A Capacity Supply Model for Virtualized Servers

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

This paper deals with determining the capacity supply for virtualized servers. First, a server is modeled as a queue based on a Markov chain. Then, the effect of server virtualization on the capacity supply will be analyzed with the distribution function of the server load.
***Keywords**: Markov chain, Weibull distribution, server, capacity supply*

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. The paper discusses a model to plan the capacity supply of physical and virtual servers. In the first step, the effect of the server load on the response time will be described. Therefore, a server is modeled as a queue based on a Markov chain. The effect of server virtualization on the capacity supply will be analyzed with the distribution function of the server load.

2 Determining the Critical Server Load

The operator of a data centre provides the customer with physical and immaterial resources. The maximum output of a resource is defined as its capacity [6]. When describing computer systems as hardware units, they can be considered as a set consisting of processor, storage and input/output devices. The operator of a data centre cannot use the complete capacity supply of a computer system because this will yield an unacceptable response time. The utilization of the capacity supply is called server load. The level of server load that does not satisfy the customer needs can be calculated using the queuing theory. Therefore, the server will be modeled as a time continuous Markov chain. The server receives requests represented as an event. These events are Poisson distributed. The arrival rate λ is defined as the number of requests to the server in a given time period. With the arrival of a new request the state i of a server is transitioning to state $i+1$. The state probability $\pi_i(t)$ represents the probability of i requests in a server at time t. The server is processing these requests. The termination of processing is represented by an event. These events are also Poisson distributed. The processing rate ν defines the number of request a server can handle in a given time period. Arrival and processing rates λ and ν are non-varying and independent from the number of requests in this model. Figure 1 [1] illustrates this context. The ratio of arrival and processing rates describes the server load ρ. The processing rate needs to be larger than the arrival rate in order to avoid an infinite growth of the unprocessed requests:

$$\rho = \frac{\lambda}{\nu} < 1 \quad (1)$$

The mean number of requests processed by a server depending on the server load ρ is [1]:

$$\overline{K} = \frac{\rho}{1-\rho} \quad (2)$$

Little's theorem [7] determines the mean number of requests processed by a server depending on the arrival rate λ and the mean response time $\overline{K}$:

$$\overline{K} = \lambda \cdot \overline{T} \quad (3)$$

By substituting Little's theorem in equation 2 the server load can be determined in dependence of the mean response time $\overline{T}$ and the

arrival rate λ:

$$\frac{\rho}{1-\rho} = \lambda \bar{T}; \quad \rho = \frac{\bar{T}\lambda}{1+\bar{T}\lambda} \quad (4)$$

For a given mean response time and arrival rate the critical server load ρ^* can be calculated by equation 4.

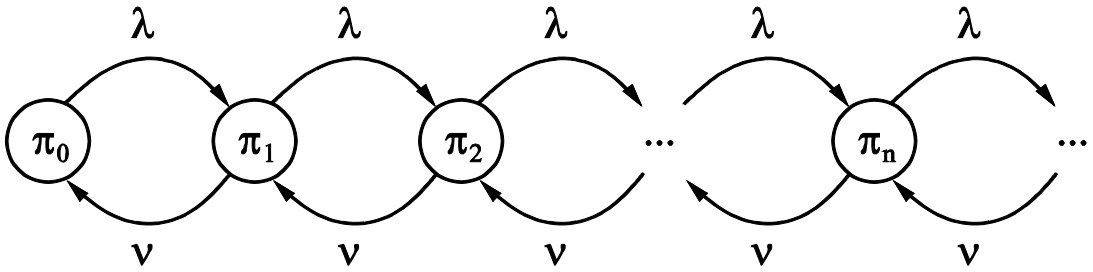

Fig. 1. Queue as Markov chain

3 Determining the Capacity Supply

A non-varying arrival rate λ was assumed for the determination of the critical server load. In reality, the arrival rate is varying over the time. It will be assumed that the arrival rate in different time periods is non-varying at different levels (see figure 2).

