3

HIGH PERFORMANCE COMPUTING & CYBER SECURTY

CSIR-4PI continues to provide High Performance Computing (HPC) facility to the computational scientists and researchers of CSIR to enable them to address Grand Challenge problems in their frontier areas of science and engineering. The HPC facility has been operational on a 24x7 basis with high uptime efficiency. Remote access from all across CSIR to the facility is provided through the high speed, redundant National Knowledge Network (NKN). The group is also actively involved in research on various aspects of cyber security and cryptography. In particular, two new Grand-in-Aid R&D projects in the area of cyber security have been initiated with funding from Ministry of Electronics and Information Technology (MeitY) and Department of Science and Technology (DST). A mission mode project funded by CSIR on Intelligent Systems has been carried out. A vehicular test-bed, which is unique in the country, has been established. The group obtained an European Union Patent granted from the European Patent Office (EPO) for one of their inventions on security aspects of next generation transport protocol.

Inside

- Map Flooding in Multipath Transmission Control Protocol
- Estimating Safe Silent Period based on the Parameters of Neighboring Vehicles that Protects Safety, Security and Privacy
- Vehicular Test-bed Based on Duckietown at CSIR-4PI
- Short Term Wind Speed Prediction Using Time Division Ensemble Technique
- Chlorophyll Prediction Using Ensemble Deep Learning Technique
- High Performance Computing

3.1 Map Flooding in Multipath Transmission Control Protocol

Multipath Transmission Control Protocol (MPTCP) is an innovative transport protocol, which is currently a proposed standard of the Internet Engineering Task Force (IETF). Its primary goal is to bring multipath capability to Transmission Control Protocol (TCP), the de facto connection oriented reliable transport protocol of the Internet. MPTCP is a complex protocol with several features including a set of newly added signaling options that needs to be exchanged regularly between MPTCP capable endpoints. Data Sequence Map (DSM) is one such crucial information being exchanged between the endpoints, typically, at packet granularity. In this work, we analyse the DSM exchange mechanism in MPTCP and identify the possibilities of some potential misbehaviors by MPTCP endpoints. In particular, we show a scenario, which we call "*Map Flooding through Map Skipping and Splitting*" which is illustrated further in Figure 3.1.

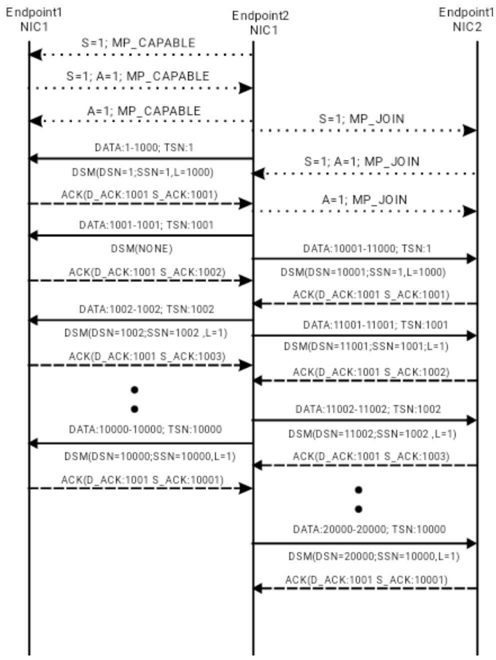


Figure 3.1 Map flooding through map skipping and splitting

In Figure 3.1, application data 1-10000 bytes is mapped for transmission through subflow1 (i.e., Endpoint2 NIC1 and Endpoint1 NIC1) and 10001-20000 bytes is mapped for transmission through subflow2 (i.e., Endpoint2 NIC1 and Endpoint1 NIC2). Except the first data segment in each subflow, all other data segments are one-byte-segments. Map skipping is performed in the first one-byte-segment in subflow1, and all other one-byte-segments are transmitted with their corresponding tiny map. The response of the map receiving endpoint to tiny maps transmitted after map skipping can be seen in the D_ACK and S_ACK of the acknowledgement segments. As data reception at subflow-level is continuous without any break or hole in the subflow sequence space, S_ACK increases. However, as map for data 1001 is skipped, this byte cannot be acknowledged at connection-level until the skipped map is transmitted. Hence, D_ACK in all acknowledgement segments is 1001. Each one-byte-segment transmitted after map skipping needs to be buffered along with its 10-byte map leading to numerous in-memory tiny maps consuming memory.

