Self-Similarity
On Network Systems With Finite Resources

Prepared by Wang, Hsu-Hui

Advisory under Prof. Po-Ning Chen

In Partial Fulfillment of the Requirements

For the Degree of
Master of Science

Department of Communications Engineering

National Chiao Tung University

Hsinchu, Taiwan 300, R.O.C.

E-mail: yukino.cm90g@nctu.edu.tw

July 9, 2003

Abstract

It has been shown recently that the modern network traffic is much more appropriately modelled by long range-dependent self-similar processes. This leads to a present research trend on network self-similarity. It has been long conjectured that heavy-tailed statistics in packet duration has a close relation with the degree of network self-similarity. Such a conjecture has recently been substantiated under the assumption that infinite network sources have been aggregated.

In this thesis, we attempt to investigate the same problem by relaxing the *infinite* sources assumption. Specifically, our thesis experiments and observes how and how much self-similarity be contributed by finite number of heavy-tailed data sources. Analysis and comparison with that obtained under infinite source assumption are addressed.

Acknowledgement

At first, I would like to thank Dr. Po-Ning Chen for his guidance. He is very patiente with his student. I especially appreciate it because I always make a lot of mistakes before achieving something small. He is tough to research, however, kind to people.

I have to thank my lab mate, Daniel. Thank you not only for helping me with LaTeX and Matlab program but also cheering me up while I am most depressed. It is very happy and fun to share your unique philosophy. Senior Jia-Long, thank you for allow me running simulation on your PC. Maybe you would think it is not worth mentioning, but my thesis cannot be completed without your generosity. Thank Senior Yong-Sheng, who graduated last year, for his helping of my early research.

Huai-Jong is my good mate. We are in the same stage of getting master degree. Both of us have to face many diffculties and frustrations during the very time. I am happy we go through it together. Thanks for your tolerance and kindness, my bear.

My parents always worry about my graduate project. They ask me, "How about your thesis?" and "Do you have any trouble on your thesis?" and even "Do you need any help for your thesis?" every time when they call me. I know they cannot do anything about a single equation or theory I am engaged and troubled with. However, it is their love that always support me.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Conventionally, teletraffic theory is based on Markovian assumptions of traffic arrival processes and of service time distributions. These conventional models have inter-arrivals or holding times that decay exponentially, which result in variance of normalized sum decaying inversely with sample size. However, by rapid development and deployment of networks, these traditional models were found to be different from what we surmised nowadays. A new look on these models becomes necessary.

1.1 Background And Problem Formation

The packets are basic units that are transmitted over the network. They have some properties that are sometimes neglected to facilitate their statistical analysis. One example is the *monopoly* property, namely, once a network link is held by a packet, other packets can never occupy the same link, and have to wait until the completion of the transmission of this packet. This monopoly property makes the usual memoryless assumption of arrival processes somewhat undue. Thus, some extensions to memoryless are brought in for a better approximation. As an example, the exponential inter-arrival duration has been substituted by a heavy-tailed counterpart in some recent researches. In concept, a heavy-tailed distribution is one that assigns large probability mass for large random value, as contrary to an exponentially decaying distribution whose probability mass over large random value decreases

exponentially fast.

It was conjectured that the heavy-tailibility is one of the key factors to establish self-similarity or long-range dependence of network arrivals. Researches have been devoted to the verification of this hypothesis. Recently, this conjecture has been substantiated under the premise that the network consists of infinite number of independent ON/OFF sources. More specifically, the conjecture is proved by assuming that the duration and number of connections are both infinite. We then note that these assumptions, although suitable for theoretical analysis, may not be applicable in practice. This observation leads to our investigation on the relation of heavy-tailibility and self-similarity if the physical links are indeed finite.

1.2 Objective Of The Thesis

As aforementioned, extension models to memoryless have been proposed for the establishment of a synthetic traffic that matches the observed self-similar real network behavior, such as M/G/1 [Box98] and superposition of infinite number of ON/OFF sources [Will97]. What we aim at in this thesis is to refine the existing self-similar models in a way that only finite number of network resources is allowed. Different combinations of finite resources are examined by simulations, and their effects are observed.

Another interesting problem as researched by [Ana96] is also investigated in this thesis. By travelling through a lot of relay nodes, the packets may experience a change in its statistics, which may converge to some limit if the relay servers are equipped with common probabilistic feature. Simulations on this issue are performed, and how the common server feature affects the ultimate traffic behavior is studied.

1.3 Plan Of The Thesis

The thesis is organized in the following fashion.

In Chapter 2, we quote the underlying definitions of long-range dependence, self-similarity and heavy-tailibility in literature. The preliminary backgrounds of these terminologies are introduced. Also in the same chapter, survey of some system models that have been used to synthesize self-similar traffic is presented. Chapter 2 is closed by the introduction of the usual analysis method for degree of self-similarity.

Chapter 3 begins with our refined models to the existing models introduced in Chapter 2. Related theories and literatures are cited to support our refinement. The simulation scenario assumed in this thesis is subsequently introduced.

Chapter 4 summarizes our simulations under various system settings, such as parallel channels with fixed utilization, parallel channels with varying utilization, serial channels with common relay nodes, etc. Remarks and conclusions based on our simulations are given in this chapter.

Conclusions are drawn in Chapter 5.

Chapter 2

Preliminary Background On Self-Similar Process And Its Determination

Self-similar processes are stochastic processes that are invariant in distribution under suitable scaling of time and space. Typically, such processes can be used to model random phenomena which have long-range dependence. The process was first applied to hydrology for estimating the flow rate of Nile, then many other fields, such as astronomy, economics (prosperity or depression, amount of deals in stock market), engineering, mathematics, and physics (especially high energy physics). Internet traffic has developed for decades and been analyzed by Poisson model. However, recent measurement studies have shown that the actual computer network traffic is long-range dependent. The new application of self-similar processes thus arises.

In incoming chapter, the basic concept of self-similarity and long-range dependence will be introduced, including slowly decay variance and autocorrelation, heavy-tailed distribution, and Hurst Parameter, etc. And we will introduce the analysis method for them.

2.1 Self-Similarity

The notation of self-similarity can be concisely seen from an exemplified figure quoted from [Stall02, pp. 224]. In Figure 2.1(a), the curves at different time-scale looks "similar" to each other. However, it can be observed from Figure 2.1(b) that at a larger scale of time, the time function is more chaotic and irregular, as contrary to a lower time-scale at which the time function has less fluctuation and higher regulation. With this introductory figure, definitions of self-similarity are given subsequently.

2.1.1 Continuous-time self-similar process

Definition 2.1 ([Will97]) *A stochastic process $X(t)$ is self-similar with Hurst parameter H, where $0.5 \leq H \leq 1$, if for any real $a > 0$, $a^{-H}X(at)$ has the same distribution as $X(t)$.*

With the above definition, the first marginal moment, second marginal moment, and autocorrelation function of $x(t)$ satisfy:

$$\begin{aligned} E[X(t)] &= \frac{E[X(at)]}{a^H} \\ \mathrm{Var}[X(t)] &= \frac{\mathrm{Var}[X(at)]}{a^{2H}} \\ R_X(t,s) &= \frac{R_X(at,as)}{a^{2H}}. \end{aligned}$$

2.1.2 Discrete-time self-similar process

One of the methods to determine self-similarity is by computer-based simulation. To avoid the high variability of observations from continuous-time self-similar processes, which is increased with time index t, a discrete-time process that is often viewed as an increment process from continuous-time self-similar processes can be used instead.

We first come to the counting process. A stochastic process $\{N(t), t \geq 0\}$ is a counting process, if $N(t)$ is the number of events that have occurred during time 0 to t. Apparently,

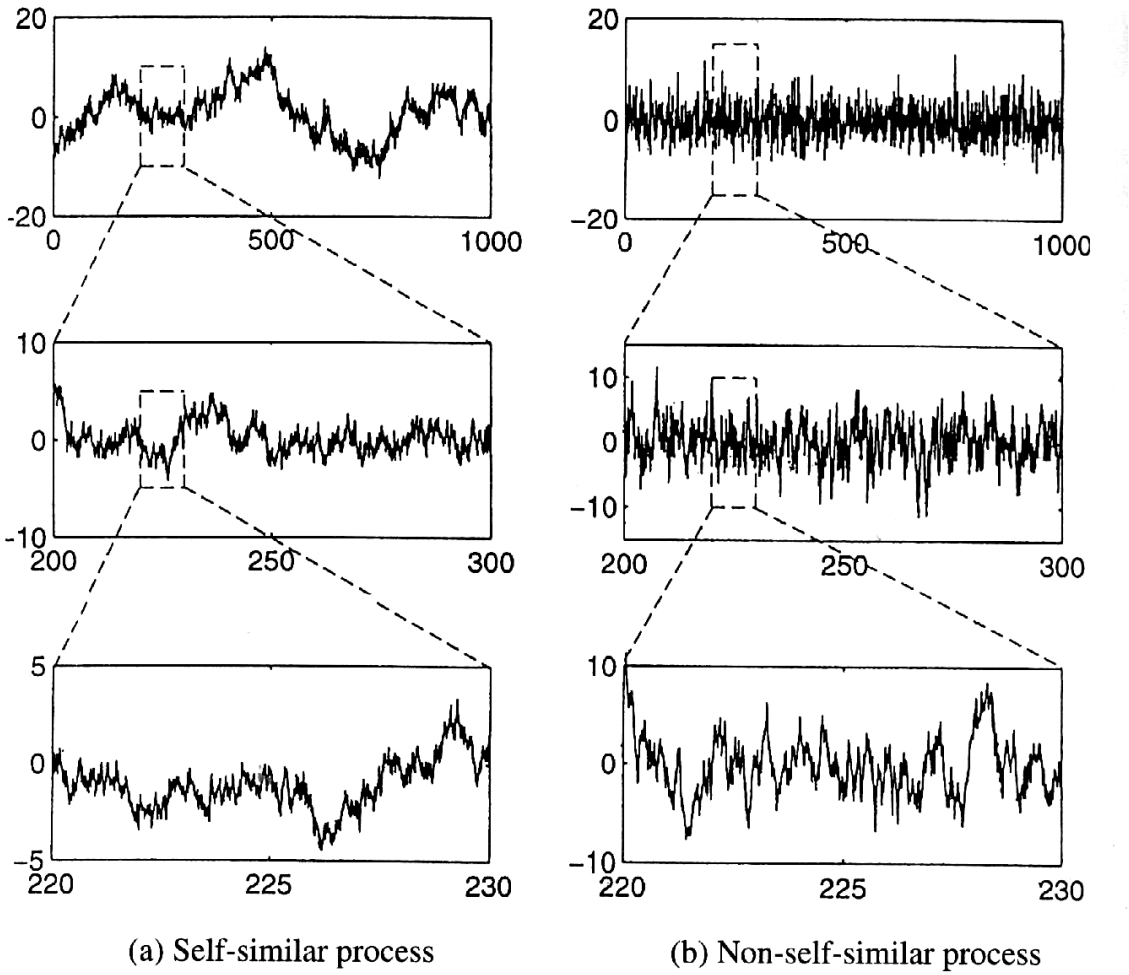


Figure 2.1: *The comparison of a self-similar and a non-self-similar processes in different time scale.*

$N(t)$ is non-decreasing in t, and takes nonnegative integer values. It is certain that the continuous-time $N(t)$ can be self-similar in nature.

Consider the discrete-time process $X_k = N(kT) - N((k-1)T)$, which is the number of events that occur in non-overlapping intervals of time. The discrete-time process $\boldsymbol{X} = \{X_k, k = 1, 2, \ldots\}$ is therefore called the increment process of $N(t)$.

As aforementioned, a continuous-time self-similar process is invariant to changes in scale. So a straightforward approach to determine the degree of self-similarity, namely the Hurst parameter, is to examine the statistics of increments with different time scales. This

introduces the notion of aggregation—the sum of m adjacent increments of the original continuous-time self-similar process. The aggregated discrete process is accordingly defined by:

$$X_k^{(m)} = \frac{1}{m} \sum_{i=km-(m-1)}^{km} X_i.$$

The statistics of different time scale can be inspected through different value of m.

Definition 2.2 ([Tsy98]) *A strictly stationary (increment) process* $\boldsymbol{X} = \{X_k, k = 1, 2, ...\}$ *is called strictly self-similar with parameter* H*, where* $0.5 < H < 1$*, if* $m^{1-H}\boldsymbol{X}^{(m)} = \boldsymbol{X}$ *(i.e., they are equal in the sense of finite-dimensional distributions) for every nature number* m*, in which* $\boldsymbol{X}^{(m)} = \{X_k^{(m)}, k = 1, 2, 3, \ldots\}$.

2.1.3 Statistical properties of self-similarity

The aggregated process $\boldsymbol{X}^{(m)}$ is a time-averaged series of the process $\boldsymbol{X}$. If $\boldsymbol{X}$ is ergodic in addition to stationary, this time average converges with probability 1 to its ensemble average. By some additional assumption of $\boldsymbol{X}^{(m)}$, such as it is uniformly dominated by some integrable random variable in m, we may use dominated convergence theorem to yield that the variance of the time average goes down to 0 as m goes to infinity. Certain observation induces the definition of *exactly second-order self-similarity*.

Definition 2.3 *A stationary process* $\boldsymbol{X}$ *is exactly second-order self-similar with parameter* β*, where* $0 < \beta < 1$*, if for every* $m = 1, 2, \ldots$,

$$\mathrm{Var}[X_1^{(m)}] = \frac{\mathrm{Var}[X_1]}{m^\beta}.$$

If the condition in the above definition only holds asymptotically, namely,[1]

$$\mathrm{Var}[X_1^{(m)}] \sim \frac{\mathrm{Var}[X_1]}{m^\beta} \text{ as } m \to \infty,$$

[1]Here, $a_m \sim b_m$ as $m \to \infty$ is brief equivalence to that $\lim_{m\to\infty} a_m/b_m = 1$ [Tsy98, pp. 1715].

then the stationary process is *asymptotically second-order self-similar*. The parameter β can be shown to be related to the Hurst parameter H as

$$H = 1 - \frac{\beta}{2}.$$

In case the increments $X_1, X_2, \ldots$ are independent to each other, the parameter β will become 1, which indicates the process is not self-similar.

2.2 Heavy-tailed distributions

In literature, heavy tails in distributions mean that the probability of exceeding a large value remains large comparatively. A formal definition is given below.

Definition 2.4 *The distribution of a random variable X is heavy-tailed, if*

$$1 - F(x) = \Pr[X > x] \geq a(x), \tag{2.1}$$

for some $a(x) \sim x^{-\alpha}$ as $x \to \infty$ and some $\alpha > 0$, where F is the cumulative distribution function of X.[2]

The exponential distribution with probability density function $\lambda e^{-\lambda x}$ for $x \geq 0$ is a simple example for a non-heavy-tailed distribution, for which

$$\Pr[X > x] = e^{-\lambda x}.$$

[2]In some literature, a heavy-tailed random variable is defined by [Stall98, pp. 190]

$$1 - F(x) = \Pr[X > x] \sim x^{-\alpha} \text{ as } x \to \infty \text{ for some } \alpha > 0.$$

This definition, although widely adopted, may arise some problem in its application. For example, suppose that the random variable X has probability density function $\log(x)/x^2$ for $x \geq 1$. Then

$$\frac{\Pr[X > x]}{x^{-\alpha}} = \frac{1 + \log(x)}{x^{1-\alpha}},$$

which does not go to 1 for any $\alpha > 0$. However, this random variable is heavy-tailed in natural, and its complementary cumulative distribution function $\Pr[X > x]$ is greater than $x^{-\alpha}$ for $\alpha = 1$; therefore, our definition can be applied.

It is obvious that $e^{-\lambda x}$ decays to 0 faster than any polynomial order.

A typical heavy-tailed distribution is the Pareto distribution, for which the distribution density function is defined as:[3]

$$f(x) = \begin{cases} 0 & \text{for } x < k; \\ \frac{\alpha}{k}\left(\frac{k}{x}\right)^{\alpha+1} & \text{for } x \geq k. \end{cases}$$

Its complementary cumulative distribution function is equal to:

$$\Pr[X > x] = \begin{cases} 1, & \text{for } x < k; \\ \left(\frac{k}{x}\right)^{\alpha}, & \text{for } x \geq k. \end{cases}$$

The Pareto distributed random variable X has the first and second moments as:

$$E[X] = \begin{cases} \frac{\alpha k}{\alpha - 1}, & \text{for } \alpha > 1; \\ \infty, & \text{otherwise}, \end{cases} \tag{2.2}$$

and

$$E[X^2] = \begin{cases} \frac{\alpha k^2}{\alpha - 2}, & \text{for } \alpha > 2 \\ \infty, & \text{otherwise}. \end{cases} \tag{2.3}$$

In this thesis, we will compare the statistical effects of two inter-arrival distributions: exponential and Pareto. In order to obtain a fair comparison, the means of both distributions are set equal, which implies that the Pareto distribution that we concern must exhibit a finite mean. The variance of the concerned Pareto distribution, however, is taken to be infinity to increase the variability of the arrival traffic. This restricts α to be $1 < \alpha \leq 2$. (cf. (2.2) and (2.3)).

[3]There is variant definition for Pareto distribution, such as [Fell71, pp. 50]

$$f(x) = \frac{\Gamma(\mu+\nu)}{\Gamma(\mu)\Gamma(\nu)} \frac{x^{\mu-1}}{(1+x)^{\mu+\nu}} \quad \text{for } x \geq 0, \mu > 0 \text{ and } \nu > 0,$$

where $\Gamma(t) = \int_0^\infty u^{t-1}e^{-u}du$ is the gamma function. The definition we adopted here has the advantage that it has a close-form formula for the cumulative distribution function.

2.3 Relationship between self-similarity and heavy-tails

2.3.1 Fractional Brownian motion

Fractional Brownian motion is a self-similar Gaussian process with stationary increments. By Definition 2.1, the distribution of $\{T^{-H}B_H(Tt), t \geq 0\}$ does not depend on T. For any $t \geq 0$ and $\Delta t > 0$, the increment $B_H(t + \Delta t) - B_H(t)$ is normally distributed with mean 0 and variance $(\Delta t)^{2H}$; thus, its density function is given by:

$$\Pr\left[B_H(t + \Delta t) - B_H(t) \leq x\right] = \frac{1}{\sqrt{2\pi}(\Delta t)^H} \int_{-\infty}^{x} e^{-y^2/[2(\Delta t)^{2H}]} dy.$$

2.3.2 ON/OFF source

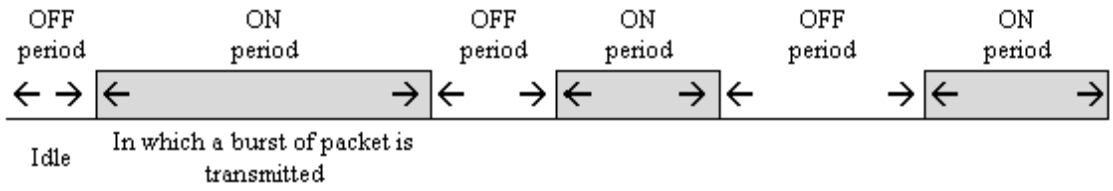


Figure 2.2: *ON/OFF source.*

The ON/OFF source can be expressed in a simple way as Figure 2.2. Conventionally, the length of ON or OFF source is characterized by finite-variance distribution, such as exponential. It is demonstrated in [Will97] that superposition of infinitely many Pareto-distributed ON/OFF sources results in a self-similar traffic with Hurst parameter

$$H = \frac{(3 - \alpha)}{2} \tag{2.4}$$

This analytical relation between α and H was, to some extent, confirmed by observations on Ethernet traffic in [Stall02]. Specifically, the authors in [Stall02] found that the individual sources with $\alpha = 1.2$ superimposed to cause an $H = 0.9$ self-similar traffic.

An alternative model of arrivals can be found in [Box98] as:

$$W(t) = \sum_{k=0}^{t} W_k \cdot I_{\left(S_0+\sum_{j=1}^{k-1} U_j,\ S_0+\sum_{j=1}^{k} U_j\right]}(t) \tag{2.5}$$

where $W_0, W_1, \ldots$ and $U_1, U_2, \ldots$ are both independent and identically distributed and are independent to each other, and $I_{\{\cdot\}}(t)$ represents the set indicator function. It is further assumed that for each j, W_j has mean 0 and finite variance, and U_j is heavy-tailed distributed with parameter $0 < \alpha < 2$. The Pareto distribution apparently satisfies the requirement of $\{U_j\}_{j\geq 1}$ in (2.5). The situation of $\alpha = 1.5$ was then derived in [Box98], and obtained again that $H = (3-\alpha)/2 = 0.75$.

