3
HIGH PERFORMANCE COMPUTING & CYBER SECURTY

High Performance Computing (HPC) is considered an indicator of national strength in scientific research. In this context, CSIR-4PI provides state-of-the-art HPC facility to the computational scientists and researchers across CSIR to address grand challenge problems in their frontier areas of science, engineering and technology. CSIR-4PI plans, implements, manages and maintains centralized CSIR HPC facility. All users are well connected through the National Knowledge Network (NKN). Broad spectrum of applications are provided for the diverse users spanning engineering, biology, chemistry, material science, physical sciences and other allied sciences. During 2016-17, steps were taken to place order for additional 125 Teraflops to be added to the existing 362 Teraflops Supercomputer. On implementation, this will enhance the much needed computing requirements as the present usage has touched more than 90% during 2016-17. CSIR-NAL, CSIR-NCL, CSIR-IGIB and CSIR-4PI were among the major users accounting to more than 60 % of the available computing power.

Cyber security research continued to gain momentum during 2016-17. The 12th Five Year Plan project (ARiEES) was successfully completed wherein Cyber Security Research and Observation (CySeRO) project was implemented. The potential of the same is being explored. Discussions with Cyber Emergency Response Team – India (CERT-IN) have led to development of further project proposals in collaboration with C-DAC. Industry has shown keen interest in cyber security and Artificial Intelligence. An MoU was signed with Cognizant Technology Solutions (CTS). CSIR-4PI is also actively participating in the Intelligent Systems mission project jointly led by CSIR-CEERI and CSIR-4PI.

In collaboration with Indian Centre for Social Transformation (Indian CST), Bengaluru, and National Productivity Council, New Delhi, the cloud services of ePashuhaat portal were migrated to CSIR-4PI.

Inside

3.1 Cyber security dynamics inference from unsolicited network traffic

It is well known that cyberspace consists of a wide variety of malicious activities. These activities typically include massively Distributed Denial-of-Service (DDoS) attacks, automated worm propagations, internet wide port-scanning, operating system finger printing, etc. In fact, cyberspace has several well-organized attack networks with millions of compromised hosts, which can be used to launch powerful attacks. Mirai is a typical example of one such current generation botnet involving millions of compromised networked devices running Linux Operating System and collectively exploited to launch DDoS attacks involving traffic rate of several hundreds of gigabits per second. In order to recruit new hosts to the attack network and thereby expand their size and geographical coverage, the compromised hosts in the attack network regularly scan the global Internet Protocol (IP) address space and infect vulnerable population.

Internet-wide scanning typically generates a special type of network traffic known as unsolicited network traffic. Unsolicited traffic is a potential resource for cyber security dynamics inference. CSIR Fourth Paradigm Institute (CSIR-4PI), Bengaluru, with its Cyber Security Research and Observation (CySeRO) team under CSIR 12th FYP Project, ARiEES, is actively researching on various aspects of unsolicited traffic for remote inference of cyber security dynamics. Towards this, we have experimentally deployed an active responder based framework for data curation, validation as well as for subsequent data analytics.

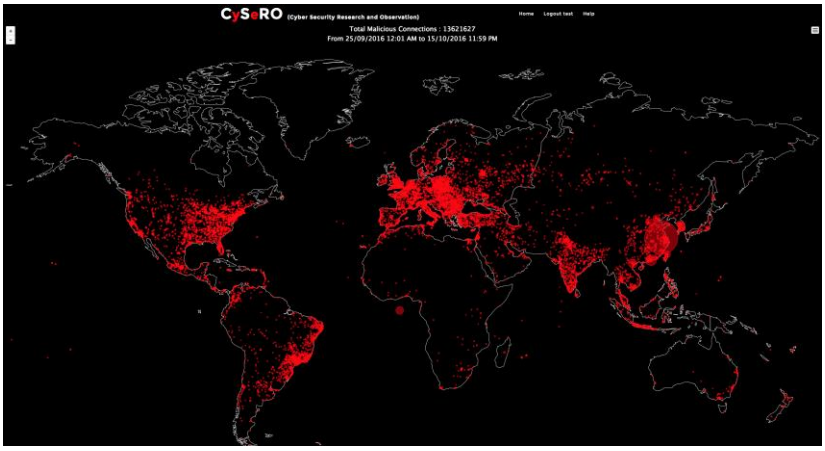

Figure 3.1 Global Map of the origin of malicious TCP connections.

