NSF CSR/NeTS Joint PI Meeting

November 4th and 5th, 2019

Abstracts of all Technical Sessions

Room Locations

Lobby Level	:	Rosslyn 1-2, Lee, Jefferson, Jackson, Madison
First Level	:	Salons 1-6, Salons A-K
Second Level	:	Alexandria, Mt. Vernon, Manassas, McLean, Fairfax

General Sessions

Monday 9:00 a.m. : Dennis Moreau
Tuesday 8:30 a.m. : Ion Stoica

The Role of the Hosting Platform in Facilitating Computing and Network Innovation

Dennis Moreau

Location: Salons 1-3

The rapid growth in the adoption of modern application development and hosting technologies has brought with it, unprecedented levels of complexity, in terms of stack decoupling, instance dynamics, and system distribution. The underlying hosting platforms readily span multiple on-premise, co-hosted and cloud-hosted sites, easily extending across geographic and regulatory boundaries. Within individual platforms there is an accompanying convergence of computation, networking and storage capabilities, realized over common resources and shared fabrics. The result is that services, applications, platforms, infrastructure and even bare metal can all be consumed on demand at incredible scale.

Systems plus ML: When sum is bigger than its parts

Ion Stoica

Location: Salons 1-3

The research at the intersection between machine learning (ML) and systems has the potential to fuel the innovation over the next decade. When used together, ML and systems can lead to a rapidly evolving positive feedback loop, in which systems accelerate ML algorithms, and ML algorithms make the systems faster and more secure. In the context of the RISELab at UC Berkeley, we have done work in both directions of this feedback loop. On one hand, we have built new systems to better support ML workloads, such as Ray, a general distributed framework that supports highly scalable ML libraries, including a reinforcement learning library (RLlib) and a Hyperparameter Search library (Tune). On the other hand, we have used Ray, RLlib and Tune to perform systems optimizations (e.g., database query optimizations, compiler optimizations), algorithm optimizations (e.g., building decision trees for packet classification), and program synthesis.

When applying ML to systems, we have developed new techniques which, instead of applying ML (e.g., deep learning) to solve the problem end to end, we apply ML to synthesize "classic" solutions for the problems at hand. For example, instead of using a deep learning (DL) network to perform packet classification, we use DL to synthesize a decision tree that performs packet classification. By doing so we develop solutions that are explainable (thus addressing a significant challenge of DL), and are easier to deploy in existing systems.

Breakout Session I

Monday 10:15 a.m. – 12:15 p.m.

Measurement support for performance and diagnostics

Neil Spring

Location: Manassas

Measurement support for performance and diagnostics: The goal of this session is to find opportunities for integrating modern kernel and systems diagnostics (ebpf, kprobes) with techniques for network debugging (traceroutes, latency monitoring) with approaches to network management (snmp), using techniques from data science and machine learning to identify faults, vulnerabilities, and opportunities for upgrades. The domain includes management of cloud applications, cloud systems, ISP networks, IoT applications, and Internet applications overall. What could go wrong, what just went wrong, and how can we fix it?

Edge Computing for Real-time ML and Cognition

Mahadev Satyanarayanan, Wei Gao

Location: Salon 5

The focus of this session is the intersection of Edge Computing and ML/Cognition, especially in latency-sensitive settings. Here are some relevant questions (not prioritized).

1. What is the kind of hardware acceleration needed for ML/Cognition at edge nodes (i.e., cloudlets)? How do we reconcile these power-hungry and heat-generating accelerators with the physical and electrical limitations of cloudlets? What does it imply for hardware design at the edge?

2. In the battle for MIMD (cores) versus SIMD (GPUS, FPGAs, etc.) at the edge, how do we strike a good balance especially in multitenant environments?

3. Is inferencing the primary focus, or is training at the edge also important? Why isn't training in the cloud good enough? When is training at the edge valuable/important?

4. 'Just-in-time learning' at the edge has been proposed as an approach to dynamically adapting classifiers to fresh incoming data (e.g. during a drone flight). How important/valuable is this? Does it only apply to a small set of use cases, or is it of broad applicability?