2.3.3 M/G/1 model

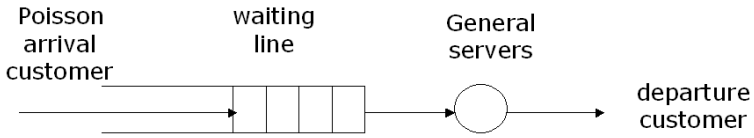


Figure 2.3: *M/G/1 queuing model.*

The M/G/1 model is essential in queuing analysis. Such a renewal process is analyzed and used to predict, e.g., the first and second moments of the waiting time, the busy period of a system, the distribution of queue length, etc. Let $\tilde{q}_n$ be the queue length at the time instance n. Then

$$\tilde{q}_{n+1} = \begin{cases} \tilde{q_n} - 1 + \tilde{v_n}, & \text{if } \tilde{q_n} > 0 \\ \tilde{v_n}, & \text{if } \tilde{q_n} = 0 \end{cases} = [\tilde{q}_n - 1]^+ + \tilde{v}_n, \tag{2.6}$$

where $\tilde{v_n}$ is the arrival sequence, and $[x]^+ = \max\{0, x\}$.

2.4 Determination of heavy-tail distributions and self-similarity

2.4.1 Heave-tail determination

For an exponential distribution, taking the logarithm of its complementary cumulative distribution function gives a straight line. However, applying the same logarithm to a heavy-tailed distribution yields a log-curve. Nevertheless, if we also take the logarithm on the observation of a heavy-tailed random variable, a log-log plot of the complementary cumulative distribution function is obtained and the curve should approximately be a straight line of negative slope $-\alpha$. This suggests a way to estimate the degree of heavy-tail for a random variable. In summary, a straight line in a log-linear plot hints a non-heavy-tail behavior like exponential, and a straight-line in a log-log plot suggests the existence of heavy tail for the concerned observations.

2.4.2 Variance-time plot

The aggregated increment process $\boldsymbol{X}^{(m)}$ of a continuous-time self-similar process follows the variance rule, as mentioned before, as:

$$\mathrm{Var}[X^{(m)}] \;=\; \frac{\mathrm{Var}[X]}{m^{\beta}}.$$

The equation can be rewritten as:

$$\log(\mathrm{Var}[X^{(m)}]) \;=\; \log(\mathrm{Var}[X]) - \beta \log(m). \tag{2.7}$$

Since $\log(\mathrm{Var}[X])$ is independent of m, the relation between $\log(\mathrm{Var}[X^{(m)}])$ and $\log(m)$ becomes a straight-line of slope $-\beta$, which can be used to determine the Hurst parameter $H = 1 - (\beta/2)$. The variance-time log-log plot therefore is a straightforward method to estimate H. Note that a straight-line of slope between -1 and 0 suggests self-similar.

Chapter 3

Investigation of System Model with Finite Resources

In the previous chapter, we introduced several traffic models that have been considered in the literature for self-similarity analysis. The basis of these models is the assumption of superposition of infinitely many resources, such as aggregations of infinitely many ON/OFF sources.

In this chapter, we will consider an unalike situation in which some of the system resources are finite, and will then investigate whether the obtained relation between the degrees of heavy-tail and self-similarity based upon the models in Subsections 2.3.2 and 2.3.3 remains. We will subsequently study how a heavy-tailed server contributes to self-similarity in traffic behavior, and how a Markov server releases the degree of traffic self-similarity.

3.1 Model of parallel servers

3.1.1 M/G/L equivalence to our system model

In our system model, the source is poured into L servers with constant service rate through a first-come-first-serve queue. The packet length is assumed to be Pareto distributed, and hence, is heavy-tailed in nature. As the single source represents an aggregated traffic, packets can be overlapped in their durations. Specifically, two consecutive packets, respectively with lengthes $\tilde{\chi}_i$ and $\tilde{\chi}_{i+1}$, arrive at times ϕ_i and ϕ_{i+1} are allowed to have $\tilde{\chi}_i > \phi_{i+1} - \phi_i$. The

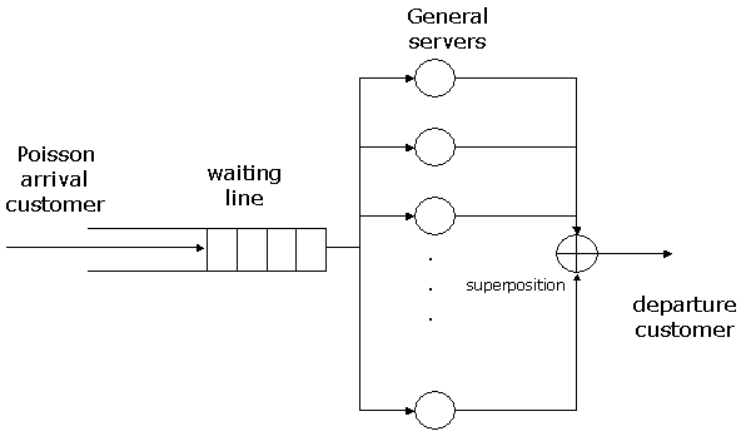


Figure 3.1: *M/G/L queueing model with a single input data stream arriving according to a Poisson process with rate λ (inter-arrival is therefore exponentially distributed). The packet length of each arrival is assumed Pareto distributed. In stead of using a single statistically-defined service-rate server to represent the aggregated behavior of many servers, we assume that there are L constant rate servers. Such a model can be equivalently viewed as that the packet length is zero, and there are L servers with i.i.d. Pareto distributed service rates.*

number of packets arrived follows the Poisson distribution. Such a system, unlike those considered in the literature, is a system with deterministic-rate finite-number servers.

As shown in Figure 3.1, such a system model can be equivalently viewed as an M/G/L model with Pareto distributed service rate in each server, and zero packet length (or impulse arrivals). We wish to investigate through this model the relation between the degrees of heavy-tail and self-similarity, as well as the number of servers, L. Intuitively, as the number of servers goes to infinity, our system will similarly reduce to the one analyzed in [Will97]; hence, for the overall output traffic, the relation between heavy-tail parameter α and self-similar parameter H follows $H = (3 - \alpha)/2$. However, a natural question to ask is that "In situation where the number of servers is finite, how does this relation change according to L?" We will try to answer this question through various simulations.

Let us first consider the simplest case where $L = 1$. As aforementioned, we can equivalently view the system as an M/G/1 model with exponential distributed inter-arrival

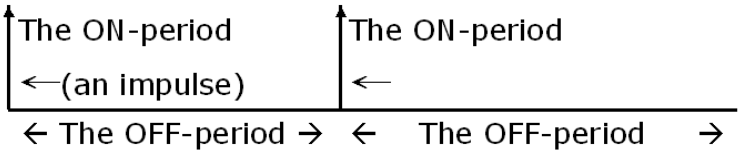


Figure 3.2: *When L reduces to 1, the system with packets of Pareto distributed length can be thought as an ON/OFF process, where the ON period is simply an impulse packet of length 0, while the OFF period is the length of inter-departures.*

of impulse packets and Pareto distributed service rate. We will interchangeably use these two equivalent views whichever is convenient

Denote the time for the first bit of the nth arrival packet to reach the queue by ϕ_n. Denote the time for the first bit of the nth departure packet to leave the queue by φ_n. Let the packet length of nth packet be $\tilde{\chi}_n$. It may be easier to analyze the system by the equivalent aspect that a sequence of impulse packets arrives at times $\phi_1, \phi_2, \ldots$, which in turns departs at times $\varphi_1, \varphi_1, \ldots$, and the sever has Pareto distributed service rate. Then $\varphi_1 = \phi_1$, since the queue is empty initially. Obviously,

$$\varphi_2 - \varphi_1 = \begin{cases} \tilde{\chi}_1, & \text{if } \varphi_1 + \tilde{\chi}_1 > \phi_2, \\ \phi_2 - \phi_1, & \text{if } \varphi_1 + \tilde{\chi}_1 \le \phi_2, \end{cases}$$

which, by $\varphi_1 = \phi_1$, implies

$$\varphi_2 = \begin{cases} \varphi_1 + \tilde{\chi}_1, & \text{if } \varphi_1 + \tilde{\chi}_1 > \phi_2, \\ \phi_2, & \text{if } \varphi_1 + \tilde{\chi}_1 \le \phi_2, \end{cases} = \max\{\varphi_1 + \tilde{\chi}_1, \phi_2\}.$$

Next,

$$\varphi_3 - \varphi_2 = \begin{cases} \tilde{\chi}_2, & \text{if } \varphi_2 = \phi_2 \text{ and } \varphi_2 + \tilde{\chi}_2 > \phi_3 \\ \phi_3 - \phi_2, & \text{if } \varphi_2 = \phi_2 \text{ and } \varphi_2 + \tilde{\chi}_2 \le \phi_3 \\ \tilde{\chi}_2, & \text{if } \varphi_2 > \phi_2 \text{ and } \varphi_2 + \tilde{\chi}_2 > \phi_3 \\ \phi_3 - \varphi_2, & \text{if } \varphi_2 > \phi_2 \text{ and } \varphi_2 + \tilde{\chi}_2 \le \phi_3 \end{cases}$$

$$= \begin{cases} \tilde{\chi}_2, & \text{if } \varphi_2 + \tilde{\chi}_2 > \phi_3 \\ \phi_3 - \varphi_2, & \text{if } \varphi_2 > \phi_2 \text{ and } \phi_2 + \tilde{\chi}_2 \le \phi_3 \\ \phi_3 - \phi_2, & \text{if } \varphi_2 = \phi_2 \text{ and } \phi_2 + \tilde{\chi}_2 \le \phi_3, \end{cases}$$

which implies

$$\varphi_3 = \begin{cases} \varphi_2 + \tilde{\chi}_2, & \text{if } \varphi_2 + \tilde{\chi}_2 > \phi_3 \\ \phi_3, & \text{if } \varphi_2 + \tilde{\chi}_2 \le \phi_3. \end{cases} = \max\{\varphi_2 + \tilde{\chi}_2, \phi_3\}.$$

Again,

$$\varphi_4 - \varphi_3 = \begin{cases} \tilde{\chi}_3, & \text{if } \varphi_3 = \phi_3 \text{ and } \varphi_3 + \tilde{\chi}_3 > \phi_4 \\ \phi_4 - \phi_3, & \text{if } \varphi_3 = \phi_3 \text{ and } \varphi_3 + \tilde{\chi}_3 \leq \phi_4 \\ \tilde{\chi}_3, & \text{if } \varphi_3 > \phi_3 \text{ and } \varphi_3 + \tilde{\chi}_3 > \phi_4 \\ \phi_4 - \varphi_3, & \text{if } \varphi_3 > \phi_3 \text{ and } \varphi_3 + \tilde{\chi}_3 \leq \phi_4 \end{cases}$$
$$= \begin{cases} \tilde{\chi}_3, & \text{if } \varphi_3 + \tilde{\chi}_3 > \phi_4 \\ \phi_4 - \varphi_3, & \text{if } \varphi_3 > \phi_3 \text{ and } \phi_3 + \tilde{\chi}_3 \leq \phi_4 \\ \phi_4 - \phi_3, & \text{if } \varphi_3 = \phi_3 \text{ and } \phi_3 + \tilde{\chi}_3 \leq \phi_4, \end{cases}$$

which implies

$$\varphi_4 = \begin{cases} \varphi_3 + \tilde{\chi}_3, & \text{if } \varphi_3 + \tilde{\chi}_3 > \phi_4 \\ \phi_4, & \text{if } \varphi_3 + \tilde{\chi}_3 \leq \phi_4. \end{cases} = \max\{\varphi_3 + \tilde{\chi}_3, \phi_4\}.$$

Continuing the derivation yields:

$$\varphi_{n+1} = \begin{cases} \varphi_n + \tilde{\chi}_n, & \text{if } \varphi_n + \tilde{\chi}_n > \phi_{n+1} \\ \phi_{n+1}, & \text{if } \phi_n + \tilde{\chi}_n \leq \phi_{n+1}. \end{cases}$$
$$= \max\{\varphi_n + \tilde{\chi}_n, \phi_{n+1}\} \quad (3.1)$$

Conclusively, we obtain that the inter-departure $\tilde{\tau}_n = \varphi_{n+1} - \varphi_n$ and the exponentially distributed inter-arrival $\tau_n = \phi_{n+1} - \phi_n$ should satisfy:

$$\tilde{\chi}_n \leq \tilde{\tau}_n = \varphi_{n+1} - \varphi_n = \max\{\tilde{\chi}_n, \phi_{n+1} - \varphi_n\} \leq \max\{\tilde{\chi}_n, \phi_{n+1} - \phi_n\} = \max\{\tilde{\chi}_n, \tau_n\}. \quad (3.2)$$

Hence, if with high probability, the Pareto distributed $\tilde{\chi}_n$ is larger than the exponentially distributed τ_n, then $\tilde{\tau}_n$ will be close to a Pareto distribution.

3.1.2 System approximation technique

As long as superposition of traffics is concerned, the approximation technique used in [Will97] is of good help.

Suppose there are M homogeneous ON/OFF sources modelled as $\{W^{(m)}(t), t \geq 0\}$, where $W^{(m)}(t) = 1$ indicates the source is turned ON, while the source is turned OFF when

$W^{(m)}(t) = 0$. Consider their superposition at time t as $\sum_{m=1}^{M} W^{(m)}(t)$. Then the aggregated accumulative packet counts in the time interval $[0, Tt)$ is given by:

$$W_M^*(Tt) \quad = \quad \int_0^{Tt} \left(\sum_{m=1}^{M} W^{(m)}(u) \right) du$$

The paper of [Will97] was interested in the statistical behavior of $\{W_M^*(Tt), t \geq 0\}$ for $M \to \infty$ and $T \to \infty$. This behavior depends on the distribution of the ON- and OFF- periods, and can be reduced to the form of $\{\sigma_{lim} B_H(t), t \geq 0\}$, where σ_{lim} is a finite positive constant and $\{B_H(t), t \geq 0\}$ is a fractional Brownian motion. In the analysis, the distributions of the ON- and OFF-periods can be different. For example, one can be Pareto distributed and the other has exponential distribution. The complementary cumulative distribution functions of the Pareto distribution assumed for the two periods are of the form:

$$1 - F_j(x) \quad \sim \quad \begin{cases} l_j \, L_j(x) \, x^{-\alpha_j}, & \text{for } 1 < \alpha_j < 2 \\ x^{-2}, & \text{for } \alpha_j = 2 \end{cases} \quad \text{as } x \to \infty, \tag{3.3}$$

where l_j is a constant, $L_j(x)$ is a locally bounded function of x at x large, and $j = 1, 2$. (For definition of locally bounded function, see [Bing87]). By defining

$$a_j = \begin{cases} l_j \dfrac{\Gamma(2 - \alpha_j)}{\alpha_j - 1}, & \text{for } \sigma_j^2 = \infty \\ \sigma_j^2, & \text{for } \sigma_j^2 < \infty \end{cases}$$

where σ_j^2 is the variance corresponding to cumulative distribution function $F_j(x)$, the authors of [Will97] established the next theorem.

Theorem 1 *For large M and T, the aggregate cumulative packet process $\{W_M^*(Tt), t \geq 0\}$ behaves statistically like*

$$TM \frac{\mu_1}{\mu_1 + \mu_2} t + T^H \sqrt{L(T) \cdot M} \cdot \sigma_{lim} B_H(t),$$

where μ_1 *and* μ_2 *are the mean of the two period distributions, and*

$$\sigma^2_{lim} = \begin{cases} \dfrac{2(\mu_2^2 a_1 \Lambda + \mu_1^2 a_2)}{(\mu_1+\mu_2)^3 \Gamma(4-\min\{\alpha_1,\alpha_2\})}, & \text{if } 0 < \Lambda < \infty; \\ \dfrac{2\ \max\{\mu_1,\mu_2\}^2\ \min\{a_1,a_2\}}{(\mu_1+\mu_2)^3 \Gamma(4-\min\{\alpha_1,\alpha_2\})}, & \text{if } \Lambda = 0; \\ \dfrac{2\ \max\{\mu_1,\mu_2\}^2\ \min\{a_1,a_2\}}{(\mu_1+\mu_2)^3 \Gamma(4-\min\{\alpha_1,\alpha_2\})}, & \text{if } \Lambda = \infty, \end{cases}$$

and

$$L(x) = \begin{cases} L_2(x), & \text{if } 0 < \Lambda < \infty; \\ \min\{L_1(x), L_2(x)\}, & \text{otherwise}, \end{cases}$$

and

$$\Lambda = \lim_{x\to\infty} \frac{1-F_1(x)}{1-F_2(x)}. \tag{3.4}$$

The author then concludes that $H = (3 - \min\{\alpha_1, \alpha_2\})/2$.

3.1.3 Simulation concern

The L Pareto distributed severs considered in our equivalent system can be assumed independent and identically distributed (i.i.d.), because the lengths of packets in the original system are i.i.d. in statistics. Hence, the only source of dependence among departures of different servers is the shared queue. This makes the ON period and OFF period of aggregated departures from L shared-queue servers becomes correlated in nature, as contrary to the i.i.d. assumption made in [Will97]. Question is whether the same relation of α (specifically $\min\{\alpha_1, \alpha_2\}$) and H as Theorem 1 is applicable to the ON/OFF-correlated aggregated departures from L shared-queue servers.

The dependence of ON/OFF periods in our L-server system model can be viewed as a hidden variable. Our aim in this research is to determine its influence on heavy-tail and self-similarity of aggregated departures.

Case of small L

In this case, it is anticipated that the aggregated traffic behavior at the L-server outputs may not be similar to that in Theorem 1. We will therefore find its trend curve among α and H and L.

Case of large but finite L

For large L, the aggregated traffic behavior at the L-server outputs may be more analogous to that in Theorem 1. What is concerned here is that "Is it possible to confirm Theorem 1 by simulations, as the rate of convergence in distribution may be slow?"

3.2 Model of tandem servers

3.2.1 Self-similarity of inter-departures at the last server

Consider a system with serially connected severs as depicted in Figure 3.3. By Burke's theorem, the sequence of inter-departure time for the M/M/1 system, in stochastic equilibrium, is exponential distributed with the same parameter as the inter-arrival distribution. Specifically, if the inter-arrival time sequence and the service time sequence are both exponential-distributed processes, so does the inter-departure at the ultimate sever output.

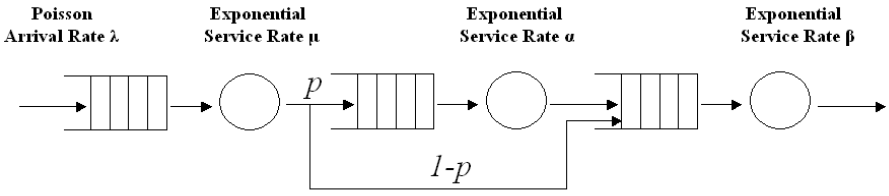


Figure 3.3: *Block diagram for simple network of tandem queues.*

Hence, a serial accumulation of exponential servers does not contribute self-similarity in datum traffic. However, does a serial accumulation of, say, heavy-tailed server, establish self-similarity? Also, if the original source is already self-similar, can this self-similarity be

released after passing through a bunch of serially connected exponential servers? These questions are the focus of the second issue examined in this thesis.

Hence, in our simulations, there will be four possible combinations of source inter-arrival and service time (cf. Figure 3.4): (1) exponential versus exponential; (2) exponential versus Pareto; (3) Pareto versus exponential; (4) Pareto versus Pareto. The statistical behavior of the first combination is already known. We will therefore concentrate on (2), (3) in our simulations with different number of servers, N. The output traffic at the very end of the serially connected server will be examined.

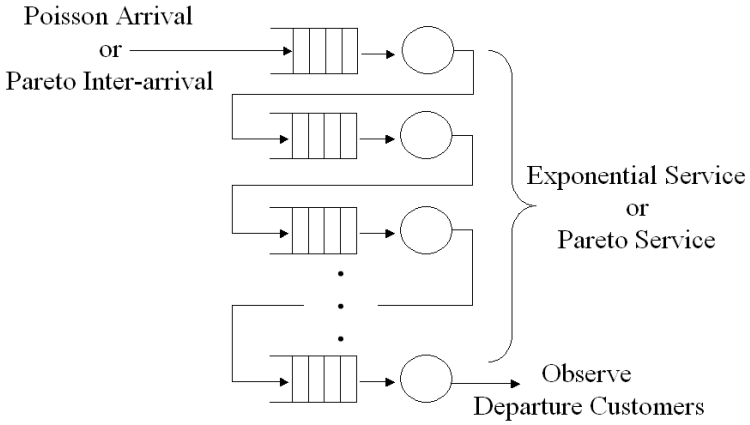


Figure 3.4: *Series of servers.*

According to [Ana96], such a serially connected network channel has memory even if the service times are independent and identically distributed. This is due to that these queues introduce so-called "back-up" effect on the packets. It was shown [Ana96] that when the service time is deterministic, the channel has infinite capacity as a timing channel. This capacity is achieved by transmitting each packet with interval farther apart than the service time. That the outputs depend on inputs in a nonlinear fashion in such a system is another "information-theoretic challenge" as mentioned in [Ana96]. This challenge hints to us that to vary N in a non-linear way is perhaps a better way to examine the system self-similarity.

