J Supercomput (2009) 47: 1–20
DOI 10.1007/s11227-008-0189-x

Workload dynamics on clusters and grids

Hui Li

Published online: 4 March 2008

Abstract This paper presents a comprehensive statistical analysis of a variety of workloads collected on production clusters and Grids. The applications are mostly computational-intensive and each task requires single CPU for processing data, which dominate the workloads on current production Grid systems. Trace data obtained on a parallel supercomputer is also included for comparison studies. The statistical properties of workloads are investigated at different levels, including the Virtual Organization (VO) and user behavior. The aggregation procedure and scaling analysis are applied to job arrivals, leading to the identifications of several basic patterns, namely *pseudo-periodicity*, *long range dependence* (LRD), and *multifractals*. It is shown that statistical measures based on *interarrivals* are of limited usefulness and *count* based measures should be trusted when it comes to correlations. Other job characteristics like run time and memory consumption are also studied. A "bag-of-tasks" behavior is empirically evidenced, strongly indicating temporal locality. The nature of such dynamics in the Grid workloads is discussed. This study has important implications on workload modeling and performance predictions, and points out the need of comprehensive performance evaluation studies given the workload characteristics.

Keywords Workload characterization · Cluster and grid computings

H. Li (✉)
Leiden Institute of Advanced Computer Science (LIACS), Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
e-mail: hui.li@computer.org

Present address:
H. Li
Department of Planning, Performance, and Quality, TNO ICT, 2612 CT Delft, The Netherlands

1 Introduction

Grid computing is rapidly evolving as the next-generation platform for system-level sciences and beyond. Scheduling in a Grid environment can be carried out at different levels, including local resource management systems on clusters, Grid-level brokering services, and virtual organization based schedulers. Consequently, performance evaluation of scheduling strategies require representative workload models at different levels. The goal of this paper is to study comprehensively the statistical properties of workloads on Grids at various levels, which serve as the basis for workload modeling and performance predictions.

Closely-related workload studies have been carried out for parallel supercomputers. On single parallel machines, a large amount of workload data has been collected,[1] characterized [9, 18, 27], and modeled [7, 18, 26]. In [7], polynomials of degree 8 to 13 are used to fit the daily arrival rates. In [18], a combined model is proposed where the interarrival times fit a hyper-Gamma distribution and the job arrival rates match the daily cycle. Time series models such as ARIMA are studied in [27], which try to capture the traffic trends and interdependencies. Other characteristics such as run time and parallelism are also investigated and models are proposed based on distribution fitting [18] or Markov chains [26]. It could be concluded that a majority of previous research results on parallel supercomputers focus mainly on marginal distributions and first order statistics while correlations and second order properties receive far less attention. The reason could be that characteristics on parallel workloads are inherently weakly autocorrelated or short range dependent (SRD). For instance, in this paper, analysis of a representative parallel workload is conducted for comparison studies. It is shown that the interarrival time process of job arrivals as well as the run time series are indeed short range dependent. Despite the fractal behavior at small scales, the job count process is also weakly autocorrelated with quickly-vanishing autocorrelation lags. Data-intensive workloads on clusters and Grids, on the other hand, exhibit pseudo-periodicity and long range dependence which are not present in parallel workloads. Therefore, second order statistics is crucial and new methodologies should be proposed for both analysis and modeling.

The contribution of this work is three-fold. First, *point process* is introduced to describe job arrivals and several representations are studied. It is shown that statistical measures based on interarrivals are of limited usefulness when it comes to autocorrelations and count based measures should be trusted instead. Secondly, the scaling analysis on job count processes enable us to understand the autocorrelation structures better. Together with the cross-correlations between characteristics, we obtain an improved understanding toward workload dynamics. Thirdly, we identify several basic patterns, such as pseudo-periodicity, long range dependence, and "bag-of-tasks" behavior. Further research on workload modeling on clusters and Grids should capture these salient properties, which could have important implications on performance evaluation studies.

The rest of this paper is organized as follows. Section 2 introduces the definition and methodology used in the analysis. Point process and its representations are

[1] Parallel Workload Archive. http://www.cs.huji.ac.il/labs/parallel/workload/.

defined. Statistical measures are presented, including marginal statistics, autocorrelation and spectrum, and cross-correlation. Topics regarding scaling, fractals, and power law behavior are treated in depth, for which various defining properties and analyzing tools are discussed. Section 3 describes workloads in a broader perspective. Related work on network traffic and cluster workloads are further reviewed and discussed. Section 4 presents a comprehensive analysis on real-world workload data. We study a variety of workloads from Grids, clusters, and a parallel supercomputer. VO and user behavior are investigated extensively. In Sect. 5, the nature and origin of workload dynamics are explained and implications on modeling and predictions are discussed. Conclusions and future work are presented in Sect. 6.

2 Definition and methodology

This section covers the statistical theories and methodologies used in workload characterization. It starts with the definition of a point process and its representations because it is the basis for analyzing job arrival processes. Statistical measures such as distributions, autocorrelation function (ACF), and periodicity are described. A large part of this section is dedicated to scaling, fractals, and power law behavior. Definitions and relationships among important notions such as long range dependence (LRD), burstiness, scaling, and wavelets are elaborated. These are the theories for understanding the temporal correlations and dynamics of the workloads presented later in this paper.

2.1 Point processes

Job traffic can be described as a (stochastic) *point process*, which is defined as a mathematical construct that represents individual events as random points at times $\{t_n\}$. There are different representations of a point process. An *interarrival time process* $\{I_n\}$ is a real-valued random sequence with $I_n = t_n - t_{n-1}$. The sequence of counts, or the *count process*, is formed by dividing the time axis into equally spaced contiguous intervals of T to produce a sequence of counts $\{C_k(T)\}$, where $C_k(T) = N((k+1)T) - N(kT)$ denotes the number of events in the kth interval. This sequence forms a discrete-time random process of nonnegative integers and it is another useful representation of a point process. A closely related measure is a normalized version of the sequence of counts, called the *rate process* $R_k(T)$, where $R_k(T) = C_k(T)/T$.

In general, forming the sequence of counts loses information because the interarrival times between events within interval T are lost. Nevertheless, it preserves the correspondence between its discrete time axis and the absolute "real" time axis of the underlying point process. The correlation in the process $\{C_k(T)\}$ can be readily associated with that in the point process. The interarrival time process, on the other hand, eliminates the direct correspondence between absolute time and the index number, thus, it only allows rough comparisons with correlations in the point process [17]. As is shown later, measures based on interarrival times are not able to reliably reveal the fractal nature of the underlying process and count based measures should be trusted instead. The different representations of a point process are illustrated in Fig. 1.