3.2 Estimating Safe Silent Period based on the Parameters of Neighboring Vehicles that Protects Safety, Security and Privacy

In order to improve the privacy of connected vehicle, use of pseudonyms rather than canonical identities and changing them in a predefined manner along with a silent period is proposed as a potential solution. IEEE (US) and ETSI (EU) standards have adopted this. However, there are safety concerns associated with the silent period.

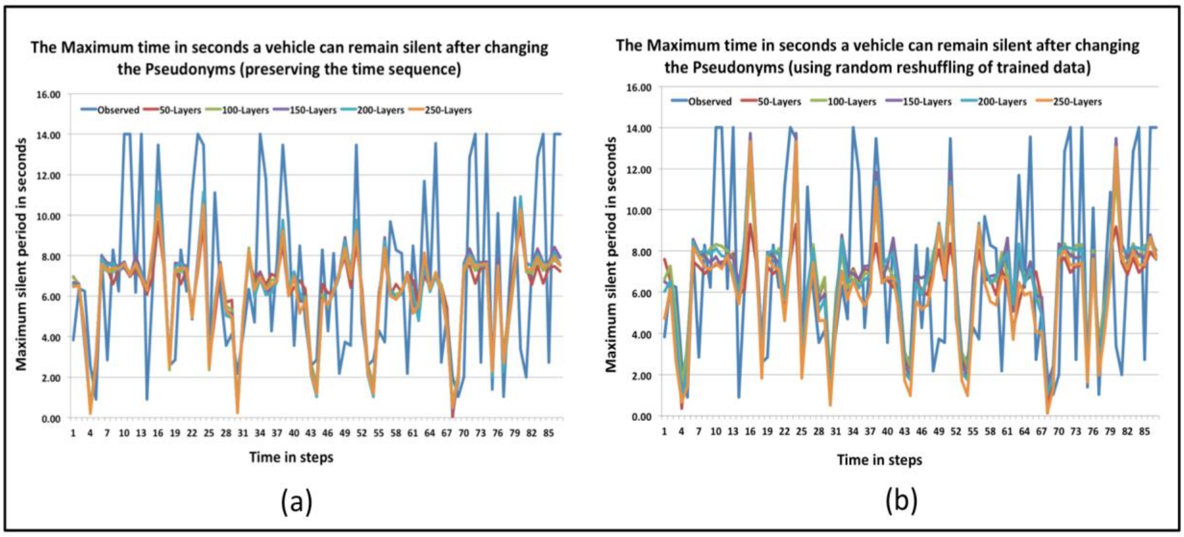


Figure 3.2 Estimation of safe silent period in seconds for vehicles in a VANET

We have developed a Deep Learning (DL) based mechanism based on Long Short Term Memory (LSTM) that considers safety as a parameter to decide on the maximum possible silent period at a particular instant of time. In the absence of real implementation, we devices a novel mechanism to generate data from the simulator. We used stacked LSTM for learning and estimating a safe silent period. In the absence of real data, we generated data for our studies in an innovative way from vehicle simulator. We used the data in two modes, one by preserving the time sequence [figure 3.2 (a)] and other by randomly shuffling [figure 3.2 (b)]. The randomly shuffled data provides a better estimation of safe silent period.