5. There is an ongoing dialectic between Tier-3 (mobile and IoT devices) and Tier-2 (cloudlets/edge computing). Over time, each new generation of Tier-3 devices has more sophisticated SoC functionality. So the optimal split between Tier-2 and Tier-3 is constantly changing. In a multi-user setting, there may be diverse splits because of old Tier-3 hardware versus new Tier-3 hardware. How do we create robust, and easily-maintained applications in the face of so much variability?

6. Splitting pipelines (e.g. DNN inferencing) across Tier-3 and Tier-2 is being actively explored by many researchers today. How successful are these efforts? What are the key challenges today? Why is it so hard?

7. Edge computing lends itself naturally to distributed learning. For example, video cameras at different cloudlets collect different parts of an overall training set for DNN training. What is the current state of the art in this effort? What makes it hard? What are the challenges?

Computing Architecture for Edge Computing (CAEC)

Yiran Chen and Yingyan Lin

Location: Alexandria

This breakout session aims exploiting the requirements of future computing architectures for edge computing. The session participants will discuss (but not limited to) the following questions

1. What are the critical requirements of computing architectures for future edge computing applications? In particular, are there any new requirements that have not been fully accommodated in the current computing architecture design?

2. What kinds of new computing units will be needed in future CAEC and are there any necessary architectural changes?

3. Can the current memory and storage systems satisfy the new requirements? If not, what are the new technologies that are worth exploring?

4. Any new computing models and algorithms that need to support?

5. How the workload to be partitioned between different layers in the context of edge computing?

6. Any new benchmarking methods required?

7. Any critical concerns and solutions on security and safety?

Heterogeneous Computing for IoT and CPS

Shuvra Bhattacharyya and Marilyn Wolf

Location: Jefferson

Heterogeneous computing offers the potential to streamline execution of key processing tasks using architectures that are better suited to those tasks than architectures composed of identical processing elements. This potential together with the increasing diversity of available processor architectures and domain-specific programming languages opens up complex new design spaces for computer systems. This breakout session will explore research opportunities and identify priority areas for research emphasis in heterogeneous computing systems. The workshop will have a specific focus on heterogeneous computing in the context of Internet of things (IoT) and cyber-physical systems (CPS), which are important, rapidly-developing areas that stand to benefit significantly from advances in heterogeneous computing.

Formal Verification

Manos Kapritsos, Carlos Varela

Location: Mount Vernon

In the last few years, formal verification has made significant inroads in the systems community. As the tools, languages and methodologies become more powerful, the promise of provably correct systems starts becoming a reality. In this session, we will discuss some of the open challenges in applying formal methods to real-world systems. In particular, we will discuss classes of applications that could benefit from formal verification, such as machine learning, cyber-physical systems, cryptographic protocols, data-driven and distributed systems. We will also talk about the various methods of formally verifying these systems, including symbolic/statistical model checking, and interactive/automated theorem proving, and their relationship to open questions on knowledge representation, modal logics, probabilistic reasoning, and uncertainty modeling, among others. Finally, we will discuss several topics that apply broadly to all aspects of formal verification, such as:

-How to scale formal verification to large systems?

-How to tame the complexity of reasoning about concurrency and distribution, including actor-, process-based, and multi-threaded programs?

-How to make formal methods modular, composable, reusable, and accessible to ordinary developers?

Data Storage

Ali Butt

Location: Salon 6

The focus of this session is to discuss and identify research directions and challenges arising in modern and emerging data and storage systems research in the face of disruptive applications such as deep learning, edge systems, and smart infrastructure. We will also discuss comprehensive techniques that provide practical solutions to the storage issues facing the information technology community. Following are some relevant questions to consider

How to evolve storage systems to meet the challenges of scale, throughput, and sustainability arising from emerging applications such as deep learning?

How to rearchitect and design storage systems to accommodate and leverage new types of memories such as NVRAM?

How to address allocation, management, privacy, performance, and multi-tenancy to meet the demands of the intense migration of data from on-premise to cloud deployments?

How to design file system APIs and higher-level yet simple-to-use interfaces for new storage systems?

How to design programming models to efficiently support innovative storage and deep storage hierarchies?

How to address pipeline issues and train the next generation of storage systems researchers?

