Forecasting Demand of Potential Factors in Data Centers

Alexander PINNOW, Stefan OSTERBURG, Lars HANISCH
Otto-von-Guericke-University, Magdeburg, Germany
{alexander.pinnow|stefan.osterburg|lars.hanisch}@iti.cs.uni-magdeburg.de

This paper deals with forecasting demand of potential factors in data centers. Firstly it will define factors themselves and their importance in the process of data centers. Then it will be shown how three standard forecasting methods can be applied for predicting capacity needs in data centers.

Keywords*: data center, forecast, capacity planning.*

1 Motivation

Nowadays, data centers are being run incident-driven. Quite often further hardware systems are installed as a reaction on new customer needs. Despite acquisition costs for further hardware systems being considered uncritical, the extension of the information infrastructure leads to higher administration, maintenance, and finally personnel costs.

For an efficient usage of the existing information infrastructure, there are concepts such as Virtual and Adaptive Computing to logically separate hard and software. Realizing these concepts allows abandoning incident-driven business structures. Therefore, the adaption of suitable processes of operative production planning and control seems to be most useful.

The planning of utilization comes along with capacity and time management. Fundament for the capacity and time management is the production planning, which can be conducted with optimization models and forecasting processes.

Factors of a data center are identified in the following chapter. Next, forecasting processes of production program planning are applied in order to determine the amount of necessary potential factors.

2. Factors

The data center as a production system produces IT-services as output, which are made available to the customers as IT-products [18]. Such a production system can be described by its productions factors. Factors which fulfill planning, controlling, and organizational activities are labeled as managerial factors. They control the combination of all production factors.

Production factors are divided into raw materials and supplies and potential factors. Raw materials and supplies are consumed during the production process by becoming part of the output or because it is their elimination which enables the production in the first place. The capacities of potential factors are put at the production process' disposal without any loss of their productive effectiveness [19].

Personnel, application systems, and fundamental systems are major parts of the information infrastructure in data centers [12]. Application systems, fundamental systems, as well as the personnel charged with operating the data center can be considered potential factors. Potential factors can be non-physical and physical. Application software and fundamental systems such as development environment, data base systems, or operational systems are non-physical potential factors. Hardware as a fundamental system and the personnel, which is charged with operating the infrastructure, is physical potential factors. The maximum output of a potential factor is described as capacity [5].

When describing computer systems as hardware units, they can be considered as a set consisting of processor, storage and input/output devices. Input/output devices are used for the communication with the computer system. The processor retrieves instructions from the storage, decodes these instructions, and finally executes them.

The hardware component storage is divided into primary and secondary storage. Primary storage holds the data during the immediate operation of application and fundamental systems [15].

Secondary storage is non-volatile storage, usually disk storage. The capacity of the secondary storage normally exceeds the primary storage several times [13].

Every hardware component has a capacity that is describable by a suitable measure. A measure to explain the processors capacity is the maximum number of operations per second that the processor can execute [8]. The evaluation of capacity is conducted in millions of instructions per second (MIPS).

The communication of input/output devices takes place via channels. A channel is the connection between the sender of information and the receiver. The transfer speed is determined through bandwidth [6], which is given in bits/s [6].

The storage capacity determines how many storage cells, which can store binary digits, there are on a medium. The capacity evaluation is conducted in bit or byte [14]. The communication with external storage systems takes place via input/output devices. Channel bandwidth needs recognition along with storage capacity.

The development of the hardware component capacity can be described through *Moore's Law*, which explains that the number of transistors in an integrated circuit doubles approximately every 18 months [15]. The hardware components' capacity C grows exponentially depending on time t. The factor C_0 describes the initial capacity. The exponent λ is the growth rate.

$$C(t) = C_0 \cdot e^{\lambda \cdot t} \tag{2.1}$$

The following is valid for the n-multiplication of capacity within the time T_n:

$$n \cdot C_0 \cdot e^{\lambda \cdot t} = C_0 \cdot e^{\lambda \cdot (t+T_n)},\ n \cdot C_0 \cdot e^{\lambda \cdot t} = C_0 \cdot e^{\lambda \cdot t} \cdot e^{\lambda \cdot T_n}$$

$$n = e^{\lambda \cdot T_n}, \ln n = \lambda \cdot T_n, \lambda = \frac{\ln n}{T_n} \tag{2.2}$$