3.2.2 Entropy rate variation of a tandem network system

"Entropy rate" is a measure of information content to a sequence of random variables. It is expected that if these random variables have memory, a smaller entropy rate will result. The definition of entropy rate is given below.

Definition 3.1 *The entropy rate for a source* $\boldsymbol{X} = \{X_1, X_2, \ldots\}$ *is given by*

$$\lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n)$$

provided the limit exists.

The Lempel-Ziv algorithm is an asymptotically optimal coder for all stationary sources. It was shown that its compressing rate ultimately approaches the entropy rate of any stationary source. The encoding rule of the Lempel-Ziv algorithm is quoted in the following.

Let $\mathcal{A}$ be the set of individual string symbols, and $\mathcal{D}$ the set of entries in a dictionary. When the number of element in $\mathcal{A}$ is equal to two, a binary string is obtained. The considered source in this thesis has $\mathcal{A} = \{a, b\}$, as shown in Fig. 3.5.

Algorithm 3.1 *Lempel-Ziv Encoding Algorithm*

1. Initialize the dictionary to contain all blocks of length one (namely, $\mathcal{D} = \{\{a\}, \{b\}\}$*).*

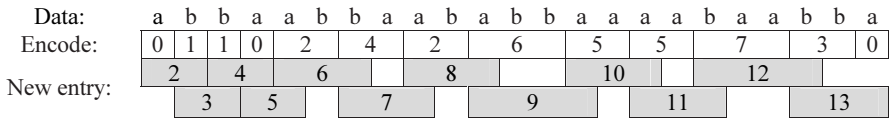
2. Search the longest block, starting from the current source position, which has appeared in the dictionary $\mathcal{D}$*.*

3. Encode W *by its index in the dictionary.*

4. Add W *concatenated by the next source symbol to the dictionary.*

5. Go to Step 2.

To facilitate the understanding of the above algorithm, an example is quoted from http://www.data-compression.com/lossless.html in Fig. 3.5.



Index	Entry	Index	Entry	Index	Entry	Index	Entry
0	a(initial)	4	ba	8	aba	12	baab
1	b(initial)	5	aa	9	abba	13	bba
2	ab	6	abb	10	aaa	-	-
3	bb	7	baa	11	aab	-	-

Figure 3.5: *A series of data string with bit element $\mathcal{D} = \{a, b\}$ being encoded. New entry at each moment is also sequentially listed.*

In this research, we wish to examine the entropy rate of the inter-departures at the last server, and investigate the influence of serially connected servers on the information content of the inter-departures.

To obtain an estimate of the entropy rate, the sequence of inter-departures will be encoded by the Lempel-Ziv algorithm, and the ultimate data rate will be recorded.

To avoid non-stationarity in departure statistics at the initial stage of the simulations, we will not begin encoding the inter-departures immediately after the simulation starts; rather, the encoding process only applied to the middle part of the output data stream (cf. Figure 3.6).

As shown in Figure 3.6, the inter-departures will be recorded into a file which in turn is Lempel-Ziv encoded. Through this, we may, to some extent, examine the degree of dependence among inter-departures.

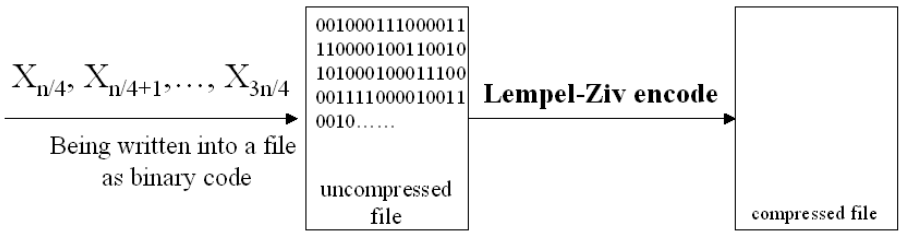


Figure 3.6: *The Lempel-Ziv simulation system for entropy rate.*

Chapter 4

Simulation Results and Analysis

This chapter examines the degree of self-similarity of the traffics generated under the system settings in Chapter 3. The analytical foundation mentioned in this chapter has been introduced in the previous two chapters.

4.1 M/G/L Systems

As shown in Fig. 3.1, the momentary queue length Q_n at time n can be described by

$$Q_n = Q_{n-1} + A_n - X_n,$$

where A_n and X_n are the arrival and departure during the nth time interval, if the server occupying time of each customer is fixed. Notably, X_n is always less than or equal to $Q_{n-1} + A_n$. Here, our main concern is the statistical behavior of $X_1, X_2, X_3, \ldots$. In case where variable server occupying time is allowed, X_n becomes a function of:

1. the queue length Q_{n-1} at time instant $n-1$;
2. the arrival count A_n between time instance $n-1$ to time instance n;
3. the server occupying time of each customer, denoted by χ_k for kth customer;
4. the remaining service time of each of L servers at time instance $n-1$, denoted respectively by $\tilde{\epsilon}^1_{n-1}$, $\tilde{\epsilon}^2_{n-1}$, ..., $\tilde{\epsilon}^L_{n-1}$.

4.1.1 System With Constant Service Mean Time

Figure 4.1 illustrates the simulated variance-time plot of M/G/L system for which $L = 4$. The input inter-arrival is exponential distributed with parameter $\lambda = 0.01$ (so the mean is $1/\lambda = 100$). The four service times for the four servers are i.id. Pareto distributed. For the Pareto service time, two mean values are investigated here: 200 and 399.9, which respectively gives a mean service rate of $\mu = 0.005$ and $\mu = 0.00250063$. The former then gives a utilization of $\rho = \lambda/(L\mu) = 0.01/(4 \times 0.005) = 0.5$,[1] while the latter presents $\rho = \lambda/(L\mu) = 0.01/(0.00250063 \times 4) = 0.99975 \approx 1$.

A) Inconsistent Slope For Different Pareto Shaping Parameters

From Fig. 4.1, we observe that the slope of the resultant variance-time curves varies for different Pareto shaping parameter α, only when $\rho \approx 1.0$. For a smaller utilization like $\rho = 0.5$, the slope of the resultant variance-time curve is always close to -1 (i.e., the estimated $\hat{H}$ from simulations is close to 0.5 as listed in Tab. 4.1). From Fig. 4.1(j) and Tab. 4.1, the slope of the obtained variance-time curve for exponential server remains -1, irrelevant to the utilization, which matches the general anticipation. As a result, the departure behavior under finite number of servers is seemingly different from the analytical result established under the assumption of infinite number of servers (cf. The right-most column of Tab. 4.1).

It needs to be pointed out that among the six simulated points in each subfigure of Fig. 4.1, which respectively correspond to $m = 1, 10, 10^2, 10^3, 10^4, 10^5$, only the first four points are used to calculate the slope of the best-fit line for Tab. 4.1. This is because that the last two points evidently belong to a line with different slope from the line decided by the first four points. To be more specific, there is a sudden slope change at $m \geq 10^3$. Such a sudden increase in slope indicates the disappearance of self-similarity at larger average

[1] There are two parameters, α and k, in Pareto distribution, and its mean is equal to $(1/\mu) = \alpha k/(\alpha - 1)$ (cf. (2.2)). In our simulation, k will be adjusted according to different α taken so that the mean $(1/\mu)$ (and hence, the utilization) remains constant.

Table 4.1: *Estimated Hurst parameter for 4 parallel Pareto servers*

α	$\hat{H}$ (utilization=0.5)	$\hat{H}$ (utilization≈1.0)	$H = (3-\alpha)/2$
2.0	0.5003	0.5722	0.5
1.8	0.5003	0.6133	0.6
1.6	0.5021	0.6617	0.7
1.5	0.5013	0.6869	0.75
1.33	0.4991	0.7282	0.835
1.25	0.4974	0.7314	0.875
1.2	0.4962	0.7262	0.9
1.125	0.4952	0.6851	0.9375
1.1	0.4953	0.66025	0.95
exp	0.5010	0.4981	N.A.

window m.

The slope change at larger average window m suggests that the dependence of two packets that are separated by m packets is prohibitively weak. Actually, if packet A enters the queue after packet B is served, the departure behaviors of packet A and packet B are statistically independent because the queue is the only place to introduce dependence in the system. By a further examination, we found that the maximum queue length that has ever been achieved during simulation is around 2000 (packets) for a moderate α lying between 1.1 and 1.33; hence, it is anticipated that for $m \geq (1/\lambda) \times 2000 = 2 \times 10^5$, the long-term dependence of the departure process should disappear.

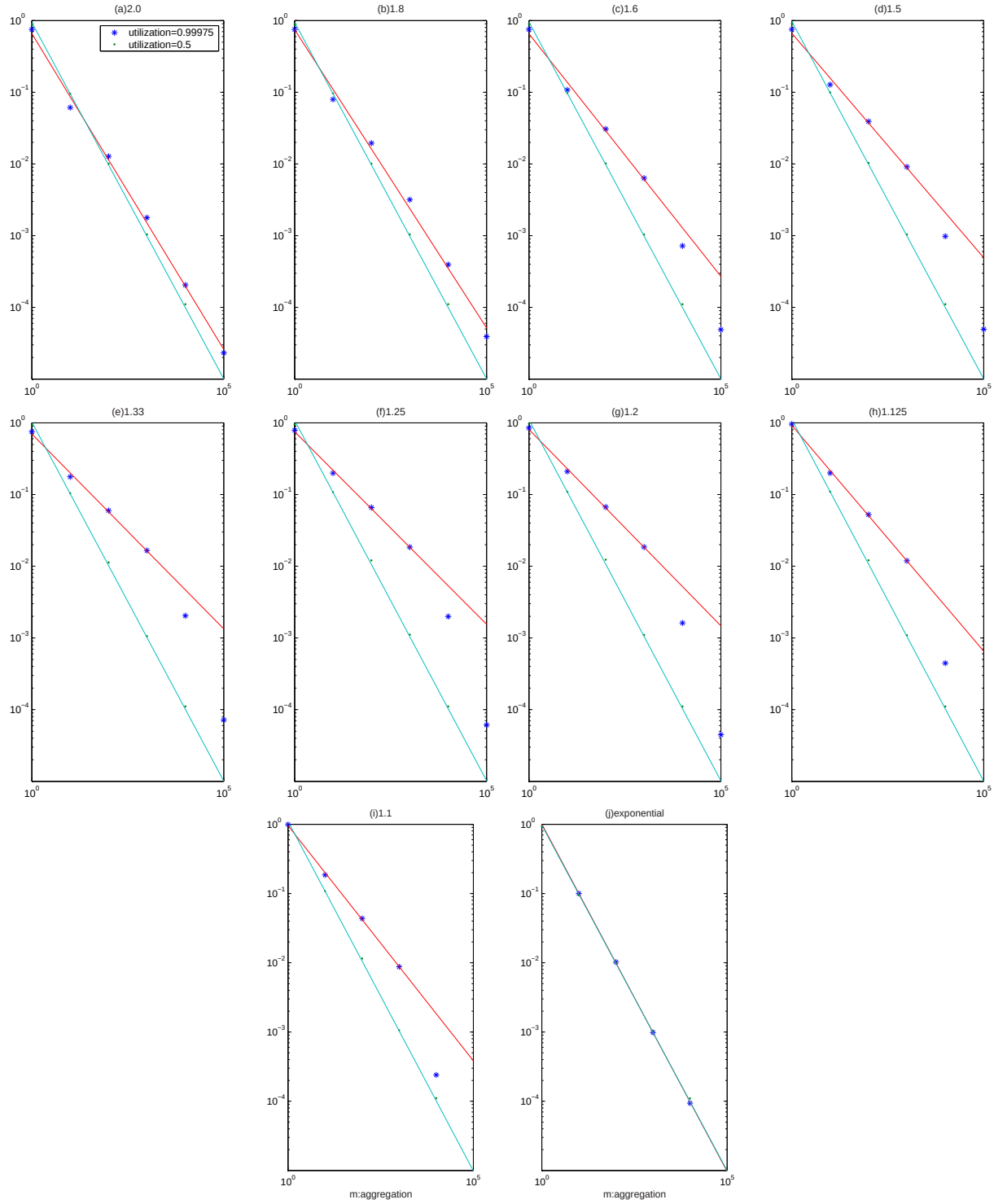


Figure 4.1: *The variance-time plot for* 4 *parallel servers* ($L = 4$). *Two utilizations are investigated—"·" for* $\rho = 0.5$ *and "∗" for* $\rho \approx 1.0$. *The number aside (a)–(i) is the shaping parameter* α *of the Pareto servers. For comparison, sub-figure (j) illustrates the exponential servers with the same mean rates as Pareto servers.*

B) Utilization And Self-Similarity

In principle, if the system utilization is far away from unity (i.e., $\lambda \ll L\mu$), the queue length is expected to be bounded. On the contrary, if utilization is greater than one (i.e., $\lambda > L\mu$), the queue length is expected to be ultimately infinite. In the previous subsection, only two utilizations are examined, which can be summarized as "$\rho \approx 1$ introduces self-similarity, while $\rho = 0.5$ implies no self-similarity." In this subsection, a further study on the relation between utilization and induced departure self-similarity is conducted.

We observe from Fig. 4.2 that there is almost no self-similarity for utilization less than 0.8. As utilization is further increased, self-similarity begins to appear at a certain range of average window m.

Figure 4.3 depicts the observed $\hat{H}$, as a function of Pareto shaping parameter α, in Tab. 4.1. We then use different order of polynomials to fit the nine points in Fig. 4.3. It can be seen that all the 2nd, 3rd and 4th order polynomial approximations peak at $\alpha \approx 1.3$. The maximum self-similar parameter for 2nd and 3rd polynomial approximations are $\hat{H} = 0.7066$ and $\hat{H} = 0.7295$. It is worth mentioning that the maximum self-similar parameter does not occur at boundary α (i.e, $\alpha = 1$ and $\alpha = 2$) but at some internal point.

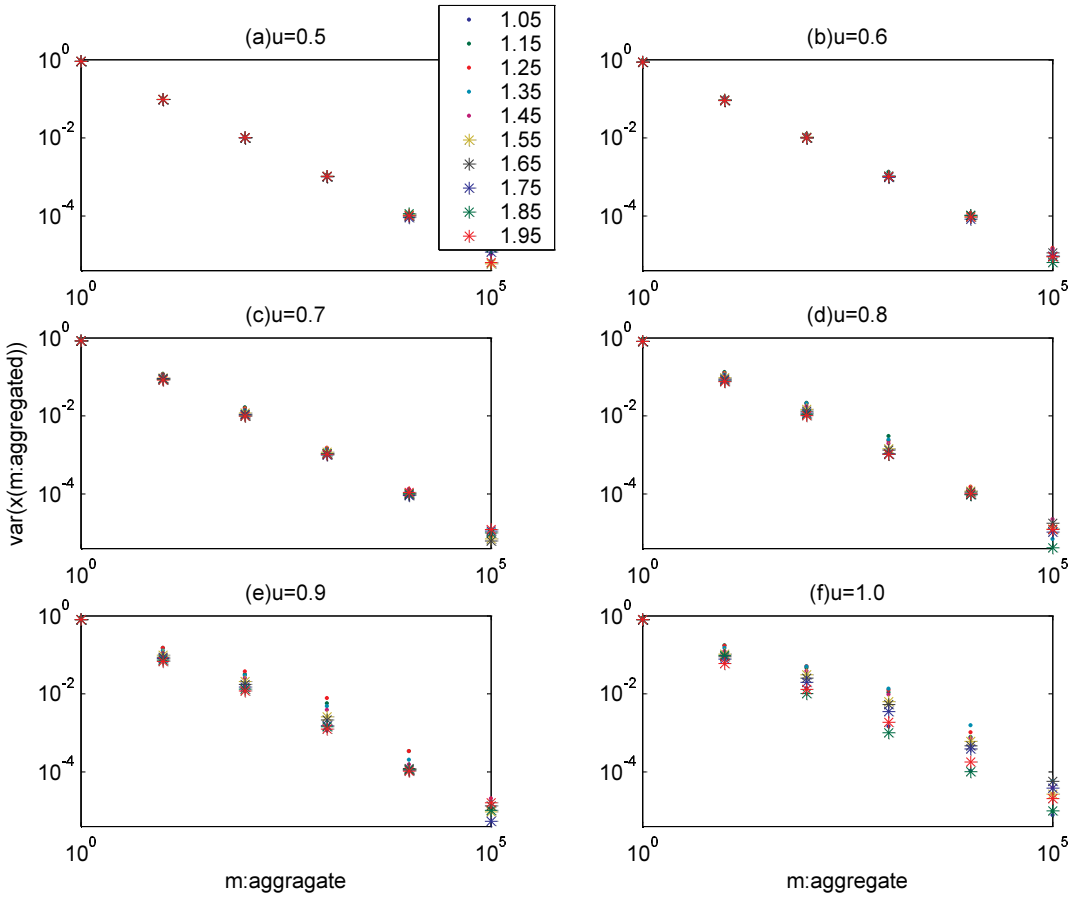


Figure 4.2: *The variance-time plot for 5 parallel Pareto servers with different utilization.*

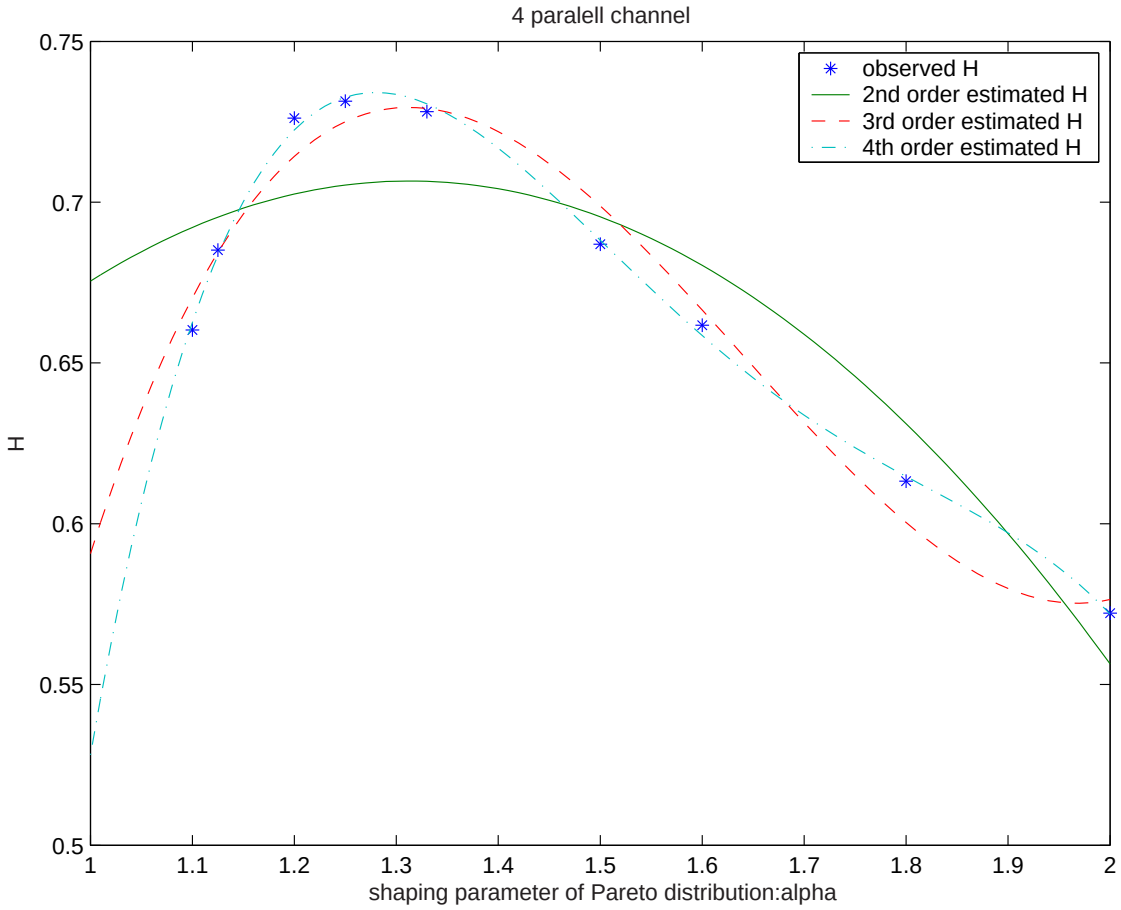


Figure 4.3: *Estimated H, as a function of α, for 4 parallel servers with* $\rho = \lambda/(L\mu) = 0.99975$.

