing in a given period of time [1, 10]. Capacity Planning is the process of measuring the amount of work that can be completed within a given time and determining the necessary physical and human resources needed to accomplish it. Capacity planning uses capacity utilization to ensure that the maximum amount of product is made and sold. The planning involves a regulation process that identifies deviations from the plan, allowing corrective action to be taken. A capacity requirements planning program can aid in the process of capacity planning.

A discrepancy between the capacity of an organization and the demands of its customers results in inefficiency, either in under-utilized resources or unfulfilled customers. The goal of capacity planning is to minimize this discrepancy. Demand for an organization's capacity varies based on changes in production output, such as increasing or decreasing

the production quantity of an existing product, or producing new products. Better utilization of existing capacity can be accomplished through improvements in overall equipment effectiveness (OEE). Capacity can be increased through introducing new techniques, equipment and materials, increasing the number of workers or machines, increasing the number of shifts, or acquiring additional production facilities.

According to Gartner [4, 5], capacity planning today is all about trying to ensure the provision of enough capacity and memory cycles to meet workload demand. Nevertheless, virtualization causes new variables to be taken into consideration and power consumption is just one among many. For IT resource planning (ITRP) there are several more elements to consider and the process must become much more strategic within an enterprise. Gartner analysts detailed the many variables that must be taken into account for appropriate enterprise ITRP. Traditional IT capacity metrics need to be considered alongside business requirements, human capital, financial metrics, facilities and power data, risk and compliance information as well as workload placement. Other considerations include configuration management, asset management, change management, event management and performance management, according to the Gartner report.

Cloud Computing Capacity Planning. The IT industry has been migrating a huge amount of computation and storage into compute clouds. A modern compute cloud allows users to share the underlying computing resources (such as CPU, memory and networking bandwidth) in an elastic manner and thus achieves an economy of scale. In a virtualization-based compute cloud, capacity planning refers to the procedure of allocating resources to virtual machines for supporting their workload. Capacity planning is a vital technology for making a cloud profitable from the cloud provider's perspective.

As data centers evolve to incorporate emerging technologies such as virtualization and cloud computing, the practice of planning for IT resources must also change. Cloud customers benefit from economies of scale such as volume purchasing, network bandwidth, operations, and administration when a cloud provider like handles these operations**. Average unit costs of computing are reduced because fixed costs are spread over more units of capacity and utilized by more users.**

Cloud capacity planning will affect cloud service performance assurance and how service level objectives/agreements will be met. To achieve elasticity and the provision of infinite computing resources available on demand, Cloud Computing providers typically rely on statistical multiplexing algorithms and load balancing mechanisms [6, 9]. Consequently, they will require various forms of «virtualization /abstraction» to mask the physical implementations of how resources are multiplexed and shared

regardless of whether commercial server virtualization technology is actually in use.

Old-fashioned capacity planning focuses on the peak usage of the application. However, now there are new capacity goals that can be summarized into the following:

- Performance (External service monitoring, Business requirements, User expectations).
- Capacity (System metrics).

One of the most advertised advantages of the cloud-computing idea is the reduction of hardware deployment and installation times. As far as capacity planning goes, cloud storage and compute instances should be viewed as just another type of resource. Just like a single server, for each instance of cloud-based computing, you have some amount of: CPU, RAM, Disk and I/O network transfer.

Each cloud resource still has its limits and costs as with any existing infrastructure. The capacity planning process is exactly the same [1, 7, 10]:

- Measuring the used resources (number of instances, CPU or storage).
- Determing the limits (needed resources to launch/stop new instance).
- Forecasting according to the past usage.

The promise of cloud computing is that capacity can be increased easily on-demand. Since every instance is essentially a purchase, many cloud providers put the control of those instances in the hands of their customers and making the decision when to launch new instances and their number can be crucially important in the face of spiking traffic. An enhanced operation that is using cloud computing might automate the process, but the automation should be tuned carefully to react not only to the load behavior of their website, but the load behavior of the cloud itself.

Clouds reduce shrink deployment time and provide for a more stringent control over capacity. Possible capacity planning principles that can be applied to cloud are as follows [1, 10]:

- Put capacity measurement into place — both metric collection and event notification systems — to collect and record systems and application statistics.
- Discover the current limits of your resources and determine how close you are to those limits.
- Use historical data not only to predict what you'll need, but to compare against what you will actually use.

Useful feature of cloud infrastructures is the ability to automatically scale an infrastructure vertically and horizontally with little or no impact to the applications running in that infrastructure.

The obvious benefit of cloud scaling is that the user pays only for the resources you use. The noncloud approach is to buy infrastructure for peak capacity. The downside of cloud scaling is that it can become a way of behavior system architects use to avoid capacity planning.

Cloud capacity planning is basically developing a strategy that guarantees a given infrastructure can support the resource demands placed on it [3, 7]:

- Knowledge of expected usage patterns as they vary in time and according to the nature of the business;
- Knowledge of how applications responds to load so that it could be identified when and what kind of additional capacity will be needed;
- Knowledge of the value of the systems towards business so you it is known when adding capacity provides value.

Capacity planning is just as important in the cloud as it is in a physical infrastructure. The general objective is to guarantee that when additional cost occurs by scaling a given infrastructure, the additional cost will be supporting the objectives for that infrastructure.