The growth rate for doubling the capacity every 1.5 periods is:

$$\lambda = \frac{\ln 2}{1.5} \approx 0.462 \tag{2.3}$$

If all q hardware units of the earliest generation were substituted by q recent hardware units in this model, then the overall capacity of the hardware components in the data center increases: the data center's overall capacity $C_{Hardware}$ in the moment t is the sum of total capacities of each hardware generation in the data center. From formula (2.1) follows:

$$C_{Hardware}(t) = \sum_{t^+=0}^{n-1} q \cdot C_0^* \cdot e^{\lambda \cdot t^+} \tag{2.4}$$

A data center consists of n generations with q hardware units. The capacity of the earliest hardware generation C_0^+ is described by formula 2.5:

$$C_0^+ = C_0 \cdot e^{\lambda(t-n-1)} \tag{2.5}$$

The resulting capacity for the data center (2.4) is:

$$\begin{aligned} C_{Hardware}(t) &= \sum_{t^+=0}^{n-1} q \cdot C_0 \cdot e^{\lambda(t-n+1)} \cdot e^{\lambda \cdot t^+} \\ &= q \cdot C_0 \cdot e^{\lambda(t-n+1)} \cdot \sum_{t^+=0}^{n-1} \left(e^{\lambda}\right)^{t^+} \end{aligned} \tag{2.6}$$

This is a geometric series and thus, can be transformed:

$$\begin{aligned} C_{Hardware}(t) &= q \cdot C_0 \cdot e^{\lambda(t-n+1)} \cdot \frac{e^{\lambda \cdot n} - 1}{e^{\lambda} - 1} \\ &= q \cdot C_0 \cdot \frac{e^{\lambda \cdot t}}{e^{\lambda(n-1)}} \cdot \frac{e^{\lambda \cdot n} - 1}{e^{\lambda} - 1} \\ &= q \cdot C_0 \cdot e^{\lambda \cdot t} \cdot \frac{e^{\lambda}}{e^{\lambda \cdot n}} \cdot \frac{e^{\lambda \cdot n} - 1}{e^{\lambda} - 1} \\ &= q \cdot \frac{e^{\lambda(n+1)} - e^{\lambda}}{e^{\lambda(n+1)} - e^{\lambda \cdot n}} \cdot C_0 \cdot e^{\lambda \cdot t} \\ &= q \cdot \frac{e^{\lambda \cdot n} - 1}{e^{\lambda \cdot n} - e^{\lambda(n-1)}} \cdot C_0 \cdot e^{\lambda \cdot t} \end{aligned} \tag{2.7}$$

Hence, the overall capacity of the data center's hardware components grows exponentially as well, yet it grows slower than the capacity of hardware components in average. This effect is due to the usage of hardware components from older generations. The following is valid for the average capacity $C^{\phi}_{Hardware}$ of a data center with n hardware generations, each equipped with q hardware units:

$$C^{\phi}_{Hardware}(t) = \frac{q}{q\cdot n}\cdot\frac{e^{\lambda\cdot n}-1}{e^{\lambda\cdot n}-e^{\lambda(n-1)}}\cdot C_0\cdot e^{\lambda\cdot t}$$
$$= \frac{e^{\lambda\cdot n}-1}{n\left(e^{\lambda\cdot n}-e^{\lambda(n-1)}\right)}\cdot C_0\cdot e^{\lambda\cdot t} \qquad (2.8)$$

If a data center employs more than one hardware generation, the average capacity $C^{\phi}_{Hardware}$ of the data center will grow slower than the common capacity *C*. The average capacity is equal to the common capacity in the special case of the data center consisting of only one current hardware generation:

$$\frac{\delta C^{\phi}_{Hardware}(t)}{\delta t} \le \frac{\delta C(t)}{\delta t} \quad \text{for } n \in N^{+}$$
$$\lambda \in \mathbb{R},\ \lambda > 0 \quad (2.9)$$
$$C_0 \in \mathbb{R},\quad C_0 > 0$$