How to design and make it easy for systems researchers to realize new applications and storage hardware?

Wireless-aware Design for Mobile Reality

Suman Banerjee and Kyle Jamieson

Location: Jackson

Mobile Reality focuses on systems that support interactive multimedia in different forms, including virtual, augmented, and mixed reality.

How to design next generation of techniques to support interactive audio and video systems, virtual, augmented and mixed reality systems.

Challenges: Extremely high data rates and low latency needs, high variability, and extensive computational needs.

Opportunities: Edge compute support, 5G networks with high speeds and low latency.

Developing a Data-Centric Ecosystem for the Big Data Revolution

Lixia Zhang and Edmund Yeh

Location: Lee

We are now in a prolific era of exploration, discovery and realization of new knowledge in many fields of data intensive science, health, and engineering. Domain experts in these fields have been generating, managing, and processing data sets that are growing dramatically in scale and complexity. At present, these respective domains face the daunting challenge of developing individual systems to handle big data and to address issues such as data storage, indexing, performance, security, and privacy. These individual solutions, while addressing similar problems, are typically developed using incremental approaches, and in isolation from each other.

We speculate that the above phenomenon is likely a direct consequence of the following basic fact: today?s computer systems/networks are generally centered around low-level primitives, such as addresses, processes, servers, and connections; so do the existing security solutions, which secure data containers and delivery pipes but not the data itself. Because big data applications focus on the data, this incongruity creates a gap between what the underlying systems provide and what the applications need, leaving individual users, who may not be computer experts, with difficult tasks of mapping their problems from application space to the underlying computing systems.

Frequently the above leads to big data researchers relying on commercially available cloud services for storage and processing. Unfortunately this does not address some of the most fundamental challenges facing big data scientists, including:

* the need to deal with lower layer device issues to collect data;

* the difficulty in navigating unstructured data collections;

* the lack of systematic solutions for security and privacy when data is outside the cloud, and the lack of auditability when the data is inside the cloud; and

* systematic support for user authentication which is needed for data access control.

We believe that harnessing the big data revolution will require a cross-disciplinary and domain-agnostic big data ecosystem that can support big data management through the whole data life cycle, starting from data production, naming, securing data directly, to scalable retrieval, and controlled data access that enables data sharing across the boundaries of different service providers, different applications, and different disciplines. Such a big data ecosystem has the potential to provide agile, integrated, interoperable, scalable, robust and trustworthy solutions for heterogeneous data-intensive domains.

One promising direction for developing this big data ecosystem is to take on a data-centric design approach, as pointed out by a recent NSF DCL emphasizing the need to 'Support Future Data-Intensive Science and Engineering Research.' The goal of this session is to articulate the need for a principled design for a data-centric ecosystem, and identify approaches towards its realization.

Machine Learning Applied to Systems

Monica Lam and Qun Li

Location: Rosslyn 2

Machine learning has fueled advances in many disciplines from medicine, science, to social science. How can machine learning advance computer systems? On the front end, artificial intelligence can significantly improve our system user interface. For example, will we be communicating with the computers in natural language? Will consumers be able to code in natural language? On the back end, how can machine learning improve our algorithms to manage systems?

Many have started to worry about the potential negative impact of AI, from algorithmic biases, loss of jobs, AI surveillance abuse, to centralization of private information. Are there system architectures that can support autonomy and privacy while letting users benefit from big data analysis?

Last but not least, machine learning has an insatiable need for computing resources. What breakthroughs can we see to support the computational demands of deep learning?

Internet of Things

Tarek Abdelzaher, Ramesh Govindan

Location: Rosslyn 1

IoT technologies will be pervasive in smart cities of the future and necessary for recovery from catastrophe and conflicts. In this session, we will explore research directions in sensing and actuation with emphasis on large-scale IoT deployments, both engineered and ad-hoc. Our session will explore emerging sensing and actuation technologies, programming models for IoT systems, data analytics systems targeted towards multimodal data streams, machine learning solutions for IoT, techniques to ensure reliability and survivability of IoT deployments in the face of extreme physical conditions, and security and privacy challenges arising from pervasive device deployments.

