Valley Times in the Spanish Academic Network

Sergio Albandea, José Luis García-Dorado, David Muelas, Jorge E. López de Vergara, Javier Aracil
High Performance Computing and Networking Research Group
Escuela Politécnica Superior, Universidad Autónoma de Madrid, Spain
Email: sergio.albandea@estudiante.uam.es, {jl.garcia, dav.muelas, jorge.lopez_vergara, javier.aracil}@uam.es

***Abstract*—The scheduling of bulk transfers such as database distribution, resources replication and security backups, and other non-time-critical tasks has a direct effect on both the performance and cost of a network. We propose to face this problem by studying the valley times—i.e., the minimum use during off-peak periods of a network—as suitable moments to carry out such tasks. To do so, we characterize them considering the valley-hour, which we define as the opposite to the well-known variable busy-hour. Our analysis, based on 6-year-long measurements from 12 points of presence in the Spanish Research and Education Network (RedIRIS), has guided us to model the valley-hour as a Gaussian process. After that, we compare its behavior in different points and detect the main factors that explain its variance, finding significant heterogeneity. With the resulting conclusions, we have proposed a system to predict valley-hours in RedIRIS with errors below an hour for most of the cases.**

***Index Terms*—Valley times; Valley-hour; Transfer scheduling; Bulk transfers; RedIRIS.**

I. Introduction

Data transfer volumes through the Internet have grown since its creation as a consequence of well-known factors such as its vast, unstoppable spread as well as the increasing traffic demands by end users. In this scenario, some tasks including database distribution, resources replication, virtual machines cloning and security backups have become recurrent and resulted in a substantial impact in both fees and efficiency of networks worldwide. All these activities are known as bulk transfers and share two main characteristics: (i) they entail large volume aggregates and (ii) do not require an accurate timing [1].

Bulk transfers may not only play a crucial role in the cost of contracting bandwidth to an Internet Service Provider (ISP), but also in the performance of any deployed network, whatever its payment procedure runs. In the former, because charges follow typically burstable billing—i.e., built on the 95th percentile of bandwidth usage method [2]—; in the latter, owing to the potential interference of bulk on time-demanding traffic—e.g., infrastructures such as academic or banking networks, or a simple dedicated point-to-point link.

In a context increasingly inclined to virtualized infrastructures, content-aware deployments and cloud environments, the importance of scheduling bulk transfers emerges. Firstly, because the number and volume of transfers increase and, secondly, because it opens the door to collaboration in the Internet arena. Actually, alliances between Content Distribution Networks and ISPs are already a fact, with remarkable success [3]. Also, ISPs and media providers agreements result in mutual benefit. As an example, Netflix proposes to ISPs the deployment of embedded Open Connect Appliance (OCA) servers—machines which store the most popular contents closer to users, reducing latency and bandwidth needs. According to Netflix public sheets, Netflix generally implements updates from 2 a.m. to 2 p.m.[1]. As a further step, the authors in [4] designed NetStitcher, a store-and-forward system designed to exploit transmission windows considered cheaper. They defined such time periods to be fixed between 3-6 a.m. local time and equal for different places in the U.S.

Alternatively, we believe that the ideal moment to carry out bulk transfers cannot be a fixed period but they should occur when the infrastructure is, in fact, having a low usage. In this manner, the impact of such bulk transfers will be attenuated as they will use the network resources minimizing its interaction with other activities. Interestingly, bandwidth profiles according to real network measurements during off-peak periods follow a concave shape with a minimum at early morning; let us say a valley shape. However, the analysis of the dynamics of such valleys has not been comprehensively carried out by the Internet community in contrast to the study of busy periods, given its immediate implication for capacity planning and charge [5]. The main objective of this paper is to shed more light on the valleys both spatially and temporally—i.e., different networks over time—as a first step towards the definition of smart bulk transfer scheduling methods.

Specifically, we approach the characterization of valley times by proposing the daily valley-hour metric—those 60 minutes per day whose aggregated traffic is minimum; in other words, the opposite to the well-known term busy-hour [6]. In such a way, the valley-hour will soft the bursty nature of bandwidth time series, but it still remains useful to state the moment when the network use is minimum. This simplification will make further comparisons and modeling more tractable.