3.3 Vehicular Test-bed Based on Duckietown at CSIR-4PI

As part of the CSIR mission project on Intelligent Systems, a unique vehicular test-bed based on Duckietown was established at CSIR-4PI. It is an inexpensive research platform for studying autonomy in complex systems. Here, each vehicle is capable of identifying the road and signal infrastructures and autonomously move and communicate with each other. This is useful for testing algorithms and protocols. This is also configured with a GYM Simulator, which provides a Direct Code Execution environment, where the algorithms can be tested and loaded on to the vehicles for implementation and validation.

Figure 3.3 Duckietown Test-bed setup at CSIR-4PI

3.4 Short Term Wind Speed Prediction Using Time Division Ensemble Technique

Wind power is one of the most popular modern and sustainable renewable energy source among other renewable energy sources like hydro power, bio-mas energy, solar energy, ocean energy etc. In power and energy sector, electric power generation from wind energy has gained a huge success as it is renewable and pollution free. Accurate prediction of wind speed and its frequency distribution help in calculating amount of electric power generation. But accurate wind speed prediction is quite challenging due to intermittent and stochastic behaviour of wind speed.

We have proposed an enhanced technique named Time Division Ensemble based on Long Short Term Memory (LSTM), a Deep Learning approach, for improving the accuracy of wind speed prediction, and have achieved substantial improvement in terms of accuracy. For our studies we have considered data during 2010 to 2014 from two towers, one tower in New Delhi in north India and the other in the city of Bengaluru in south India. We have trained our proposed model using data for the period 2010 to 2013 and have tested the model with data for the period 2014. We have used the four meteorological variables i.e. temperature (t), pressure (p), humidity (h) and wind speed (v) available at 30 minutes averaged interval. The number of hidden neurons, hidden layers and number of epochs are tuned continuously for arriving at the least error possible. We have compared the results of the improved technique with normal LSTM, two other data driven approaches i.e. Extreme Learning Machine (ELM) and Support Vector Machine (SVM).