Table 4.2: *List of observed Hurst parameter and Pareto shaping parameter in Fig. 4.4.*

α	$L = 8$	$L = 16$	$L = 32$	$L = 64$	$H = (3 - \alpha)/2$
1.01	0.4955	0.498	0.493	0.5038	0.995
1.11	0.589	0.5065	0.50115	0.50425	0.945
1.21	0.6538	0.60685	0.56165	0.54605	0.895
1.31	0.69075	0.66425	0.64135	0.7132	0.845
1.41	0.6867	0.6642	0.67285	0.72665	0.795
1.51	0.6684	0.661	0.6721	0.7213	0.745
1.61	0.66855	0.6685	0.6868	0.68515	0.695
1.71	0.63515	0.63475	0.6671	0.67105	0.645
1.81	0.6162	0.61195	0.63905	0.6357	0.595
1.91	0.60225	0.60215	0.62205	0.6111	0.545

C) Impact Of Server Number On Self-Similarity

Next, we examine how the number of Pareto servers affects the relation between Pareto shaping parameter α and observed self-similar parameter $\hat{H}$. The results are summarized in Fig. 4.4.

All these figures have the same curve trend. In other words, the system begins with non-self-similarity at small α, and reaches its maximum self-similarity at certain α, and then returns to non-self-similarity for further increasing α.

Comparing Figs. 4.3 and 4.4 with the analytical equation $H = (3-\alpha)/2$ obtained for infinite servers, we note that only when α is moderately large can the departure self-similar behavior be close to departure behavior under infinite servers.

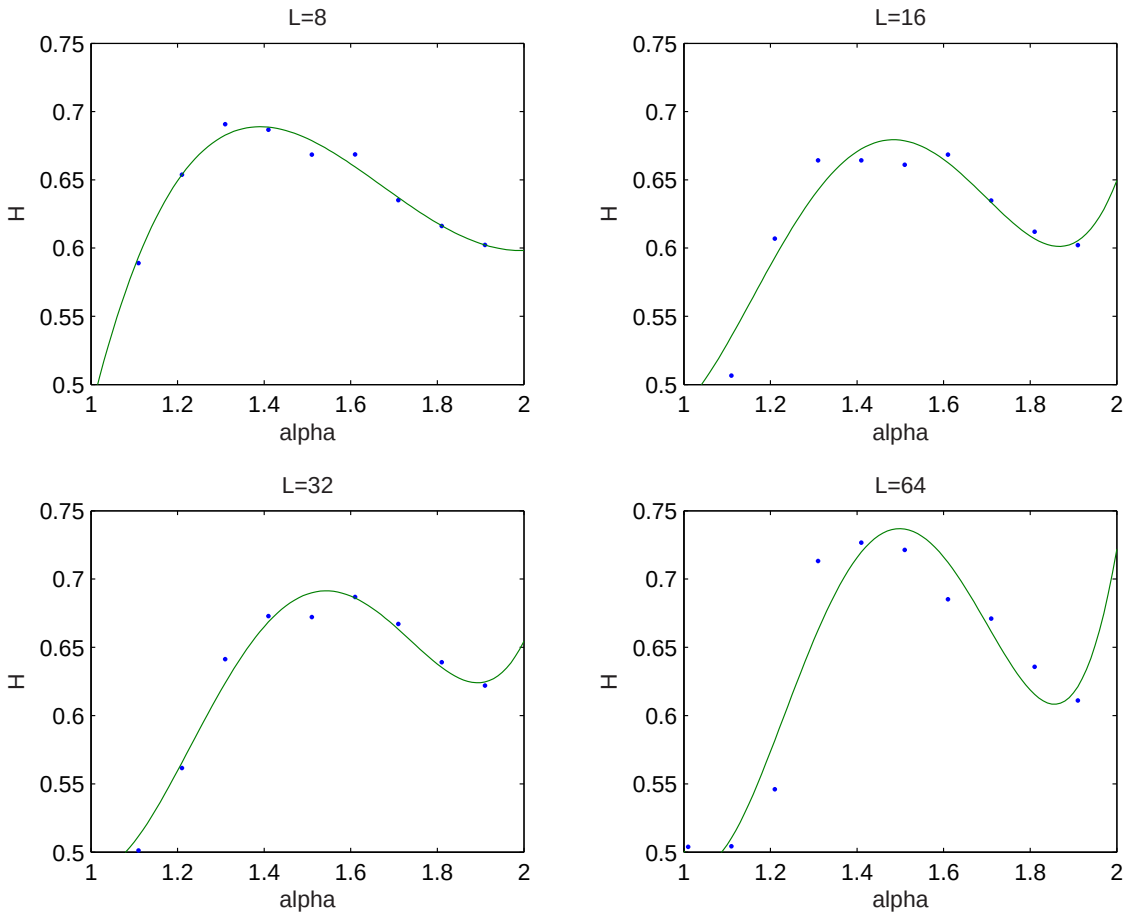


Figure 4.4: *Observed Hurst parameter versus* α *under* $\rho = \lambda/(L\mu) = 0.975$. *Fourth-order polynomial approximation is also provided. The variance-time plots corresponding to each server number are depicted in Figs. 4.5, 4.6, 4.7 and 4.8, respectively.*

D) Interpretation Of Non-Self-Similarity At Small α

In all the above simulations, the Pareto mean service time is chosen according to the required utilization, and remains fixed while the Pareto shaping parameter α varies. This is done by properly adjusting k in the Pareto mean $k\alpha/(\alpha-1)$ (cf. (2.2)).

An aftereffect of this approach is that when α is close to 1, k must be very small, and deviates largely from the mean service time. This indicates that there will be a certain number of "very short" packets (or equivalently, a certain number of packets requiring "very short" service time).

Recall that in Section 4.1.1-A), we mention that if packet A enters the queue after packet B is served, the departure behaviors of packet A and packet B are statistically independent because the queue is the only place to introduce dependence in the system. We then proceed to state that the maximum queue length that has ever been achieved during simulation is around 2000 (packets) for a moderate α lying between 1.1 and 1.33; hence, it is anticipated that for $m \geq (1/\lambda) \times 2000 = 2 \times 10^5$, the long-term dependence of the departure process should disappear. Now, as a certain number of packets requires much less service time at small α close to 1, the long-term dependence of the departure process should disappear at smaller average window m. This explains why in our simulation (system setting), the degree of self-similarity decreases, rather than increases as expected by the analytical result under infinite number of servers, when α reduces to 1.

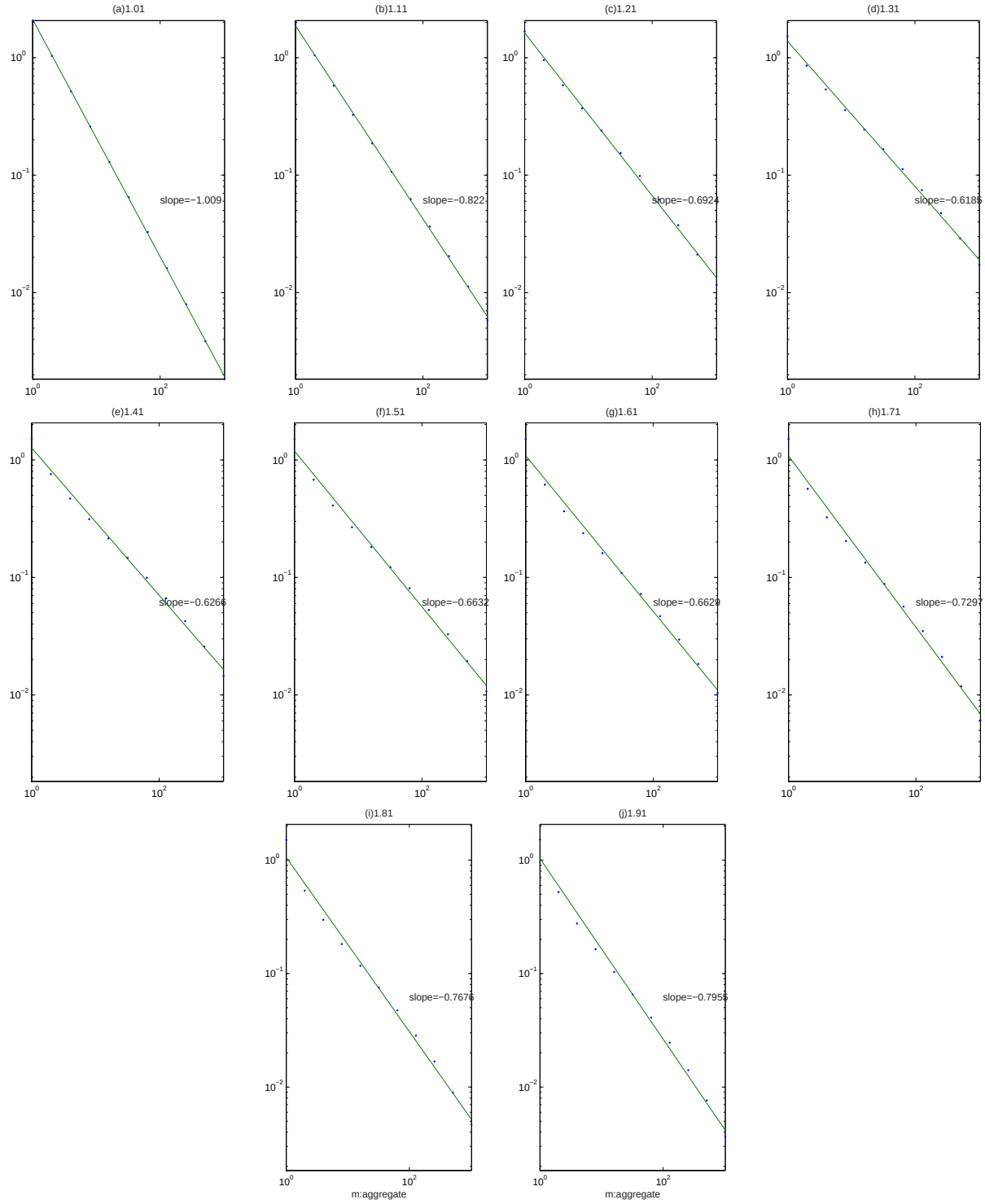


Figure 4.5: *The variance-time plot for 8 parallel Pareto servers with* $\rho = \lambda/(L\mu) = 0.975$. *The number aside (a)–(j) is the shaping parameter* α *of the Pareto servers.*

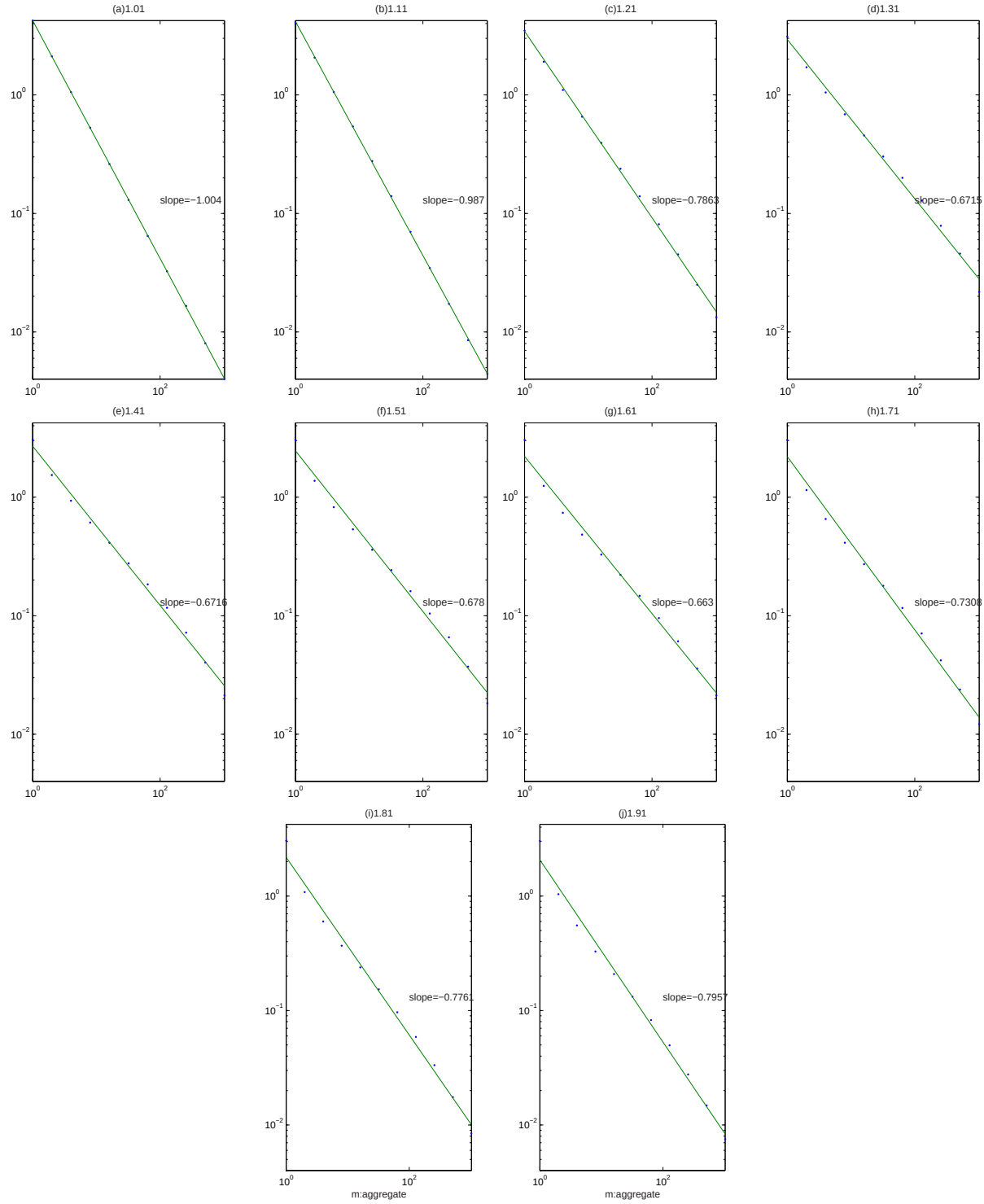


Figure 4.6: *The variance-time plot for 16 parallel Pareto servers with* $\rho = \lambda/(L\mu) = 0.975$. *The number aside (a)–(j) is the shaping parameter* α *of the Pareto servers.*

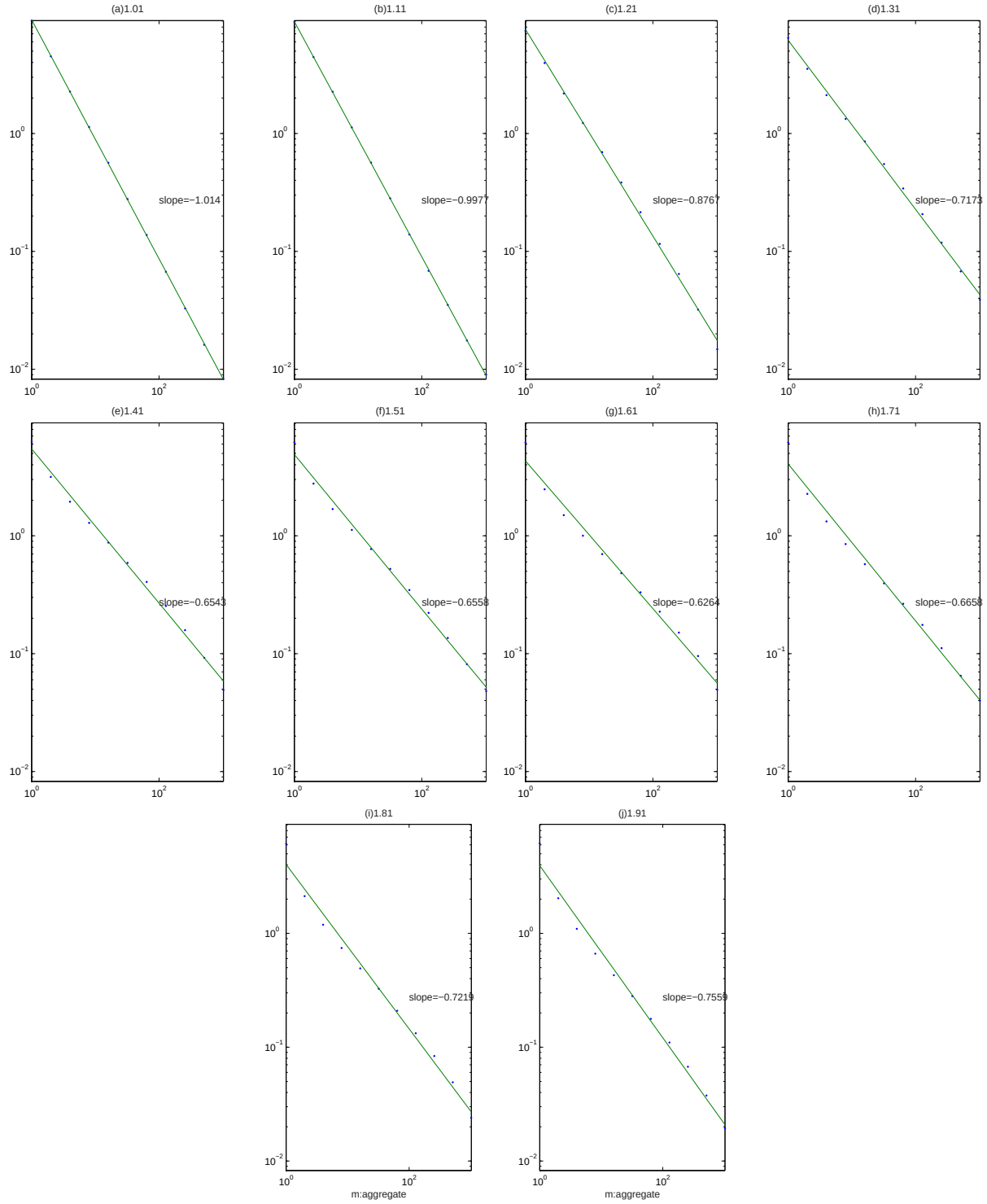


Figure 4.7: *The variance-time plot for 32 parallel Pareto servers with* $\rho = \lambda/(L\mu) = 0.975$. *The number aside (a)–(j) is the shaping parameter* α *of the Pareto servers.*

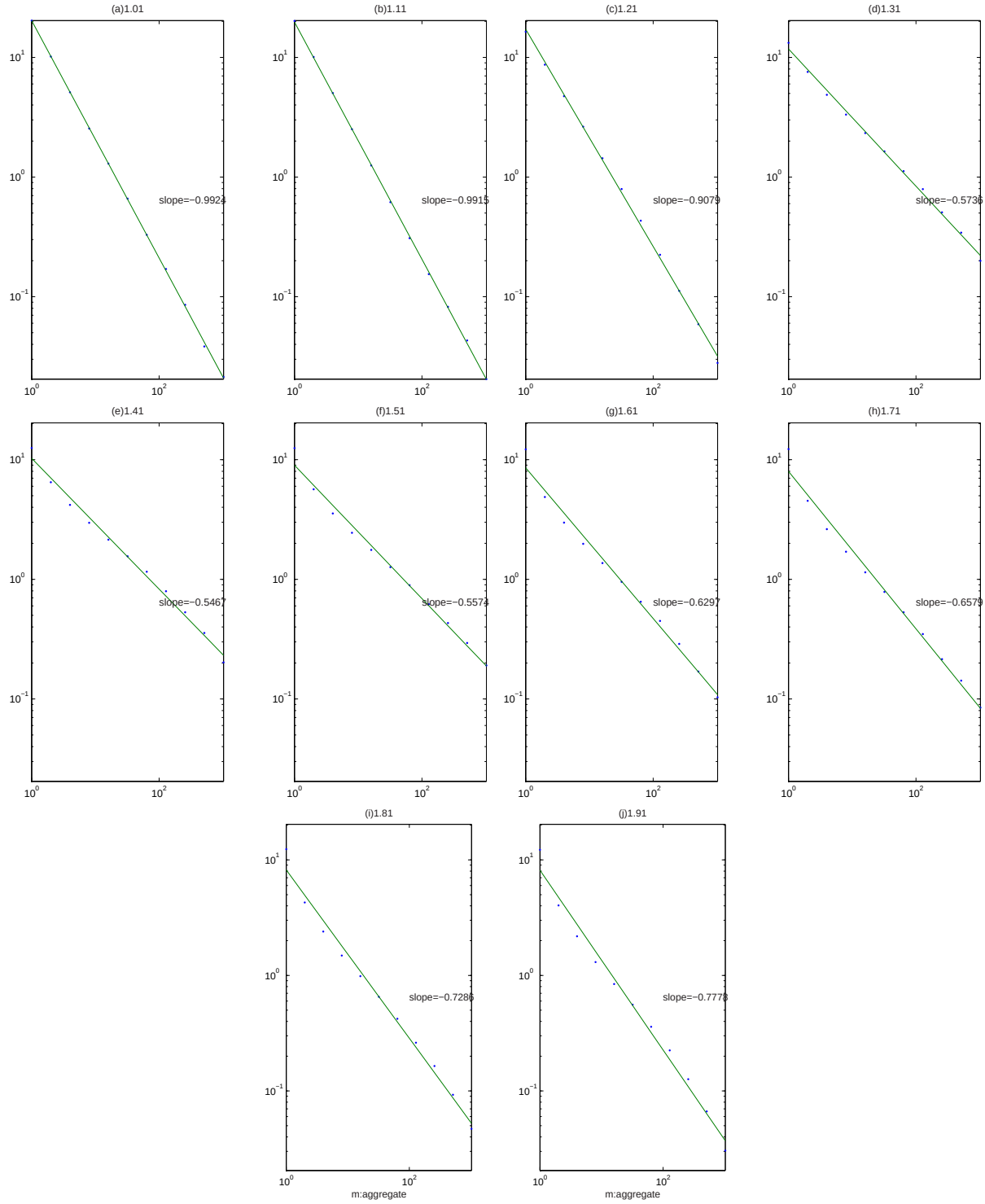


Figure 4.8: *The variance-time plot for 64 parallel Pareto servers with* $\rho = \lambda/(L\mu) = 0.975$. *The number aside (a)–(j) is the shaping parameter* α *of the Pareto servers.*

4.1.2 System With Varying Service Mean Time

In the simulations illustrated in the previous section, a smaller α close to unity gives an $\hat{H}$ close to 0.5, which is contrary to the theoretical result obtained under infinite servers. As explained in Section 4.1.1-D), this is because we reduce k as α decreases in order to have a fixed service mean time (or utilization). Based on this explanation, one way to obtain the theoretically anticipated trend at small α is to fix the parameter k in the Pareto distribution and the ratio of λ/L, and relax the restriction of constant utilization $\rho = \lambda/[L(\alpha-1)/(k\alpha)] = k \times (\lambda/L) \times [\alpha/(\alpha-1)]$.