The global map shown in Figure 3.1 is generated from our data collected for a period of three consecutive weeks from 25 September 2016 to 15 October 2016. It indicates that security dynamics in cyberspace invariably triggers huge amount of data. Arguably, it is one of the potential sources of 'big data' with high Velocity, Volume and Veracity. In this particular case, it consists of about thirteen millions of validated malicious TCP (Transmission Control Protocol) connections. Location of these malicious connection originators are identified using IP2 location mapping. The map provides a good visual perception on how threats are evolving on the cyberspace.

3.2 Multi path transmission control protocol: Early deployment and experimentation on cyber security testbed at CSIR-4PI

Multi Path Transmission Control Protocol (MPTCP) is an innovative next generation transport protocol being standardized by the Internet Engineering Task Force (IETF). One of major design objectives of MPTCP is to eliminate an inherent limitation of Transmission Control Protocol (TCP) that it is a single path protocol. A single path protocol, like TCP, cannot use multiple paths between

a pair of end-points in parallel for data transfer associated with a particular user session. In order to benefit from many multi-homing devices (e.g. mobile phones with WiFi and 3G/4G, servers in datacenters, etc.) through resource pooling, deployment of MPTCP is crucial. It is expected that the millions of TCP flows in the current Internet will be replaced with MPTCP flows in the near future.

With an objective of performing security and performance analysis of MPTCP in a controlled environment, the protocol is being deployment in the CSIR-4PI cyber security testbed established under CSIR 12th FYP Project ARiEES. Selected nodes in the testbed are augmented with MPTCP protocol. The augmentation includes modification of the TCP/IP module of the Linux kernel, recompilation of the kernel and setting up of multiple physical paths between nodes, enabling coupled congestion control, etc. MPTCP-aware performance monitoring tools like iperf, mptcptrace and data collection tools like 'tcpdump' and wireshark are deployed for tracking packets from MPTCP capable connections. MPTCP features like MP_CAPABLE, MP_JOIN, MP_FASTCLOSE, ADD_ADDR, REMOVE_ADDR, etc. are being analyzed from a security perspective.

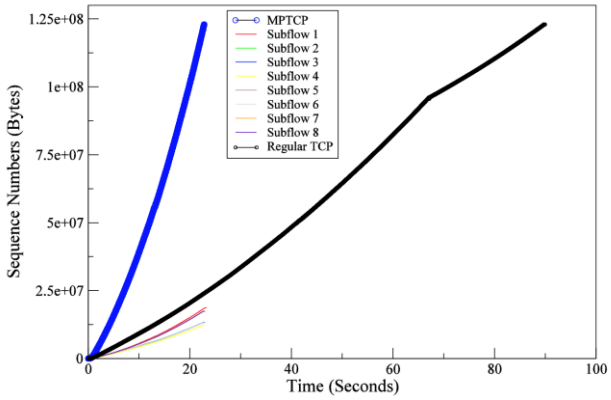

Figure 3.2 TCP and MPTCP flow dynamics in network testbed

Figure 3.2 shows one representative result in which a single MPTCP connection is established between two end-points and the end-points open x sub-flows, all belonging to the same MPTCP. Figure summarizes the results from two different experiments using the same pair of end-points. In the first experiment the end-points used iperf over single path TCP for the data transfer and in the second experiment iperf over MPTCP is used for the data transfer. The MPTCP in the second experiment consists of 8 sub-flows performing the data transfer on behave of the MPTCP and the graph labeled MPTCP is the cumulative data transfer done by all the subflows. The testbed used for the experimentations is a highly reconfigurable and observable environment consisting of 60 nodes. The testbed is hosted in a self-contained data center

3.3 Filling gaps in ocean color observation using deep learning techniques

Ocean color plays a major role in the global environmental problems related to fisheries, algal bloom detection etc. Gaps in the observed data are one of the major obstacles for carrying out research in these areas. We have attempted to use deep learning techniques using neural network for filling the gaps in the ocean color data in the Indian Ocean. We have used ocean color data obtained from Sea-viewing Wide Field-of-view Sensor (SeaWiFS) on satellite, developed by NASA. In our Neural Network (NN) model, we have taken these parameters as input and processed through multiple hidden layers each of them having 60 neurons and estimated the values of chlorophyll-a concentration (Chl-a). We have used five years of ocean color data for the training purpose and have estimated the data for the next two years. We have observed that the

accuracy of the estimated values increase with increase in the number of data points. Also, increasing the number of hidden layers up to a threshold, improves the accuracy of the estimation. However beyond the threshold any further increase does not affect the accuracy. We believe that deep learning techniques can provide requisite solution for environmental scientists to address the problem of gap filling in ocean color data.