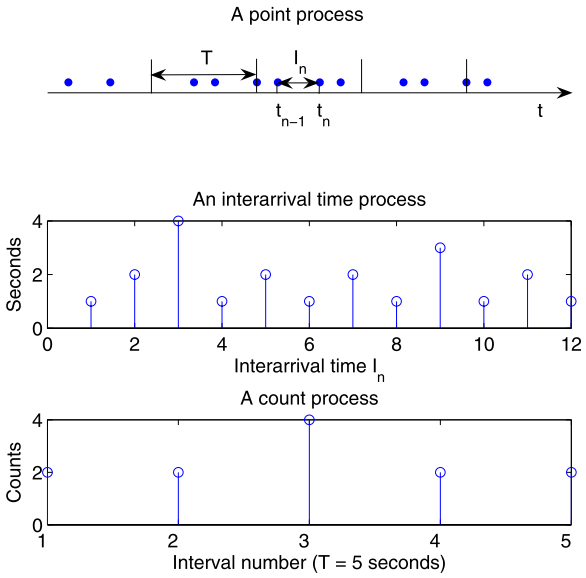

Fig. 1 An example of a point process and its two representations: an interarrival time process and a count process

2.2 Statistical measures

No single statistic is able to completely characterize a point process and each provides a different view and highlights different properties. A comprehensive analysis toward an improved understanding requires many such views. In this section, the statistical measures used throughout this paper are defined. These measures apply to both interarrival time and count (rate) representations, although their usefulness depends heavily on the analytic context.

The first set of measures focus on the marginal properties of a process, including *mean* (μ), *variance* (σ^2), *probability density function* (PDF), and *cumulative distribution function* (CDF). In practice, the sample mean ($\overline{X}$) and sample variance (S^2) are used to estimate mean and variance, respectively. The so-called complementary cumulative distribution function (CCDF) is commonly used to show probability distributions. Histogram, a graph that shows the frequency of data in successive equal-size numerical intervals, is used to estimate the probability density function. The reader is referred to [25] for a detailed treatment on these basic statistical measures.

2.2.1 Autocorrelation and spectrum

The autocorrelation function (ACF) of a process X describes the correlations between different points in time. If X is *second order stationary*, i.e., mean μ and variance σ^2 do not change over time, the autocorrelation function depends only on lag k[2] and it can be defined as

$$R(k) = \frac{E[(X_i - \mu)(X_{i+k} - \mu)]}{\sigma^2}, \tag{1}$$

[2]For a discrete time series of length n, k is the difference in time and there is $0 \leq k < n$.

where E is the expected value (mean) operator. It should be noted that in *signal processing* the above definition is often used without normalization, namely, without subtracting the mean and dividing by the variance.

For the interarrival time process, there is no direct relationship between the lag k and time t, so the ACF $R_I(k)$ as well as other interarrival based measures have limited usefulness, especially in the scaling analysis. The count autocorrelation proves to be a valuable measure as it provides information about the second-order properties. For distinction count ACF is denoted as $R_C(k, T)$ for the inclusion of the count interval T.

Fourier transforming the autocorrelation function (ACF) yields the power spectral density (PSD, or power spectrum) $S(f)$

$$S(f) = \sum_k R(k) e^{-i2\pi kf}, \tag{2}$$

where f is the frequency. Autocorrelation and power spectrum are commonly-used measures for studying the correlation structures and second-order properties of a single process. Like the autocorrelation, the count-based ($S_C(f, T)$) and rate-based spectrums ($S_R(f, T)$) prove to be useful in the identification of fractal behavior. An estimate of power spectrum can be derived via methods such as periodogram. *Discrete Fourier Transform* (DFT) is used exchangeably to show the frequency components of the signal.

2.2.2 Periodicity

From the theory of Fourier analysis, it is known that periodicity shows up as peaks in the frequency domain. Real world data, however, seldom exhibits perfectly periodic behavior. In most situations, *pseudo-periodic* signals are observed instead, potentially arising from various sources of noises and the time-varying nature of the generation scheme. From this perspective, it is necessary to use quantitative methods to measure the degree of periodicity in the data. Periodicity in a process can be detected and quantified using power-spectrum based methods. The first measure P_f is defined as the normalized difference of the sum of the power spectrum values at the highest amplitude frequency and its multiples, and the sum of the power spectrum values at the halfway-between frequencies [21]. The *total spectrum entropy* (TSE) calculates the entropy for the whole power spectrum while the *saturated spectrum entropy* (SSE) excludes the first one or two "big" power spectrum values, which represent the total energy of the signal. All measures have values between 0 and 1. Higher P_f and lower entropy correspond to stronger periodicity in the signal. These measures are used to study pseudo-periodic job arrivals.

2.2.3 Cross-correlation

Besides studying how events of the same process are correlated with each other, it is also important to reveal the correlations between events of distinct random processes. The simplest way of investigation is to plot samples of both variables and visually identify if any pattern exists. A common alternative is the *scatter plot*, which displays

the sample values of X and Y jointly in a two-dimensional figure. Simply plotting the data gives us lots of information of the underlying correlation structures.

Nevertheless, visual information cannot be used to give definite answers and quantitative measures are needed for identifying correlations in practice. In statistics, a simple and common measure is called *correlation coefficient*, which indicates the strength and direction of a linear relationship between two random variables. The best known coefficient is the *Pearson product-moment correlation coefficient* and it is obtained by dividing the covariance of the two variables by the product of their standard deviations. It is formulated as

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y}. \tag{3}$$

A more advanced version is referred as *Spearman's rank correlation coefficient*, which does not require any assumptions of linear relationship or the distributions of variables.

2.3 Scaling and power law

Physical processes can be observed from a vast range of *scales*, in other words, *multi-resolution*. For instance, in network traffic studies one can represent the traffic as number of bytes or packets at the level of milliseconds, seconds, and even minutes. On clusters and Grids, the number of job arrivals can be aggregated and averaged every second, every minute or even every hour. Scaling or *scale invariance* means the lack of any special characteristic scale, namely all scales have equal importance. Scaling leads to power law dependencies in the scaled quantities as $f(as) = g(a)f(s)$. It is shown in [17] that the only nontrivial solution of this scaling function for real functions and arbitrary a and s is $f(s) = bs^c$, for some constants b and c. In some contexts, c is referred as the *scaling component*. *Self-similar* and *long range dependent* (LRD) processes are two most important classes of general scaling processes and LRD is relevant in the context of this paper.

A process $X(t)$ is said to be *long range dependent* (LRD) if either its autocorrelation function or power spectrum satisfies the following conditions

$$R(k) \sim c_r k^{\alpha-1}, \quad k \to \infty, \quad \text{or} \quad S(f) \sim c_f f^{-\alpha}, \quad f \to 0, \tag{4}$$

where c_r, c_f are constants. The autocorrelation function $R(k)$ decays so slowly that $\sum_{k=-\infty}^{\infty} R(k) = \infty$ and $S(0) = \infty$. Frequency-domain characterization of LRD also leads to a class of so-called $1/f$*-like processes* ($1/f$ noise) [31].

