Malicious activity against an HPC service environment exhibits a power-law-like frequency distribution

Jae-Kook Lee, Sung-Jun Kim, Taeyoung Hong, and Minsu Joh
Div. of Supercomputer Service,
Korea Institute Science and Technology Information
Daejeon, KOREA
Email:{jklee, sjkim, tyhong, msjoh}@kisti.re.kr

Huiseung Chae
Center for Computational Science Platform,
Korea Institute Science and Technology Information
Daejeon, KOREA
Email:hschae@kisti.re.kr

***Abstract*—Many researchers often use the HPC (High Performance Computing) service at KISTI via advanced high-speed network technology. In particular, the 4th Supercomputer at KISTI provides HPC services through an internet connection via 10-Gbps high-speed networks. However, there has been an increasing number of cyberattacks targeting the supercomputing services every year. In this study, based on the collected data, we show that the number of cyberattacks against KISTI supercomputing services is increasing every year. In addition, we show that this malicious activity appears to follow a power-law distribution by categorizing the source IP addresses of these attacks based on the originating country and analyzing the frequency of these attacks.**

I. Introduction

The Supercomputing Service Center (KSC) at the Korea Institute of Science and Technology Information (KISTI) provides access to its facilities and services to as many as two hundred domestic and international research institutes, government agencies, universities, and companies through a variety of programs such as grand challenge and strategic research projects. Since 2009, the 4th supercomputer at the center, consisting of a cluster-type TACHYON 2 and large capacity memory-type SINBARAM, with a performance of 360 TFLOPS, has been in service [1]. In this study, to provide more secure supercomputing infrastructure and services, we collect and analyze access logs from login servers at the system level as well as security events from firewalls on the network level. Based on these analyses, if abnormal behavior is detected, we consider them as cyberattacks and accordingly deny the agents involved from accessing the supercomputer services [2]. In particular, we categorize the blocked IP addresses based on country codes and conduct probability or statistical analysis to determine the characteristics of these IP addresses. In general, the measurements made by scientists are of a typical size or scale; these individual measurements are largely based on the concepts described in [3]. In probability theory, normal distribution is the most widely known and used among all distributions; because normal distribution appropriately approximates several natural phenomena, it has developed into a standard of reference for many probability problems. Furthermore, a random variable with a Gaussian distribution is said to be normally distributed; such a variable is referred to as a normal deviate; for example, the heights of people and speeds of cars have found to follow normal distribution. Another representative probability distribution is the power-law distribution, which can also be observed in an extraordinarily diverse range of phenomena; for example, the populations of cities, computer file sizes, frequency of word usage in any human language, frequency of the occurrence of personal names in most cultures, number of papers scientists write, number of hits on web pages, sales of books, numbers of species in biological taxa, people's annual incomes, and a host of other variables all have been observed to follow a power-law distribution [3]. The remainder of this paper is organized as follows: Section 2 summarizes the related works about detection methods for abnormal behavior that are used for the KISTI supercomputing service environment; in addition, we briefly describe the power-law distribution in Section 2 as well. In Section 3, we describe the results of our analysis regarding the probability distribution of malicious IP addresses classified based on their country code. Finally, we provide concluding remarks and directions for future work in Section 4.

II. Related works

A. Malicious behavior detection

In order to detect abnormal behavior directed against the supercomputer, we use the following two methods. The first method involves parsing log files of all login systems of the supercomputer [2]; in particular, the method used is proposed in [2], and involves clustering source IP address or user accounts based on failed access logs of all login systems in real-time. If the total count clustered by a source IP address or user account is over a threshold value, as indicated by Equation 1, then this IP address or user account is detected as a cyberattack source, whose access to the supercomputer is then blocked. The second method involves analyzing all events recorded by firewalls in a supercomputing service network. In general, as a first line of defense, firewalls monitor and control irregular and malicious traffic flows

TABLE I
NUMBER OF IP ADDRESSES DENIED ACCESS BASED ON DETECTION ANALYSIS FROM 2013 TO 2016 ON THE SUPERCOMPUTING SERVICE ENVIRONMENT

Year	2013	2014	2015	2016
# of IPs	7,618	8,233	10,574	22,084

the from internal to external network or vice versa. In the cases of IP or Port scanning attacks, attackers try to access IP addresses or ports that are not currently serviced; these abnormal packets are then dropped by the firewall, allowing for these blocked events to be grouped by the source IP address. As in the case of the first method described above, if the number of objects in such a firewall-based group is over a designated threshold, then the events are considered as cyberattacks. In Equation 1, Δt is size of the time window and $f(e_i)$ is failed access logs associated with the same user account or blocked packet events associated with the same IP address. There can be multiple different $Threshold$ values depending on the Δt condition.

