RISC 2016: Research Symposium of Department of CSE, IIT Bombay

Abstracts

2nd April, 2016

Welcome

We welcome you to RISC 2016, the research symposium organized by Department of Computer Science and Engineering of IIT Bombay, India. This abstract booklet contains abstracts of all researchers who will be presenting their work at the symposium.

The abstracts are divided across two chapters: sprint thesis talks, and posters. These abstracts were requested in two categories: early RS (Research scholars at an early stage of their research) and senior RS (Research scholars at an advanced stage of their research).

We are positive that you will find these abstracts interesting. We are happy to have you here, and urge you to engage with the RS and their supervisors.

RISC 2016 Organizing Team
Prof. Sharat Chandran
Meghshyam Prasad
Aditya Joshi
Preeti Gopal
Anshuman Dhuliya
Thyagarajan Radhakrishnan

Index

Sprint Thesis Talks

NLP Resource Creation and Enrichment using Deep Learning
Scheduling Trains on a Railway Network
Computational Sarcasm
Optimization of Database Application Programs by Rewriting to SQL
Optimizing Remote Calls in Parallel Data Management Systems
Generating Program Analyzers (GAP)
Towards Precise Software Verification
Power Considerations in Virtualized Enviroments
Framework for Optimal NFV Architecture in Cellular Networks
Large scale Multi Armed Bandit and MDP
Flow and Context Sensitive Points-to Analysis using Higher Order Reachability
Design and Implementation of an Espionage Network for Cache-based Side Channel Attacks on AES
Mosaicing Scenes with a Quadcopter
TCP Download Performance in Dense WiFi Scenarios: Analysis and Solution
Process Edification
Complimentary Usage of GPUs in Virtualized Systems
Graceful Degradation of QoS in Smart Grid
Theory of Dynamic Evolution in Petri Net Models of Workflow Processes
Improving Performance of a Network Function Virtualization (NFV) System
Segment Slicing
Nested Containers: Containers within a Virtual Machine
Multi-slice Tomographic Reconstruction: To Couple or Not to Couple
Colocalization Estimation using Statistical Analysis and Modeling of Images
A Finitary Analogue of the Downward Lowenheim-Skolem Property
Heap Abstractions for Static Analysis
Decomposition of automata
Mitigating BootStorms in Virtualized Datacenters
Smart Energy Management in Buildings
Design Robust local Image Descriptor Using ConvNet
Evolutionary Design of Robotics
Using Machine Translation Evaluation Techniques to Evaluate Text Simplification Systems

Posters

A Statistical Model for Smooth Shapes in Kendall Shape Space

Multilingual Named Entity Recognition using Deep Learning
Tracking Unsafe Stopping Behaviour of Buses
JalTantra: Tool for Optimization of Piped Water Networks
A Comparison of Some Direct 2D Reconstruction Methods using Discrete Projective Mathematical Transforms
A Computational Approach to Automatic Prediction of Drunk-Texting
Optimizing Remote Calls in Parallel Data Management Systems
Design and Implementation of an Espionage Network for Cache-based Side Channel Attacks on AES
Mosaicing Scenes with a Quadcopter
Realizability and Games in Distributed Systems
How hard can it be? The E-Score - A scoring metric to assess the complexity of text
A Finitary Analogue of the Downward Lowenheim-Skolem Property
Substring-based unsupervised transliteration with phonetic and contextual knowledge
Scanpath Complexity: A Measure of Cognitive Load for Text Reading and Annotation
Civique: Using Social Media to Detect Urban Emergencies
A Logical Characterization for Dense-Time Visibly Pushdown Automata
Noun Compound Interpretation

Chapter 1

Sprint Thesis Talks

1. **Title**: NLP Resource Creation and Enrichment using Deep Learning
   Collaborators: **Kevin Patel**, Prof. Shivaram Kalyanakrishnan, Prof. Pushpak Bhattacharyya

   Recent advances in the field of Deep Learning have made it possible to learn features automatically, sparing one from tedious feature engineering. Deep Learning based solutions hold the current state of the art in many tasks in the field of Natural Language Processing (NLP). Traditionally, NLP has relied significantly on linguistic resources. The quality of the resource being used decides the bounds of the performance of any solution for a particular task. These resources are hand-crafted by expert linguists, and their creation and maintenance is a costly activity. In our work, we explore the possibility of using deep learning to enrich existing resources, and creating new resources that can aid NLP. We demonstrate certain investigations such as detecting novel sense of word in running text (that can be added to dictionaries), handling multiword expressions (with focus on noun compounds), linking of multilingual wordnets, etc.

2. **Title**: Scheduling Trains on a Railway Network
   Collaborators: **Apoorv Garg**, Prof. A.G. Ranade

   Train scheduling and re-scheduling are complex problems which are routinely faced in railway operations. The latter problem, also called 'train-dispatching', is more challenging due to its real-time nature. Due to lack of efficient computer algorithms, in practice it is solved manually in a distributed manner by a hierarchy of human experts called 'dispatchers'. For a theoretical analysis of these problems, we study the analogous 'packet scheduling problem with bounded buffers'. We examine the problem on a simple network topology: a unidirectional line (path graph) with some intermediate nodes being 'junction nodes' where trains/packets can enter or leave the line. Efficient algorithms or complexity results are not yet available even for such simple cases. We present some preliminary results, including a 2-approximate algorithm for the flow-time objective.

3. **Title**: Computational Sarcasm
   Collaborators: **Aditya Joshi**, Pushpak Bhattacharyya, Mark Carman with Vinita Sharma, Anupam Khattri, Anoop Kunchukuttan

   Sarcasm is a form of verbal irony that is intended to express contempt or ridicule. Computational sarcasm refers to approaches pertaining to sarcasm: sarcasm generation, detection and sarcasm understanding in humans.

4. **Title**: Optimization of Database Application Programs by Rewriting to SQL
   Collaborators: **K. Venkatesh Emani**, Sudarshan S.