The study of valley times over 12 points of presence (PoP) in the Spanish Research and Education Network (RedIRIS[2]) shows that the valley-hour moment in each PoP can be modeled as a Gaussian process, which makes a formal comparison between them possible. Such comparison, supported by the ANOVA methodology, shows that each PoP behaves somehow differently and that these differences can be partially related to the month and day of the week/month when measurements

[1] https://openconnect.netflix.com/
[2] https://www.rediris.es/

were gathered. By applying these conclusions to a system in order to forecast valley times, we have found errors of less than an hour for most of the cases, highlighting the applicability of the system to schedule future transfers.

II. Valley times

We first describe the available dataset, formally define the metric daily valley-hour and how it can be modeled over time.

A. RedIRIS dataset

The data used in this work have been obtained from RedIRIS and span from January, 2008 to December, 2013. RedIRIS comprises more than a million users spread across 350 institutions, especially universities and research centers, but also hospitals and governmental buildings. RedIRIS features Internet exchange points with the European Research and Education Network (GEANT), national points, and Cogent and Level3 backbone providers, among others. Its topology is formed by different PoPs spread across the country. This way, the comparison among different PoPs, *ergo* different set of users, during 6 years will provide temporal and spatial conclusions as general as possible [7].

The monitoring system that gathers all this vast amount of data is based on NetFlow v5. NetFlow is a passive and cheap method for monitoring high-speed networks, achieving a size reduction in the order of 2000x [8] versus entire packet capture. Particularly, we have been provided with NetFlow records from 12 PoPs, which were generated with a sampling rate of 1 out of 200 packets. Assuming that during the life of a flow its bandwidth is uniformly distributed over time [9], the used bandwidth time series are specifically constructed at a minute grain. Afterwards, an extensive set of characteristics such as busy and valley hours, flow concurrence, activity per IP address, among other characteristics, are estimated.

B. Daily Valley-hour

A representative example of how daily bandwidth typically varies in RedIRIS [10] is shown in Fig. 1. Although the specific times dips/peaks occurrence can change, the shape is shared by most of the academic networks [11], and even commercial ones [12]. As remarkable characteristics, a clear concave shape at night—namely, a valley—and one or two peaks at noon or mid-afternoon during daytime. The study of the former is the key of this work as it is, intuitively, the ideal time to schedule bulk/non-time-critical tasks.

As a mechanism to synthesize, let us define formally the valley-hour as those consecutive 60 minutes during which the exchanged traffic volume is minimum. Then, we define the valley-hour moment as the half point of this interval. Finally, we have the daily valley-hour process as the succession of valley-hour moments over time in a per-day basis. We believe that such metric is a good approximation to the valley-time phenomenon.

TABLE I. Summary of averaged measurements in Mb/s

	Band-width	Valley-hour bandwidth	Valley Time	Busy-hour bandwidth	Busy Time	Dickey-Fuller
PoP_1	185	46	06:13	393	12:35	✓
PoP_2	210	86	06:01	369	12:59	✓
PoP_3	105	30	06:09	216	13:11	✓
PoP_4	1154	520	06:11	1822	13:20	✓
PoP_5	410	91	05:18	774	12:40	✓
PoP_6	240	54	06:00	528	12:38	✓
PoP_7	104	25	06:17	226	13:08	✓
PoP_8	101	29	06:14	212	12:51	✓
PoP_9	115	24	05:49	239	12:22	✓
PoP_{10}	785	380	05:53	1078	13:13	✓
PoP_{11}	1324	647	05:47	1823	13:17	✓
PoP_{12}	853	475	06:04	1177	12:40	✓

The use of an interval of one hour is borrowed from the typical planning metric busy-hour; but, importantly, note that it is a simple mechanism to identify the minimum of the bandwidth curve, avoiding that finer-scale intervals pollute the study due to the random bursts and dips that bandwidth experiments. For transfers longer than an hour, the accurate minimum of the curve will depend on both its slopes before and after the minimum. But still in this case, the valley-hour will be a good approximation to the phenomenon.

Table I summarizes the measurements gathered over the period of capture for the set of PoPs under study—labeled as PoP_n for privacy reasons. Specifically, it includes the average bandwidth in total and during both busy and valley hours, as well as the averaged valley and busy moments. Additionally, we have studied the stationarity of the daily valley-hour process. That is, during the period of measurement, we have calculated the valley moment—both per PoP and on a daily basis—and applied the Augmented Dickey-Fuller test in temporal windows. Evidences to reject the null hypothesis of non-stationarity process were found in all PoPs.