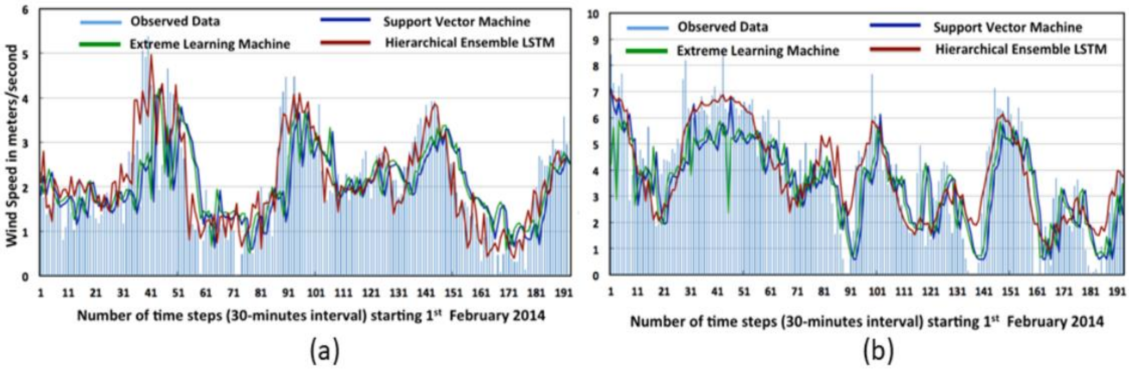


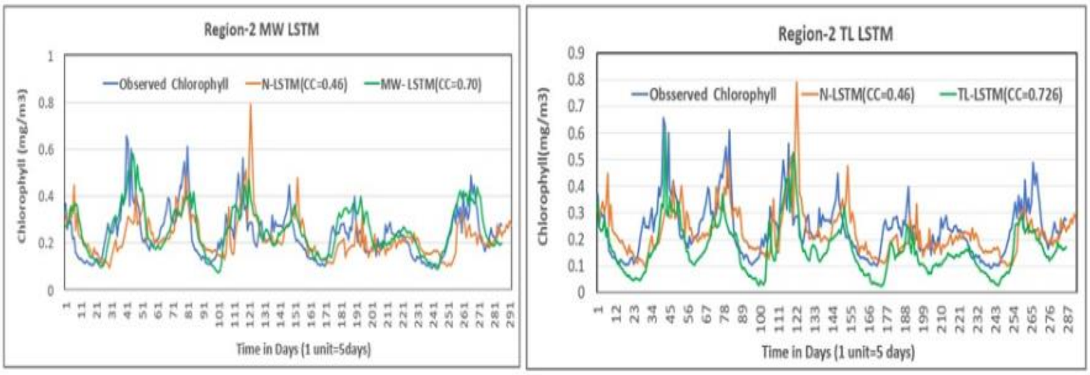
Figure 3.4 Comparison of predicted wind speed from Time Division LSTM, normal LSTM, ELM, and SVM vs. observed wind speed (a) at New Delhi and (b) at Bangalore.

As shown in Figure 3.4, the result using proposed method (Time Division LSTM) has performed better than other methods like normal LSTM and Extreme Learning Machine (ELM) and Support Vector Machine (SVM).

3.5 Chlorophyll Prediction Using Ensemble Deep Learning Technique

Understanding the various use? cases of oceanic chlorophyll, like its capability in identification of Potential Fishing Zone (PFZ), in providing a reliable, timely advisory on the potential zones of fish aggregation which will benefit the fishing community, etc. and recent development of Deep Learning technique in the domain of Artificial Technology, we have proposed two ensemble methods which are enhancement of Long Short-Term Memory

(LSTM), a Deep Learning approach, to predict oceanic chlorophyll with two months lead time. Our proposed two algorithms named Moving Window LSTM and Time Lagging LSTM are based on ensemble technique prepared by distributing input sequences. In Moving Window LSTM, ensemble data sets are prepared by dividing the input data set into fixed sized multiple moving windows. In Time Lagging LSTM, ensemble data sets are built upon by considering different starting point of input data. For this experiment, a portion of the Arabian Sea (65E:72E, 12.5N:15N) has been selected as the study area for the duration 2004 to 2016. As per AI requirement, the whole dataset is divided into training set which consists of data collected during 2004 to 2012 and the data collected during 2013 to 2016 is considered as testing set. The data are collected in five-day intervals. In both the proposed models, Sea Surface Temperature (SST), Sea Surface Salinity (SSS), Sea Surface Heights (SSH), Chlorophyll are considered as input and output is two months chlorophyll.



(a) (b)

Figure 3.5 Comparison of predicted vs. observed chlorophyll from Moving Window LSTM (a) and Time Lagging LSTM (b).

It is observed from Figure 3.5 that prediction from Moving Window LSTM and Time Lagging LSTM are better than prediction from Normal LSTM.

3.6 High Performance Computing

CSIR centralised High Performance Computing facility; the “Ananta” Supercomputer continues to serve the computational scientists working across CSIR for a record 7th year successfully (Figure 3.6). This is still the major computing workhorse for solving compute intensive scientific problems in areas of Aerospace, Biology, Computational Chemistry, Environmental Science etc.

Figure 3.6 CSIR centralised 489 TF High Performance Computing Facility.

The HPC system with a combined computing capability of 489×10^{12} floating-point operations per second (TF) comprises of two generations of systems. One based on the older generation processor from Intel namely Sandy Bridge, with a peak computing capability of

362 TF and the other based on Intel Skylake, a newer generation processor, with a peak computing capability of 127 TF. The High Performance LINPACK (HPL) benchmark, normally used for getting an idea about application performance, is 334 TF and 84 TF for the Sandy Bridge and Skylake based system respectivey. In addition, the Institute continues to host the SGI Altix-ICE medium range HPC along with a hierarchical storage infrastructure, which has been is use for more than 10 years.