   Optimizing the performance of database applications is an area of practical importance, and has received significant attention in recent years. In our work we present an approach to this problem which is based on extracting a concise algebraic representation of (parts of) an application, which may include imperative code as well as SQL queries. The algebraic representation can then be translated into SQL to improve application performance, by reducing the volume of data transferred, as well as reducing latency by minimizing the number of network round trips. Our techniques can be used for performing optimizations of database applications that techniques proposed earlier cannot perform. Our experiments indicate that the techniques we present are widely applicable to real world database applications, in terms of successfully extracting algebraic representations of application behavior, as well as in terms of providing performance benefits when used for optimization.

5. **Title**: Optimizing Remote Calls in Parallel Data Management Systems
   Collaborators: **Bikash Chandra**, S. Sudarshan

   Consider a parallel data processing framework running an application on a cluster. This framework could be a MapReduce pipeline in Hadoop or Spark, or a stream processing application running on Storm, Muppet or S4. Many such applications require stored data to be accessed from a parallel data store. For some of these applications, it may also be necessary to perform computations based on the accessed data. Accessing remote data synchronously for each input data may lead to poor performance in terms of throughput. We present novel runtime techniques for frequency based caching and dynamic load balancing, to dynamically optimize data access and function call execution in a parallel data processing framework which could be based on stored or streaming data. The load balancing takes into account CPU, network and I/O costs as well as the load at clients and servers. We also describe how to extend the APIs of the MapReduce and the Muppet stream processing platform to incorporate (possibly batched) prefetching. We implement our optimization techniques, using HBase as the parallel data store. Our experiments show that our optimization techniques provide up to 5 times improvement in throughput.

6. **Title**: Generating Program Analyzers (GAP)
   Collaborators: **Anshuman Dhuliya**, Uday Khedkar

   Program Analysis has a rich set of abstractions that facilitate the design of sound analysis. It can be used for various purposes like code optimization and software debugging. Although the abstractions help us design sound analysis, the lack of effective algorithms to automatically transform these analysis into an efficient analyzer, forces manual labour. This not only makes it painful on the part of the designer but also the resulting analyzer becomes error prone. There have been attempts to develop automated tools that bridge the gap between the abstractions in Program Analysis and a working analyzer. We present some of these analyzer generators here and propose a new framework of analyzer generator called SpecDFA, that tries to solve some fundamental flaws in the existing ones.

7. **Title**: Towards Precise Software Verification
   Collaborators: **Divyesh Unadkat**, Prof. Supratik Chakraborty with Prof. Ashutosh Gupta

   Software is ubiquitous. From aviation systems to autonomous cars, entertainment systems to home appliances, software is an integral part. A bug in such systems can result in loss of millions of dollars in business-critical systems and human lives in safety-critical systems. Rigorous analysis before deployment is unavoidable as advocated by several standards such as DO178B and ISO26262 for safety-critical systems. Static program analysis discovers program properties without executing the program and can prove the absence of bugs from various classes such as division-by-zero, array-index-out-of-bounds, null-dereference and so on. However, a large number of false warnings are reported due to imprecision in the analysis, hampering practical applicability. In recent years, there has been considerable progress in the development of analysis techniques that harness the reasoning power of tools in a modular way. A prime example is that of Satisfiability Modulo Theories (SMT) solvers in which multiple decision procedures co-operate to decide the satisfiability of a formula in a combination of theories. At the heart of such solvers lie complex procedures and algorithms for effective learning and exchange of information. Inspired by the success of SMT solvers, we develop static analyzers that can harness the power of existing but complementary analyzers. By effectively learning and exchanging information between the analyzers through an appropriate logical framework, we make each one of them more precise than otherwise possible. We present illustrative example that motivates this idea. It remains to check if the technique scales to large enough programs.

8. **Title**: Power Considerations in Virtualized Enviroments
   Collaborators: **Swetha PT Srinivasan**, Umesh Bellur

   Energy takes upto half the operational expenses of datacenters making

energy conservation a critical goal. Advances in processor technology provide fine-grained control over operating frequency of processors and this control can be used to reduce power at the cost of performance degradation. Placing tasks on a set of servers is often formulated as a bin packing problem where the aim is to pack tasks in the least number of servers. However, our investigations reveal that tightest packing is not the most energy efficient due in part to the fact that the duration of task allocation and higher frequencies needed for tighter packing are not taken into consideration. In this talk, we explore how to provision tasks in an energy optimal manner. We show that the provisioning and placement problem is really one of variable-sized bin packing and then explore heuristic solutions for the same. We analyze energy efficiency of 14 classical bin packing heuristics and their variations for different distributions of task deadline requirements. Our heuristics saves as much as 45

9. **Title**: Framework for Optimal NFV Architecture in Cellular Networks
   Collaborators: **Akanksha Patel**, Mythili Vutukuru

   Cellular networks need several middleboxes such as MME, S-gateway and P-gateway to serve users. However, these middleboxes come as expensive and proprietary hardware that require specially trained personnel for deployment and maintenance, hence incur high CAPEX and OPEX, and are hard to scale. Network function virtualization (NFV) is emerging as a promising technology to reduce the overall cost for telecom operators by replacing the hardware middleboxes with software middleboxes, referred to as Virtualized Network Functions (VNFs). Each such VNF comprises of several processing functions, and each processing function accesses specific user states. Existing works try to figure out the optimal placement of VNFs or processing functions at possible locations while minimizing the associated cost. But, each VNF comprises of same processing functions as that of its hardware counterpart. We want to leverage the possibility of revisiting this split by deciding which processing functions and user states shall be co-located in one component. Our framework revisits the architecture of VNFs, redefines interfaces to decide the optimal placement of processing functions and user states in telecom operator's infrastructure with the goal of minimizing the cost of communication among processing functions, and between processing functions and user states.