$$\frac{\sum_{i=1}^{n} f(e_i)}{\Delta t} > Threshold \tag{1}$$

B. Power-law distribution

A power-law is a functional relationship in which a relative change in one quantity gives rise to a proportional relative change in the other quantity, independent of the initial size of those quantities [4]. Thus, if a distribution satisfies Equation 2, then it is said to follow a power-law (or heavy-tailed distribution); here, α is the $exponent$ of the power-law, while C is a constant(in general, constant C is an uninteresting value)[3].

$$p(x) = Cx^{-\alpha} \ . \tag{2}$$

If plotted on a logarithmic scale, the histogram for a power-law distribution is as a straight line. Aside from those mentioned earlier, other examples of power-law distribution include the cumulative distribution of the number of times that words occur in a typical text of English or other languages, number of ”hits” received by web sites during a single day, total wealth of the richest people in the United States, size of the human population of US cities, and so on [3], [5]-[8].

III. FREQUENCY ANALYSIS OF REQUESTS FROM IP ADDRESSES

In our study, we used the detection techniques described in the previous section to detect and block IP attacks that seemed suspicious. In particular, we applied these detection algorithms to the supercomputing service environment. Table I lists the number of denied IP address from 2013 to 2016; it is clear that the number of IP addresses that were denied access owing to abnormal activity steadily increased every year.

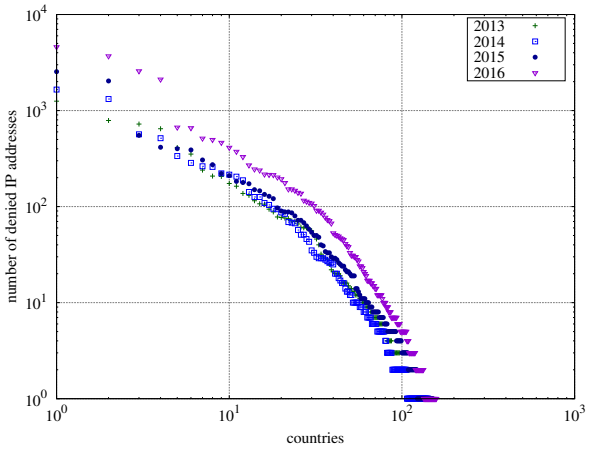

Fig. 1. Number of denied IP addresses by country from 2013 to 2016

TABLE II
TOP 10 COUNTRIES IN TERMS OF DENIED IP ADDRESS FREQUENCY FROM 2013 TO 2016

Year	2013			2014		
rank	country	count	ratio	country	count	ratio
1	CN	1253	0.164	CN	1653	0.201
2	TH	788	0.103	US	1319	0.160
3	US	723	0.095	DE	566	0.069
4	MX	647	0.085	KR	516	0.063
5	BR	412	0.054	TW	336	0.041
6	RU	352	0.046	RU	285	0.035
7	UA	240	0.032	MX	263	0.032
8	DE	208	0.027	FR	259	0.031
9	KR	206	0.027	GB	222	0.027
10	IN	174	0.023	JP	215	0.026
Year	2015			2016		
rank	country	count	ratio	country	count	ratio
1	CN	2539	0.241	CN	4587	0.208
2	US	2033	0.193	BR	3691	0.167
3	BR	552	0.052	US	2579	0.117
4	DE	415	0.039	RU	2114	0.096
5	KR	401	0.038	DE	668	0.030
6	RU	389	0.037	KR	659	0.030
7	FR	306	0.029	IN	513	0.023
8	IN	273	0.026	VN	497	0.023
9	NL	221	0.021	FR	464	0.021
10	GB	210	0.020	IT	412	0.019

In order to depict the distribution of these IP addresses, first, we converted the source IP address to a country code using an IP-to-country converter. Then, we measured the frequency of the IP address per country and plotted it on a log-log scale graph; this graph is shown in Figure 1

We compiled a ranking of countries based on the number of IP addresses that were denied from those countries. In Table II, we list the summarized results of the frequency of denied IP addresses from rank 1 to 10.

Figure 2 shows the cumulative distribution with the frequency of denied IP addresses.

In Figure 2, the x-axis represents the number of blocked IP

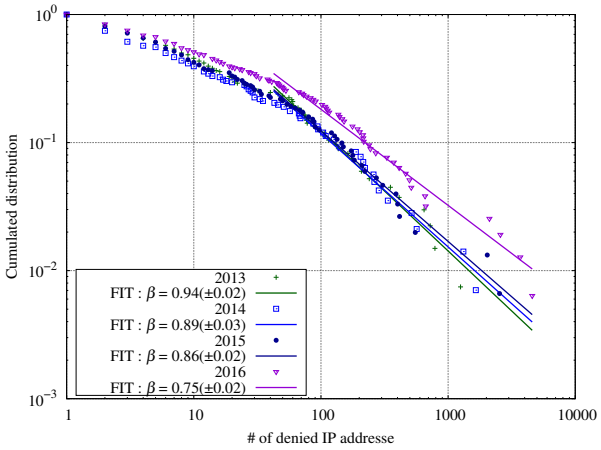

Fig. 2. Cumulative distribution of the number of denied IP addresses(2013~2016)