Notably, taking fixed k and λ/L may result in a greater-than-unity utilization at small α, which in principle should ultimately yield a simulation-non-attainable infinite queue length. In addition, simulations can only be performed for a finite duration, and hence, the theoretical assumption of considering the system time approaching infinity in [Will97] can never be satisfied. As a result, the system may still be in a non-stationary state at the end of our simulation. Nevertheless, we can still observe the trend through adequately long simulations, and examine how much deviation of our simulation (performed within a finite duration under a finite-resource system setting) to the theoretical result (obtained under an infinite-resource assumption).

It can be seen from Figs. 4.9–4.15 that the anticipated trend that the system is more self-similar at small α (specifically, $\alpha = 1.01$) is observed, although the system now quickly behaves non-self-similar by fairly small α (e.g., $\alpha = 1.11$ for $L \geq 16$).

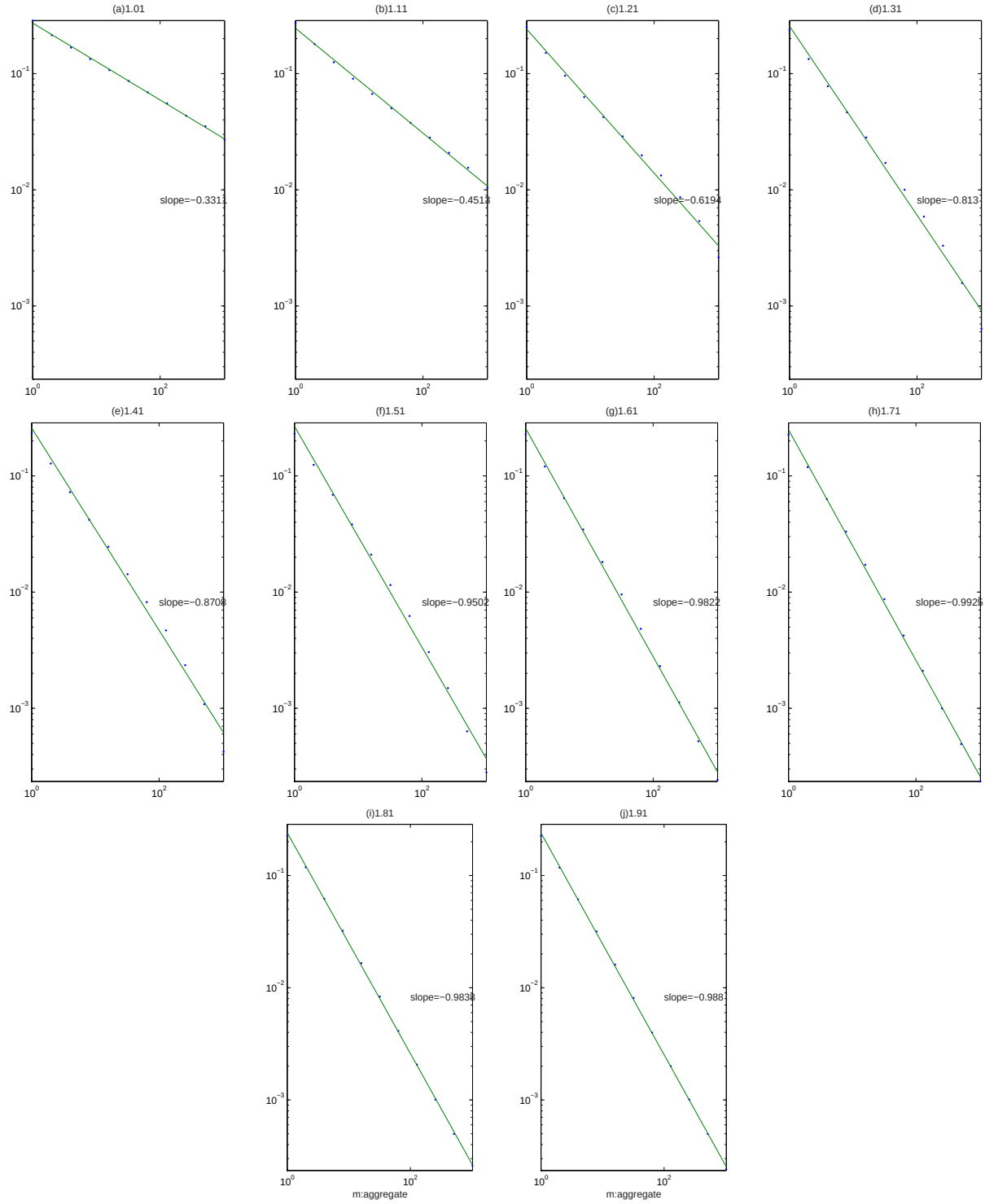


Figure 4.9: *The variance-time plot for single server with fixed k. The number aside (a)–(j) is the shaping parameter α of the Pareto servers. In these simulations, $k = 40$ and $\lambda/L = 0.0025$.*

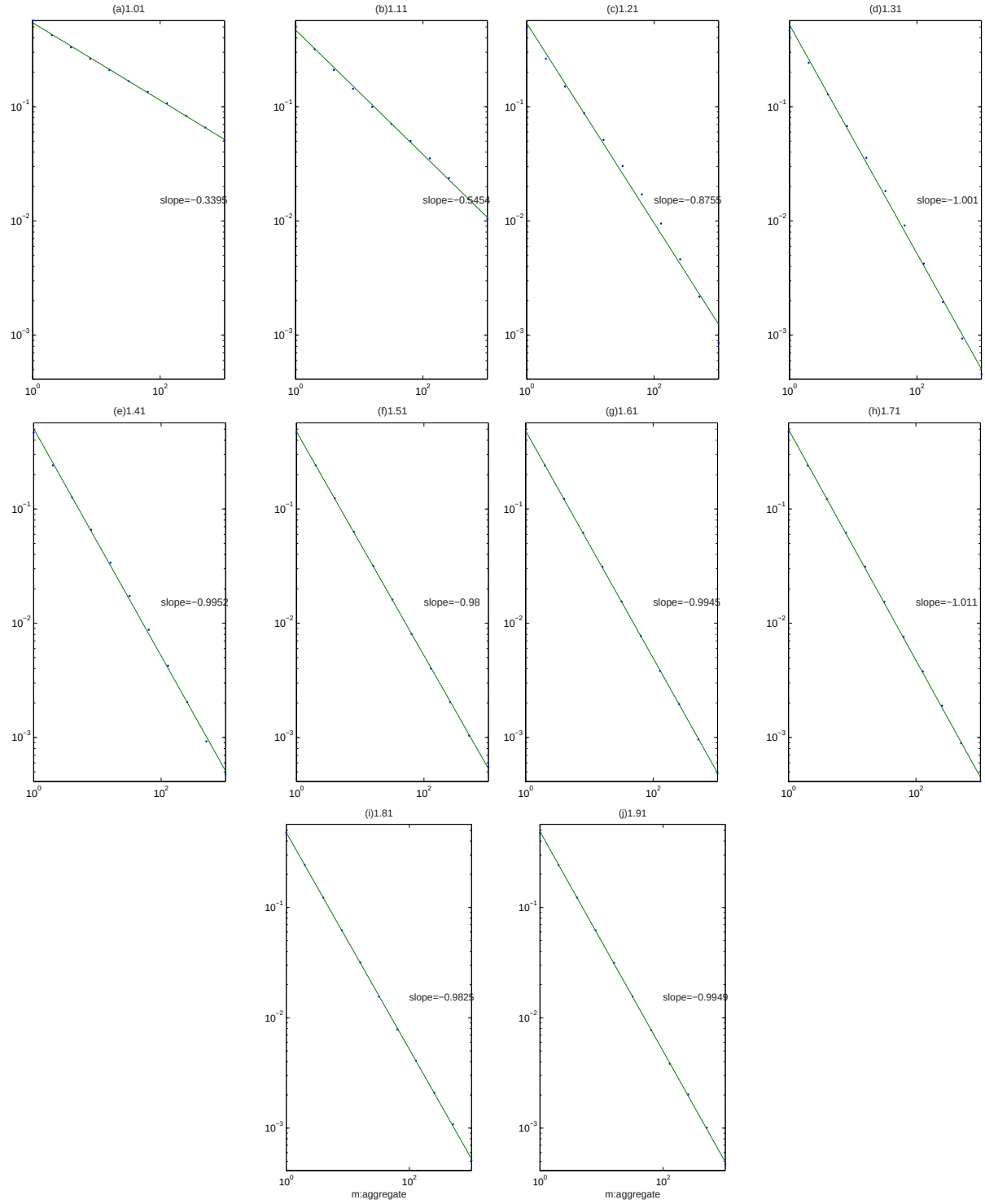


Figure 4.10: *The variance-time plot for 2 parallel servers with fixed k. The number aside (a)–(j) is the shaping parameter α of the Pareto servers. In these simulations, $k = 40$ and $\lambda/L = 0.0025$.*

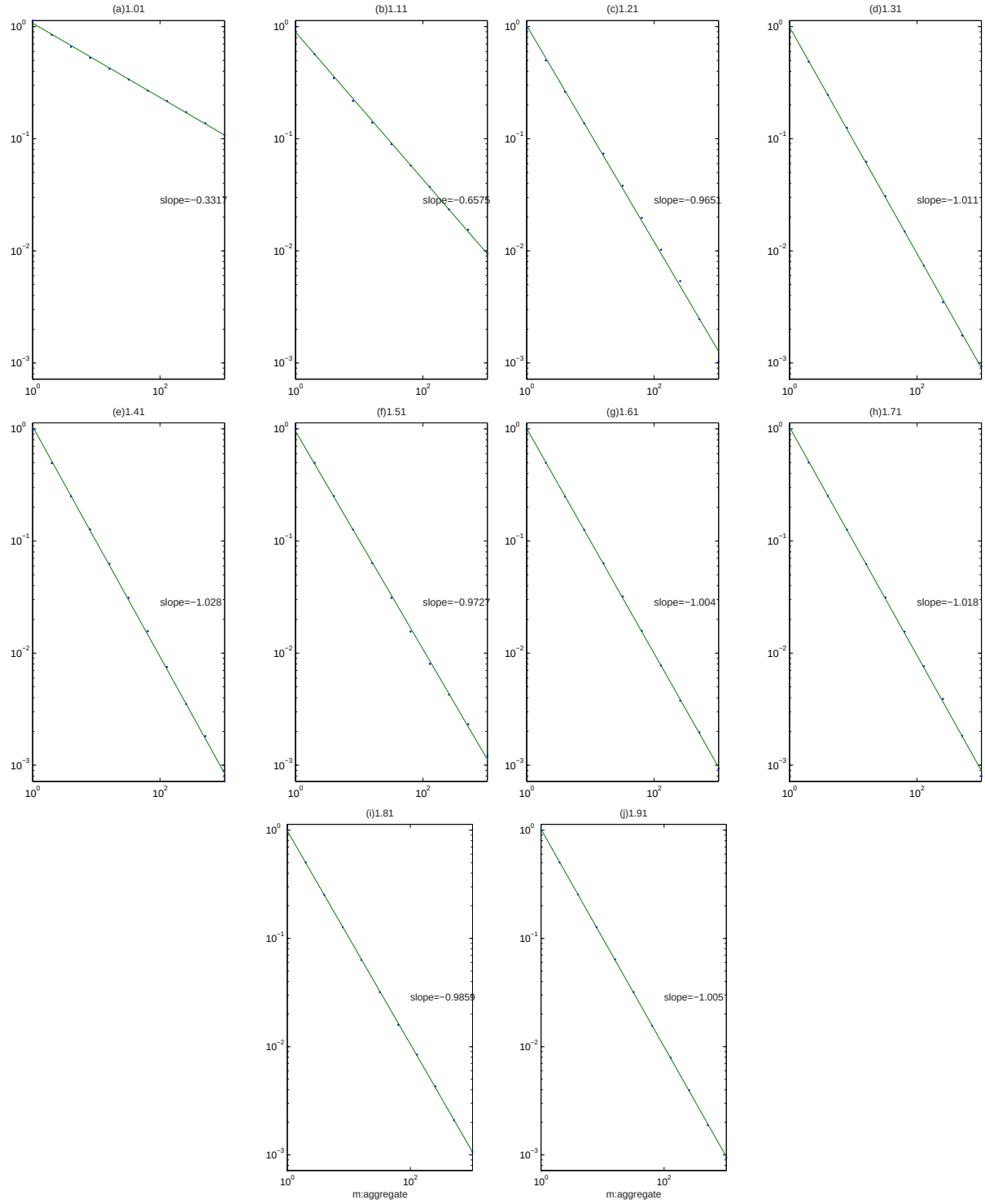


Figure 4.11: *The variance-time plot for 4 parallel servers with fixed k. The number aside (a)–(j) is the shaping parameter α of the Pareto servers. In these simulations, $k = 40$ and $\lambda/L = 0.0025$.*

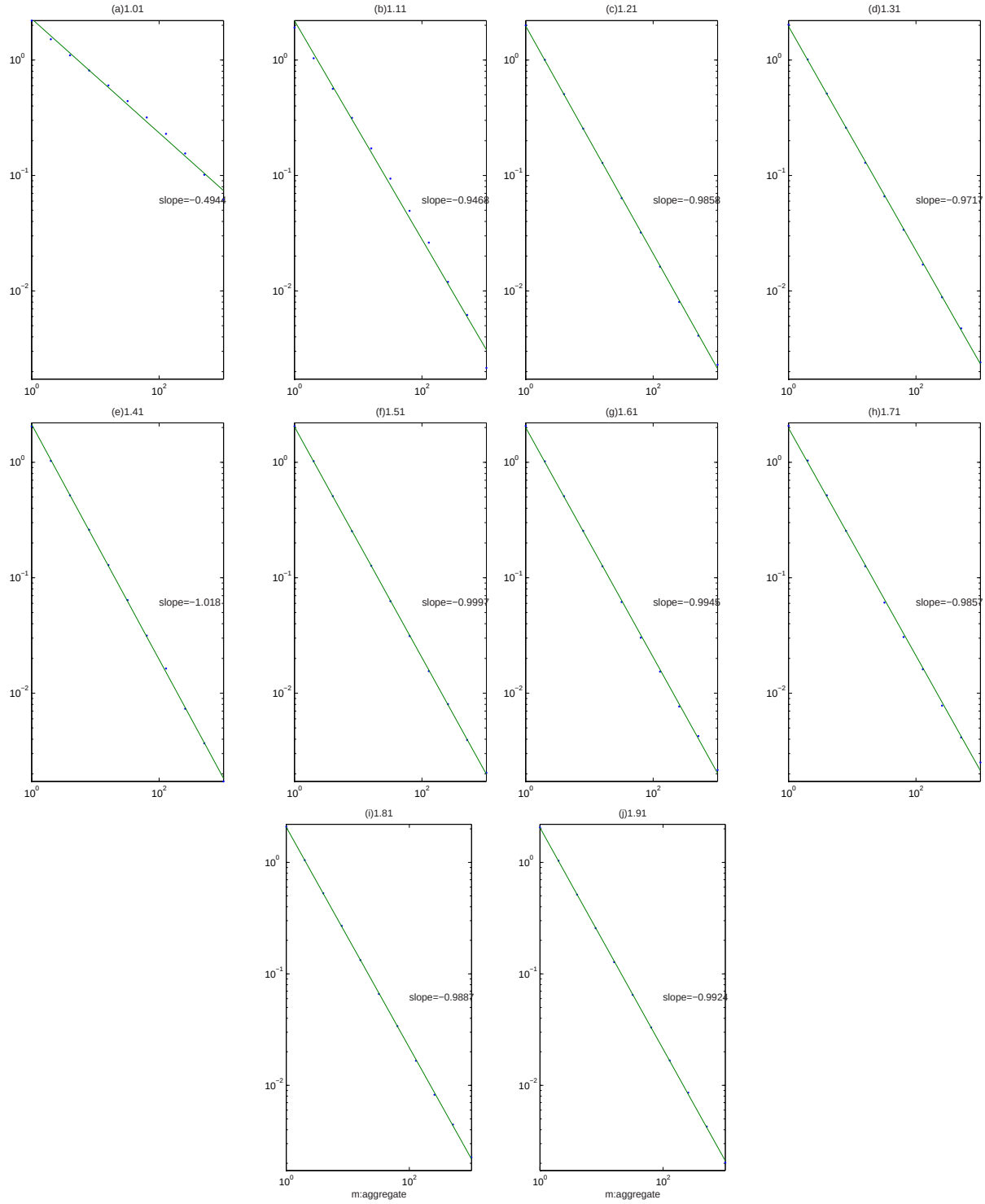


Figure 4.12: *The variance-time plot for 8 parallel servers with fixed k. The number aside (a)–(j) is the shaping parameter α of the Pareto servers. In these simulations, $k = 40$ and $\lambda/L = 0.0025$.*