C. Modeling daily valley-hours

The comparison and prediction of the ideal daily valley-hour process may benefit from the development of a model that summarizes the behavior in few parameters. From Fig. 2(a), which illustrates the process for a given PoP, it becomes apparent that the variation is fairly bounded and that distri-

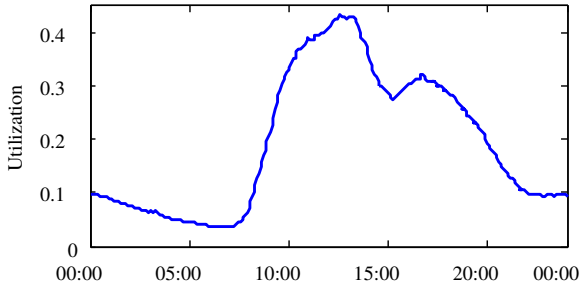

Fig. 1. Typical bandwidth daily pattern

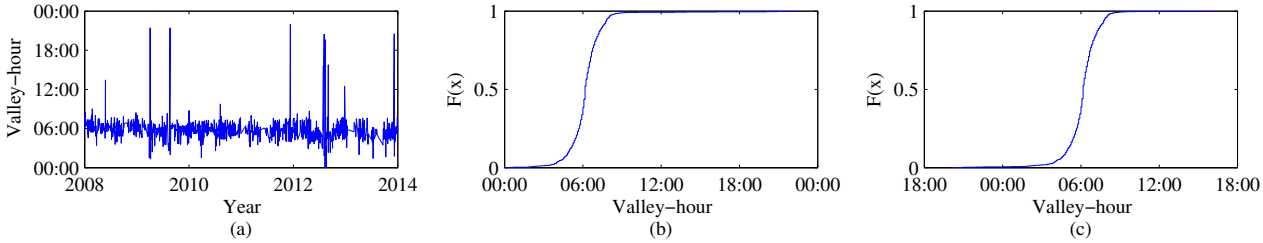

Fig. 2. Examples for the daily valley-hour of a given PoP: (a) over time, (b) CDF and (c) CDF after translation

butions such as Gaussian and uniform ones are promising candidates to characterize it. Testing analytical distributions on network measurements over long periods of time is still an open research question, as best-known statistical tests were formulated for smaller datasets and not for measurement campaigns of years of duration. In this regard, the authors in [13] pointed to a traffic-aware test based on the classical QQ-plot. In a QQ-plot, the order statistics of the empirical sample are depicted as a function of the percentiles of the other distribution—e.g., Gaussian or uniform distributions. If the data follow such distribution, then it nearly fits a straight line. Such traffic-aware test checks whether the linear correlation coefficient, R, in the QQ-Plot gives a relatively high value; say 0.9 for an equivalent power of Kolmogorov-Smirnov test at 95% confidence interval.

The second and third columns in Table II show the goodness-of-fit coefficients of the daily valley-hour process per PoP computed for both Gaussian and uniform distributions. Coefficients are not good enough to consider neither normality nor uniformity. Both tests failed because of an excess of mass in the tail of the distribution (Fig. 2(b)). By paying attention to the rationale behind this, we noticed that this is caused because of assuming a random variable that assigns midnight with the value 0, 1 to the first minute, and so on up to the last possible value (1440), the number of minutes in a day. This human vision of when a day begins and ends is—although natural—certainly arbitrary and breaks the symmetry of the distribution, since valley-hours close to midnight are artificially far from the average at early-morning.

Searching for symmetry, we define the beginning of the day at 6 p.m. and, after such translation (Fig. 2(c)), the assumption of Gaussianity can be fairly accepted in all PoPs. The new results of both tests are available in the last two columns of Table II. It is worth remarking that the uniform approximation is not very far from being a real option, suggesting some kurtosis in the sample.