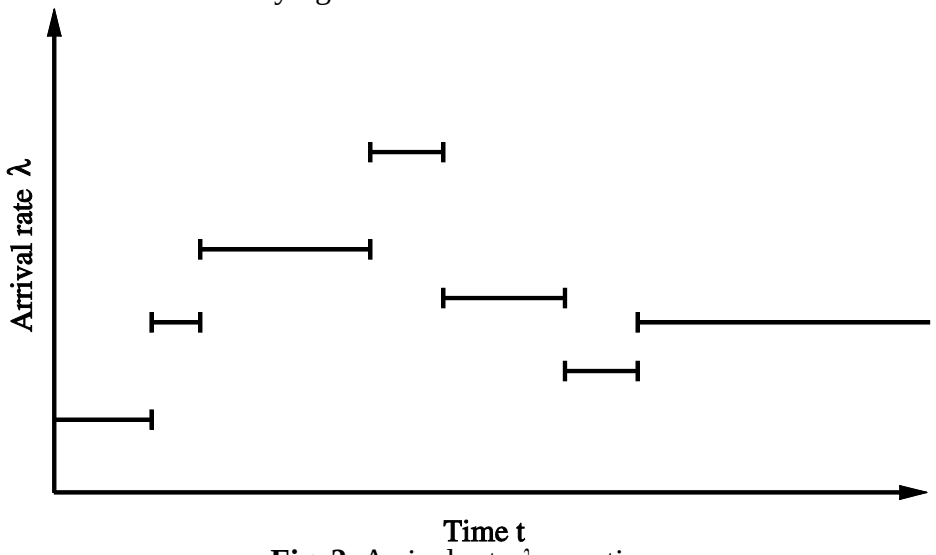

Fig. 2. Arrival rate λ over time

In such a time period a steady state is reached. The arrival rate λ used to determine the critical server load is the maximum arrival rate a server must be able to handle to ensure a noncritical response time. The levels of the arrival rate λ are subject to a distribution function. In the model, a normal distribution will be assumed. From equation 1 follows that the server load ρ is also normally distributed. The server capacity supply c_S, describes the maximum output of the resource. The requests to a server cause the capacity demand c_D. The server load ρ can be described as a ratio of capacity supply c_S and capacity demand c_D:

$$\rho = \frac{c_D}{c_S} \quad (5)$$

Because of the distribution function of the server load ρ the capacity demand c_D is also normally distributed. The quota of server requests that do not exceed a defined capacity demand can be determined by the distribution function F of the normal distribution:

$$F(c_D) = P(C \le c_D) \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{c_D} \exp\left(-\frac{(c_D-\mu)^2}{2\sigma^2}\right) \quad (6)$$

The integral of the distribution function for the capacity demand c_D cannot be calculated

analytically and described by a known function in a closed form [3]. Therefore, the normal distribution of the capacity demand will be approximated by the Weibull distribution. The distribution function F of the Weibull distribution is [8]:

$$F(x;\Theta) = 1 - \exp\left[-\left(\frac{x-\gamma}{\alpha}\right)^{\beta}\right] \quad for\ x \geq \gamma \quad (7)$$

The parameters $\Theta = \{\alpha, \beta, \gamma\}$ of the distribution function are defined as scale, shape and location parameters. The density function is given as [8]:

$$f(x) = \frac{\beta}{a}\left(\frac{x-\gamma}{\alpha}\right)^{\beta-1} - \exp\left[-\left(\frac{x-\gamma}{\alpha}\right)^{\beta}\right] \quad for\ x \geq \gamma \quad (8)$$

The expectation of the three parameter Weibull distribution is determined by the scale, shape and location parameters [8]:

$$\mu = E(X) = \gamma + \alpha\,\Gamma\left(1 + \frac{1}{\beta}\right) \quad (9)$$

Scale and shape parameters define the variance of the three parameter Weibull distribution [8]:

$$\sigma^2 = \alpha^2\left[\Gamma\left(1+\frac{2}{\beta}\right) - \left[\Gamma\left(1+\frac{1}{\beta}\right)\right]^2\right] \quad (10)$$