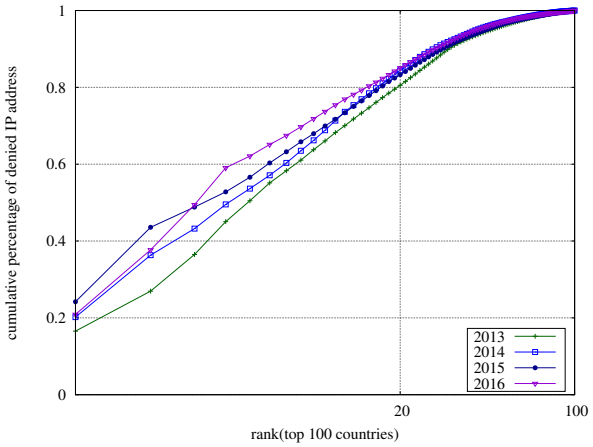

Fig. 3. Cumulative percentage of denied IP addresses collected by the KSC(2013~2016)

addresses, while the y-axis is the cumulative distribution value from 2013 to 2016. The cumulative distribution value $D(t)$ is inversely proportional to frequency of IP addresses t, which can be represented as follows:

$$D(t) \propto t^{-\beta} \ , \tag{3}$$

where $\beta = \alpha - 1$.

In Figure 2, the straight lines are obtained by performing least-square fitting using Equation 4. Through fitting, exponent α for the distribution is found to be $1.75 \pm 0.02 \geq \alpha \geq 1.94 \pm 0.02$ (because exponent β of cumulative plot in Figure 2 is well-known to satisfy $\beta = \alpha + 1$ [3]). Based on our obtained results, we can confirm that frequency of denied IP addresses follows a power-law distribution; furthermore, it is clear that the countries from which the most number of attacks are recorded, are becoming more aggressive as time elapses in that the number of attacks originating from these countries is steadily increasing.

$$D(t) = b \times t^{-\beta} \ . \tag{4}$$

The cumulative percentages of the number of denied IP addresses of the top 100 countries in 2013, 2014, 2015, and 2016 is shown in Figure 3. Though there has been a change in the rankings of countries, it can be observed that the number of cyberattacks from each country has increased every year. In addition, more than 80% of the attacks on our system originate from the top 20 countries.

IV. Conclusion

In this study, we classified malicious IP addresses attacking the KISTI supercomputing services based on country code as well as analyzed the frequency of attacks. Based on our analysis results, we confirmed that the frequency of attacks follow a power-law distribution. In addition, we observed that the top 20 countries in terms of the source of attacks were found to account for more than 80 percent of the attacks against our system. Furthermore, the rate of attacks from these top ranked countries is steadily increasing every year. In the future, KISTI is going to build a 5th supercomputer with bandwidths of 40 Gbps and 100 Gbps; as the network bandwidth increases, the number of cyberattacks might steadily increase. Therefore, to enhance the security of the supercomputer, we will continue studying these attacks to better understand the frequency of attack occurrence based on country.

Acknowledgment

This research was supported by Korea Institute of Science and Technology Information(KISTI)

References

[1] Jae-Kook Lee, Sung-Jun Kim, and Taeyoung Hong, "An Analysis of Intrusion Prevention Data against HPC Services in KISTI," *Int'l Conf. Security and Management (SAM'17)*, Jul. 2017.

[2] Jae-Kook Lee, Sung-Jun Kim, Chan Yeol Park, Taeyoung Hong, and Huiseung Chae, "Heavy-Tailed Distribution of the SSH Brute-Force Attack Duration in a Multi-user Environment," *Journal of the Korean Physical Society*, Vol. 69, No. 2, July 2016, pp.253-258

[3] Newman, Mark EJ., "Power laws, Pareto distributions and Zipf's law," *Contemporary physics*, Vol. 46, No. 5, 2005, pp.323-351.

[4] Yaneer Bar-Yam "Concepts:Power laws," *New England Complex Systems Institute*, Retrieved, 18 August 2015.

[5] G. K. Zipf, "Human Behaviour and the Principle of Least Effort," *Addison-Wesley*, Reading, MA, 1949.

[6] B. Gutenberg and R. F. Richter, "Frequency of earthquakes in California," *Bulletin of the Seismological Society of America*, 34 (4), 1944.

[7] G. Neukum and B. A. Ivanov, "Crater size distributions and impact probabilities on Earth from lunar, terrestrial planet, and asteroid cratering data," *Hazards Due to Comets and Asteroids*, University of Arizona Press, Tucson, AZ, 1994

[8] E. T. Lu and R. J. Hamilton, "Avalanches of the distribution of solar flares," *Astrophysical Journal*, 380, 1991