TABLE II. Results for goodness-of-fit tests

	NO TRANSLATION		TRANSLATION	
	Gaussian	Uniform	Gaussian	Uniform
PoP_1	0.7076	0.6113	0.8988	0.8171
PoP_2	0.7629	0.6870	0.9617	0.9157
PoP_3	0.8337	0.7573	0.9551	0.8961
PoP_4	0.7518	0.6625	0.9235	0.8543
PoP_5	0.7213	0.6249	0.9399	0.8761
PoP_6	0.6819	0.5802	0.8620	0.7829
PoP_7	0.7607	0.6709	0.9164	0.8379
PoP_8	0.7994	0.7279	0.9432	0.8898
PoP_9	0.6992	0.6013	0.9213	0.8499
PoP_{10}	0.7594	0.6637	0.9188	0.8359
PoP_{11}	0.7959	0.7035	0.9044	0.8200
PoP_{12}	0.7223	0.6216	0.8841	0.7978

III. Applicability to RedIRIS

In this section, we characterize the factors that determinate valley-hours in the PoPs under study. After that, we leverage the conclusions to propose a system to predict valley times.

The Gaussian modeling allows us to apply to the sample one of the most used statistical methodologies, the analysis of variance (ANOVA) [14]. ANOVA is a statistical technique to analyze and quantify the effects of a set of factors on a given response variable. Its use requires some assumptions: normal distribution, stationarity, independence of observations and homoscedasticity—i.e., equal intra-group variances. On the one hand, the stationarity and normality were previously shown, and a simple test of runs proves marginal dependence between samples. On the other hand, homoscedasticity can be relaxed with a large number of samples as it is our case.

This way, ANOVA allows us to formally check if the valley-hour—the response variable—is homogeneously distributed over the full set of PoPs. The factors that are considered in this study are the PoP, day of the month, month, day of the week, and bank holiday, referred to an academic calendar. Hence, following a main effects approach, the resulting system for the response variable (daily valley time, VT) is:

$$VT_{pdmwb} = Mean + PoP_p + Day_d + Month_m + Day\ of\ the\ week_w + Bank\ holiday_b + \epsilon$$

where an observation, VT_{pdmwb}, results from the addition of the total mean, a term depending on what PoP (p) observation belongs to, the day (d, e.g., 1...31) and month (m, e.g., January...December) when the sample was gathered, the day of the week (w, e.g., Sunday...Saturday), if this day was a bank holiday or not (b, e.g., yes/no), and finally an error term that accounts for the rest of the unexplained variance.

Table III shows the results after applying ANOVA. According to them, the null hypothesis that supports the homogeneity of means can be rejected for all the factors but $Bank\ holiday$ (last column, Sig. or p-value). That means that the rest of the factors significantly contribute with some explanation of

TABLE III. ANOVA table for the valley moment (in thousands).
$\bar{R}^2_{factors} = 0.137$ $\bar{R}^2_{system} = 0.933$

Source of variation	Sum of squares	df	Mean square	F (units)	Sig.
Mean	1170561	1	1170561	127193.59	0.000
PoP	5289	11	481	52.24	0.000
Day	495	30	17	1.79	0.005
Month	877	11	80	8.66	0.000
Day of the Week	5277	6	880	95.57	0.000
Bank holiday	0.2	1	0.2	0.02	0.890
Error	105328	11445	9		
Total	1566150	11505			

the phenomenon variance. While the determination coefficient for the factors ($\bar{R}^2_{factors}$) is only 0.137, the same coefficient for the system ($\bar{R}^2_{system}$) is high. The former represents the fraction of explained variance for the factors with respect to the total mean, whereas the latter represents such fraction with respect to noise.

All this illustrates that valley-hours tend to occur at the same time—close to the mean—and also that the studied factors are significant and must be considered in the estimation of valley-hours, although the non-negligible remaining unexplained variance suggests that some other significant factors may be missed. For example, some important unknown factors are the classes of users being part of each PoP—i.e., hospitals may behave different from universities—, low-duration network outages, and that any PoP might already have applied mechanisms to carry out bulk transfers at nighttime.

Once we have defined a factorial system which reduces the uncertainty on the valley-hour occurrence, we assess the accuracy—in minutes—of its predictions. To evaluate this, we work out the parameters for the ANOVA system by using the three most recent years of our measurement campaign, two of them for training purposes and assessing the results in the following year. Fig. 3 depicts the difference between the ANOVA estimates for the valley-hour moment and when it really occurs.

The results show that for most of the cases the error is below an hour, and for about 90% of cases below two hours. This is coherent with the previous results, which showed that significant variance was not properly captured by the general approach. In other words, the system can be considered adequate to evaluate the most common cases—and so being general—but, on the other hand, some anomalous days appear to not be taken into account by the system.