It is of crucial importance to recognize the usefulness of different representations of processes. In network traffic, both interarrival and count based measures prove to be useful in analyzing the scaling behavior [1, 24]. However, for job arrivals on clusters and Grids, measures based on interarrivals fails to reveal the fractal behavior of the underlying process and only count/rate based measures can be trusted. This problem is discussed with greater detail in a more theoretical treatment [17].

The scaling behavior introduced so far has one single exponent, thus it can be called *monofractal*. There are cases in which a range of fractal behaviors exist within

one process, or the scaling exponent is time-dependent. The process is then called *multifractal*. A complete presentation of multifractal formalism is referred to [22]. The later mentioned biscaling is a very simple form of multifractals.

2.3.1 Wavelets and scaling

Due to its inherent multi-scale/resolution properties, wavelets provide a natural framework for analyzing the scaling behavior. Like the Fourier transform that analyzes signals with sinusoidal functions, the wavelet transform projects the signal onto the so-called *wavelets* [8, 28]. A *wavelet function* $\psi(t)$ is a bandpass function that can be scaled and shifted

$$\psi_{j,k}(t) = 2^{-j/2}\psi\left(2^{-j}t - k\right). \tag{5}$$

There also exists a *scaling function* $\phi(t)$, which is a lowpass function that can be scaled and shifted as well. A discrete wavelet transform (DWT) of a signal can be executed by passing the signal recursively through a set of lowpass and bandpass filters [28]. As a result, the signal is decomposed into a sum of weighted scaling functions and wavelet functions

$$X(t) = \sum_{k} c(j_0, k)\phi_{j_0,k} + \sum_{j \le j_0} \sum_{k} d(j,k)\psi_{j,k}(t), \tag{6}$$

where $c(j_0, k)$ are referred as *scaling coefficients* (or approximations) and $d(j,k)$ as *wavelet coefficients* (or details).

A very attractive feature of wavelet analysis lies in the fact that the long range dependent, nonstationary original process turns into stationary, nearly uncorrelated, or short range dependent wavelet coefficients $d(j,k)$. In the case of scaling, the energy of these coefficients is power law dependent of scale j, denoted by

$$\frac{1}{n_j}\sum_{k=1}^{n_j}\left|d(j,k)\right|^2 \propto 2^{j\alpha}. \tag{7}$$

This property leads to a wavelet-based scaling exponent estimation tool called the *logscale diagram* [3]. Compared with other power law based estimators like aggregated variance and periodogram, this technique is shown to have better statistical and computational properties [4].

As has been explained and formulated in [3], if $\alpha \in (0, 1)$ and the range of scales is from some initial scale j_1 to the largest scale, then scaling could be related to LRD with a scaling exponent of measured α. It is also highly possible that real world data have more than one alignment region within a single logscale diagram, which is referred as *biscaling*. Biscaling can be regarded as different scaling exponents at small and large scales, respectively. A natural generalization of logscale diagram beyond second order can be denoted as $\mu_j^{(q)} = 1/n_j \sum_k |d(j,k)|^q$, where q is of real value. It is shown in [3] that $E[\mu_j^{(q)}] \sim 2^{j(\zeta(q)+q/2)}$. For monofractals such as exact self-similar processes, there is $\zeta(q) = qH$, meaning that self-similarity can be identified

by testing the linearity of $\zeta(q)$. If on the other hand $\zeta(q)$ is nonlinear, then multifractal scaling is detected. The so-called *multiscale diagram* is a realization of this result. The qth order scaling exponent $\alpha_q = \zeta(q) + q/2$ can be estimated in the qth order logscale diagram for multiple q values. The multiscale diagram consists of the plot of $\zeta(q) = \alpha_q - q/2$ against q along with the confidence intervals. A lack of linearity in the multiscale diagram suggests multifractal behavior, therefore, it becomes a useful tool for identifying multifractal processes.

3 Workloads in a broader perspective

Studies on network traffic are reviewed because it includes a rich collection of advanced statistic tools for analyzing and modeling self-similar, long range dependent, and fractal behavior. The self-similar nature of Ethernet traffic is discovered in [14], and consequently, a set of exact self-similar models such as fractional Brownian motion and fractional Gaussian noise are proposed as traffic models [20, 29]. Network traffic is also shown to be long range dependent, exhibiting strong temporal burstiness [2, 23]. Both self-similar and LRD processes are most well-known examples of general scaling processes, characterized by the scaling and power law behavior [3]. Due to its inherent multi-resolution nature, wavelet is proposed as an important tool for analysis and synthesis of processes with scaling behavior [1, 3, 30]. Multifractal models and binomial cascades are proposed for those processes with rich fractal behavior beyond second-order statistics [11, 24]. Recent advances include a more general Infinitely Divisible Cascade (IDC) process [6]. These methodologies enable the scaling analysis on job arrivals and the identification of important patterns.

Workload characterization on clusters with marginal statistics can be found in [12, 16, 19]. In [19] an ON-OFF Markov model is proposed for modeling job arrivals, which is essentially equivalent to a two-phase hyperexponential renewal process. The major modeling drawback using renewal processes is that the autocorrelation function (ACF) of the interarrival times vanishes for all nonzero lags so they cannot capture the temporal dependencies in time series [13]. A more sophisticated n-state Markov modulated Poisson process is applied for modeling job arrivals at the Grid and VO level [15], making a step forward toward capturing autocorrelations. Nevertheless, only limited success is obtained by MMPP because of the rich behavior and patterns hidden in Grid workloads at different levels. This paper identifies and characterizes those salient workload patterns on clusters and Grids.

4 Application to workload data

The workload data under study are collected from real production clusters and Grids. Table 1 presents a summary of workload traces used in this paper. *LCG1* and *LCG2* are two traces from the LHC Computing Grid.[3] The LCG production Grid consists

[3]LCG is a data storage and computing infrastructure for the high energy physics community that will use the Large Hadron Collider (LHC) at CERN. http://lcg.web.cern.ch/LCG/.

Table 1 Summary of workload traces used in the experimental study (NIK–NIKHEF)

Trace	Location	Arch.	Scheduler	CPUs	Period	#Jobs
LCG1	Grid wide	data Grid	Grid Broker	∼30 k	Nov. 2005	188,041
LCG2	Grid wide	data Grid	Grid Broker	∼30 k	Dec. 2005	239,034
NIK05	NIK, NL	PC cluster	PBS/Maui	288	Sep.–Dec. 2005	63,449
RAL05	RAL, UK	PC cluster	PBS/Maui	1,000	Oct.–Nov. 2005	332,662
LPC05	LPC, FR	PC cluster	PBS/Maui	140	Feb.–Apr. 2005	71,271
SBH01	SDSC, US	IBM SP	LoadLeveler	1152	Jan.–Dec. 2001	88,694

of approximately 180 active sites with around 30,000 CPUs and 3 petabytes storage (Dec. 2005), which is primarily used for high energy physics (HEP) data processing. There are also jobs from biomedical sciences running on this Grid. Almost all the jobs are independent, computationally-intensive tasks, requiring one CPU to process a certain amount of data. The workloads are obtained via the LCG Real Time Monitor[4] for two periods: *LCG1* consists of jobs of eleven consecutive days from November 20–30 in 2005, while *LCG2* is from December 19–30 in the same year. These two traces carry valuable information about the user behavior at the Grid level.