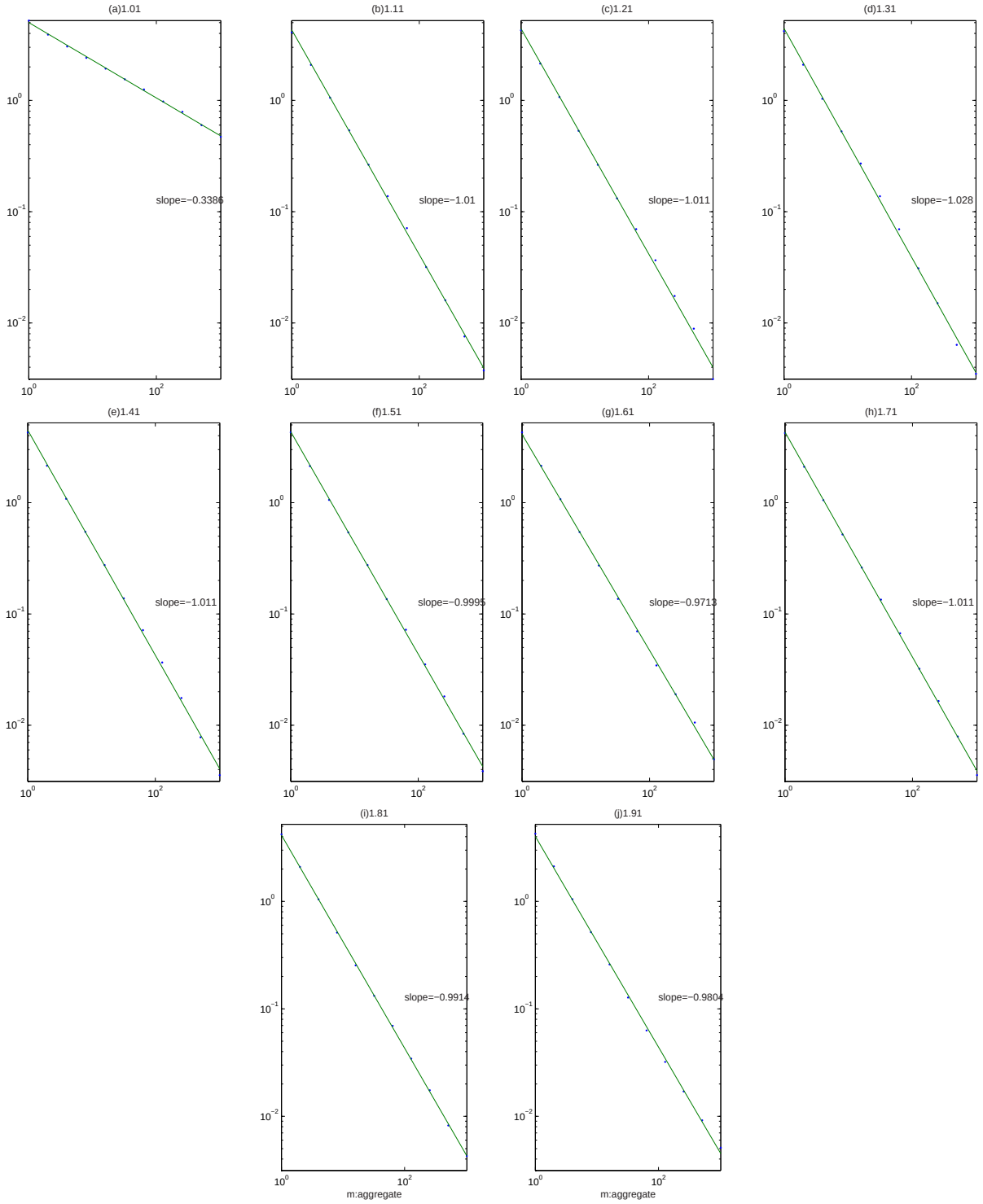


Figure 4.13: *The variance-time plot for 16 parallel servers with fixed k. The number aside (a)–(j) is the shaping parameter α of the Pareto servers. In these simulations, $k = 40$ and $\lambda/L = 0.0025$.*

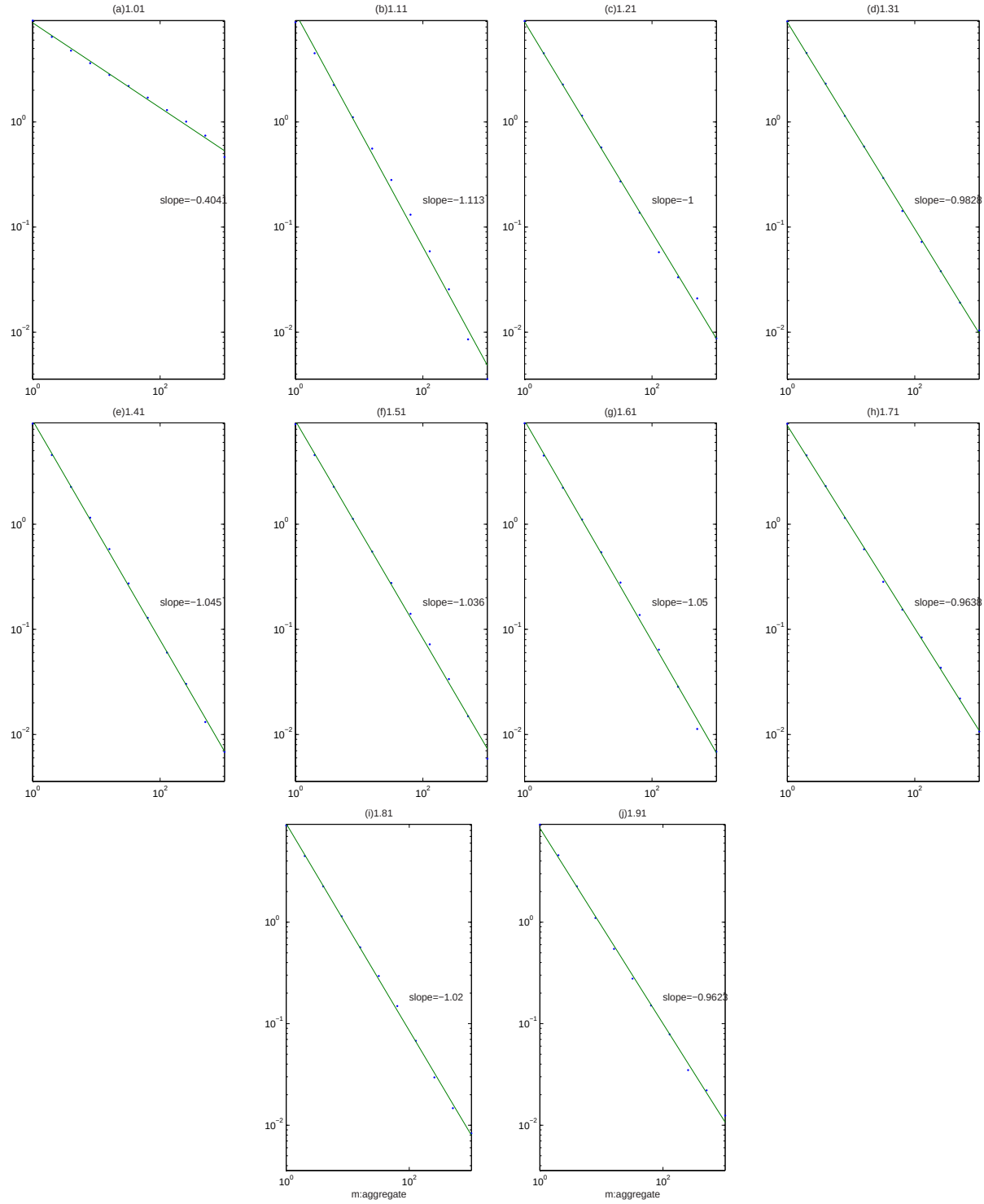


Figure 4.14: *The variance-time plot for 32 parallel servers with fixed k. The number aside (a)–(j) is the shaping parameter α of the Pareto servers. In these simulations, $k = 40$ and $\lambda/L = 0.0025$.*

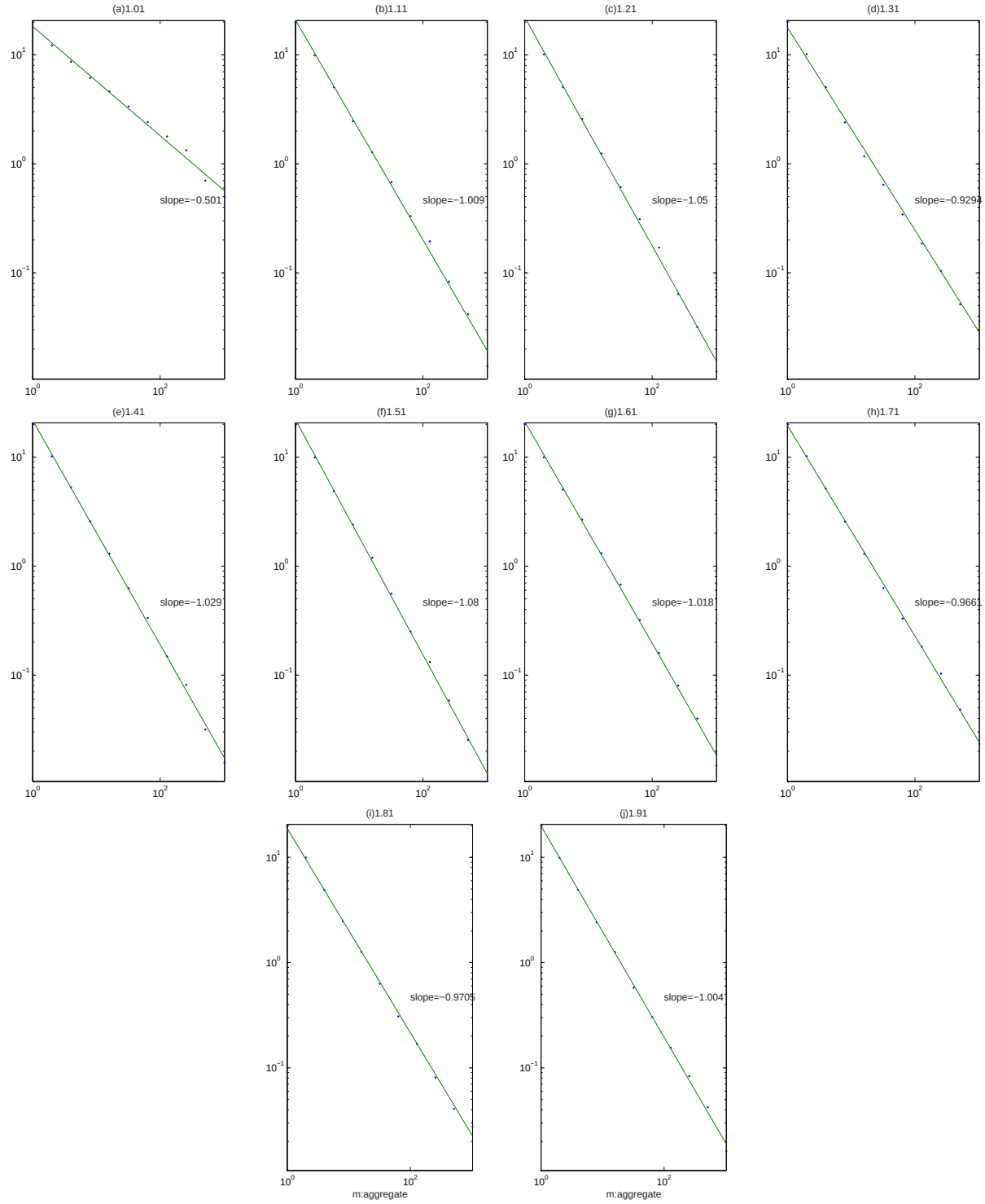


Figure 4.15: *The variance-time plot for 64 parallel servers with fixed k. The number aside (a)–(j) is the shaping parameter α of the Pareto servers. In these simulations, $k = 40$ and $\lambda/L = 0.0025$.*

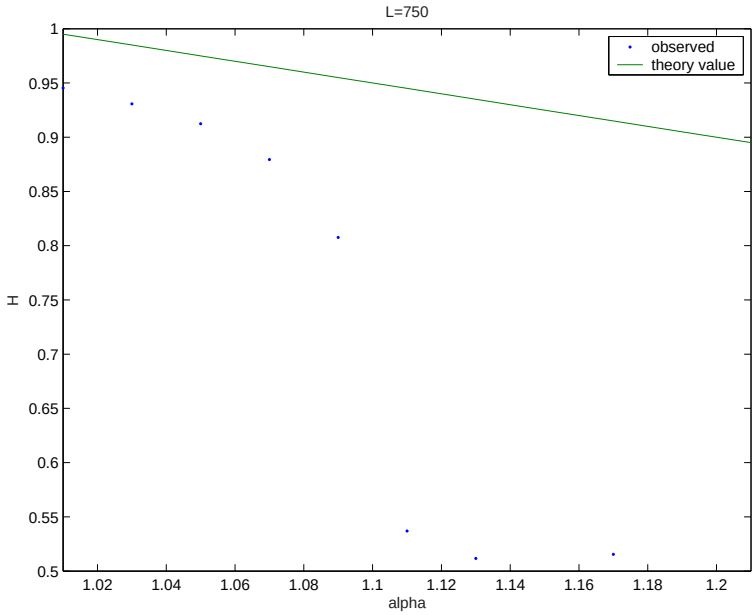


Figure 4.16: *The H-versus-α plot for $L = 750$, and the curve of $H = (3 - \alpha)/2$. In these simulations, $k = 100$ and $\lambda/L = 0.0001$.*

By further increasing L beyond 64 as shown in Figs. 4.16–4.21, we observe that the relation between the observed $\hat{H}$ and α is getting closer to the theoretical curve $H = (3-\alpha)/2$. We can even observe a more self-similar behavior at certain α than that suggests by the theoretical formula at very large L.

Again, in Figs. 4.16–4.21, the points obtained from simulations can be divided into two groups, which respectively belong to two lines with different slopes. The first group forms a line with a slope characterizing as self-similarity, while the second group belongs to a line that has a "much-less-self-similar" slope. We found that the "intersection" (or the turning point) of the two lines grows with L (cf. Fig. 4.22); hence, we may expect that as L approaches infinity, the second group will disappear, and somehow confirm that by aggregating infinite number of independent sources with heavy-tailed duration (service-time), a self-similar departure process can be observed.

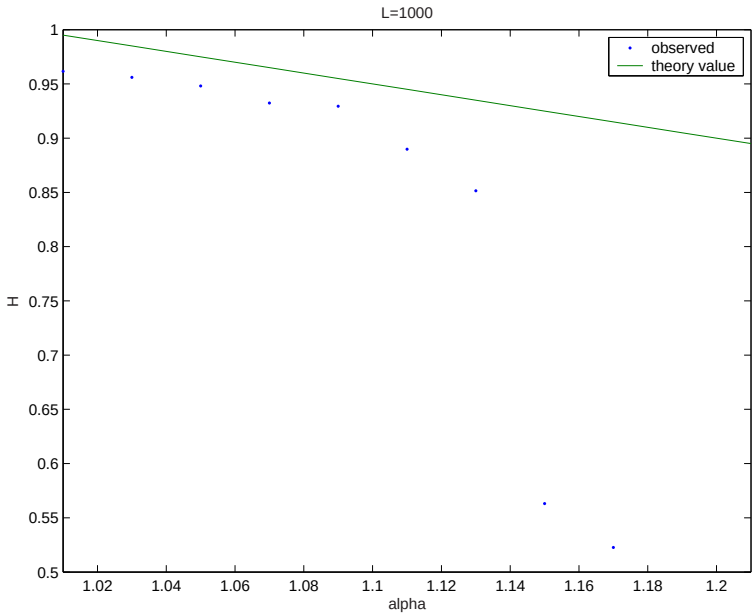


Figure 4.17: *The H-versus-α plot for $L = 1000$, and the curve of $H = (3 - \alpha)/2$. In these simulations, $k = 100$ and $\lambda/L = 0.0001$.*

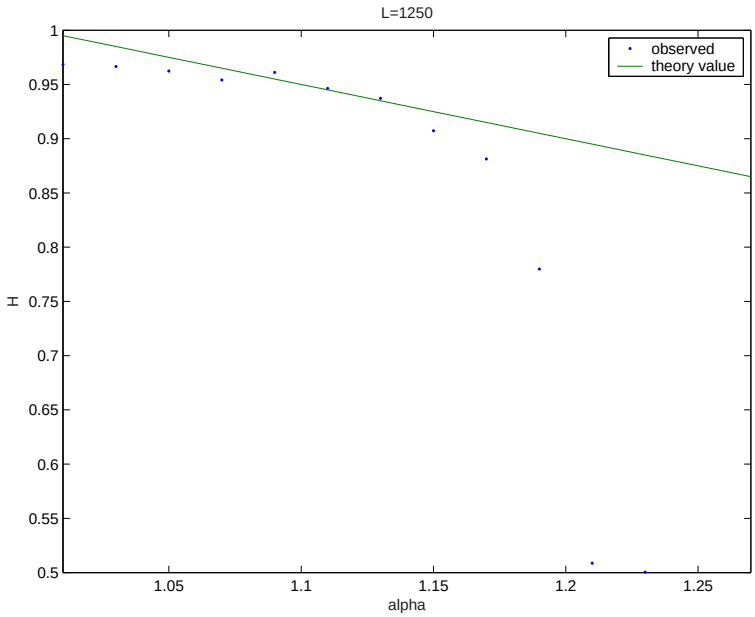


Figure 4.18: *The H-versus-α plot for $L = 1250$, and the curve of $H = (3 - \alpha)/2$. In these simulations, $k = 100$ and $\lambda/L = 0.0001$.*

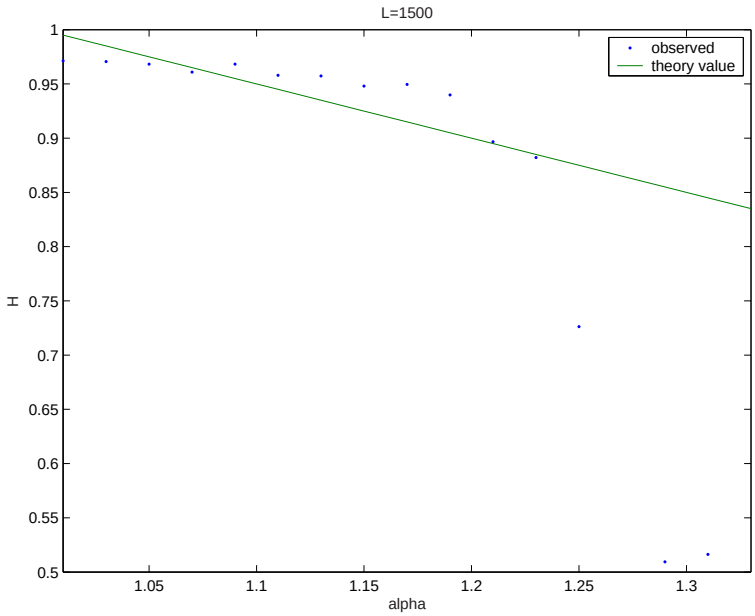


Figure 4.19: *The H-versus-α plot for $L = 1500$, and the curve of $H = (3 - \alpha)/2$. In these simulations, $k = 100$ and $\lambda/L = 0.0001$.*

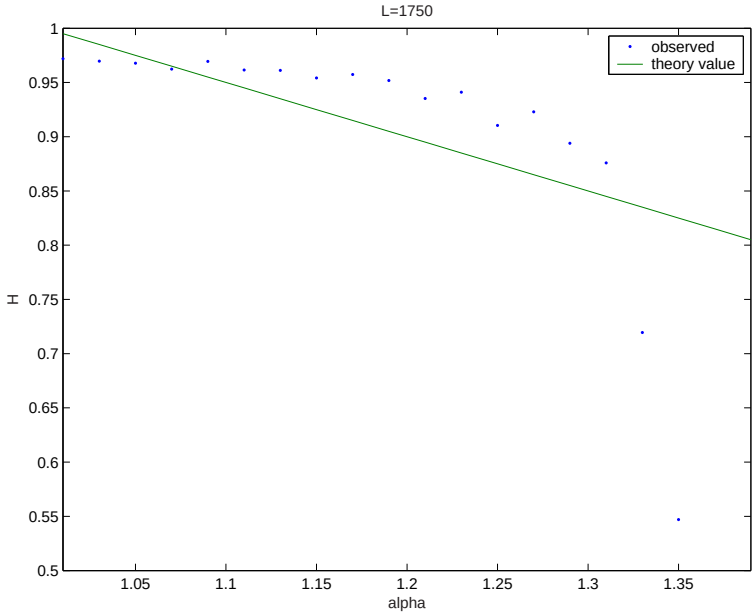


Figure 4.20: *The H-versus-α plot for $L = 1750$, and the curve of $H = (3 - \alpha)/2$. In these simulations, $k = 100$ and $\lambda/L = 0.0001$.*