The result of the first derivative is:

$$\frac{e^{\lambda\cdot n}-1}{n\left(e^{\lambda\cdot n}-e^{\lambda(n-1)}\right)}\cdot C_0\cdot\lambda\cdot e^{\lambda\cdot t} \le C_0\cdot\lambda\cdot e^{\lambda\cdot t} \qquad (2.10)$$
$$\frac{e^{\lambda\cdot n}-1}{n\left(e^{\lambda\cdot n}-e^{\lambda(n-1)}\right)} \le 1$$

It is proven by mathematical induction, that:

$$e^{\lambda\cdot n}-1 \le n\left(e^{\lambda\cdot n}-e^{\lambda(n-1)}\right) \qquad (2.11)$$

Inequation 2.11 holds for $n = 1$:

$$e^{\lambda\cdot 1}-1 \le 1\left(e^{\lambda\cdot 1}-e^{\lambda(1-1)}\right) \qquad (2.12)$$
$$e^{\lambda}-1 \le e^{\lambda}-1$$

The inductive step for inequation 2.11 results to:

$$e^{\lambda(n+1)}-1 \le (n+1)\left(e^{\lambda(n+1)}-e^{\lambda\cdot n}\right) \qquad (2.13)$$

The exponential function is $e^{\lambda} > 1$ under the condition of $\lambda > 0$, hence it must be true that:

$$e^{\lambda\cdot n}-1 \le n\left(e^{\lambda\cdot n}-e^{\lambda(n-1)}\right) \le e^{\lambda}\cdot n\left(e^{\lambda\cdot n}-e^{\lambda(n-1)}\right)$$
$$e^{\lambda\cdot n}-1 \le n\cdot e^{\lambda(n+1)}-n\cdot e^{\lambda\cdot n}$$
$$e^{\lambda(n+1)}+e^{\lambda\cdot n}-1 \le n\cdot e^{\lambda(n+1)}+e^{\lambda(n+1)}-n\cdot e^{\lambda\cdot n}$$
$$e^{\lambda(n+1)}-1 \le n\cdot e^{\lambda(n+1)}+e^{\lambda(n+1)}-n\cdot e^{\lambda\cdot n}-e^{\lambda\cdot n}$$
$$e^{\lambda(n+1)}-1 \le (n+1)\,e^{\lambda(n+1)}-(n+1)e^{\lambda\cdot n}$$
$$e^{\lambda(n+1)}-1 \le (n+1)\left(e^{\lambda(n+1)}-e^{\lambda\cdot n}\right)$$

q.e.d. (2.14)

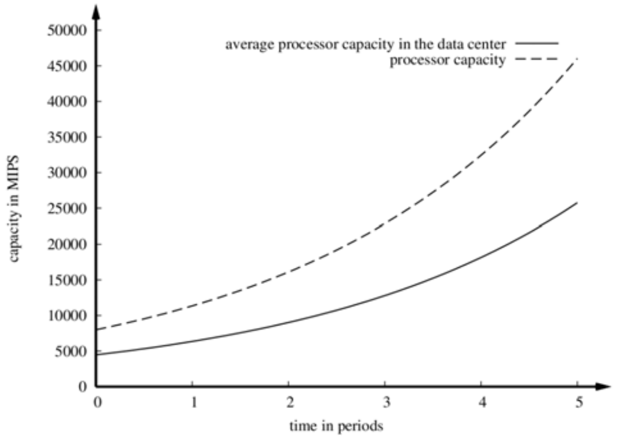

Fig. 1. Development of processor capacity

Fig. 1 shows the development of processor capacity in a data center under the assumptions that the capacity doubles every two periods, a hardware generation is exchanged after five periods, and the initial capacity is 8000 MIPS.

The exponentially increasing capacity of hardware components are also demanded by the customers. The compensation of technological progress through higher consumption of resources is known as the *Rebound-effect* [11]. The increase in consumption of resources is called *primary Rebound-effect* [7]. The capacity of potential factors held available does not represent the actual demand. A demand prognosis is necessary, in order to reveal the actual demand of potential factors, which will be discussed in the following chapter.

3. Demand Forecast

Coming along with the production planning, forecasting procedures are used with the goal to predict the future demand of raw materials and supplies based on their historic demand [16]. In case of regular need one distinguishes between forecasting procedures for constantly, trendily, and seasonally fluctuating demand.

Procedures for trendy demand can always be used instead of forecasting methods for constant demand. In the special case of constant demand the trend comes to zero [1]. Forecasting methods for seasonally fluctuating demand are most sensibly used in seasonal cycles.