The Grid sites consists of computing clusters and storage systems. Each cluster runs its local resource management system and defines its own sharing policies. It is also important to analyze the workloads at the cluster level. Traces are obtained from three data-intensive clusters, which are named *NIK05*, *RAL05*, and *LPC05*. They are located at the HEP institutes in the Netherlands, UK, and France, respectively, and all of them participate in LCG. The clusters are made of commodity components, and deploys similar cluster software suite (e.g., PBS/Maui) and Grid middleware from LCG. It should be noted that these clusters are involved in multiple collaborations simultaneously and have their own local user activities. Grid jobs from LCG only account for a portion of the whole workloads, depending on the level of involvement and local policies. The trace *SBH01* is from a SDSC parallel supercomputer and it is included for comparison studies.

Workloads typically have certain structures. Jobs come from different groups and users. In Grids, *Virtual Organization* (VO) is an important concept and one can consider a VO as *a collection of entities (users, resources, etc.) that belong to multiple organizations but have common goals or shared policies*. In LCG, VOs are mostly named after major HEP experiments and scientific disciplines, such as *lhcb*, *atlas*, or *biomed*. It is observed that a small number of top VOs and users often dominate the workload, as is shown in Fig. 2. These type of patterns can also be empirically found in many social and physical phenomena, such as database transactions and Unix file sizes [5, 9]. By analyzing the main VOs and users, a good understanding of the whole workload can be obtained. Moreover, patterns emerge by simply using

[4]The Real Time Monitor is developed by Imperial College London and it monitors jobs from all major Resource Brokers on the LCG Grid, therefore, the data it collects are representative at the Grid level. A Resource Broker (RB) is a service to receive and schedule jobs from Grid users. http://gridportal.hep.ph.ic.ac.uk/rtm/.

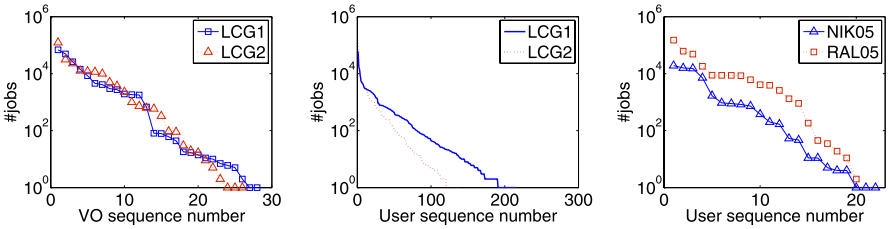

Fig. 2 Distributions of number of jobs by VOs and users on clusters and Grids

Table 2 Different levels and characteristics under study for the Grid, the cluster, and the supercomputer (SC) traces

Category	Traces	Levels	Characteristics to study
Grid	LCG1/LCG2	Grid/VO	Arrival, Run time
Cluster	NIK05/RAL05/LPC05	Site/VO/User	Arrival, Run time, Memory
SC	SBH01	Site/User	Arrival, Run time, Parallelism

Table 3 Names for different VOs or users in experimental studies. *lhcb*, *atlas*, and *cms* are major HEP experiments in the LCG Grid. *dteam* is a VO mostly consisting of software monitoring and testing jobs in the Grid. *hep1* is a HEP collaboration between institutes in UK and US, part of which is also involved in LCG. *biomed* is the VO with biomedical applications and it contributes to ∼65% of *LPC05* jobs. *com1* is a company partner with NIKHEF, which runs medical-related data-intensive jobs. *user45*, *user328*, and *user272* are the top three users on SDSC Blue Horizon with most of the job submissions

Trace	VO or user names under study
LCG1	lhcb, atlas, cms, dteam
LCG2	lhcb, atlas, cms, dteam
NIK05	lhcb, atlas, com1
RAL05	hep1, atlas
LPC05	biomed
SBH01	user45, user328, user272

the nominal VO names for categorization without applying sophisticated clustering techniques. From a performance evaluation perspective, it is also desirable to include VO or users in the models since most of the policy rules are based on their names. Given these many motivations, the analysis in this paper focuses on the VO level. User level experiments are carried out for *SBH01* because the VO/group information is not available. The levels and the different VO/user names under study are shown in Tables 2 and 3, respectively.

Table 2 shows the job characteristics at different levels. Different characteristics are investigated for each level based on their usage and availability. For data-intensive Grids job, *arrivals* and *run times* are being analyzed. On clusters, job *memory* consumption becomes available for study. In both cases, *parallelism* need not to be con-

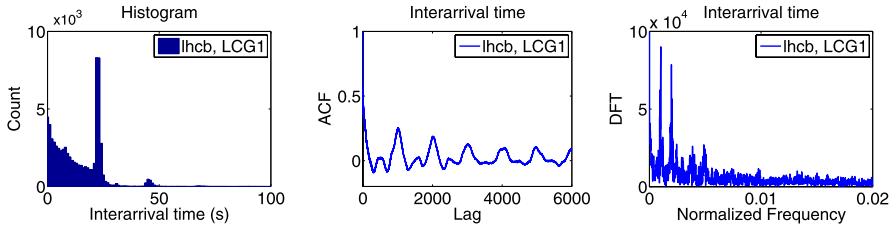

Fig. 3 The histogram plot, autocorrelation function (ACF), and discrete Fourier transform (DFT) for the interarrival time process of *lhcb, LCG1*

sidered because of its equality to one. On the supercomputer, however, parallelism becomes an important characteristics so it is included in the study.

The analysis is to apply the statistical measures discussed in Sects. 2.2 and 2.3 to each level of workloads for different characteristics. This has generated a large number of data and figures. The interest point, however, is to discover some basic pattern or patterns of the workload characteristics. Therefore, the presentation of results is categorized by the discovered patterns and only representative figures of each pattern are shown. In the following sections, the job arrival patterns is analyzed first, followed by run time, memory, and parallelism. Cross-correlations between characteristics are then examined.

4.1 Job arrival process

There are three basic patterns identified for job arrivals on clusters and Grids: *pseudo-periodicity*, *long range dependence* (LRD), and *(multi)fractals*, which are presented subsequently in the following sections. Short range dependence is also observed for cluster workloads. It is not included here in the characterization, but will be investigated in the performance studies of workload correlations.