10. **Title**: Large scale Multi Armed Bandit and MDP
    Collaborators: **Arghya Roy Chaudhuri**, Prof. Shivaram Kalyanakrishnan

    Markov Decision Process(MDP) and Reinforcement Learning(RL) has been subject of interest for almost a century. At the middle of last century the advent of Dynamic Programming has shown promising directions to tackle the combinatorial optimizations to solve general class of problems in this field. But for solving a large scale MDP and RL problems this dynamic programming approaches are not sufficient due to their intractability of state-action space. It is shown that even some toy problems in MDP

may lead to exponential running time of dynamic algo on such instances. Thanks to the research on Approximate Dynamic Programming(ADP) that works remarkably well despite these negative results. Multi Armed Bandit(MAB) is a branch of RL in which the objective is to output a strategy that maximizes the gain for given a set of slot machines. Considerable amount of research has been done on settings like Regret Minimization, PAC etc. The applicability of MAB ranges from Online advertisement, Drug testing, Automated game playing to scenarios where exploration and exploitation tradeoff plays the main role to optimize the objective. Our current interest of research deals with infinite armed bandit problem which can be exemplified by a set of infinite number of slot machines. In practical domain often such scenario is encountered where number of options is too high to manipulate all of them to solve the problems in real time/near real time. Our algorithm is capable of performing being independent of the number of arms. We are able to find a "good" arm in a pool of large/infinite number of arms by sampling efficiently.

11. **Title**: Flow and Context Sensitive Points-to Analysis using Higher Order Reachability
Collaborators: **Pritam Manohar Gharat**, Prof. Uday Khedker

    The bottom up interprocedural approaches construct summary flow functions for procedures and use them in the place of calls. They have been ef- fectively used for many analyses except for flow and context sensitive points-to analysis which require representing indirect accesses of pointees defined in the callers. This is conventionally handled by using placeholders and creating cus- tomized call-specific versions. However, their sizes are not bounded by the num- ber of pointer variables. We propose a bounded representation of summary flow functions for points-to analysis called the higher order reachability graph (HRG). The conventional graph reachability based program analyses relate variables but not their pointees whereas HRGs relate the (transitively indirect) pointees of a variable with those of another variable in terms of indirection levels. They are expressive and capture the relevant information in a compact form without over- approximating. A simple algebra on indirection levels is sufficient to relate the indirect pointees defined in the callers obviating the need of placeholders. HRGs are context independent, and hence suitable for context sensitive analysis; they retain enough information for flow sensitivity and also enable strong updates. Our empirical measurements on SPEC benchmarks show that most summary flow functions are compact and are used multiple times. We have been able to scale flow and context sensitive points-to analysis to 158 kLoC using HRGs. Thus, this is a promising direction for further investigations in efficiency and scalability of points-to analysis without compromising on precision.

12. **Title**: Design and Implementation of an Espionage Network for Cache-based Side Channel Attacks on AES
Collaborators: **Bholanath Roy**, Prof. Bernard Menezes with Ravi Prakash Giri, Ashokkumar C

We design and implement the espionage infrastructure to launch a cache-based side channel attack on AES. This includes a spy controller and a ring of spy threads with associated analytic capabilities – all hosted on a single server. By causing the victim process (which repeatedly performs AES encryptions) to be interrupted, the spy threads capture the victim's footprints in the cache memory where the lookup tables reside. Preliminary results indicate that our setup can deduce the encryption key in fewer than 30 encryptions and with far fewer victim interruptions compared to previous work. Moreover, this approach can be easily adapted to work on diverse hardware/OS platforms and on different versions of OpenSSL.

13. **Title**: Mosaicing Scenes with a Quadcopter
Collaborators: **Meghshyam G. Prasad**, Sharat Chandran, Michael S. Brown

    This paper focuses on a method of constructing panoramas from a quadcopter, and a new mosaicing sub-problem when the scene contains significant regions of vacant spaces. These vacant spaces yield little to no features to match input images and hence challenge existing mosaicing techniques. We describe a framework that is able to handle this unique input by leveraging the availability of the inertial measurement unit (IMU) data from the quadcopter. Specifically, our method uses the imprecise IMU data accompanying a video to select a subset of images that contain interesting scene content. When the scene is such that this subset contains no vacant space, an appropriate panorama is effected; however, with featureless spaces, existing mosaicing methods do not work. In this paper, the subset is partitioned into multiple clusters. These subsets can now be stitched into a series of mini-panoramas, but a complete mosaic is not yet available. The gaps between these mini-panoramas represent regions of featureless spaces in the scene. Therefore, we once again use the IMU data together with a novel stereo reconstruction to determine appropriate portions of the images to complete the panorama. We demonstrate the efficacy of our approach on a number of input sequences that cannot be mosaiced by existing methods.

14. **Title**: TCP Download Performance in Dense WiFi Scenarios: Analysis and Solution
Collaborators: **Mukulika Maity**, Prof. Bhaskaran Raman and Prof. Mythili Vutukuru

    How does a dense WiFi network perform, specifi- cally for the common case of TCP download? While the empirical answer to this question is 'poor', analysis and experimentation in prior work has indicated that TCP clocks itself quite well, avoiding contention-driven WiFi overload in dense settings. This paper focuses on measurements from a real-life use of WiFi in a dense scenario: a classroom where several students use the network to download quizzes and instruction material. We find that the TCP download performance is poor, contrary to that suggested by prior work.

Through careful analysis, we explain the complex interaction of various phenomena which leads to this poor performance. Specifically, we observe that a small amount of upload traffic generated when downloading data upsets the TCP clocking, and increases contention on the channel. Further, contention losses lead to a vicious cycle of poor interaction with autorate adaptation and TCP's timeout mechanism. To reduce channel contention and improve performance, we propose a modification to the AP scheduling policy to improve the performance of large TCP downloads. Our solution, WiFiRR, picks only a subset of clients to be served by the AP during any instant, and varies this set of "active" clients periodically in a round-robin fashion over all clients to ensure that no client starves. We have done extensive evaluation of WiFiRR in simulation and in real settings. By reducing the number of contending nodes at any point of time, WiFiRR improves the download time of large TCP flows upto 3.5x of our classroom scenario. We also compare WiFiRR with state-of-the-art prior work WiFox, WiFiRR improves download time by 2.25x over WiFox.

15. **Title**: Process Edification
Collaborators: **Vrinda Yadav**, Rushikesh K. Joshi (IIT Bombay), Chris Ling (Monash University)

    Evolutionary development in an architecture-centric process involves architectural follow-ups from decision making. The architectural impact of decision-based informatics makes its systematic elicitation necessary in a model-driven traceable process. A systematic way called Process Edification is presented for the elicitation of decision-based informatics during modification and/or composition of business processes. The decision alternatives become edifiers, which can be plugged into a business process as per the architectural needs of evolution. Edification results in making decision-based informatics evident through traceable models and the specifications generated thereafter. The technique applies to decisions driven by non-functional requirements in the context of BPMN-based process architectures.