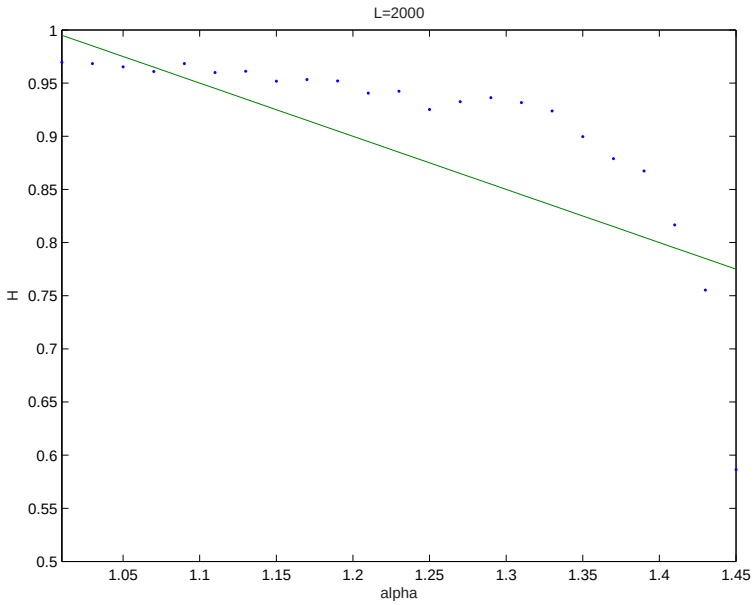


Figure 4.21: *The H-versus-α plot for $L = 2000$, and the curve of $H = (3 - \alpha)/2$. In these simulations, $k = 100$ and $\lambda/L = 0.0001$.*

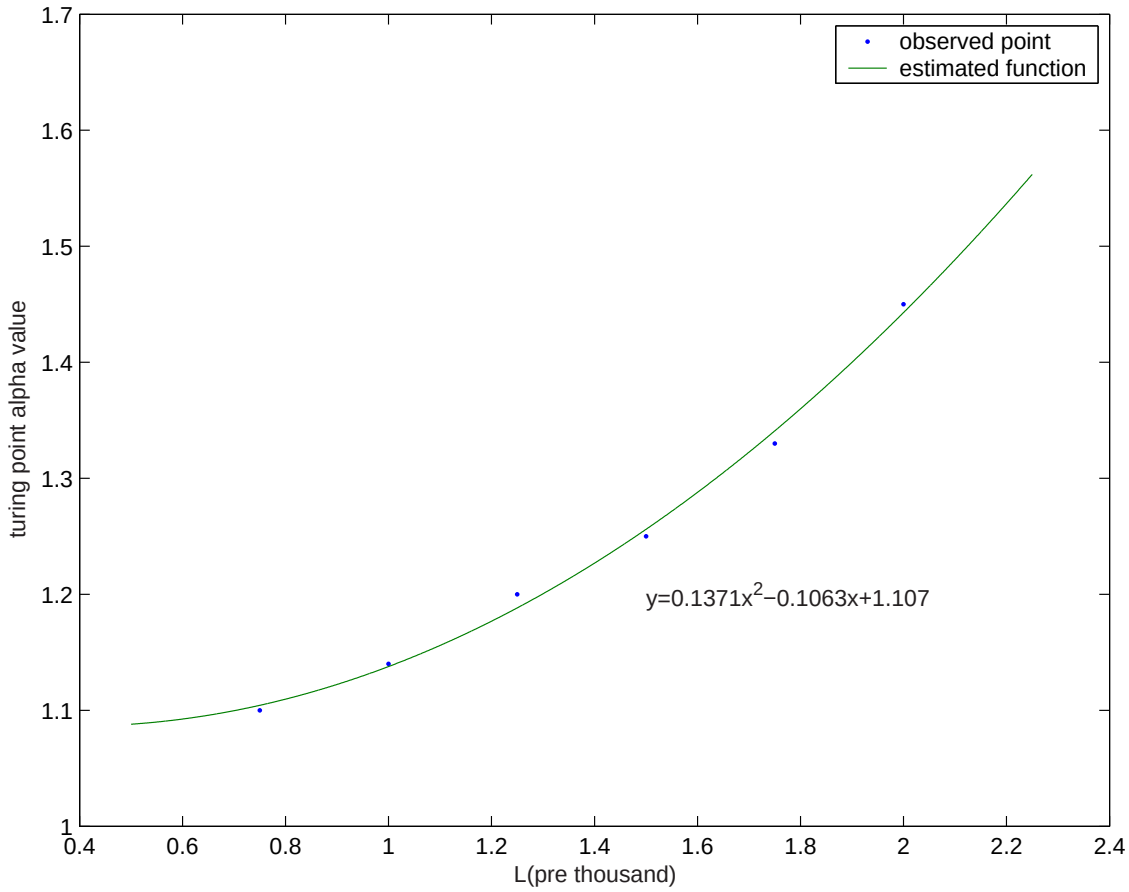


Figure 4.22: *The turning points of α after which the degree of self-similarity drops.*

4.2 Tandem Server Systems

As what we have done in parallel server system, we will perform the variance-time analysis on the departure process that is collected at the last output of serially connected servers. Similar to that used in parallel server system, the original input is Poisson distributed, and each server has Pareto-distributed service time. A seemingly difference between the serial and parallel systems is that only a single queue is presumed in the parallel server system, while each server has its own queue in the serial server system. This difference makes the dependence between packets behaves differently in parallel and serial server systems.

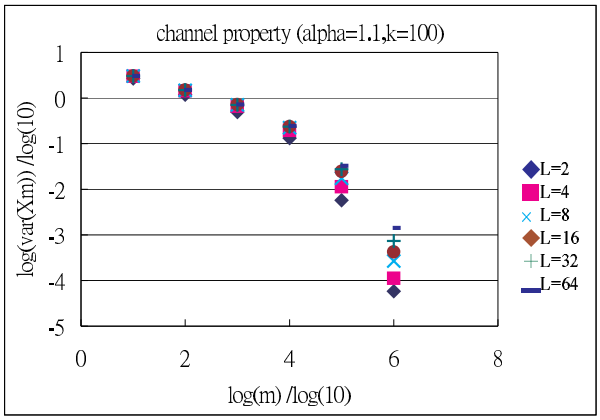
Recall that the queue is the only place to introduce dependence in our system. So, if one of the tandem queues has length zero, it introduces no dependence for consecutive packets. This leads to the following analysis.

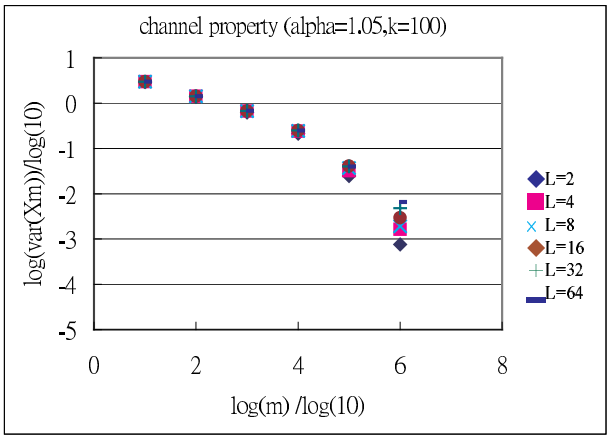
Let the queue lengths of tandem server system at time instant n be respectively denoted by Q_n^1, Q_n^2, ..., and Q_n^L. Denote by j_1, j_2, ... the queues that has length zero, namely, $Q_n^{j_1} = Q_n^{j_2} = \cdots = 0$. Then if there are $Q_{\max}$ packets in-between packet A and packet B, then packet A behaves independently of packet B, where

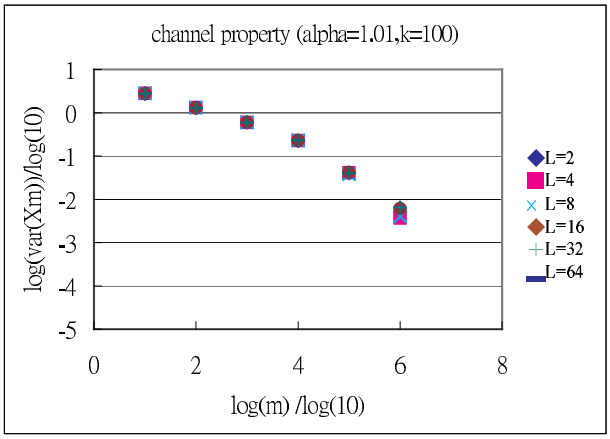
$$Q_{\max} = \max\left\{\sum_{j=0}^{j_1-1} Q_n^i, \sum_{j=j_1}^{j_2-1} Q_n^i, \sum_{j=j_2}^{j_3-1} Q_n^i, \ldots\right\}. \tag{4.1}$$

4.2.1 Variance-Time Analysis

¿From Fig. 4.23, the variance-time analysis of the departures of the tandem server system does not appear to be straight line, nor to be a combination of two lines of different slopes as what obtained in parallel server system. The degree of dependence gradually decreases with the observation range.







(Continue on the next page.)

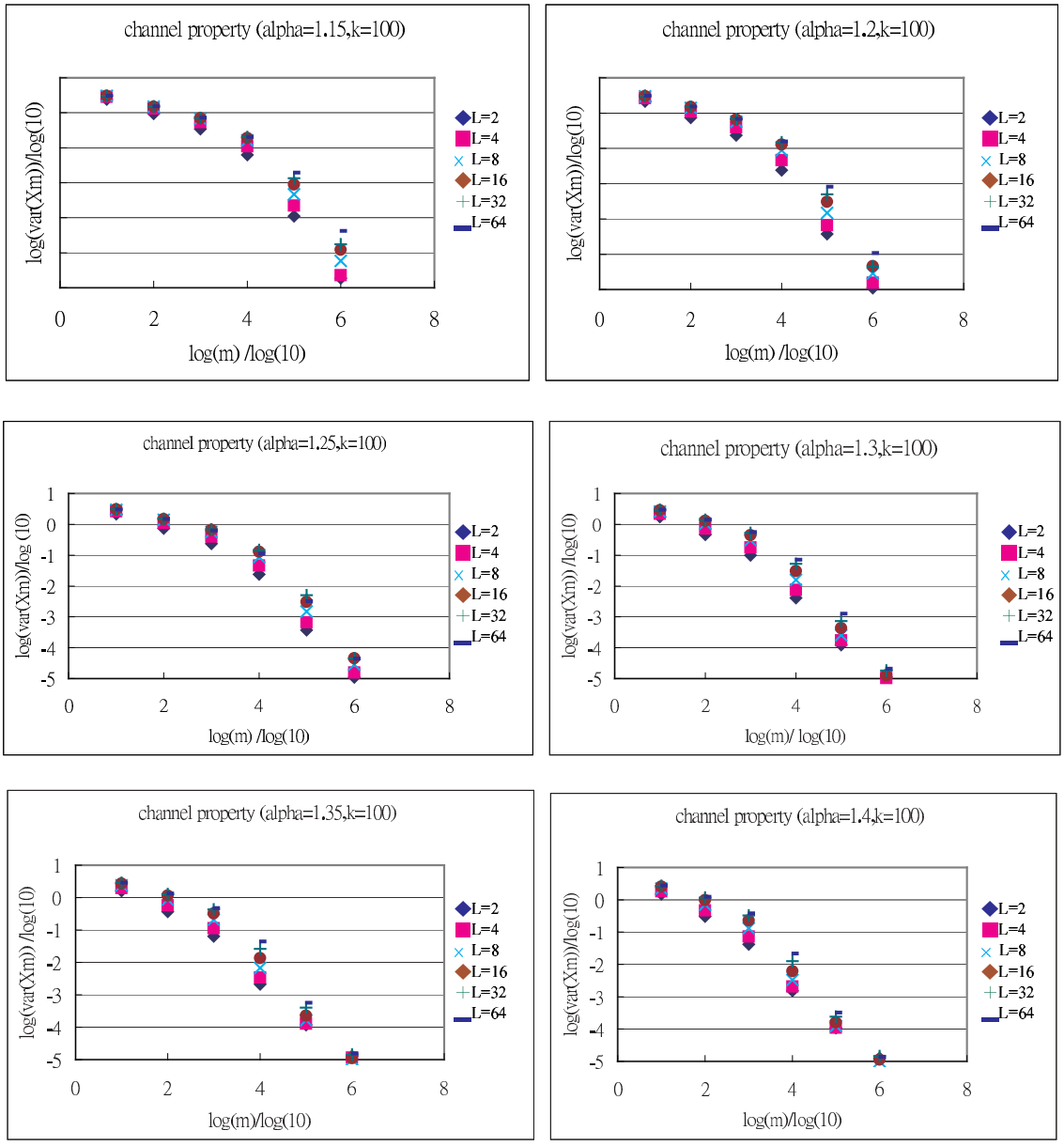


Figure 4.23: *The variance-time plots of the departures for tandem L-server system. The originated Poisson arrival has mean* $\lambda = 0.001$. *The Pareto parameters,* α *and* k, *used for each Pareto server are indicated in each subfigure.*

Another observation that can be made from Fig. 4.23 is that the variance-time curves diverge at large average window m for different number of tandem servers, when α is small (or equivalently, when utilization $\rho = \lambda k\alpha/(\alpha - 1) = 0.001 \times 100\alpha/(\alpha - 1) = 0.1\alpha/(\alpha - 1)$ grows large). Specifically, for $\alpha = 1.01$, 1.05 and 1.1, the variance of m-average departure process scatters more at $m = 10^6$ than at $m = 10$ and $m = 10^2$.

A special note should be taken at the case of $\alpha = 1.01$, in which all the variance-time plots for $L = 2$—64 are in great agreement to each other. An interpretation of this result is that when α is much less than $10/9 \approx 1.1111$, the utilization is far greater than unity. It can be expected that with utilization much greater than 1 (such as $\rho = 10.1$ for $\alpha = 1.01$), the first tandem queue grows very fast, and its length will ultimately approaches infinity. As a consequence, $Q_{\max}$ will be dominated by the length of the first tandem queue, which explains why the variance-time plots of different L coincides to each other in situation where α is very close to 1.

For smaller utilization less than 1 (namely, $\alpha > 1.1111$), Fig. 4.23 shows two kinds of behaviors for different regions of m—for some m, the variances of the m-average departure process diverge for different L, while for some other m, they actually converges. However, it is hard to tell at which m the curves corresponding to different L disperses, and at which m the curves gather.

A final remark from Fig. 4.23 is the serial connected server system shows a certain degree of self-similarity even if the system utilization is below 0.5 (or equivalently, $\alpha \geq 1.25$). This is contrary to the situation of parallel server system or the case of $L = 1$ for serially connected server, where utilization of 0.5 shows no self-similarity in departure process. This phenomenon can be explained as that in serially connected server system, the packet dependence can be enhanced by adding a queue of non-zero length. Hence, the tandem system can easily create very long-range dependence if the serially adjacent queues are all non-empty.

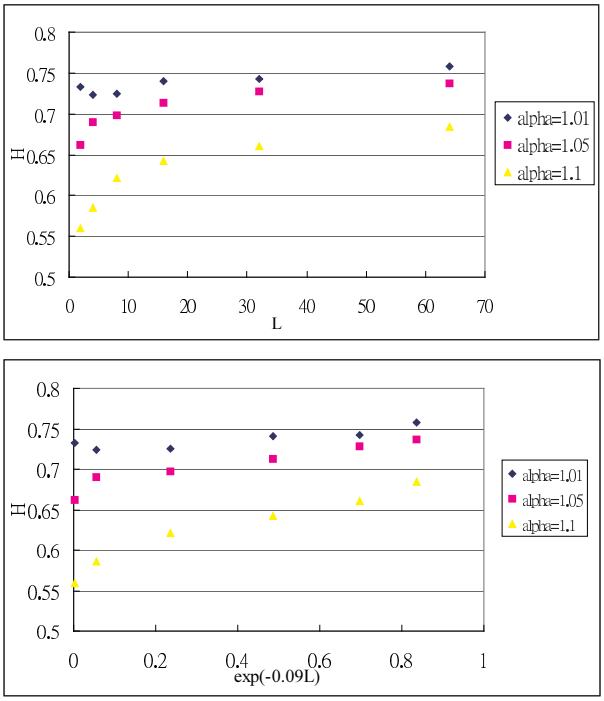


Figure 4.24: *The observed Hurst parameter $\hat{H}$ obtained by finding the best-fit lines to Fig. 4.23. All six points corresponding to $m = 10, 10^1, 10^2, 10^3, 10^4, 10^5, 10^6$ are used in the determination of the slope of the best-fit line. The upper subfigure gives a linear scale for L, and the lower subfigure presents with a logarithmic scale for L.*

A) Relation between H, L and α

Next, we depict the estimated Hurst parameter with respect to different L in Fig. 4.24. We then found an interesting result. That is, the observed Hurst parameter in serially connected server system may be an exponential function of L. Specifically, the lower subfigure in Fig. 4.24 suggests that

$$H \approx f_1(\alpha)e^{-a(\alpha)\cdot L} + f_2(\alpha), \tag{4.2}$$

where $f_1(\cdot)$, $f_2(\cdot)$ and $a(\cdot) > 0$ are some positive functions of α.

In term of different presentation of our simulation results in Fig. 4.25, we found that H can be approximated by a linear function of α, namely,

$$H \approx c - g(L)(\alpha - 1), \tag{4.3}$$

where $g(\cdot)$ is some function of L, and c is a constant. We conjecture from Eq. (4.2) and (4.3) that the self-similar parameter can be approximated by:

$$H \approx c - (\alpha - 1)(c_1 - e^{-c_2 L}),$$

hence, as $L \to \infty$, H approaches $c - c_1(\alpha - 1)$. Notably, the theoretical analysis under the assumption of aggregating infinite number of ON/OFF sources gives that $c = 1$ and $c_1 = 1/2$.

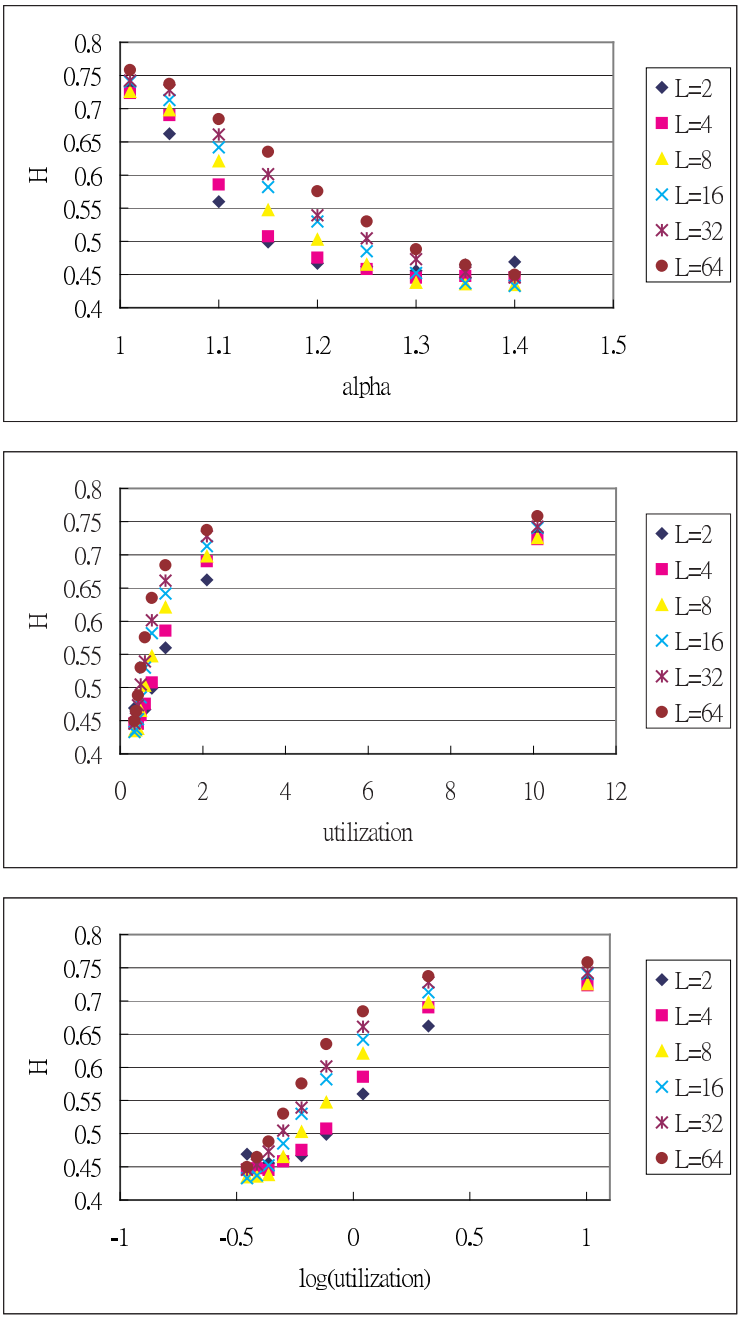


Figure 4.25: *The Hurst parameter versus α and utilization.*

4.2.2 Complement Cumulative Distribution Function Analysis for Different System Parameters

In this subsection, we analyze the complementary cumulative distribution function (CCDF) of the departure processes, and examine its degree of heavy-tailibility.