4.1.1 Pseudo-periodicity

There are a number of VOs at the Grid and the cluster level which exhibit pseudo-periodic patterns and *lhcb* on *LCG1* is used as the example here. Figure 3 shows the first and second order statistics of job interarrival times of *lhcb, LCG1*. A strong deterministic component of around 20 seconds is observed in the histogram plot. As to the second-order properties, certain periodicity is detected in the ACF and DFT plot. The decaying peaks in the ACF plot correspond to the two main spikes in the low frequency domain of the DFT. Nevertheless, periodicity for interarrival times does not hold for all processes belonging to this pattern. This is in accordance with the fact that interarrival based measures eliminate the direct relation with the real time axis and count based measures should be examined.

The next step naturally goes to the aggregation procedure which uses count based measures. Figure 4 plots the count process together with its ACF and power spec-

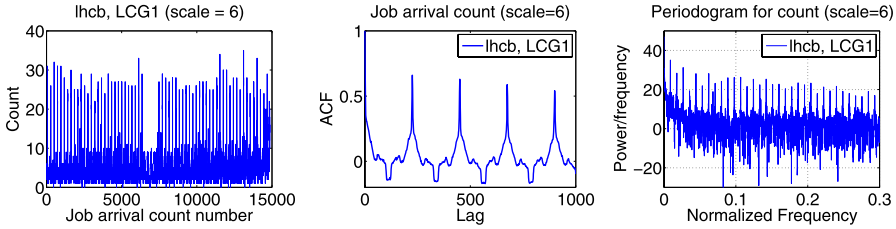

Fig. 4 The sequence plot, autocorrelation function (ACF), and power spectrum via periodogram for the count process of *lhcb, LCG1*

Table 4 Periodicity measures. TSE—total spectrum entropy, SSE—saturated spectrum entropy. P_f—the periodicity measure as defined in Sect. 2.2.2

Trace	TSE	SSE	P_f
lhcb, LCG1 (scale = 6)	0.40	0.74	0.84
lhcb, LCG2 (scale = 6)	0.18	0.72	0.78
dteam, LCG1 (scale = 6)	0.69	0.71	0.94
dteam, LCG2 (scale = 6)	0.68	0.70	0.95
com1, NIK05 (scale = 8)	0.79	0.80	0.89
all, NIK05 (scale = 8)	0.88	0.91	0.79
biomed, LPC05 (scale = 8)	0.63	–	0

trum for scale[5] $= 6$. Periodicity is clearly detected by the equally-spaced peaks in the ACF plot and the multiple harmonics in the power spectrum. The quantitative measures for periodicity are shown in Table 4. SSE values should be used to examine the strength of periodicity and its results are consistent with those of P_f: lower SSE values correspond to higher P_f values, which indicate stronger periodic behavior. It is observed that all listed processes except *biomed, LPC05* show quite strong periodicity. As a comparison *biomed, LPC05* shows no periodicity at all and it is long range dependent.

4.1.2 Long range dependence (LRD)

biomed, LPC05 is used as a representative example for illustrating long range dependence. As is shown in Fig. 5, the interarrival time distribution is heavy-tailed and amplitude burstiness is observed. The ACF of interarrival times, on the other hand, has quickly decaying lags and shows short range dependence. This is in accordance with the scaling exponent estimate $\alpha = 0.164$ in the logscale diagram in Fig. 5. For the logscale diagram of count based measures, the scaling region is from the octave 8 (corresponding to scale 10 in the variance plot) up to the largest scale with an estimated scaling exponent $\alpha = 0.96$. This type of scaling strongly suggests long range dependence behavior [3]. Plotting the count processes from several scales and their second order statistics further confirm LRD. It is shown in Fig. 5 that the ACF and

[5] A dyadic scale is used so scale j means $T = 2^j$ seconds in the count process. This applies to all the scales in the count based measures used throughout this paper.

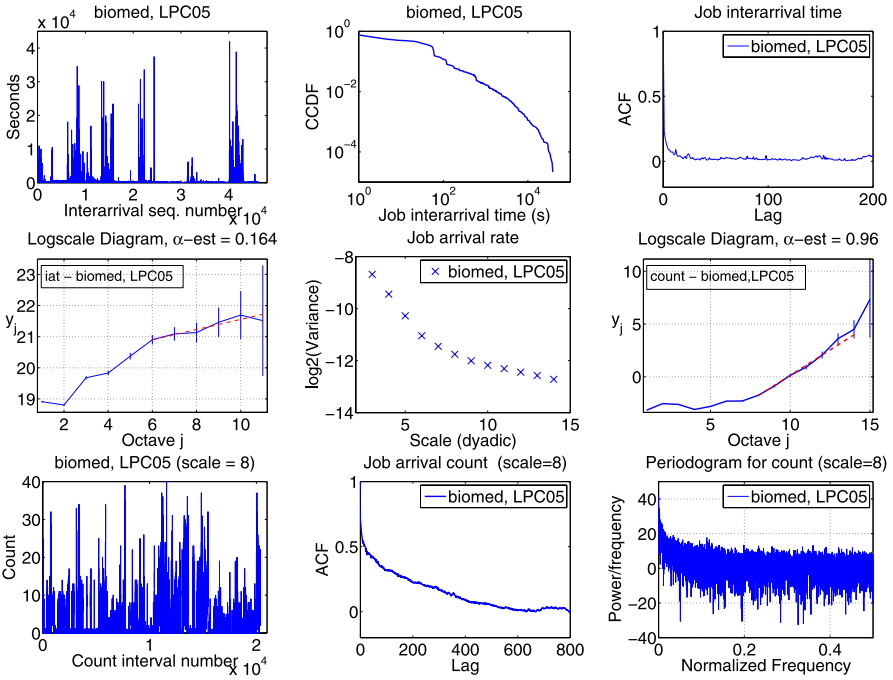

Fig. 5 Plots of the first, second order statistics and scaling analysis for both interarrival and count processes of *biomed, LPC05*. The *dash line* in the logscale diagram is the linear fit for estimating the scaling exponent α

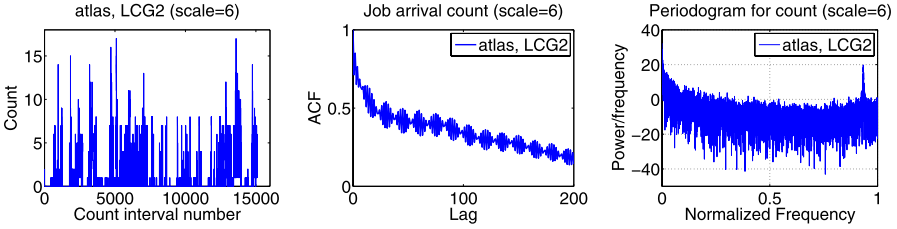

Fig. 6 The sequence plot, autocorrelation function (ACF), and power spectrum via periodogram for the count process of *atlas, LCG2*

the spectrum of scale 8 decay very slowly. It should be noted that the scaling and LRD behavior has a certain lower bound beyond which scaling is not obeyed.