Linear regression is the best-known demand forecasting method for a trendy demand [10], and so is *Brown's Linear Exponential Smoothing* [2] as well as *Holt's method* [4]. The mentioned forecasting models rely on the assumption that the level of demand linear trend over time. This means that [16]:

$$y_k = b_0 + b_1 \cdot k + \varepsilon_k \quad (3.1)$$

The variable b_0 refers to the axis' section of the trend line. Factor b_1 describes the slope of the trend line. The independent variable k represents the chronological course. Random fluctuations are displayed via ϵ_k. In formula 3.1 ϵ_k is normal-standard distributed with an anticipation term of $E_\epsilon = 0$ [9].

Further forecasting procedures exist for trend-like courses of higher order. Linearization is necessary in advance if these measures are employed [1]. The following linear transformation results in case of exponential demand functions [17]:

$$y_k = b_0 \cdot \varepsilon^{b_1 \cdot k}$$

$$\ln y_k = \ln b_0 + b_1 \cdot k \cdot \ln \varepsilon \qquad (3.2)$$

$$\ln y_k = b_0^+ + b_1 \cdot k \mid b_0^+ = \ln b_0$$

Production planning should adapt forecasting procedures as a method to determine the demand of potential factors. The average capacity demand of hardware units shall be estimated using of this method. The application of forecasting procedures is useful because the past demand of capacities can be ascertained with respective tools. Usually, only one third or less of the hardware capacities are employed. If a potential factor is not fully utilized, then its degree of utilization reflects its actual demand. Hence, the development of average utilization of capacity can serve as data base for forecasting procedures. The prognosis for each hardware unit is conducted separately.

The average utilization of capacity of each hardware component per hardware unit is determined via the respective measuring tools. The average capacity demand c is determined by the average utilization of capacity of hardware components η per hardware unit i:

$$c_i = \eta_i \cdot C_i \quad (3.3)$$

The average capacity demand of a hardware component c results from the sum of the single demand of capacity:

$$c = \sum_{i=1}^{n} c_i \quad (3.4)$$

The average processor capacities displayed in table 1 serve as data sample. The capacity demand for potential factors is not free from seasonal fluctuations. Thus, the prognosis period is dimensioned so that it encompasses a complete seasonal cycle because hardware units cannot be installed and de-installed with seasonal dependency. The shown forecasting procedures are applicable because the growth of capacity demand follows an exponential trend and seasonal fluctuations must not be considered.

Table 1. Measured average processor capacity demand

Period t	Capacity c_t in MIPS	ln c_t
1	95 200	11.4637
2	90 800	11.4164
3	151 600	11.9290
4	249 600	12.4276
5	257 800	12.4599
6	502 200	13.1268
7	713 500	13.4779
8	836 900	13.6375

In order to predict the average processor capacity demand, a linear regression, *Brown's Linear Exponential Smoothing*, and the *Holt's method* are conducted. The development of demand follows an exponential trend. Linearizing the exponential values according to formula 3.2 is compulsory with the intention of applying the described forecasting procedures. The linearization results shown in table 1 as well. Fig. 2 shows the development of the average demand of processor capacity for data centers and the linearization results.

Suitable values for parameter α have to be estimated with the goal of achieving satisfying forecast results with the Brown's Linear Exponential Smoothing. For the estimation of the smoothing constant of first order, having n samples, the following rule is suggested [3]:

$$\alpha_1 = \frac{2}{n+1} \quad (3.5)$$

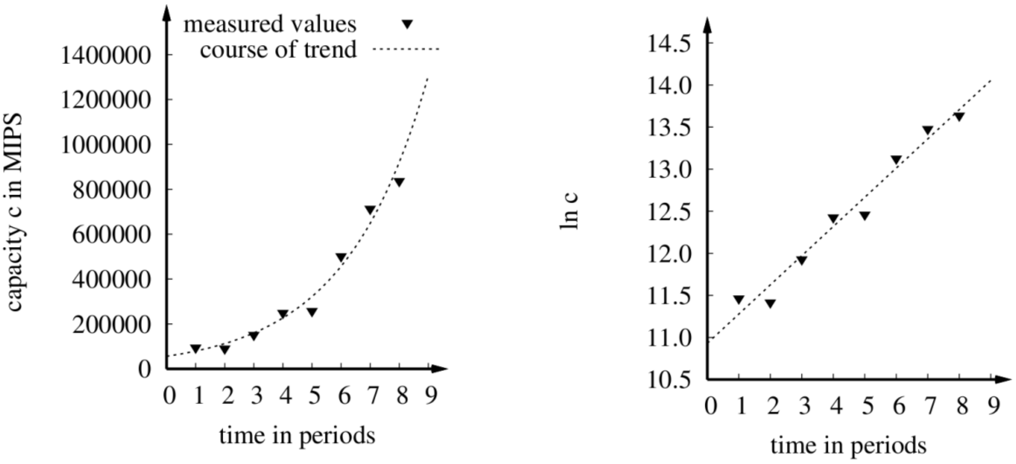

Fig. 2. Measured values of processor capacity

The smoothing constant of *m*-th order is related to the smoothing constant of first order [2]:

$$1+\alpha_1 = (1-\alpha_m)^m \quad (3.6)$$

Smoothing constant α is estimated on α = 0.1181 for the eight present samples. α is reused for the *Holt's method*. Due to simplification reasons, smoothing parameter β is considered as a parameter of the next higher order and estimated at β = 0.0804.

The prognosticated values of the last four periods of the respective forecasting procedure are compared to the relevant sample values with the aim of evaluating the forecasting procedure's results. The sample values of the first four periods only serve as a data base.

The prediction of the values for one period is only conducted at hand of the preceding period's sample values. In table 2 one can see the forecasting results $\hat{y}$ as well as the residual values ε of the forecasting procedures out of the last four periods. Table also shows the standard deviation values of the residue $\sigma_{\varepsilon t}$, which were calculated in an ex-post-analysis. *Holt's method* has the smallest standard deviation value of the residue. Hence, the procedure delivers the best results for the present sample values in cooperation with the given smoothing parameter.

Table 2. Residue values of the forecasting procedures

			Linear regression		Brown's Linear Exponential Smoothing		Holt's method	
t	α	β	$\hat{y}_t$	ε_t	$\hat{y}_t$	ε_t	$\hat{y}_t$	ε_t
5	0.2254	0.1566	12.6602	-0.2003	12.6706	-0.2107	12.6603	-0.2003
6	0.1835	0.1264	12.8404	0.2863	12.8408	0.2860	12.8405	0.2862
7	0.1548	0.1061	13.3317	0.1463	13.3375	0.1404	13.3316	0.1463
8	0.1340	0.0914	13.7565	-0.1191	13.7625	-0.1250	13.7564	-0.1191
9	0.1181	0.0804	14.0539		14.0555		14.0540	
$\bar{\varepsilon}$			0.0283		0.0227		0.0283	
$\sigma_{\varepsilon t}$			0.2269		0.2306		0.2263	

The average processor capacity demanded $\hat{c}$ in period 9 is:

$$\hat{c}_9 = e^{\hat{y}_9} = e^{14.0540} \approx 1269000 MIPS \quad (3.7)$$

4. Conclusions

The capacity of most hardware components per data center grows exponentially. The *Rebound-effect* describes the circumstance that the extra capacities are consumed by the customer as well.

Linear regression, Brown's Linear Exponential Smoothing, and Holt's method can be employed in order to predict the average capacity demand of hardware components in data centers. The sample data of average utilization of capacity of preceding periods serves as data base. The hardware components capacities are usually dimensioned so that their utilization of capacity is lower than one-third. Thus, the utilization of capacity of hardware components is suitable as data base. The data collection is conducted with the necessary measuring tools.

The result of the forecast is the expected average demand of capacity for the succeeding period. The demand of capacity is not constant over the period but is explained via a distribution function. Capacity shortages can be prevented because one can configure the capacity of hardware components with the help of statistical procedures that are based on the average utilization of capacity. Hence, the shown forecasting procedures are a cornerstone for the actual capacity planning.

References

[1] R. G. Brown. Materials management systems: a modular library. Malabar, Fla.: Krieger, 1984, pp. 81-83.

[2] R. G. Brown and R. F. Meyer. The fundamental Theorem of Exponential Smoothing. Operations Research: the journal of the Operations Research Society of America, Vol. 9, No. 5, 1960, pp. 673-685.