16. **Title**: Complimentary Usage of GPUs in Virtualized Systems
Collaborators: **Anshuj Garg**, Purushottam Kulkarni

    GPU's increasing computing power has led to their wide usage in solving general purpose compute intensive problems. The model for GPU computing is to use CPU and GPU together in a heterogeneous co-processing computing model. The sequential part of the application runs on the CPU and the computationally intensive part is accelerated by the GPU. The architecture of GPUs is such that they are most suitable for running the programs having SIMD (Single Instruction Multiple Data) nature. Several research work have reported the speed up of 2x to 100x or more when applications with SIMD nature are executed on GPUs. A number of applications are being developed that rely on GPUs for accelerating their computing e.g. AutoCAD, Adobe Photoshop etc. Also there are research works which have proved the utility of GPUs for accelerating the system

softwares (compilers, disk encryption etc), network processing (routers, firewalls etc) etc. In our work we address the problems related to the GPU resource and virtualization. We are trying to exploit the presence of GPUs in virtualized hosting platforms by using GPUs to compliment optimization task for resource management in virtualized systems. The question that we are trying to address here is assuming that if the virtualized system has access to GPUs, can the computation capability of GPUs be leveraged for hypervisor managed tasks. There two main advantages of offloading computation to GPUs: (i) Some computation tasks like hashing, encryption, compression etc are performed periodically in virtualized systems and they require CPU cycles. CPU is shared among virtual machines in virtualized systems. Offloading these tasks to GPU will save CPU cycles and improve the overall CPU throughput of the system. (ii) GPUs are designed to perform the SIMD computation faster. Offloading SIMD part of computation of virtualized systems to GPU will not only save CPU cycles but also speed up the computation.

17. **Title**: Graceful Degradation of QoS in Smart Grid
Collaborators: **Rohit Gupta**, Krithi Ramamritham

    Power grid is the main source of energy for all sort of consumers including individual homes, hospitals, educational institutes, industries etc. Consumers are highly dependent on the grid to satisfy their power needs. Power grid commit to provides power whenever a customer switches on his appliance, but when the requirement of power exceeds the production capacity, power grids are not in a position to satisfy all its customers. Power grids tackle this problem by removing customer of a particular area from the grid, which is termed as blackout. This is done to provide guarantee of power to others. Power grids imposes the blackout in round robin fashion for the maximum fairness they can provide. Better distribution of power could be achieved by imposing different quality of service for different appliance. Instead of imposing blackout, where none of the appliance can work, only critical appliance should be allowed to work. Shutting down non critical appliance will create the demand supply balance. Power grid should introduce systems to enable critical appliance to work during overload period thereby providing graceful degradation of QoS across different appliances. This can be achieved by establishing a proper sensing, analyzing and actuation cycle.

18. **Title**: Theory of Dynamic Evolution in Petri Net Models of Workflow Processes
Collaborators: **Ahana Pradhan**, Prof. Rushikesh K. Joshi

    Facilitating dynamic changes in workflow systems have been an active topic of research in the Business Process management community over the past decade. Though there are various approaches addressing specific case types, the theory and understanding of various scenarios and possibilities offer new research problems, especially in the case of dynamic evolution of workflows. In this thesis, we explore various problems and

work on them addressing a few fundamental issues involved in dynamic change, namely, the notion of consistency, algorithms for dynamic instance migration, change regions and handing dynamic evolution in the context of distributed workflows. For the purpose of our thesis, process models based on the formalism of Petri nets are used.

19. **Title**: Improving Performance of a Network Function Virtualization (NFV) System
Collaborators: **Priyanka Naik**, Prof. Mythili Vutukuru

    Network Function Virtualization (NFV) is a new trend in networking, where network functions are moving from custom hardware appliances to software implementations running on virtual machines (VMs) hosted on commodity hardware. NFV is of particular interest to telecommunication service provider networks that have a large number of networking elements. While the benefits of NFV such as cost reduction and increased agility are well understood, doubts still exist on whether a software implementation can match up to the high performance that hardware appliances deliver. In this context, network operators would benefit from frameworks that monitor performance and identify bottlenecks in Virtual Network Function (VNF) implementations obtained from vendors. We are building a performance monitoring and bottleneck detection tool for NFV. The tool is provided with the configuration file specifying the basic architecture of the VNF. The tool generates the per-hop throughputs and delays by sniffing the VM-to-VM communication and using the details provided in the configuration file. These black box measurements are used to identify the performance bottlenecks in real time, without requiring any instrumentation in the NFV system. The flexibility of scaling that the virtualized environment provides can be leveraged to make the NFV system scalable to handle the ups and downs in the traffic. We are working on a generic library to handle the scaling by interfacing between the NFV component and the cloud infrastructure. As in case of any distributed system, scaling a NFV system would require a load balancer which would distribute the load between the VMs handling a network function, a data store and a methodology to perform the scale when an overload is detected. The various components of the library would handle these functionalities in a generic manner. The developer of the NFV system just needs to provide the network function logic. So, the tool and library put together would help improve the NFV performance.

20. **Title**: Segment Slicing
Collaborators: **Omkarendra Tiwari**, Rushikesh K. Joshi

    Legacy softwares are developed over many years and across multiple teams around the globe. Therefore, such softwares are hard to understand, debug, and maintain. These problems may arise due to high LOC (lines of code), poor style of writing code, badly written code, etc. Modularity in code, is one of the possible solutions for its maintenance. It also improves readability of code making it more understandable and helps in debugging

errors if any. Modularizing the code manually is time taking and cumbersome for the team. New and unwanted errors can also creep in if this process is followed. A new graph based algorithm called Segment Slicing is presented which aims at automating this process of code modularization. As of now, this technique targets only intra-procedural code.