3.4 Security analysis and improvement to permutation parity machine

Neural Cryptography is a non-number theoretical approach to derive a common key through the process of mutual learning between two participants. It makes use of light-weight Permutation Parity Machine (PPM) to generate secret key in a public channel. More details can be found in literature. However, this mechanism is prone to probabilistic attack. Probabilistic attack employs Monte-Carlo approach in the learning process during the inner rounds of PPM.

In order to mitigate probabilistic attack, a feedback mechanism is incorporated into the PPM after a few rounds which keep part of input as private in the synchronization process. This improves the percentage of overlap of legitimate participants than the percentage of overlap of attackers and makes the attacker inefficient to impersonate the synchronization process. Figure 3.3 explains the overlap of genuine participants and overlap of attacker with genuine participant as a function of outer rounds in the proposed techniques. It is observed that the attacker's overlap does not reach the point of synchronization while the participants synchronize their PPM.

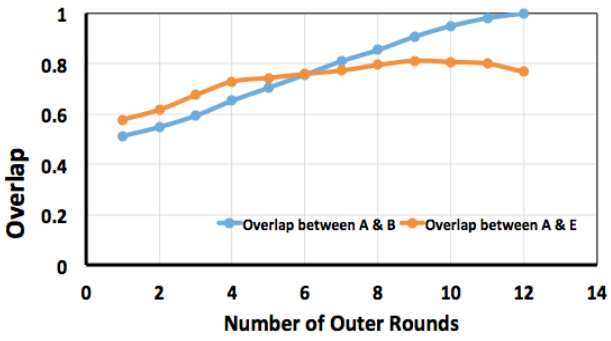

Figure 3.3 Overlap between an attacker and participants as a function of outer rounds with feedback input for the parameter G=128, K=2 and N=16 averaging the similar outer rounds

3.5 Application of homomorphic encryption to genomics

In this work we have applied Homomorphic Encryption (HE) (that allows computations to be carried out on cipher text) to Genomic Application. We have proposed an evaluation algorithm for secure computation of the Hamming distance between two encrypted DNA (Deoxy-Ribo Nucleic Acid) sequences.

Homomorphic Encryption H is a set of four functions as follows:

1. Key Generation: Client will generate pair of keys public key $p_k$and secret key s_k for encryption of plaintext.
2. Encryption: Using public key p_k client encrypt the plain text *PT* and generate E_{sk} *(PT)* and along with public key p_k this cipher text *CT* will be sent to the server.
3. Evaluation: Server has a function *'f'* for doing evaluation of cipher text *CT* and performed this as per the required function using p_k.

4. Decryption: Generated *Evalf (PT)* will be decrypted by client using its s_k and it gets the original result.

The task is to privately compute the Hamming distance between the encrypted genome sequences. Suppose that two participants have Variation Call Format (VCF) files, which summarize their variants compared with the reference genome (e.g., insertion, deletion, or substitution at a given position of a given chromosome). Variants are stored along with Reference Genome. Two VCF file format consisting of Genomes are compared, and Hamming distance is found. This VCF file is used for Secure Sequence comparison of Genome datasets. Thus Secure DNA sequence comparison is achieved by performing Homomorphic Encryption.

3.5 High Performance Computing

CSIR centralized High Performance Computing facility has been the main lifeline of the computational scientists of CSIR for last decade. “Ananta” the 360TF supercomputer is the largest supercomputer of CSIR and is hosted in CSIR-4PI and being used by more than 200 scientists working across CSIR. In addition the center hosts an ALTIX-ICE medium range HPC along with a hierarchical storage infrastructure.

Figure 3.4 CSIR centralized 360TF High Performance Computing Facility

The Ananta supercomputer (Figure 3.4) has a peak computing power of 360TF and a sustained High Performance LINPACK (HPL) of 334TF. This is currently listed as the 6th fastest system in the country. The highlight of the supercomputer has been the heavy average utilization (more than 85%) during 2016-17 and an uptime efficiency of more than 99%.

The supercomputer Ananta, is a cluster of 1088 computing nodes, distributed over 17 numbers racks. The inter-node communication is powered by high speed FDR in finiband (providing a dedicated 56 Gbps interconnect bandwidth) in a FAT tree topology. Each node is having 4GB memory per core, which results in of about 68TB of distributed memory for the total system. The supercomputer also has an online storage using LUSTRE parallel file system of about 2.1 Peta Bytes and is capable of providing a minimum of 20 Gbps simultaneous read and write performance.

Figure 3.5 shows the intra-day maximum usage in terms of number of nodes for the period 1st April 2016 till 31st March 2017 indicating a heavy requirement of computational powers by CSIR scientists.

In addition to Ananta system, the Altix ICE cluster has been of great utility (utilization by different CSIR laboratories) for running smaller to medium size computing problems. The PBSPro work load manager ensures efficient usage of the system.