The shape parameter β specifies the skewness of the Weibull distribution. The normal distribution can be approximated by the Weibull distribution with a special shape parameter [4]:

$$\beta = 3.60232 \quad (11)$$

For the shape parameter approximating the normal distribution the following values for the gamma function can be calculated [2]:

$$\Gamma\left(1+\frac{1}{\beta}\right) = \Gamma\left(1+\frac{1}{3.60232}\right) \approx 0.90072$$
$$\Gamma\left(1+\frac{2}{\beta}\right) = \Gamma\left(1+\frac{2}{3.60232}\right) \approx 0.88964 \quad (12)$$

The scale parameter α depends on the variance σ^2 for a shape parameter β approximating the normal distribution:

$$\alpha = \sqrt{\frac{\sigma^2}{0.07834}} \quad (13)$$

The location parameter γ depends on the variance σ^2 and the expectation value μ for a shape parameter β approximating the normal distribution:

$$\gamma = \mu - 0.90072 \cdot \sqrt{\frac{\sigma^2}{0.07834}} \quad (14)$$

Therefore, the distribution function describing the capacity demand by the three parameter Weibull distribution is:

$$F(c_D) = 1 - \exp\left[-\left(\frac{c_D - \gamma}{\alpha}\right)^{\beta}\right]$$

$$= 1 - \exp\left[-\left(\frac{c_D - \mu - 0.90072 \cdot \sqrt{\frac{\sigma^2}{0.07834}}}{\sqrt{\frac{\sigma^2}{0.07834}}}\right)^{3.60232}\right] \quad (15)$$

The quota of server requests exceeding a defined capacity demand c_D is the difference between the total probability and the distribution function of the capacity demand c_D. For a given capacity demand with a normal distribution approximated by the three parameter Weibull distribution, the quota of critical server requests κ exceeding the capacity demand c_D is:

$$\kappa = P(C > c_D) = 1 - F(c_D) \quad (16)$$

$$\kappa = \exp\left[-\left(\frac{c_D - \gamma}{\alpha}\right)^{\beta}\right]$$

According to equation 5, the capacity demand c_D can be substituted by the critical server load ρ^*.

$$\kappa = \exp\left[-\left(\frac{\rho^* \cdot c_S - \gamma}{\alpha}\right)^{\beta}\right] \quad (17)$$

From equation 17, the capacity supply c_S can be determined depending on the expectation value μ and the variance σ^2 of the capacity demand c_D and the critical server load κ:

$$\exp\left[-\left(\frac{\rho^* \cdot c_S - \gamma}{\alpha}\right)^{\beta}\right] = \kappa$$

$$-\left(\frac{\rho^* \cdot c_S - \gamma}{\alpha}\right)^{\beta} = \ln(\kappa) \quad (18)$$

$$\frac{\rho^* \cdot c_S - \gamma}{\alpha} = \left[-\ln(\kappa)\right]^{\frac{1}{\beta}}$$

$$c_S = \frac{\alpha\left[-\ln(\kappa)\right]^{\frac{1}{\beta}} + \gamma}{\rho^*}$$

Replacing the shape, scale and location parameters with values for the approximated normal distribution, the needed capacity supply c_S of a server with a quota of κ allowed critical server requests is:

$$c_S = \frac{\sqrt{\frac{\sigma^2}{0.07834}}\left[-\ln(\kappa)\right]^{\frac{1}{3.60232}} + \mu - 0.90072 \cdot \sqrt{\frac{\sigma^2}{0.07834}}}{\rho^*} = \frac{\sqrt{\frac{\sigma^2}{0.07834}}\left(\left[-\ln(\kappa)\right]^{\frac{1}{3.60232}} - 0.90072\right) + \mu}{\rho^*} \quad (19)$$

Hence, the capacity supply is determined by the expectation value and the variance of the capacity demand, the number of server requests processed in a given time period, and the mean response time of a server request.