21. **Title**: Nested Containers: Containers within a Virtual Machine
Collaborators: **Chandra Prakash**, Prof. Umesh Bellur, Prof. Purushottam Kulkarni

    Derivative cloud is an intermediate layer existing between native cloud (such as Google Cloud Platform, Amazon EC2, Windows Azure or Private Cloud) and user. It works as a cloud broker which buys servers from native cloud providers and then re-sells those servers to the users. Native cloud provides infrastructure in the form of virtual machines and Nested virtualization is used to create a cloud service on top of a virtual machine or cloud service. One of the major drawback of the nested virtualization is two times virtualization overhead, one at native cloud layer and another at derivative cloud layer. This overhead leads to performance degradation of the virtual machine hosted on the derivative cloud. One of the possible solution to overcome this overhead is to run containers within the virtual machines. Instead of running virtual machines within virtual machines, it would be better to run Linux Containers inside the virtual machines provided by the native cloud. A major problem associated with such scenario is that hypervisor will not be aware of the containers running inside the virtual machine and the assumption that a single application is running inside a single virtual machine will not be valid. Hypervisor will treat virtual machine as a single entity and will not able to perform fair or prioritized container aware provisioning of resources. Containers running inside virtual machine can be impacted potentially in a negative way by the management actions performed by the hypervisor to manage the resources of the host machine. In our work we are trying to address this problem and trying to come up with a solution to overcome it.

22. **Title**: Multi-slice Tomographic Reconstruction: To Couple or Not to Couple
Collaborators: **Preeti Gopal**, Ajit Rajwade, Sharat Chandran, Imants Svalbe (Monash)

    -

23. **Title**: Colocalization Estimation using Statistical Analysis and Modeling of Images
Collaborators: **Thyagarajan Radhakrishnan**, Prof. Suyash Awate

    In microscopy imaging, colocalization between two biological entities (e.g., protein-protein or protein-cell) refers to the (stochastic) dependencies between the spatial locations of the two entities in a biological specimen.

Colocalization studies help reveal details of possible interactions between the entities in complex biochemical processes that are crucial in understanding the mechanisms of several diseases, including cancer. Measuring colocalization relies on fluorescence imaging of the specimen using two fluorescent chemicals, each of which indicates the presence / absence of one of the entities at any pixel location, followed by image analysis. Our proposed framework employs a probabilistic graphical image modeling and variational Bayes inference scheme for estimating all model parameters, including colocalization, directly from corrupted image data. Our results demonstrate improved performance over the heuristics-based state of the art approach.

24. **Title**: A Finitary Analogue of the Downward Lowenheim-Skolem Property
Collaborators: **Abhisekh Sankaran**, Supratik Chakraborty, Bharat Adsul

We present a model-theoretic property of finite structures, that can be seen to be a finitary analogue of the classical downward Löwenheim-Skolem property from model theory. We call this the *L-equivalent bounded substructure property*, denoted L-EBSP(S, k), where S is a class of finite structures, k is a natural number, and L is either first order logic or monadic second order logic. We show that many classes of structures of interest in computer science, satisfy L-EBSP(·, k); examples include the classes of words, trees (unordered, ordered, or ranked), nested words, graph classes of bounded tree-depth, graph classes of bounded shrub-depth and n-partite cographs. Further, any class of structures that is well-quasi-ordered under embedding satisfies L-EBSP(·, 0). We show that L-EBSP(·, ·) is closed under operations that are implementable using quantifier-free translation schemes, such as disjoint union and various products (cartesian, tensor, strong, lexicographic). We also prove that L-EBSP(·, k) entails generalizations of two classical theorems from model theory, namely the Łoś-Tarski theorem and the homomorphism preservation theorem. It turns out any hereditary class of graphs satisfying L-EBSP(·, k) for k ¿= 2, must have bounded induced path lengths, motivating the question of a structural characterization of L-EBSP(·, k).

25. **Title**: Heap Abstractions for Static Analysis
Collaborators: **Vini Kanvar**, Prof. Uday P. Khedker

Study shows that the most common challenge faced by novice C/C++ programmers is the management of dynamic memory (heap). Understanding the concept of pointers in itself is non trivial. Without this understanding, poorly written programs have memory leaks that impact the performance of the programs. Such programs use unnecessarily large system resources, and worst of all they fail due to out-of-resource problems. As a consequence, analysis of heap memory is becoming increasingly important. My PhD work involves creating a new heap abstraction that yields results that are both scalable to large sized programs and precise enough to produce

useful results.

26. **Title**: Decomposition of automata
Collaborators: **Saptarshi Sarkar**, Bharat Adsul

Krohn-Rhodes theorem, announced in the 1960s, is one of the fundamental results of automata and semigroup theory. From the automata theoretic point of view, it states that any finite automaton can be decomposed into a cascaded product of several simpler finite automata. The theorem only applies to finite state automata. We propose scenarios like distributed automata, games etc where extending the theorem can be helpful. We also show a theoretical application of the theorem.

27. **Title**: Mitigating BootStorms in Virtualized Datacenters
Collaborators: **Durgesh**, Prof. Umesh Bellur

Large scale concurrent virtual machine (VM) deployment is a very slow process. This is primarily due to three reasons. Firstly, a large number of concurrent instantiation requests, lead to the formation of long IO request queues at the central repository server storing the VM image files. Secondly, it requires a long time to transfer several GBs of these VM image files, across the heavily over-subscribed network links of the datacenter. Finally; the CPU, disk and network usage of the existing co-located VMs, slows down the process of reception of these VM image files and the subsequent boot-up of VMs, off them. As a consequence of all of the above, the end users have to endure a long wait, from the time they sent their instantiation request, to the time they are able to log in and use their VMs. When hundreds of VMs are to be instantiated off large image files, this waiting time may be in the order of a few hours. For companies which provide virtual desktops to their employees, it is extremely common to encounter such a large number of instantiation requests every morning, when their employees turn in for work. The widespread occurence of this phenomenon has earned it a moniker, namely boot storm. In addition to the above, there are also situations where a large number of VMs are required to be deployed instantly, in order to absorb a spike in the workload. In such a scenario, long deployment times are unacceptable. As a result reducing the same is of paramount importance. In the thesis, we will explore different ways of mitigating boot storms in virtualized data centers

28. **Title**: Smart Energy Management in Buildings
Collaborators: **Anshul Ajay Agarwal**, Prof. Krithi Ramamritham