Spectrum Management in 5G and beyond

Rose Hu, Swarun Kumar

Location: McLean

In wireless networks, radio resource and spectrum management have always been among the key research topics in delivering network capacity and QoS requirements. When it comes with 5G, especially with requirements such as high node density, high heterogeneity, massive connectivity, low latency, high capacity (with mmWave or even Tera Hz bandwidth), the complexity of resource management goes up tremendously. Traditional centralized resource management and computing may soon reach their imitations, due to concerns on complexity, latency, communications bandwidth, privacy, security, etc. Further motivated by the increasing computational capacity of wireless local devices as well as the ever increasing concerns on sharing data due to privacy and security, next-generation communications/computing networks will encounter a paradigm shift from conventional cloud computing to edge computing, which largely deploys computational power to the network edges/fog nodes to meet the needs of applications that demand very high computations and low latency. As a result, 5G and beyond will fully exploit the combination of a centralized and decentralized computing architecture for efficient resource allocation. Furthermore, 5G goes beyond meeting evolving consumer mobile demands by also delivering various industry vertical sectors such as autonomous vehicles, e-health, etc. Spectrum needs to be sensed effectively and shared appropriately. Deployment paradigms such as licensed spectrum, shared spectrum, and unlicensed spectrum at different possible bands such as sub-6 Ghz and mmWave can be supported. Complicating resource optimization further is that fact that 5G will see a heterogenous mix of user devices ranging from energy starved and low power devices, to highly capable high bandwidth devices (e.g. augmented reality headsets) and high mobility contexts (vehicles, UAVs, high speed rail, etc.). This session will focus on 5G and beyond resource management and spectrum management as well as how to leverage the edge computing to enable distributed resource allocation and spectrum management. Specifically, Machine Learning has become a powerful tool to enable mining the valuable information from big data and autonomously uncover relationships that would have been too complex to extract by traditional approaches. We would also like to discuss how cloud/edge computing platforms facilitate the research and development of distributed machine learning for radio resource and spectrum management in 5G and beyond.

Breakout Session II

Monday 1:30 p.m. – 3:30 p.m.

Network Function Virtualization

Sonia Fahmy, K. K. Ramakrishnan

Location: McLean

1. What use cases are most important for NFV, e.g., 5G, IoT, VR/AR?

2. What are the different challenges with different types of VNFs: middleboxes, services, micro-services with multiple components?

3. How are stateful VNFs handled when scaling out or migration takes place? How can load be balanced among multiple instances of a VNF?

4. What do different types of deployments on VMs, containers, or serverless instances that use external data sources, entail?

5. What types of challenges does deploying and orchestrating VNFs in the edge entail, versus deploying them in the core, or a joint deployment?

6. What challenges do service function chains (SFCs) introduce in terms of performance bounds, scaling up or down, in or out? How do we guarantee various performance goals for an SFC with SDN stitching the VNFs? Are there any VNF interoperability problems?

7. Are new attack surfaces being introduced with NFV? How can we achieve built-in security and resilience? Does VNF placement have to consider security, privacy, and trust goals?"

Crowd Sensing

Christine Bassem

Location: Jackson

In Mobile Crowd Sensing (MCS) platforms, the power of crowds is leveraged to assist incompleting spatio-temporal sensory tasks. Thus expanding the pool of resources foundin typical sensor systems to include already roaming devices and the humans controlling them. This new area opens up various research topics which are of interest to both theCSR and NeTS program (though maybe more to CSR).

Topics of discussion may include: Building MCS and IoT platforms with humans in the loop MCS and edge computing Privacy concerns for participants in such platform Incentive mechanisms for participants in such platforms Mobility analysis and coordination

Systems for AR/MR/VR

Felix Lin and Karthik Dandu

Location: Jefferson

This session will focus on OS/networking challenges raised by augmented reality (AR), mixed reality (MR), and virtual reality (VR), e.g. those backed by 360-degree or volumetric videos. We will identify key research questions and discuss potential solutions. Sample topics include:

- New IO technologies for spanning AR/MR/VR experiences over multiple devices

- New OS abstractions and facilities for storing and AR/MR/VR contents, e.g. for space efficiency and fast retrieval

- New frameworks for rendering AR/MR/VR with high quality and low delay comparable to human vision

- Domain-specific power management for AR/MR/VR apps

- New OS facilities for sharing resources and amortizing costs among co-running AR/MR/VR apps

- New network mechanisms for delivering large AR/MR/VR streams

- New facilities for trustworthy AR/MR/VR experience

Networking and Machine Learning

Xiaowei Yang

Location: Alexandria

This breakout session aims to discuss the relationship between networking research and ML techniques and applications. Specifically, we hope to discuss the following questions:

What are your experiences of using ML techniques in your research? What have you found useful?