4 Determining the Capacity Supply in Case of Virtualization

In the next step, the effect of virtualization on the capacity supply will be analyzed. A physical server can be split into different virtual servers. In the model, a virtualization technique with a dynamic allocation of the capacity supply of a physical server to the virtual servers is assumed. The capacity supply can be used completely by the virtual servers. Thus, virtualization overhead will be ignored in the model. The capacity demands of the virtual servers are random variables. These random variables are correlated if a statistical dependency between them exists. The capacity demands of the virtual servers should be uncorrelated. The covariance matrix [5] is:

$$Cov(x) = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{pmatrix} \quad with \ \sigma_{ij} = \begin{cases} 0, & i \neq j \\ \sigma_i, & i = j \end{cases} \quad (20)$$

The expectation value of the physical server $E(C_{PD})$ is the sum of the expectation values of the virtual servers $E(C_{V\,Di})$ [3]:

$$E(C_{PD}) = E\left(C_{V\,D1} + C_{V\,D2} + \ldots + C_{V\,Dn}\right) = \sum_{i=1}^{n} E(C_{V\,Di}) = \sum_{i=1}^{n} \mu_{V\,Di} \quad (21)$$

The variance of the capacity demand of the physical server $Var(C_{PD})$ for two virtual servers is in general [3]:

$$Var(C_{PD}) = Var\left(C_{V\,D1} + C_{V\,D2}\right) = Var\left(C_{V\,D1}\right) + Var\left(C_{V\,D2}\right) + 2\,Cov\left(C_{V\,D1}, C_{V\,D2}\right) \quad (22)$$

The distribution of the capacity demand should be uncorrelated. In this case, the variance of the capacity demand of a physical server for *n* virtual servers is:

$$Var(C_{PD}) = \sigma_{PD}^2$$

$$= Var\left(C_{V\,D1} + C_{V\,D2} + ... + C_{V\,Dn}\right)$$

$$= \sum_{i=1}^{n} Var(C_{V\,Di}) \quad (23)$$

$$= \sum_{i=1}^{n} \sigma_{V\,Di}^2$$

Ignoring the virtualization overhead, the capacity supply c_{PS} of a physical server is the sum of the capacity supplies c_{VSi} of the virtual servers:

$$c_{PS} = \sum_{i=1}^{n} c_{VSi} \quad (24)$$

In the model, the physical server is operating *n* identical virtual servers. The expectation values and the standard deviations are equal for all virtual servers. The expectation value $E(C_{PD})$ and the variance $Var(C_{PD})$ of the capacity demand of the physical server are:

$$E(C_{PD}) = \mu_{PD} = n\mu_{VD}$$

$$Var(C_{PD}) = \sigma_{PD}^2 = n\sigma_{VD}^2 \quad (25)$$

For *n* identical servers operated by a physical server the required capacity supply c_{PS} of the physical server is:

$$c_{PS} = nc_{VS} \quad (26)$$

Equation 19 describes the quota of critical server requests to a physical server without virtualization:

$$c_{PS} = \frac{\sqrt{\frac{\sigma_{PD}^2}{0.07834}}\left(\left[-\ln(\kappa)\right]^{\frac{1}{3.60232}} - 0.90072\right) + \mu_{PD}}{\rho^*} \quad (27)$$

With equation 25, the expectation value and variance of the physical server can be substituted by the expectation value and variance of the virtual servers. The required capacity supply c_{PS} on a physical server with *n* identical virtual servers depending on the critical server load ρ^*, the quota of allowed critical server requests κ, the expectation value of the capacity demand of a virtual server μ_{VD} and the variance of the capacity demand of a virtual server σ_{VD}^2:

$$c_{PS} = \frac{\sqrt{\frac{n\sigma_{VD}^2}{0.07834}}\left(\left[-\ln(\kappa)\right]^{\frac{1}{3.60232}} - 0.90072\right) + n\mu_{VD}}{\rho^*} \quad (28)$$