A building management system (BMS) tracks and controls available energy, environmental parameters (e.g., temperature, humidity), occupancy status and count, etc. But a BMS should also be able to help reduce and optimize power consumption (PC), monitor the status and health of the appliances in the building, profile energy consumption of different areas, and identify zones with anomalous PC. Such tasks require the BMS

to sense various facets, like PC, temperature and occupancy status. A straightforward approach is to deploy a network of sensors to sense these values in all parts of the building. But this leads to huge capital cost for initial deployment, upgradation cost as the system gets used, and maintenance cost with respect to fault handling. To address these problems, we are working towards the development of a Smart Energy Management System founded on three tenets which lead to a drastic minimization of the number of sensors deployed. 1. Replace hard/physical sensors (PS) with soft sensors (SS) 2. Use building structure to reduce number of sensors 3. Use correlation between different facets observed and sensors available

29. **Title**: Design Robust local Image Descriptor Using ConvNet
Collaborators: **Rahul Mitra**, Dr. Sharat Chandran with Dr. Arjun Jain

    Our proposal is to devise a convolutional network (ConvNet) based algorithm for efficient computation of robust local image descriptors. For an input image patch centered on a pixel, we would learn an invariant low-dimensional non-linear embedding in feature space. This descriptor should not rely on any scale or affine parameters from the detection stage and should be highly invariant to a wide variety of geometric and photometric changes including scale, rotation, viewpoint change, image blur, complex illumination changes and compression artifacts. We also introduce a new dataset (
    emphDeep Patch Dataset), which is a collection of different scenes (collection of photos of a place or object) scraped from the internet. The proposed descriptor would be trained on this
    emphDeep Patch dataset. We also use an training architecture which enables us to efficiently train on large training data . Such descriptor can then be used in other Computer Vision related tasks like object recognition, finding stereo correspndences, solving SLAM problems.

30. **Title**: Evolutionary Design of Robotics
Collaborators: **Abhishek Chakraborty**, Siddhartha Chaudhuri

    Customized robots can perform numerous specialized tasks such as rescue, transportation, and medical operations. The design of robots tailored to tasks, which has traditionally been manual, can be enhanced and accelerated by computational assistance. The proposed work employs machine learning techniques and evolutionary algorithms to obtain optimal robotic designs matching high-level specifications of the designer.

31. **Title**: Using Machine Translation Evaluation Techniques to Evaluate Text Simplification Systems
Collaborators: **Sandeep Albert Mathias**, Pushpak Bhattacharyya

    We look at techniques to find out ways to evaluate automated text simplification systems, based on the grammaticality and simplicity of the output,

as well as the meaning preserved in the output, from the input text, and the overall quality of simplification of the system. Two of the techniques are already used in machine translation, to check for grammaticality and meaning preservation. We make use of a new technique for text complexity analysis to assess the quality of the text simplification system.

Chapter 2

Posters

1. **Title**: A Statistical Model for Smooth Shapes in Kendall Shape Space
   Collaborators: **Saurabh Shigwan**, Suyash Awate with Akshay Gaikwad

   This paper proposes a novel framework for learning a statistical shape model from image data, automatically without manual annotations. The frame- work proposes a generative model for image data of individuals within a group, relying on a model of group shape variability. The framework represents shape as an equivalence class of pointsets and models group shape variability in Kendall shape space. The proposed model captures a novel shape-covariance structure that incorporates shape smoothness, relying on Markov regularization. Moreover, the framework employs a novel model for data likelihood, which lends itself to an inference algorithm of low complexity. The framework infers the model via a novel expectation maximization algorithm that samples smooth shapes in the Riemannian space. Furthermore, the inference algorithm normalizes the data (via similarity transforms) by optimal alignment to (sampled) individual shapes. Results on simulated and clinical data show that the proposed framework learns better-fitting compact statistical models as compared to the state of the art

2. **Title**: Multilingual Named Entity Recognition using Deep Learning
   Collaborators: **Rudra Murthy**, Pushpak Bhattacharyya

   Named Entities do not change labels across languages. Existing algorithms take advantage of this information either in isolation or along with annotated parallel text to induce label agreements between the two NER systems. This technique has been shown to improve the performance of both the systems. Also, this is beneficial for many languages which have little or no resources. The system trained on resource rich knowledge can then transfer the knowledge to resource scarce language. Alternatively, both the systems can be trained jointly thereby providing features which help each other. We explore a Deep Learning based Bilingual Named Entity Recognition (NER) system for closely-related languages. Instead

of having a separate NER system trained for each language, we have a single system trained jointly for any language pairs. This enables sharing of knowledge across languages implicitly and thereby improving the NER performance for both the languages.

3. **Title**: Tracking Unsafe Stopping Behaviour of Buses
   Collaborators: **Ravi Bhandari**, Prof. Bhaskaran Raman with Venkat Padmanabhan

   Road safety is a critical issue. Road accidents cause an estimated 1.3 million fatalities each year, placing it in the top 10 leading causes of death in the world. We believe that mobile devices can play a positive role in this context by detecting dangerous conditions and providing feedback to enable timely redressal of potential dangers. We focus on a specific problem that is responsible for many accidents in India: the stopping behaviour of buses especially in the vicinity of bus stops. Buses could come in to a bus stop at a high speed, could continue rolling forward instead of coming to a complete halt, or could even choose to stop some distance away from the bus stop, possibly even in the middle of a busy road.

4. **Title**: JalTantra: Tool for Optimization of Piped Water Networks
   Collaborators: **Nikhil Hooda**, Om Damani

   Multi village piped water schemes are projects designed to provide water to a number of villages from a common source of water. They consist of several components requiring many choices to be made regarding their sizing and service. These choices impact the cost of the scheme which is the major factor in deciding whether a scheme is implemented or not. One of the most important aspects in the design of these projects is the choice of pipe diameters from a discrete set of commercially available pipe diameters. In general, each link (connection between two nodes) can consist of several pipe segments of differing diameters. Larger the pipe diameters, better the service (pressure), but higher is the capital cost. The branched piped water network cost optimization problem is the selection of pipe diameters that minimize the system cost while providing the requisite service (pressure at demand points). The aim of our research is to provide a system to the designer of such projects, which aids him/her in the design process and helps him/her make a more informed choice. To achieve this we have created the software JalTantra which will provide an optimal and scalable solution.