A) Poisson Arrival and Pareto Server

The CCDF in Fig. 4.26 corresponds to the system with Poisson arrival and Tandem Pareto servers. First, we note from the three subfigures in the left-column of Fig. 4.26 that a larger interdeparture time can be observed for large L. Secondly, the slope of the CCDFs in log-log scale, which should equal the negation of Pareto shaping parameter α, is getting smaller (namely, decaying faster) for large L. These indicate that the interdeparture process does have a heavy tail.

B) Pareto Inter-Arrival and Tandem Exponential Servers

Figure 4.27 also presents a tendency of larger interdeparture time for larger L. In other words, the mean of interdeparture time grows with L. As anticipated, the tail probability is much less heavy in nature (cf. The right column in Fig. 4.27), if it is compared with those obtained in Fig. 4.26. It needs to be pointed out that it is reasonable to expect that the interdeparture behaves the same as the interarrival, if all tandem queues are empty. Therefore, since the interdeparture becomes more like a non-heavy-tailed exponential distribution, we can expect that at least some of the queues are non-empty during the simulation.

C) Summary of A) and B)

In summary, we know from Eq. (3.2) that the inter-departure will trend to be the larger one between the interarrival and the service time. This does reflect to the CCDF we observed.

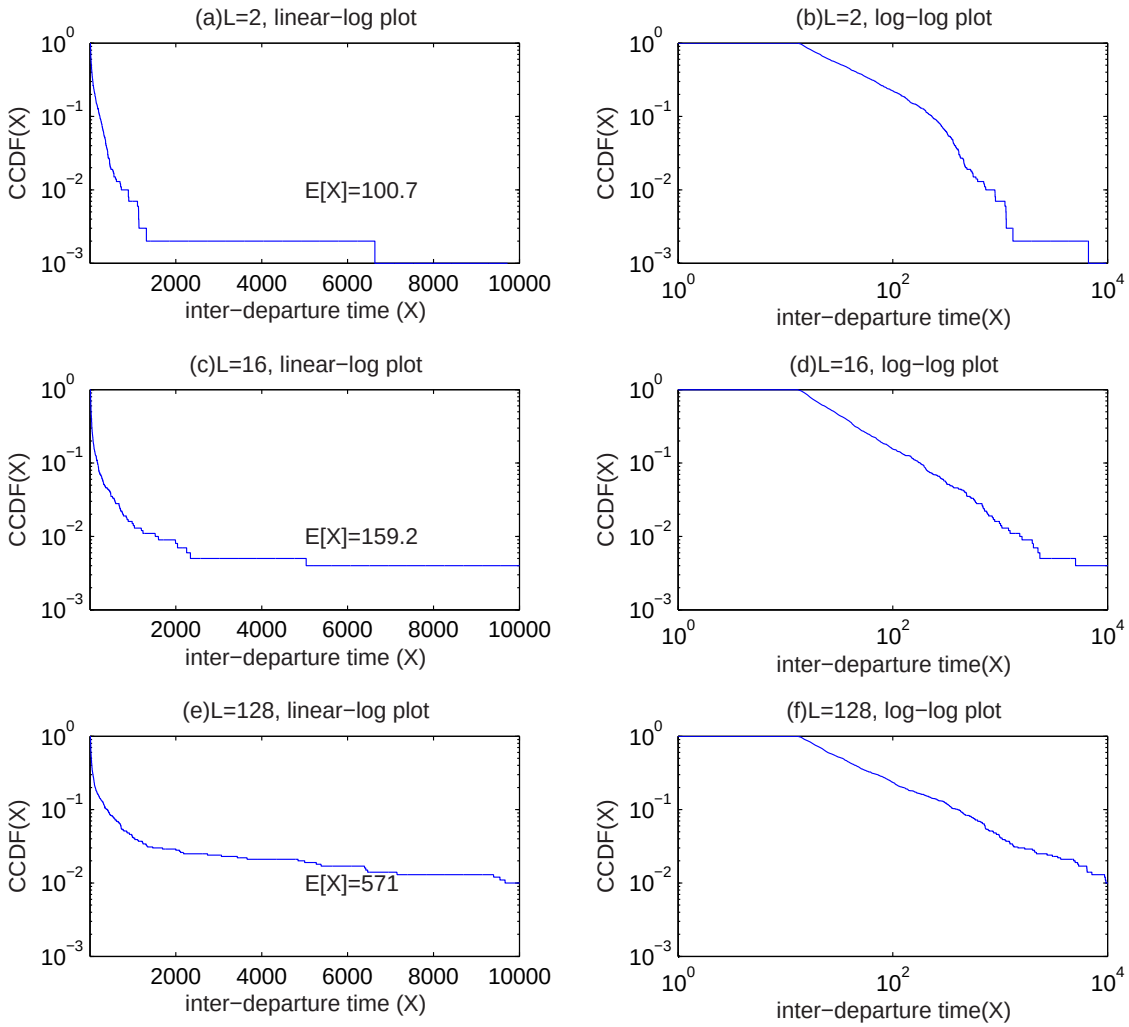


Figure 4.26: *The CCDF of inter-departure time obtained by passing Poisson arrivals through tandem Pareto servers. The mean of Poisson arrival is* $\lambda = 0.01$*, and the distribution parameters for Pareto are taken to be* $\alpha = 1.25$ *and* $k = 12$*.*

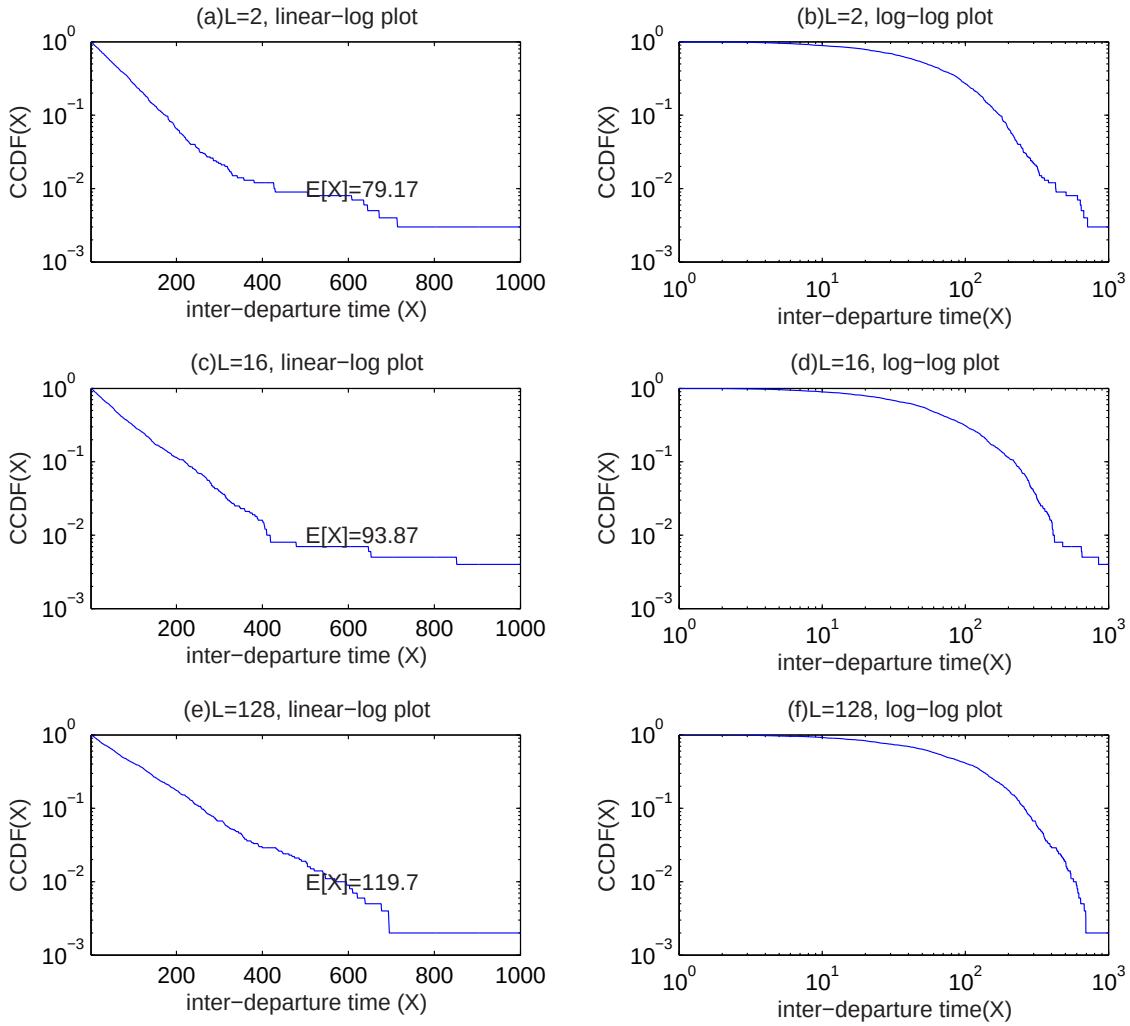


Figure 4.27: *The CCDF of inter-departure time obtained by passing Pareto inter-arrival process through tandem exponential servers. The distribution parameter for exponential distribution is* $\lambda = 0.0167$*, and the distribution parameters for Pareto are taken to be* $\alpha = 1.25$ *and* $k = 20$*.*

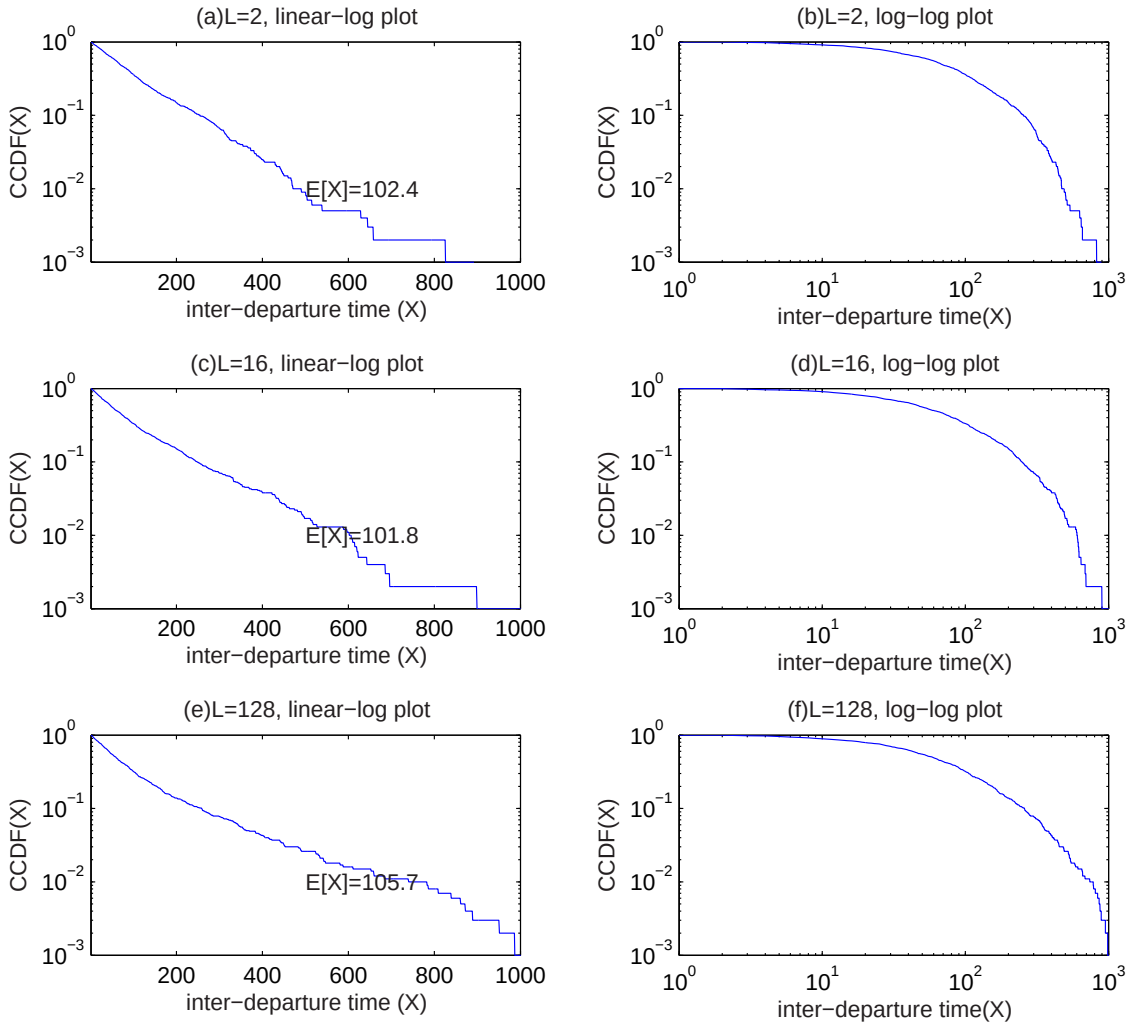


Figure 4.28: *The CCDF of inter-departure time obtained by passing Poisson arrivals through tandem exponential servers. The mean of Poisson arrival is* $\lambda_{\text{arrival}} = 0.01$*, and the distribution parameters for exponential server is* $\lambda_{\text{server}} = 0.0167$.

4.2.3 Entropy Rate Of The Interdeparture Process

In this subsection, we examine the entropy rate (upper bound, actually) of the interdeparture process of the tandem server system. The results are summarized in Tabs. 4.3–4.5. The results match what has been concluded in [Ana96], where the entropy rate is maximized for Poisson departure processes (cf. Tab. 4.5). Our results also hint that Pareto servers may generate an interdeparture process consisting of the most redundancy among those cases we considered.

Table 4.3: *The sizes of Lempel-Ziv coded and uncoded files of interdeparture time with different number of tandem servers. The interarrival is Pareto distributed with* $\alpha = 1.25$ *and* $k = 20$*, and the exponential service rate has distribution parameter* $\lambda = 0.0167$*.*

L	2	32	256
original file size	400,000 Bytes	400,000 Bytes	400,000 Bytes
coded file size	130,552 Bytes	137,445 Bytes	146,173 Bytes
compression rate	0.32638	0.343613	0.362933

Table 4.4: *The sizes of Lempel-Ziv coded and uncoded files of interdeparture time with different number of tandem servers. The interarrival is exponential distributed with* $\lambda = 0.01$*, and the Pareto service rate has distribution parameters* $\alpha = 1.25$ *and* $k = 12$*.*

L	2	32	256
original file size	400,000 Bytes	400,000 Bytes	400,000 Bytes
coded file size	101,775 Bytes	97,643 Bytes	94,236 Bytes
compression rate	0.254438	0.244108	0.23559

Table 4.5: *The sizes of Lempel-Ziv coded and uncoded files of inter-departure time with different number of tandem servers. The interarrival is exponential distributed with* $\lambda_{\text{arrival}} = 0.01$*, and the exponential service rate has distribution parameters* $\lambda_{\text{server}} = 0.0167$*.*

L	2	32	256
original file size	400,000 Bytes	400,000 Bytes	400,000 Bytes
coded file size	392,137 Bytes	392,584 Bytes	391,873 Bytes
compression rate	0.980343	0.98146	0.979683

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Recently, self-similarity becomes one of the important issues for network study. To ride on this trend, we examine the degree of self-similarity of network departure under finite resource assumption in this thesis. It is our hope that our work can be a basis for a comprehensive understanding of the cause of self-similarity observed from measurement of network traffics.

In this thesis, we emphasize on the relationship between the heavyness of tail probability and the degree of self-similarity. A known result on their relationship was recently made under the assumption that a self-similar traffic can be established by aggregating an infinite number of ON/OFF sources with heavy tail probability. We therefore query whether this relationship is still valid when the number of sources is finite. Such a finite resource assumption is perhaps more close to what occurs in the real world. As a result, an additional parameter L, which represents the number of servers (or equivalently sources), is introduced.

Along this research line, we examine the M/G/L system, which we called *parallel server system*. We then investigate the long-range dependence of the aggregated departure. Our simulations show that when the system utilization is fixed, adjusting the shaping parameter α (used in the Pareto server) does not make the inter-departure traffic as self-similar as $H = (3 - \alpha)/2$ suggests. We realize from this simulation that the reduction of k neutralizes the anticipated degree of self-similarity at small α. This leads to the subsequent simulations

in which k is kept fixed. By fixing k rather than the mean in Pareto server, increasing of α gives a monotone decreasing of self-similar parameter H as anticipation. A further increasing of L, which is the number of parallel Pareto servers, grows the degree of departure self-similarity.

For the tandem server system, we perform the CCDF analysis on the interdeparture. We observed an inconsistence on the CCDF behavior for different combinations of arrival and service time distributions. Only the Pareto servers present apparent heavy probability tail. The probability tail for exponential server, due to either to Pareto interarrival or exponential interarrival, becomes not so heavy (or light) in probability mass.

Finally, even with similar CCDFs for both Pareto and exponential interarrivals passing to exponential servers, they have different (Lempel-Ziv bound of) entropy rates. Our result suggests that Pareto servers may generate an interdeparture process consisting of the most redundancy among those cases we considered.

5.2 Future Work

Our simulated M/G/L parallel server has a queue that allows infinite in length. A more practical assumption could be a queue with finite length assumption. Other consideration, such as the usage of a non-First-Come-First-Server queue, may also be incorporated.

Another future work which is possibly very hard is to provide a theoretical basis for our simulation results. Notably, the current analytical results are all established under infinite resource assumption. A counterpart theory based upon finite resource assumption may be an important but hard future work.

Bibliography

[Addie99] R. G. Addie, T. D. Neame and M. Zukerman, "Modeling superposition of many sources generating self similar traffic," *IEEE International Conference on Communication*, vol. 1, pp. 387–391, 1999.

[Ana96] V. Anantharam and S. Verdu, "Bits through queues," *IEEE Trans. Inform. Theory*, vol. 42, no. 1, pp. 4–18, January 1996.

[Bing87] N. H. Bingham, C. M. Goldie and J. L. Teugels, *Regular Variation*, Cambridge, New York, Melburn: Cambridge Univ. Press, 1987.

[Box98] O. J. Boxma and J. W. Cohan, "The M/G/1 queue with heavy-tailed service time distribution," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 5, pp. 749–763, June 1998.

[Fell71] W. Feller, *An Introduction to Probability Theory and Its Applications (Volumn II)*, 2nd edition, New York: John Wiley & Sons, 1971.

[Leland94] W. E. Leland, M. S. Taqqu, W. Willinger and D. V. Wilson, "On the self-similar nature of ethernet traffic (extended version)," *IEEE/ACM Trans. Networking*, vol. 2, no. 1, pp. 1–15, February 1994.

[Stall98] W. Stallings, *High-Speed Networks: TCP/IP And ATM Design Principles*, Prentice Hall, 1998.

[Stall02] W. Stallings, *High-Speed Networks and Internets: Performance and Quality of Service*, 2nd edition, Prentice-Hall, 2002.

[Tsy98] B. Tsybakov and N. D. Georganas, "Self-similar processes in communications networks," *IEEE Trans. Inform. Theory*, vol. 44, no. 5, pp. 1713–1725, September 1998.

[Will97] W. Willinger, M. S. Taqqu, R. Sherman and D. V. Wilson, "Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level," *IEEE/ACM Trans. Networking*, vol. 5, no. 1, pp. 71–86, February 1997.

Vita

Hsu-Hui Wang

Date of Birth

March 5, 1979

Place of Birth

Taipei, Taiwan, R.O.C.

Education

- **M.S.-Communication Engineering**

  Department of Communication Engineering

  National Chiao Tung University (NCTU), Taiwan , R.O.C. -2003

  Advisor: Professor Po-Ning Chen

- **B.S.-Electronic Engineering**

  Department of Electronic Engineering

  National Chiao Tung University (NCTU), Taiwan , R.O.C. -2001

- **High School**

  Provincial Wu Ling Senior High School, Taiwan , R.O.C. -1997

  (National now)

Teaching Experience

- **Teaching Assistant**

  Queueing Theory

Leadership

- **Financial President, Soprano Leader**

  Chorus Club, NCTU

  Feb. 1998 - Feb. 1999

- **Chief Executor**

  Union Chorus Concert

  13 Dec. 1998

Other Activies

- **Personal Website**

  "*Snow Covered Field*" http://yukino.24cc.com

- **Personal E-paper**

  "*The Light in Midnight*"

- **Self-Published Works**

  "*Wandering in September*" Feb. 2002

  "*Come, Sweet Death*" Dec. 2002