Because the physical server operates *n* identical virtual serves, the required capacity supply of the physical server can be distributed to the virtual servers:

$$\frac{c_{PS}}{n} = c_{VS}$$

$$= \frac{\sqrt{\frac{n\sigma_{VD}^2}{0.07834}}\left(\left[-\ln(\kappa)\right]^{\frac{1}{3.60232}} - 0.90072\right) + n\mu_{VD}}{\rho^* n} \quad (29)$$

$$= \frac{\sqrt{\frac{\sigma_{VD}^2}{0.07834n}}\left(\left[-\ln(\kappa)\right]^{\frac{1}{3.60232}} - 0.90072\right) + \mu_{VD}}{\rho^*}$$

Claim: If less than half of the server requests cause a critical capacity demand ($\kappa < 0.5$), the required capacity supply per virtual server decreases with increasing number of vir-

tual servers.

Proof: If this is true, equation 29 must be a sloping function for $\kappa < 0.5$:

$$\frac{d\frac{c_{VS}}{n}}{dn} = \underbrace{-\frac{\sqrt{\frac{\sigma_{VD}^2}{0.07834}}}{2n^{\frac{3}{2}}\rho^*}}_{Term1} \cdot \underbrace{\left(\left[-\ln(\kappa)\right]^{\frac{1}{3.60232}} - 0.90072\right)}_{Term2} < 0 \quad (30)$$

For a positive number of virtual servers n and a positive server load ρ term 1 is always negative. Hence, term 2 is:

$$\left[-\ln(\kappa)\right]^{\frac{1}{3.60232}} - 0.90072 > 0$$

$$\left[-\ln(\kappa)\right]^{\frac{1}{3.60232}} > 0.90072$$

$$-\ln(\kappa) > 0.90072^{3.60232} \quad (31)$$

$$\ln(\kappa) < -0{,}68615$$

$$\kappa < e^{-0.68615}$$

$$\kappa < 0.50351 \approx 0.5 \quad q.e.d$$

From the capacity planning point of view, it is useful to operate as many as possible virtual servers on a physical server. The required capacity supply of an infinite large physical server operating an infinite number of virtual servers would be the ratio of the expectation value of the capacity demand and the critical server load:

$$\lim_{n\to\infty} \frac{c_{PS}}{n} = \frac{\mu_{VD}}{p^*} \quad (32)$$

Because of the limitation of the capability of a physical server such a machine cannot be implemented. If a physical server operates two identical virtual servers, the impact of the correlation on the capacity supply can be shown. The equation to calculate the capacity supply of two correlated virtual servers is:

$$c_{PS} = \frac{\sqrt{\frac{2\sigma_{VD}^2 + 2Cov(C_{VD1}, C_{VD2})}{0.07834}}\left(\left[-\ln(\kappa)\right]^{\frac{1}{3.60232}} - 0.90072\right) + 2\mu_{VD}}{\rho^*} \quad (33)$$

In case of correlation, the required capacity supply should be less than in the uncorrelated case. Therefore, two identical physical servers operating two identical virtual servers will be compared. The two virtual servers on one physical server are correlated the others are not. The following inequation describes this case:

$$\frac{\sqrt{\frac{2\sigma_{VD}^2}{0.07834}}\left(\left[-\ln(\kappa)\right]^{\frac{1}{3.60232}} - 0.90072\right) + 2\mu_{VD}}{\rho^*} \quad (34)$$

$$> \frac{\sqrt{\frac{2\sigma_{VD}^2 + 2Cov(C_1, C_2)}{0.07834}}\left(\left[-\ln(\kappa)\right]^{\frac{1}{3.60232}} - 0.90072\right) + 2\mu_{VD}}{\rho^*}$$

Less than half server requests should cause a critical capacity demand. Applying equation 31 for $\kappa < 0.5$ results to:

$$\left[-\ln(\kappa)\right]^{\frac{1}{3.60232}} - 0.90072 > 0 \quad (35)$$

The server load ρ is always positive. Hence, the inequation of the capacity supplies can be written as follows:

$$\sqrt{\frac{2\sigma_{VD}^2}{0.07834}} > \sqrt{\frac{2\sigma_{VD}^2 + 2Cov(C_1, C_2)}{0.07834}}$$

$$\sqrt{\sigma_{VD}^2} > \sqrt{\sigma_{VD}{}^2 + Cov(C_{VD1}, C_{VD2})} \quad (36)$$