Data from real production systems is highly complex and different patterns can be observed within one process. Long range dependence, for instance, can be mixed with periodic components. There are two types of periodic components added to a LRD process. The first type is *LRD plus high-frequency periodic components*. Figure 6 shows the count process of *atlas, LCG2*. A slowly-decaying ACF lag indicates the presence of long range dependence. There is also a high frequency periodic component observed in the power spectrum. As is shown in the ACF plot, the periodic

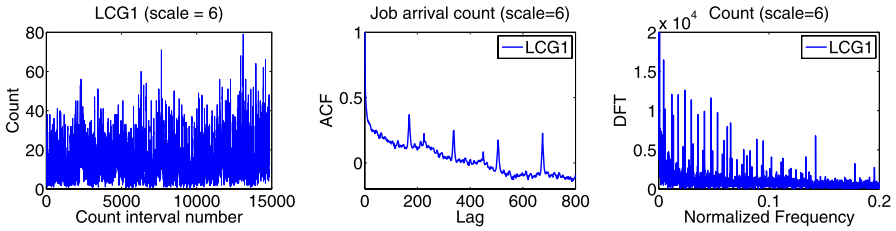

Fig. 7 The sequence plot, autocorrelation function (ACF), and discrete Fourier transform (DFT) for the count process of *LCG1*

fluctuations are nicely aligned with the power law decaying lags. The high frequency component can be related to some of the deterministic job submissions from this virtual organization.

The second type of periodic behavior contains multiple components, mostly concentrated in the lower frequency domain. This type is usually found in the aggregated whole trace with mixed deterministic and stochastic components. The Grid level *LCG1* and *LCG2* are examples of this pattern and *LCG1* is shown in Fig. 7. The count process (scale = 6) is LRD along with multiple low frequency peaks. These peaks can be related to the behavior of main VOs. By cross-referring the ACF plot of *lhcb, LCG1*, it can be found that the 240-minute peak is contributed by *lhcb*. This indicates that the count/rate processes at the Grid level are formed by aggregations of the VO processes.

4.1.3 Multifractals

Figure 8 shows *hep1, RAL05* as an example for multifractals. The interarrival time process is short range dependent. The logscale diagram of the count process exhibits biscaling (see Sect. 2.3.1). The scaling concentrated at the lower scales indicates the fractal nature of the sample path. The alignment at higher scales, on the other hand, resembles that of a stationary SRD process. This is further visualized for scale = 6 with quickly vanishing ACF lags and a white-noise like spectrum. For testing multifractality, the multiscale diagram of the count process is plotted ("blue circle", middle-right in Fig. 8). A simulated fractional Gaussian noise (fGn) with $H = 0.8$ is also shown as reference of monofractals ("red cross" in the figure). It is shown that the ζ_q of fGn (star-dotted line) is linear to q while the *hep1-RAL05* count process (circle-dashed line) is nonlinear, indicating multifractal scaling. This corresponds to the plot on the right: the h_q of the count process departs heavily from the horizontal line-like fGn. A multifractal model is needed to capture the scaling behavior of such patterns [24].

Table 5 shows that different levels of traces as categorized by the arrival patterns. It is concluded that most of the data-intensive traces are either pseudo-periodic, long range dependent or the combination of the two, whether it is at the cluster, the Grid, or the VO level. Certain VOs and clusters exhibit multifractal behavior (e.g., *RAL05*) and at larger scales their count processes turn to be short range dependent (SRD). For the supercomputer trace, multifractal or SRD patterns are observed, excluding long

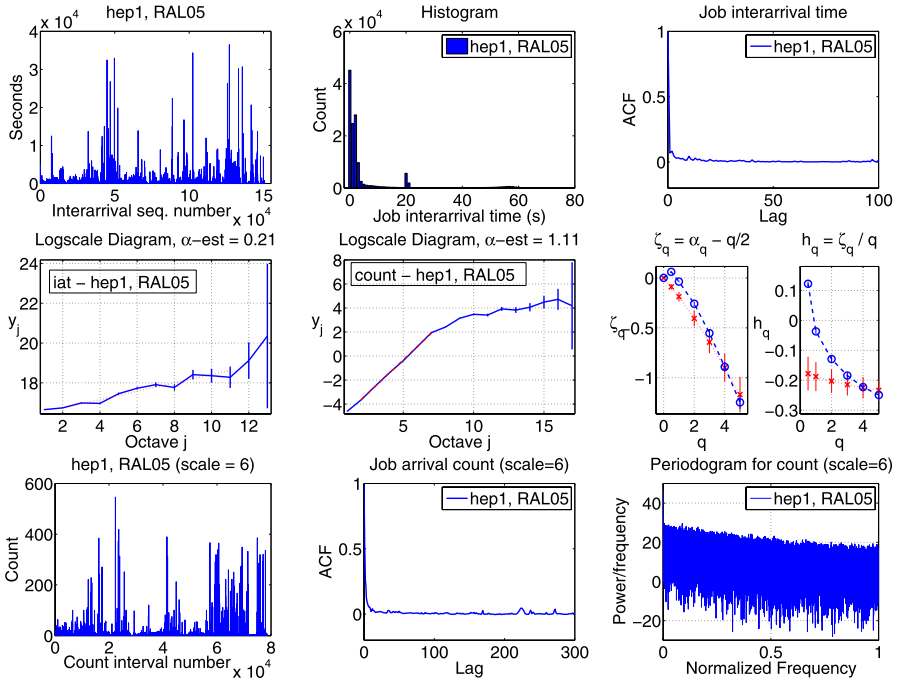

Fig. 8 Plots of the first, second order statistics and scaling analysis for both interarrival and count processes of *hep1, RAL05*. The *right figure* in the *middle row* is the multiscale diagram as explained in Sect. 2.3

Table 5 Different levels of workload traces are categorized according to job arrival patterns

Arrival Patterns	Level names
Pseudo-periodic	lhcb-LCG1, lhcb-LCG2, dteam-LCG1, dteam-LCG2, NIK05, com1-NIK05, lhcb-NIK05
LRD	atlas-LCG1, cms-LCG1, biomed-LPC05, atlas-NIK05, atlas-RAL05
LRD + Periodic	LCG1, LCG2, atlas-LCG2, cms-LCG2
Multifractals	RAL05, hep1-RAL05, SBH01, user45-SBH01
SRD	user328-SBH01, user272-SBH01

range dependence. The nature and origin of different arrival patterns are discussed in depth in Sect. 5.

4.2 Run time, memory, and parallelism

This section focuses on the workload characteristics such as run time and memory. The data is ordered ascendantly by the job arrival times and the autocorrelation function is used to examine temporal correlations in the sequence of data.