5. **Title**: A Comparison of Some Direct 2D Reconstruction Methods using Discrete Projective Mathematical Transforms
   Collaborators: **Preeti Gopal**, Imants Svalbe (Monash University), Ajit Rajwade (IITB), Sharat Chandran (IITB)

   Tomographic acquisitions can be described mathematically as discrete projective transforms. Direct reconstruction methods aim to compute an

accurate inverse for such transforms. We assemble a limited set of measurements and then apply the inversion to obtain a high-fidelity image of the original object. In this work, we compare the following direct inversion techniques for sets of discrete projections: Radon-i(inverse)Radon, a least squared error method and filtered back-projection for Mojette inversion. We observe that filtered back-projection is the best of these methods, as the reconstruction errors that arise using this method depend least strongly on the image structure. We aim to improve results for the filtered back-projection method by optimizing the design of the regularizing filter and here present work towards eliminating the regularization threshold that is used as part of this method.

6. **Title**: A Computational Approach to Automatic Prediction of Drunk-Texting
   Collaborators: **Aditya Joshi**, Pushpak Bhattacharyya, Mark Carman with Abhijit Mishra, Balamurali AR

   Alcohol abuse may lead to unsociable behavior such as crime, drunk driving, or privacy leaks. We introduce automatic drunk-texting prediction as the task of identifying whether a text was written when under the influence of alcohol. We experiment with tweets labeled using hashtags as distant supervision. Our classifiers use a set of N-gram and stylistic features to detect drunk tweets. Our observations present the first quantitative evidence that text contains signals that can be exploited to detect drunk-texting.

7. **Title**: Optimizing Remote Calls in Parallel Data Management Systems
   Collaborators: **Bikash Chandra**, S. Sudarshan

   Consider a parallel data processing framework running an application on a cluster. This framework could be a MapReduce pipeline in Hadoop or Spark, or a stream processing application running on Storm, Muppet or S4. Many such applications require stored data to be accessed from a parallel data store. For some of these applications, it may also be necessary to perform computations based on the accessed data. Accessing remote data synchronously for each input data may lead to poor performance in terms of throughput. We present novel runtime techniques for frequency based caching and dynamic load balancing, to dynamically optimize data access and function call execution in a parallel data processing framework which could be based on stored or streaming data. The load balancing takes into account CPU, network and I/O costs as well as the load at clients and servers. We also describe how to extend the APIs of the MapReduce and the Muppet stream processing platform to incorporate (possibly batched) prefetching. We implement our optimization techniques, using HBase as the parallel data store. Our experiments show that our optimization techniques provide up to 5 times improvement in throughput.

8. **Title**: Design and Implementation of an Espionage Network for Cache-

based Side Channel Attacks on AES
Collaborators: **Bholanath Roy**, Prof. Bernard Menezes with Ravi Prakash Giri, Ashokkumar C

We design and implement the espionage infrastructure to launch a cache-based side channel attack on AES. This includes a spy controller and a ring of spy threads with associated analytic capabilities – all hosted on a single server. By causing the victim process (which repeatedly performs AES encryptions) to be interrupted, the spy threads capture the victim's footprints in the cache memory where the lookup tables reside. Preliminary results indicate that our setup can deduce the encryption key in fewer than 30 encryptions and with far fewer victim interruptions compared to previous work. Moreover, this approach can be easily adapted to work on diverse hardware/OS platforms and on different versions of OpenSSL.

9. **Title**: Mosaicing Scenes with a Quadcopter
Collaborators: **Meghshyam G. Prasad**, Sharat Chandran, Michael S. Brown

This paper focuses on a method of constructing panoramas from a quadcopter, and a new mosaicing sub-problem when the scene contains significant regions of vacant spaces. These vacant spaces yield little to no features to match input images and hence challenge existing mosaicing techniques. We describe a framework that is able to handle this unique input by leveraging the availability of the inertial measurement unit (IMU) data from the quadcopter. Specifically, our method uses the imprecise IMU data accompanying a video to select a subset of images that contain interesting scene content. When the scene is such that this subset contains no vacant space, an appropriate panorama is effected; however, with featureless spaces, existing mosaicing methods do not work. In this paper, the subset is partitioned into multiple clusters. These subsets can now be stitched into a series of mini-panoramas, but a complete mosaic is not yet available. The gaps between these mini-panoramas represent regions of featureless spaces in the scene. Therefore, we once again use the IMU data together with a novel stereo reconstruction to determine appropriate portions of the images to complete the panorama. We demonstrate the efficacy of our approach on a number of input sequences that cannot be mosaiced by existing methods.

10. **Title**: Realizability and Games in Distributed Systems
Collaborators: **Nehul Jain**, Bharat Adsul

The realizability problem for a specification asks if there exists a matching implementation in the presence of an adversarial environment. Realizability problem of distributed open reactive systems has been considered. The communication between processes might be synchronous or asynchronously. The problem can be described in terms of an equivalent game of incomplete information, where one of the players in a two player game has a distributed nature. We look at a variation of such a game. We

consider one such game where one of the players, the system consists of just two processes, and see what new issues arise due to the non sequential nature of the game. It is expected that, this would help to solve the general distributed synthesis problem.

11. **Title**: How hard can it be? The E-Score - A scoring metric to assess the complexity of text
Collaborators: **Sandeep Albert Mathias**, Pushpak Bhattacharyya

    We present the E-Score, an evaluation metric that that utilizes structural complexity of sentences and language modelling of simple and normal English to come up with a score that tells us how simple / complex a piece of text is. We gather gold standard human data, and use it to evaluate our system against a pair of popular existing systems - the Flesch Reading Ease Score (FRES), and the Lexile Framework.