[3] S. A. DeLurgio. Forcasting principles and applications. Boston, Mass. a.o: McGraw-Hill, 1988, p. 156.

[4] C. C. Holt. Forecasting seasonals and trends by exponentially weighted moving averages. In: International journal of forecasting, vol. 20, 2007, pp. 5-10.

[5] W. Kern, W. Industrielle Produktionswirtschaft. Poeschel: Stuttgart, 1992, pp. 21.

[6] H. Klimant, R. Piotraschke and D. SchönfeldInformations- und Kodierungstheorie. Stuttgart: Teubner, 2006, pp. 75, 86.

[7] R. Kuhlen. Wissensökologie, in: Kuhlen, R./Seeger, T./Strauch, D. : Grundlagen der praktischen Information und Dokumentation. München: Saur, 204, pp. 108.

[8] H. Langendörfer, H, Leistungsanalyse von Rechensystemen: Messen, Modellieren, Simulation. München: Hanser, 1992, pp. 10

[9] S. Makridakis, S. C. Wheelwright and V. E. McGee. Forecasting: methods and applications. a.o. New York : Wiley, 1983, p. 220.

[10] J. Neter et al. Applied linear statistical models. a.o. Boston, Mass.: McGraw-Hill, pp. 198-200.

[11] F. J. Rademacher. Building the Infor-

mation Society: Labour Market Pressures, Globalisation, and the Political Goal of Sustainability as Challenges to the Regions in Europe. In: Sturm, R., Weimann, G.: The information society and the regions in Europe: a British-German comparison. Baden-Baden: Nomos-Verl.-Ges, 2000, p. 232

[12] C. Rautenstrauch. Effizienter Einsatz von Arbeitsplatzsystemen -- Konzepte und Methoden des Persönlichen Informationsmanagements. a.o. Bonn: Addison-Wesley Longman, 1997, pp. 13.

[13] A. Heuer and G. Saake, Datenbanken: Implementierungstechniken. Bonn: MITP, 1999, pp. 44.

[14] H. J. Schneider. Lexikon Informatik und Datenverarbeitung. a.o. München: Oldenbourg, 1998, p. 802.

[15] A. S. TanenbaumStructured computer organization. a.o. Englewood Cliffs: Prentice-Hall Internat, 1999, p 25, 113.

[16] H. Tempelmeier. Material-Logistik: Modelle und Algorithmen für die Produktionsplanung und -steuerung in Advanced-Planning-Systemen. a.o. Berlin: Springer, 2006, pp. 36, 50

[17] K. Weber. Wirschaftsprognostik. München: Vahlen, 1990, p. 70.

[18] R. Zarnekow, W. Brenner, U. Pilgram. Integrated Information Management: Applying Successful Industrial Concepts, in IT. Berlin, Heidelberg: Springer, 2006, p. 16.

[19] G. Zäpfel, Grundzüge des Produktions- und Logistikmanagements. München, Wien: Oldenbourg, 2001, pp 16.

Alexander PINNOW studied Business and Computer Science until 2003. Then he worked as software engineer specialized in financial markets. Since 2006 he is scientific assistant at the Very Large Business Applications Lab of the Otto-von-Guericke-University in Magdeburg, Germany. His area of research is the data center as a production facility for IT-services with focus on capacity management in adaptive and virtualized data centers. He published several papers in this field of study. The research in this area is a cooperation of the VLBA-Lab and the Germany based IT-service provider T-Systems.

Stefan OSTERBURG studied Computer Science until 2001 and passed his diploma with distinction. He had worked as a software engineer and consultant for Microsoft ERP software (Dynamics AX) for several years, before in 2006 he became scientific assistant at the Very Large Business Applications (VLBA) Lab of the Otto-von-Guericke-University in Magdeburg. His area of research is the data center as a production facility for IT-services with focus on availability management in adaptive and virtualized data centers. The research in this area is a cooperation of the VLBA-Lab and the Germany based IT-service provider T-Systems.

Lars HANISCH has been studying International Management at the Otto-von-Guericke University Magdeburg since 2003. During his studies he worked for Volkswagen in Germany and China consulting the management of the Technical Development department. Furthermore, he assisted the research of the Otto-von-Guericke-University in Magdeburg, focusing on Enterprise Resource Planning Systems for IT service providers. His area of research interests are international cooperation and international supply chain management.