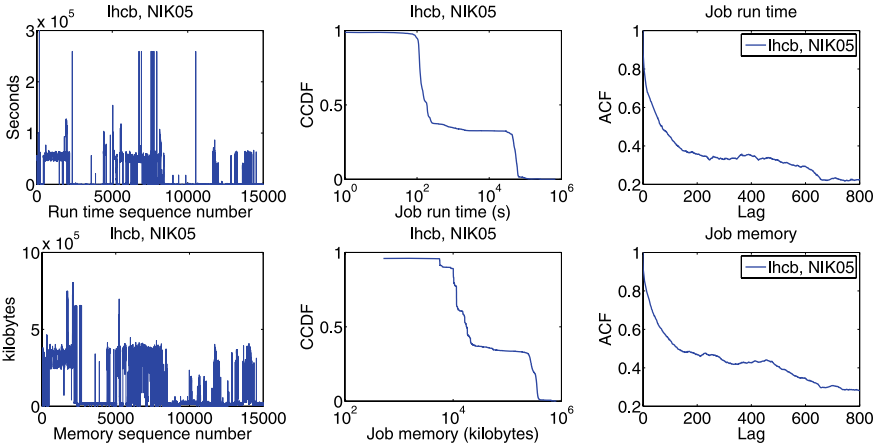

Fig. 9 The sequence plot, complementary cumulative distribution function (CCDF), and autocorrelation function (ACF) for run time and memory of *lhcb, NIK05*

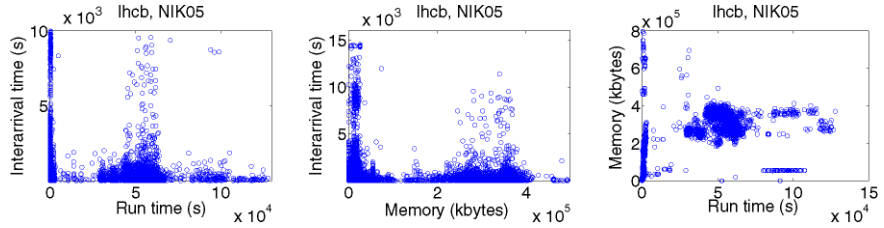

Fig. 10 Scatter plots of interarrivals, run time, and memory of *lhcb, NIK05*

4.2.1 Clusters and grids

Figure 9 plots the marginal distributions and autocorrelations for job run time and memory of *lhcb, NIK05*. The distributions of run times are highly multimodal, meaning that applications within one VO are more similar to each other with specific values of run times. Similar results are observed for memory consumption. Run times and memories with similar values also turn to appear subsequently in time, which is evidenced by the fluctuating horizontal lineups in the sequence plot. It is not surprising to see the strong autocorrelations in the sequences of run times or memories. One explanation of these observations is that the computing environment at the cluster level is more homogeneous compared to the Grid so less variations are expected on job run times and memories. The "bag-of-tasks" behavior and similarity resulted by VO categorization lead to a strong degree of temporal locality [10].

It is also interesting to see how the interarrival times are jointly distributed with the sequences of job attributes. This helps to correlate arrivals and run times (memories) and identify the "bag-of-tasks" phenomenon on data-intensive environments. Figure 10 shows the scatter plots of run times and memories against interarrival times of *lhcb, NIK05*. It is observed that job run times and memories are heavily clustered in the range of small interarrival times. This suggests that not only similar values

Table 6 Results of Pearson's and Spearman's rank correlation coefficients (CC, defined in Sect. 2.2.3) for run times vs. memories on clusters, and for run times vs. parallelism on the supercomputer

Trace	Characteristics	Pearson's CC	Spearman's Rank CC
biomed-LPC05	Run time, Memory	0.173	0.695
lhcb-NIK05	Run time, Memory	0.756	0.826
hep1-RAL05	Run time, Memory	0.013	0.456
SBH01	Run time, Parallelism	0.100	0.430

Fig. 11 Plots of run time and parallelism for a parallel supercomputer trace *SBH01*

appear in a sequence, but also times between arrivals in a sequence are relatively small. Figure 10 also contains a scatter plot of run time against memory indicating strong correlations. Correlation coefficients calculated by *Pearson's* as well as by *Spearman' rank* are given in Table 6. Among the three VOs *lhcb, NIK05* shows the strongest correlation between run time and memory. For the other two VOs, weak to moderate correlation coefficients are obtained, however, correlation coefficients are used only in combination with other measures due to their inherent limitations (especially Pearson's). It can concluded that temporal locality and "bag-of-tasks" behavior are clearly evidenced for workloads on clusters and Grids.

4.2.2 Parallel supercomputers

Figure 11 shows the statistical properties of run times and parallelism for the parallel supercomputer *SBH01*. No multimodality is detected and there is moderate to weak autocorrelations in the sequence of run times. For parallelism, a power-of-two phenomenon is clearly observed as reported in the parallel workloads literature. In this case, a power-of-eight pattern is prominent, mostly because the IBM SP has nodes with eight processors. The cross-correlation between run time and parallelism has shown diverse results [16, 18], and there is no correlations for the parallel workload under study.

5 The nature of grid workload dynamics

The focus of this paper is on production Grid environments whose workloads consist of flows of independent, computationally-intensive tasks. By looking at the cur-

rent workload structure, together with the booming factor of computing-based solutions to system-level sciences such as physics and biology, it can be envisioned that computationally-intensive applications contribute to a main part of workloads running on current and future Grids. This type of applications are also well suited to run on a heterogeneous Grid environment because of its loosely-coupled and data-parallel nature. Real parallel applications such as those on traditional supercomputers, on the other hand, are more tightly-coupled with heavy inter-process communications. Based on the different properties of applications and architectures, it is expected that cluster and Grid workloads possess structures and patterns that are different from those on parallel supercomputers. The quest starts with the origin of job arrival dynamics.

There are three patterns that are identified for data-intensive job arrivals. The first one exhibits strong periodicity, which suggests certain deterministic job submission mechanisms. *lhcb* is a large HEP experiment in the LCG Grid with the largest portion of jobs. By taking into account that close to 90% of *lhcb* jobs (around 60,000) are from a single "user" during eleven consecutive days in *LCG1*, it can be assumed that scripts are made to submit those jobs, which are deterministic in nature. It can also be interpreted that automated tasks need to be implemented to process such a huge amount of scientific data. Periodicity can also comes from testing and monitoring jobs in the Grid such as those from *dteam*. *dteam* stands for "deployment team" and it is dedicated for a continuously functioning and operating Grid. Mostly testing and monitoring jobs are initiated automatically by software in a periodic fashion. The periodic pattern is also observed for VOs at the cluster level. It is considered as a basic pattern that originates from automated submission schemes. The second pattern is long range dependent (LRD) and it applies to many production VOs. It can be partially explained by the repetitive executions of multiple specific applications. A typical user would submit sequences of tasks with a heavy-tailed inter-submission time. This behavior shows temporal burstiness, which is argued in [5] that it essentially originates from a priority selection mechanism between tasks and nontasks waiting for execution. LRD forms the second basic pattern that characterizes job arrivals on clusters and Grids. By combining periodicity and LRD, some interesting patterns emerge. The process can be long range dependent with high frequency oscillations, rooting from the short-period repetitions of job arrival rates at small time scales. The process can also be LRD with multiple lower frequency components, which is mainly due to the addictive nature of aggregation at the Grid level.