12. **Title**: A Finitary Analogue of the Downward Lowenheim-Skolem Property
Collaborators: **Abhisekh Sankaran**, Supratik Chakraborty, Bharat Adsul

    We present a model-theoretic property of finite structures, that can be seen to be a finitary analogue of the classical downward Löwenheim-Skolem property from model theory. We call this the *L-equivalent bounded substructure property*, denoted L-EBSP(S, k), where S is a class of finite structures, k is a natural number, and L is either first order logic or monadic second order logic. We show that many classes of structures of interest in computer science, satisfy L-EBSP(·, k); examples include the classes of words, trees (unordered, ordered, or ranked), nested words, graph classes of bounded tree-depth, graph classes of bounded shrub-depth and n-partite cographs. Further, any class of structures that is well-quasi-ordered under embedding satisfies L-EBSP(·, 0). We show that L-EBSP(·, ·) is closed under operations that are implementable using quantifier-free translation schemes, such as disjoint union and various products (cartesian, tensor, strong, lexicographic). We also prove that L-EBSP(·, k) entails generalizations of two classical theorems from model theory, namely the Łoś-Tarski theorem and the homomorphism preservation theorem. It turns out any hereditary class of graphs satisfying L-EBSP(·, k) for k ¿= 2, must have bounded induced path lengths, motivating the question of a structural characterization of L-EBSP(·, k).

13. **Title**: Substring-based unsupervised transliteration with phonetic and contextual knowledge
Collaborators: **Anoop Kunchukuttan**, Prof. Pushpak Bhattacharyya with Mitesh Khapra

    Transliteration is a key building block for multilingual and cross-lingual NLP since it is useful for user-friendly input methods and downstream

applications like machine translation and cross-lingual information retrieval. The best performing solutions are supervised, discriminative learning methods which learn transliteration models from parallel transliteration corpora. However, such corpora are available only for some language pairs. It is also expensive and time-consuming to build a parallel corpus. This limitation can be addressed in three ways: (i) training a transliteration model on mined parallel transliterations, (ii) transliterate via a bridge language, or (iii) unsupervised transliteration. We explore unsupervised transliteration in this work, addressing shortcomings in previous work. We propose an unsupervised approach for substring-based transliteration where we use novel methods for incorporating two new sources of knowledge to guide the learning process: (i) context by learning substring mappings, as opposed to single character mappings, and (ii) rich phonetic features which capture cross-lingual character similarity via prior distributions. We show that these phonetic features can be extracted from the orthographic representation alone for Indian languages, which were the focus of our experiments. For other languages, a phoneme dictionary or grapheme-to-phoneme converter is required. Our approach is a two-stage iterative, boot-strapping solution, which vastly outperforms a state-of-the-art unsupervised transliteration method and gives an improvement in top-1 accuracy of up to 50

14. **Title**: Scanpath Complexity: A Measure of Cognitive Load for Text Reading and Annotation
Collaborators: **Abhijit Mishra**, Prof. Pushpak Bhattacharyya with Kuntal Dey, Seema Nagar

    Measuring cognitive load, the degree of difficulty perceived during reading / annotation, is useful for practical purposes such as optimizing learning material design, bettering annotation pricing schemes and modelling text comprehensibility. We propose multiple "easy-to-implement" methods that quantify cognitive load by modelling Scanpath Complexity: complexity of eye-movement patterns of readers and annotators. Scanpath complexity is modelled as a function of fixational and saccadic complexity, fixational complexity capturing fixation duration and saccadic complexity capturing the degree of uncertainty of saccadic transitions. We demonstrate the effectiveness of our scanpath complexity measure by showing that it correlates with different measures of lexical and syntactic complexity and readability metrics (Flesch Kincaid, Gunning-Fog and SMOG indices), better than baselines that consider fixation and saccadic properties alone for both general-reading and annotation tasks.

15. **Title**: Civique: Using Social Media to Detect Urban Emergencies
Collaborators: **Diptesh Kanojia**, Krithi Ramamritham with Vishwajeet Kumar

    The use of social media has seen a tremendous surge in developing nations such as India, in the past few years. We present the Civique system for

emergency detection in urban areas by monitoring micro blogs like Twitter. The system detects emergency related events, and classifies them into appropriate categories like "fire", "accident", "earthquake". We classify Twitter posts (or "tweets") in real time, visualize the ongoing event on a map interface and present users with an alert notification, along with options to contact relevant authorities, both online and offline. Users have access to this functionality using both a web interface and an Android application. We demonstrate Civique with both tweet detection, and visualization of the incident on a map.

16. **Title**: A Logical Characterization for Dense-Time Visibly Pushdown Automata
Collaborators: **Devendra Bhave**, Krishna S., Ashutosh Trivedi with Ramchandra Phawade, Vrunda Dave

    Two of the most celebrated results that effectively exploit visual representation to give logical characterization and decidable model-checking include visibly pushdown automata (VPA) by Alur and Madhusudan and event-clock automata (ECA) by Alur, Fix and Henzinger. VPA and ECA – by making the call-return edges visible and by making the clock-reset operation visible, respectively – recover decidability for the verification problem for pushdown automata implementation against visibly pushdown automata specification and timed automata implementation against event-clock timed automata specification, respectively. In this work we combine and extend these two works to introduce dense-time visibly pushdown automata that make both the call-return as well as resets visible. We present MSO logic characterization of these automata and prove the decidability of the emptiness problem for these automata paving way for verification problem for dense-timed pushdown automata against dense-timed visibly pushdown automata specification.

17. **Title**: Noun Compound Interpretation
Collaborators: **Girishkumar Ponkiya**, Pushpak Bhattacharyya

    Noun compounds (or noun sequences) are a productive, continuous sequence of more than one nouns. Most of the noun compounds appear only once in a large corpus. These characteristics of noun compounds make them a special case and demand special treatment. Here, the problem is to find noun compounds from text, parse them if required, and extract semantic relation between components of the noun compound. A task of extracting an abstract relation between components of the noun compound (e.g., apple pie: MadeOf), or paraphrasing noun compound using verb and prepositions (apple pie : "a pie made of apple" or "a pie with apple flavor"), is known as interpretation of noun compound (or noun compound interpretation). For our work, we use a set of predefined abstract labels as semantic relations. Following are major bottlenecks in current system, and our approaches to solve the same: (a) There is no acceptable inventory of semantic relations. We are trying to come up with a data driven approach which will help in refining the current inventories.

(b) Inspite of millions of noun compounds in large corpora, there is no sufficiently large annotated dataset for supervised training. We are planning to use semisupervised approach to tackle this. (c) The present datasets have annotation for each noun compound without context. But, inference of semantic relation between components of a noun compound is difficult, and it requires information about how the components can be related in sentences. We are planning to use web-extracted information to bring the context in play.