This inequation is always true for negative covariance between the capacity demands of the virtual servers.

$$Cov(C_{VD1}, C_{VD2}) < 0 \quad (37)$$

In the uncorrelated case, for $\kappa < 0.5$, a larger capacity supply is required than in the case of negative correlation. The random variables C_{VD1} and C_{VD2} represent the capacity demand of the virtual servers. Negative correlation means that C_{VD1} and C_{VD2} have a inverse linear coherence [3]. Positive correlation will raise the required capacity supply. Thus, it is reasonable to aggregate virtual servers with negative correlation on a physical server to minimize the required capacity supply.

5 Example

A data centre is planning the capacity supply of four identical servers. A server should be able to process in mean 100 requests per second. The arrival and the process rates are Poisson distributed. The mean response time needs to be 0.03 seconds. This requirement should be satisfied for 95% of the requests. The utilization of the capacity supply is normally distributed. The normal distribution will be approximated by the Weibull distribution. The expectation value of the capacity demand is 1000 MIPS with a standard deviation of 200 MIPS. Figure 3 illustrates the distribution and density functions of the capacity demand of a server. The capacity demands of the servers are uncorrelated.

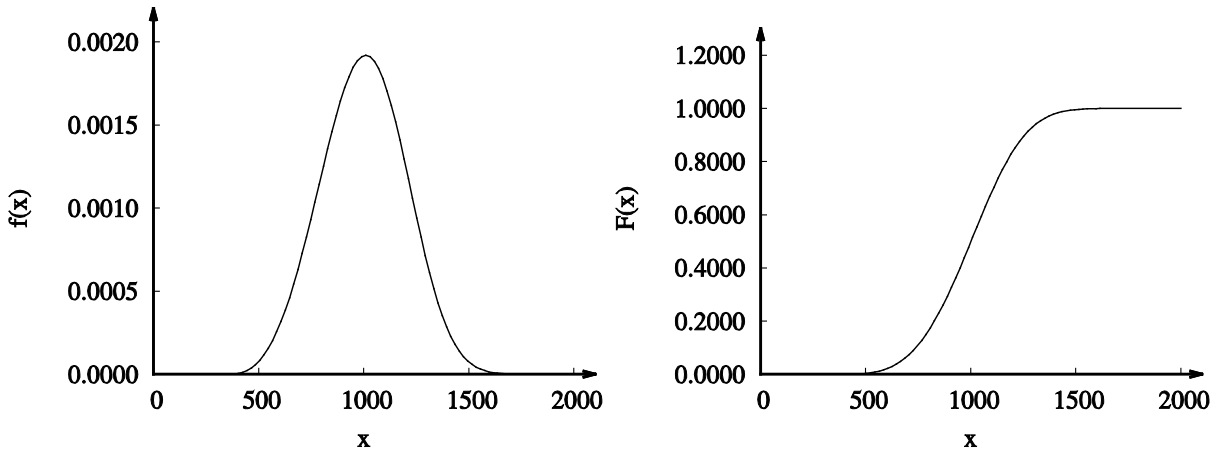

Fig. 3. Distribution and density functions

First, the capacity supply is determined for the four servers realized as physical servers. Furthermore, the capacity saving will be determined if four virtual servers are operated by one physical server with a dynamic allocation of the capacity supply.

The arrival rate λ of the Poisson distributed server requests is 100 requests per second:

$$\lambda = 100 \quad (38)$$

The mean response time $\overline{T}$ to process a request is 0.03 seconds:

$$\overline{T} = 0{,}03 \quad (39)$$

From equation 4, the critical server load ρ^*

can be determined by the arrival rate λ and the mean response time $\overline{T}$. The critical server load must not be exceeded to comply with the mean response time specification:

$$\rho^* = \frac{\overline{T}\lambda}{1+\overline{T}\lambda}$$

$$= \frac{0.03 \cdot 100}{1+0.03 \cdot 100} \quad (40)$$

$$= 0.75$$

To achieve a mean response time of 0.03 seconds with 100 requests per second, a server load of 75% must not be exceeded.