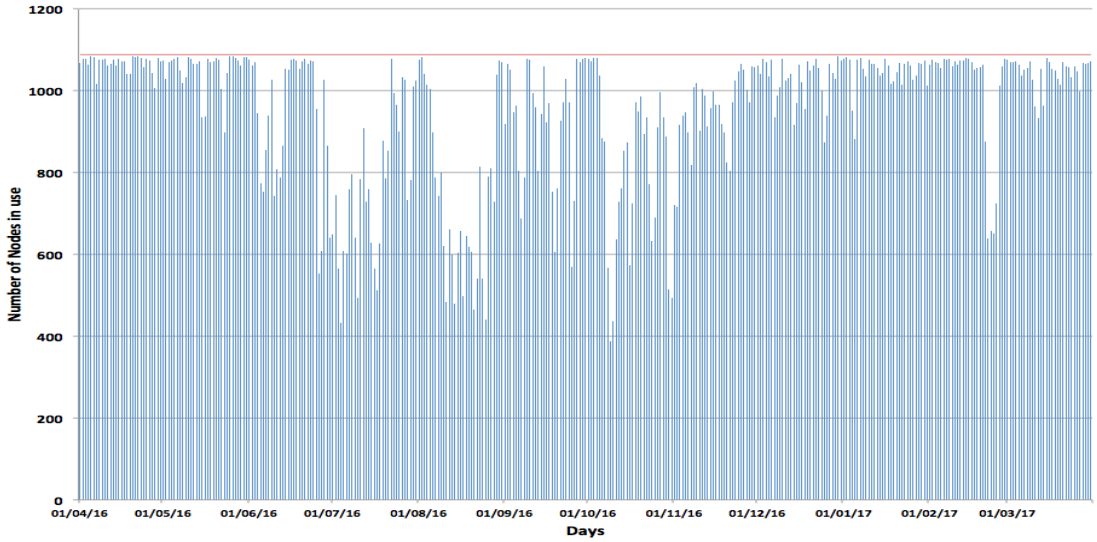

Figure 3.5 Intra-day maximum node used since the 1st April 2016 till 31st March 2017

To store and archive the results, an archival system based on a high performance 3-tired storage SAN (Storage Area Network) is established and upgraded regularly to support the growing need of storage.

The Ananta supercomputer is located in a Tier-3 equivalent state-of-the-art data center efficiently supported bya state-of-the-art energy farm. The most noteworthy component of the datacenter is the water based cooling mechanism called Rear Door Heat Exchangers (RDHx). Due to this the datacenter is one of the high density and high power efficient datacenter (Power Usage Efficiency (PUE) of less than 1.5) in the country. An energy farm consisting of two numbers of redundant compact substations of 1.25 MVA each and for ensuring 24x7 power supply to the datacenter three numbers of diesel generators, an underground diesel yard (15000 liters), three numbers of 400 KVA UPS with battery backup supports the data centre.

Steps have been taken to enhance the computing power of Ananta by placing a purchase order for 48 nodes of Intel Skylake processors with Infiniband EDR interconnect at 100 Gbps. Each node consists of two numbers of Intel Xeon 6100 series 18 core processors, 192 GB DDR4 RAM, EDR Infini band interface and will be integrated to the existing system. These nodes are expected to add an additional 125 TF computing capacity to the Ananta supercomputer making it a 485 TF system.

ePashuhaat portal

The ePashuhaat portal (epashuhaat.gov.in) of Ministry of Agriculture and Farmers Welfare, Government of India, under the Department and Animal Husbandry, Dairying and Fisheries (DADF) is migrated and hosted by CSIR-4PI under an MoU with Indian Centre for Social Transformation (Indian CST) and in collaboration with National Productivity Council. This portal is a part of the National Mission on Bovine Productivity of Government of India.

Network Facilities

The E-mail services of CSIR-4PI were migrated to cloud based NIC mail service and currently accessible through "*mail.gov.in*". An Unified Threat Management (UTM) system ensures protection of the CSIR-4PI networks as well HPC system from multiple security threats through both the NKN and ERNET links.

Other Technical Services

Technical supports were provided to a large number of users from CSIR labs including CSIR-4PI & CSIR-NAL. The team also has provided technical support for establishing HPC facilities at other CSIR laboratories. This includes the HPC system at CSIR CIMFR, Dhanbad under the collaborative research project Deep Coal. The team also has audited the OneCSIR portal and has provided recommendation to improve the security. In addition, several students from academic institutions across the country have availed the computing services as part of their academic work at CSIR-4PI under the SPARK program. Technical advices and consultancies were provided to various institutions within and outside CSIR.