Certain principles must be followed for efficient cloud capacity planning [1, 7, 10]:

- Plans must be developed for an infrastructure to support expected loads.
- Recognize when actual load is diverging in a meaningful way from expected load.
- Understand the impact of changing application requirements on your infrastructure.
- Extensive support for optimizing virtual machines and hosts helps cut costs and consolidate operations without compromising service quality. Automatically identifies underutilized virtual machines and safe consolidation candidates.
- Broad support for virtualization environments from VMware, Microsoft, Citrix, Sun, HP and IBM, in addition to comprehensive support for modeling hardware, OS and other components.
- Reporting and automated dashboards engage business users and convey complex data in an interactive format, allowing users and IT stakeholders to quickly and intuitively identify cost saving opportunities. Business data such as hardware costs and power consumption can be correlated into reports for informed decisions.
- Solution kits help users successfully navigate common business problems such as server consolidations, new functionality rollouts.

The future cannot be automatically foreseen. Cloud capacity planning is not to eliminate unexpected peaks in demand, but to help plan for the expected, recognize and react to the unexpected appropriately to the deviation.

Conclusion. Using cloud computing it is possible to simply add capacity as needed. Since it offers a pay-as-you-use model, the cloud capacity can very efficiently planned using the clearly defined steps for planning main resources of the cloud infrastructure:

- Storage capacity (GB)
- Server processing (CPU cycles) & RAM capacity (GB)
- Network bandwidth (Gbps)
- Database transactions per second (TPS)
- Storage input/output operations per second (IOPS).

Literature:

1. Allspaw J. The Art of Capacity Planning, O'Reilly Media, Inc., 2008. — 154 p.
2. Capacity Planning, http://en.wikipedia.org/wiki/Capacity_planning.
3. Cohen R. Cloud Computing Infrastructure Capacity Planning, http://cloud-computing.sys-con.com/node/1114572.
4. Dubie D. Are you ready for IT resource planning, http://www.networkworld.com/newsletters/nsm/2009/022309nsm1.html.
5. Gardner D. HP's Cloud Assure for Cost Control allows elastic capacity planning to better manage cloud-based services, http://www.zdnet.com/blog/gardner/hps-cloud-assure-for-cost-control-allows-elastic-capacity-planning-to-better-manage-cloud-based-services
6. Hess K. Has Virtualization Made Capacity Planning Obsolete, http://www.linux-mag.com/id/7423.
7. Langley K. Things to Consider When Planning Your Application System and Software Architecture for Scalability Over Time, http://www.productionscale.com/home/2008/10/24/things-to-consider-when-planning-your-application-system-and.html.
8. Reese J. Cloud Application Architectures, O'Reilly Media, Inc., 2009, 206 p.
9. Sayegh E., Cloud Economics, http://www.rackspacecloud.com/blog/2009/02/20/cloud-economics/.
10. Shum A. A Measured Approach To Cloud Computing Capacity Planning and Performance Assurance, http://www.bsmreview.com/bsm_cloudcomputing.html.

Термин cloud (облако) используется как абстракция Интернет инфраструктуры. Инфраструктура облака (cloud infrastructure) предоставляет компьютерную инфраструктуру под платформой для виртуализации (virtualization) как сервис (обслуживание). Поставщики облачных вычислительный (cloud computing) предоставляют в онлайн режиме общие бизнес приложения, доступ к которым осуществляется web-браузером, в то время как программное обеспечение и данные сохранены на серверах.

Одно из наиболее рекламируемых преимуществ облачных вычислений — сокращение времен развертывания и установки аппаратных средств. Сохранение данных и вычислительные операции в облаке

должны быть рассмотрены как другой тип ресурса. Так как каждая вычислительная операция является облачным сервисом, оплачиваемым пользователями, многие поставщики облачных сервисов помещают их контроль в руках своих клиентов. Решение о запуске нового сервиса и их число может быть кардинально важным по отношению наращивания трафика транзакций и вычислительных способностей, что в пиковых ситуациях могли бы закончить неэффективностью в возможностях облака.

Настоящая статья пытается предложить набор принципов для эффективного планирования возможности облачных вычислений.

Ключевые слова: *планирование, облачные вычисления, ресурсы.*

Отримано 25.05.10

УДК 681.3.057:518.12:621.314.6:537:312.62

А. А. Верлань, канд. техн. наук
НТУУ «КПИ», г. Киев.

ОБ ОДНОМ ПОДХОДЕ К РАСЧЕТУ ПЕРЕХОДНЫХ ПРОЦЕССОВ В СЛОЖНЫХ ПОЛУПРОВОДНИКОВЫХ И СВЕРХПРОВОДНИКОВЫХ ВЕНТИЛЬНЫХ ПРЕОБРАЗОВАТЕЛЯХ

Описаний та проаналізований підхід до розв'язування задач динаміки складних напівпровідникових та надпровідникових вентильних перетворювачів на основі декомпозиції схеми, що розраховується, і застосування інтегральних рівнянь як динамічних моделей утворених лінійних підсхем. Показана ефективність даного підходу, що обумовлена згладжуючими властивостями інтегральних моделей та обчислювальною економічністю відповідних чисельних алгоритмів.

Ключові слова*: комп'ютерне моделювання, вентильні перетворювачі, динамічні моделі, інтегральні рівняння, декомпозиція*

Введение. Задачи анализа динамики устройств преобразовательной техники в современных энергетических полупроводниковых (ПП) и сверхпроводниковых (СП) объектах и системах является одной из наиболее сложных. Безальтернативным путем к эффективному решению данной задачи является применение методов и средств компьютерного моделирования. Компьютеризация научно-исследовательских и проектно-конструкторских работ все больше внедряется в практику разработки новых ПП и СП преобразователей, оказывает большое влияние на совершенствование методов научных исследова-