The expectation value μ of the capacity demand c_D of a server is 1000 MIPS:

$$\mu = E(C) = 1000 \quad (41)$$

The standard deviation σ of the capacity demand c_D of a server is 200 MIPS:

$$\sigma = 200 \quad (42)$$

In 95% of the cases an additional request should not exceed a server load of 75%. The quota of critical server requests κ is:

$$\kappa = 0.05 \quad (43)$$

From equation 19 the capacity supply c_S complying with these requirements can be determined:

$$c_S = \frac{\sqrt{\frac{200^2}{0.07834}}\left(\left[-\ln(0.05)\right]^{\frac{1}{3.60232}} - 0.90072\right) + 1000}{0.75} \quad (44)$$

$$= 1767.14995$$

The capacity supply of a server is 1768 MIPS. In the next step, the capacity supply of a physical server operating four identical virtual servers will be determined. The capacity supply c_S of the physical server can be determined by equation 28:

$$c_{PS} = \frac{\sqrt{\frac{4 \cdot 200^2}{0.07834}}\left(\left[-\ln(0.05)\right]^{\frac{1}{3.60232}} - 0.90072\right) + 4 \cdot 1000}{0.75} \quad (45)$$

$$= 6200.96656$$

The required capacity supply is 6201 MIPS. The capacity saving by virtualization is the difference between the capacity supply of the four physical servers without virtualization and the physical server operating four virtual servers:

$$4 \cdot 1768 - 6201 = 871 \quad (46)$$

The capacity saving for operating four virtual servers on a physical server is 871 MIPS.

6 Conclusion

The capacity supply describes the maximum output of a server. The utilization of the capacity supply is the capacity demand. The server load is the ratio of the capacity demand and the capacity supply. The application of a Markov chain demonstrated that it is impossible to utilize the full capacity supply. For a given arrival rate λ of server requests the critical server load to achieve a required response time $\overline{T}$ can be determined.

The level of the arrival rate λ is not fixed in the course of time but is statistically varying. For a known distribution function of the levels of the arrival rate the capacity demand not exceeded by a defined quota of server requests can be determined by the expectation value and the variance. The capacity demand can be used for determining the capacity supply.

Assuming a normal distribution approximated by the Weibull distribution for the level of the arrival rate λ demonstrates that the use of virtualization saves capacity. Negative correlation of the capacity demand of two virtual servers leads to more capacity savings than in the uncorrelated case.

References

[1] G. Bolch, *Queueing networks and Markov chains: modelling and performance evaluation with computer science appli-*

cations, Wiley, New York, 1998.
[2] I. N. Bronshtein, K. A. Semendyayev, G. Musiol and H. Muehlig, *Handbook of Mathematics*, Springer, Berlin, 2007.
[3] F. M. Dekking, C. Kraaikamp and H. Paul, *A Modern Introduction to Probability and Statistics*, Springer, London, 2005.
[4] S. Y. D. Dubey, "Normal and Weibull distributions," *Naval Research Logistics Quarterly 14 (1967)*, pp. 69-79.
[5] W. Härdle and Z. Hlávka, *Multivariate Statistics: Exercises and Solutions*, Springer, New York, 2007.
[6] W. Kern, *Industrielle Produktionswirtschaft*, Poeschel, Stuttgart, 1992.
[7] J. D. C. Little, "A proof for the queuing formula: L = λW," *Operations Research: The journal of the Operations Research Society of America 9 (1961)*, pp. 383-387.
[8] D. N. P. Murthy, M. Xie and R. Jiang, *Weibull models*, Wiley-Interscience, Hoboken, NJ, 2003.

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 centre 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.