IV. Conclusions

Valley time is an important metric for network management given its direct impact on the scheduling of non-time-critical tasks, such as bulk transfers. Along the paper, we have developed the idea of leveraging the valley-hour as a good synonym for valley times. We have reported daily valley-hour dynamics in an extensive set of PoPs in the Spanish Research and Education Network and modeled them over time as a Gaussian process after a translation in time. Thanks to this, we

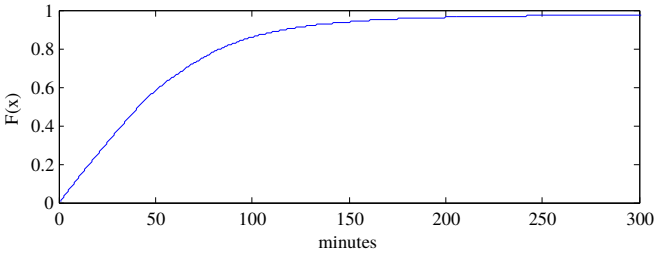

Fig. 3. CDF for errors of system's predictions in RedIRIS

have applied ANOVA as a mechanism to both formally express the variation of the response variable—the valley-hour—and describe it as an addition of per-factor terms. While we found significant heterogeneity between factors, the resulting system has been proven useful as a promising tool to predict valley-hour moments with limited excursions. As future work, we plan to include more factors, interactions and multi-hop paths, pay more attention to anomalous days, and consider other statistical methods to draw further inferences.

Acknowledgment

This work was partially supported by the Spanish Ministry of Economy and Competitiveness and by the European Regional Development Fund under the project TRÁFICA (MINECO/FEDER TEC2015-69417-C2-1-R).

References

[1] N. Laoutaris *et al.*, "Delay-tolerant bulk data transfers on the Internet," *IEEE/ACM Transactions on Networking*, vol. 21, no. 6, pp. 1852–1865, 2013.
[2] X. Dimitropoulos *et al.*, "On the 95-percentile billing method," in *Proc. International Conference on Passive and Active Network Measurement*, 2009, pp. 207–216.
[3] B. Frank *et al.*, "Pushing CDN-ISP collaboration to the limit," *ACM Computer Communication Review*, vol. 43, no. 3, pp. 34–44, Jul. 2013.
[4] N. Laoutaris *et al.*, "Inter-datacenter bulk transfers with Netstitcher," in *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 4. ACM, 2011, pp. 74–85.
[5] V. Gupta, "What is network planning?" *IEEE Communications Magazine*, vol. 23, no. 10, pp. 10–16, Oct. 1985.
[6] J. L. García-Dorado *et al.*, "Characterization of the busy-hour traffic of IP networks based on their intrinsic features," *Computer Networks*, vol. 55, no. 9, pp. 2111–2125, 2011.
[7] S. Floyd and V. Paxson, "Difficulties in simulating the Internet," *IEEE/ACM Transactions on Networking*, vol. 9, no. 4, pp. 392–403, 2001.
[8] R. Hofstede *et al.*, "Flow monitoring explained: From packet capture to data analysis with Netflow and IPFIX," *IEEE Communication Surveys and Tutorials*, vol. 16, no. 4, pp. 2037–2064, 2014.
[9] R. Sommer and A. Feldmann, "Netflow: Information loss or win?" in *Proc. ACM SIGCOMM Workshop on Internet measurement*, 2002, pp. 173–174.
[10] F. Mata *et al.*, "Detection of traffic changes in large-scale backbone networks: The case of the Spanish academic network," *Computer Networks*, vol. 56, no. 2, pp. 686–702, 2012.
[11] P. Velan *et al.*, "Network traffic characterisation using flow-based statistics," in *Proc. IEEE/IFIP Network Operations and Management Symposium*, 2016, pp. 907–912.
[12] F. Mata *et al.*, "Anomaly detection in diurnal data," *Computer Networks*, vol. 60, pp. 187–200, 2014.
[13] R. Van De Meent *et al.*, "Gaussian traffic everywhere?" in *Proc. IEEE International Conference on Communications*, 2006, pp. 573–578.
[14] O. J. Dunn *et al.*, *Applied statistics: analysis of variance and regression*. Wiley, 1974.