The Ananta Supercomputer, is a cluster of 1088 computing nodes, each with two numbers of eight core Intel Xeon E5-2670 processor and 48 numbers of newer generation nodes, each with two numbers of eighteen core Intel Xeon Gold 6140 processors, distributed over eighteen water cooled racks. The older nodes have 64 GB memory per node, while the newer nodes have 192 GB memory per node. The inter-node communication for the older system is powered by fully redundant high speed FDR infiniband switches, with a dedicated 56 Gbps inter-connect bandwidth in a FAT tree topology and the newer nodes are powered by EDR infiniband switches of 100 Gbps inter-connect bandwidth. The total memory of the system is about 77 TB. The supercomputer is supported by an online storage using LUSTRE parallel file system of about 3 Petabyte of raw capacity and is capable of providing a minimum of 20 Gbps simultaneous read and write performance. PBSPro workload manager ensures efficient usage of the system.

Usage of Sandy Bridge processor based nodes is about 8.8 million nodes hours in 2019-20, which amounts to more than 90% usage on 24x7 basis. Usage of Skylake nodes is about 1.74 lakh node hours in the reporting period. The usage pattern of the Sandy Bridge nodes is given in the Figure 3.7

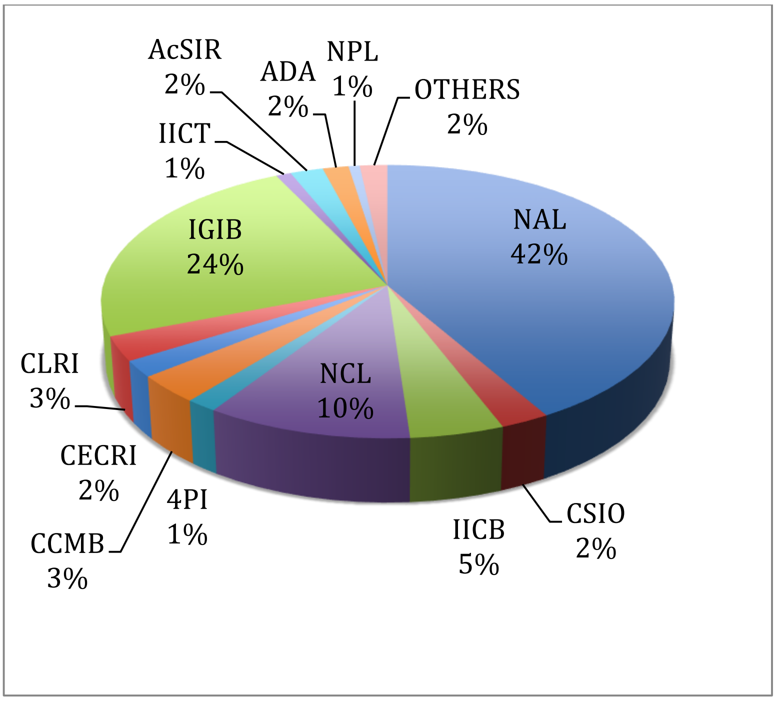


Figure 3.7 Usage pattern of Sandy Bridge nodes in 2019-20

The support infrastructure plays an important role in smooth running of the system. Tier-3 equivalent state-of-the-art data center, supported by the state-of-the-art energy farm is key for the system being in production mode still after 7 years. The most noteworthy component of the datacenter is the water based cooling mechanism called Rear Door Heat Exchangers (RDHx) which has resulted in providing one of the best power and space efficient datacenter (Power Usage Efficiency (PUE) of less than 1.5) in the country. The energy farm consists of two numbers of redundant compact substations of 1.25 MVA, three numbers of 1010 KVA diesel generators, an underground diesel yard (more than 15000 litres) and three numbers of 400 KVA UPS with battery backup for ensuring 24x7 power supply to the datacenter.