When more characteristics such as run time and memory are taken into account, "bags-of-tasks" behavior is empirically evident for data-intensive workloads. The marginal distributions for run time and memory are highly multimodal. Certain numeric values not only occur subsequently, but also turn to appear within certain bursty periods. This is because of the nature of data-intensive applications. On conventional parallel supercomputers, on the other hand, such behavior is not present in the workloads [7, 16, 18].

6 Conclusions and future work

In this paper, a comprehensive statistical study was carried out for workloads on clusters and Grids, with an emphasis on the correlation structures and the scaling behavior. It was shown that statistical measures based on interarrivals are of limited usefulness and count based measures should be trusted instead when it comes to correlations. Pseudo-periodicity, long range dependence, and "bag-of-tasks" behavior with strong temporal locality are important characteristic properties of workloads on clusters and Grids, which is not present in traditional parallel workloads. Future work naturally extends to workload modeling that tries to capture the correlation structures and patterns obtained in this paper. Experimental performance evaluation studies using simulations are needed to investigate the impact on scheduling and how to improve it under such workload patterns.

Acknowledgements The LCG Grid traces are provided by the HEP e-Science group at Imperial College London. *NIK05* and *RAL05* traces are provided by colleagues at NIKHEF (NL) and RAL (UK), respectively. *LPC05* and *SBH01* traces are obtained from Parallel Workload Archive. We want to express our gratitude to all who graciously provide us with the data.

References

1. Abry P, Veitch D (1998) Wavelet analysis of long-range dependent traffic. IEEE Trans Inf Theory 44(1):2–15
2. Abry P, Baraniuk R, Flandrin P, Riedi R, Veitch D (2002) The multiscale nature of network traffic: discovery, analysis, and modelling. IEEE Signal Process Mag 19(3):28–46
3. Abry P, Taqqu MS, Flandrin P, Veitch D (2000) Self-similar network traffic and performance evaluation. In: Park K, Willinger W (eds) Wavelets for the analysis, estimation, and synthesis of scaling data. Wiley, New York
4. Abry P, Veitch D, Flandrin P (1998) Long-range dependence: revisiting aggregation with wavelets. J Time Ser Anal 19(3):253–266
5. Barabasi A-L (2005) The origin of bursts and heavy tails in human dynamics. Nature 435:207–211
6. Chainais P, Riedi RH, Abry P (2005) On non-scale-invariant infinitely divisible cascades. IEEE Trans Inf Theory 51(3):1063–1083
7. Cirne W, Berman F (2001) A comprehensive model of the supercomputer workload. In: Proceedings of IEEE 4th annual workshop on workload characterization
8. Faubechies I (1992) Ten lectures on wavelets. BMS-NSF reg. conf. series in applied math. SIAM, Philadelphia
9. Feitelson DG (2002) Workload modeling for performance evaluation. In: Lecture notes in computer science, vol 2459. Springer, Berlin, pp 114–141
10. Feitelson DG (2006) Workload modeling for computer systems performance evaluation. draft version 0.7
11. Feldmann A, Gilbert AC, Willinger W (1998) Data networks as cascades: Investigating the multifractal nature of Internet WAN traffic. In: SIGCOMM, pp 42–55
12. Iosup A, Dumitrescu C, Epema D, Li H, Wolters L (2006) How are real grids used? The analysis of four grid traces and its implications. In: Proceedings of 7th IEEE international conference on grid computing (Grid'06)
13. Jagerman DL, Melamed B, Willinger W (1996) Stochastic modeling of traffic processes. In: Frontiers in queueing: models, methods and problems

14. Leland W, Taqqu M, Willinger W, Wilson D (1994) On the self-similar nature of Ethernet traffic (extended version). IEEE/ACM Trans Netw 2(1):1–15
15. Li H, Muskulus M (2006) Analysis and modeling of job arrivals in a production grid. In: ACM SIGMETRICS performance evaluation review, Dec 2006
16. Li H, Groep D, Wolters L (2005) Workload characteristics of a multi-cluster supercomputer. In: Lecture notes in computer science, vol 3277. Springer, Berlin, pp 176–193
17. Lowen SB, Teich MC (2005) Fractal-based point processes. Wiley, New York
18. Lublin U, Feitelson DG (2003) The workload on parallel supercomputers: modeling the characteristics of rigid jobs. J Parallel Distrib Comput 63(11):1105–1122
19. Medernach E (2005) Workload analysis of a cluster in a grid environment. In: Proceedings of 11th workshop on job scheduling strategies for parallel processing
20. Paxson V (1997) Fast, approximate synthesis of fractional Gaussian noise for generating self-similar network traffic. Comput Commun Rev 27(5):5–18
21. Polana R, Nelson R (1993) Detecting activities. In: Proceedings of IEEE CVPR
22. Riedi RH (2002) Long range dependence: theory and applications. In: Doukhan, Oppenheim, Taqqu (eds) Multifractal processes. Birkhauser, Basel, pp 625–715
23. Riedi RH, Willinger W (2000) Self-similar network traffic and performance evaluation. In: Park K, Willinger W (eds) Toward an improved understanding of network traffic dynamics. Wiley, New York
24. Riedi RH, Crouse MS, Ribeiro VJ, Baraniuk RG (1999) A multifractal wavelet model with application to network traffic. IEEE Trans Inf Theory 45(3):992–1019
25. Ross SM (2003) Introduction to probability models, 8th edn. Academic Press, San Diego
26. Song B, Ernemann C, Yahyapour R (2004) Parallel computer workload modeling with Markov chains. In: Lecture notes in computer science, vol 3277. Springer, Berlin, pp 47–62
27. Squillante MS, Yao DD, Zhang L (1999) The impact of job arrival patterns on parallel scheduling. ACM SIGMETRICS Perform Eval Rev 26(4):52–59
28. Strang G, Nguyen T (1996) Wavelets and filter banks. Wellesley-Cambridge Press, Cambridge
29. Vehel JL, Riedi R (1997) Fractional Brownian motion and data traffic modeling: The other end of the spectrum. In: Fractals in engineering. Springer, Berlin, pp 185–202
30. Veitch D, Abry P (1999) A wavelet based joint estimator of the parameters of long-range dependence. IEEE Trans Inf Theory 45(3):878–897. Special issue on "Multiscale statistical signal analysis and its applications"
31. Wornell GW (1993) Wavelet-based representations of the 1/f family of fractal processes. Proc IEEE 81(10):1428–1450

Hui Li obtained his M.Sc. and Ph.D. degree in Computer Science, Leiden University, Netherlands, in 2003 and 2008 respectively. His research interests include performance modeling, prediction, and evaluation of computer systems and services. He has co-authored over 20 papers in peer-reviewed conferences and journals. Hui Li is a member of IEEE Computer Society.