What have you found challenging? What lessons can you share?

What challenges in the field of networking are made easier by ML techniques? Traffic engineering, traffic anomaly detection, capacity planning, congestion control, âĂę?

How can networks better support ML applications?

Looking forward, how do you envision networking research shapes ML applications or vice versa?

Edge Computing: The where, the why, and the how

Misha Rabinovich, Aruna Balasubramanian

Location: Salon 5

We are going through a burst of activity around edge computing. From industrial consortia to new academic conferences, there is a staggering amount of effort devoted to shaping and getting this area off the ground. A number of diverse applications in such areas as IoT, sensor networks, pervasive computing, gaming, streaming, virtual and augmented reality, are re-casted as edge applications. However, there is no commonly accepted system model that properly characterizes the edge computing environment, and this has become a significant handicap in placing the myriad of technical innovation in proper context. Where is the ?edge? situated between the end-devices and the central cloud? Will edge-native or automatically partitioned applications drive edge computing deployments? Where would the edge computing occur ? at the impoverished ubiquitous servers at the cell towers, the lighting polls, etc., or more powerful computing facilities in metro-area data centers? Do we foresee edge deployments to be as ubiquitous as that of WiFi routers? Will we see one in each home and coffee shop, or do we assume that the edge will be deployed at universities and enterprises but will not be ubiquitous? What should our assumptions be about the edge capabilities -- are they compute, network, or memory bottlenecked? And ultimately, what is the killer app for edge computing, which would not merely benefit but be enabled by this environment? This session will discuss these and other questions to help providing a focus for edge computing research.

Internet-scale Distributed Systems and Services

Ramesh Sitaraman

Location: Manassas

This session pertains to challenges that arise in the context of large distributed networks, such as Content Delivery Networks, Cloud Systems, etc (Think half-a-million edge servers in hundreds of locations). Future research challenges pertain to novel edge services, web/video/application performance, security, sustainability, and operations.

Example Challenges:

1) 360 Video Delivery: Imagine delivering the soccer world cup in real-time to a billion users who are wearing VR headsets for an immersive experience. The combination of high bitrate, scale, and small latency (20 ms of motion-to-photon latency to avoid cyber sickness) is much beyond the realm of what video delivery networks can do today.

2) Edge Security: The CDN edge is the perimeter behind which much of the worldâĂŹs major online assets are deployed (both commercial and government). How do we defend this perimeter in a distributed and real-time fashion to prevent DDoS and other attacks.

3) Sustainability: Internet-scale networks consume large amounts of energy, more than some mid-sized countries. How do we decrease the energy consumption of these large networks without sacrificing performance, reliability, and security?

4) Operations: The problems that arise in operating highly-distributed networks of this scale are very challenging and poorly researched. Key issues include provisioning, deployment, dynamically adapting policies to changes induced by billions users, thousands of ISPs, thousands of content providers, and constantly-evolving workload characteristics.

Scalability, reliability, and systems challenges in IoT

Prashant Shenoy, Saurabh Bagchi

Location: Rosslyn 1

As the popularity of IoT devices and systems continues to grow, future IoT systems and applications, particularly for smart cities, will see deployments of large numbers of IoT devices, resulting in new challenges in systems and network design, This session will discuss the following research questions. For each, we will survey from the participants what is the state-of-the-art and state-of-the-practice, discuss challenging technical problems and what are promising directions being explored today.

1. What system challenges need to be addressed in designing IoT systems with very large numbers of devices? How can these systems scale from a performance standpoint when they are composed of millions or billions of devices?

2. What are the reliability challenges that emerge when systems are run at such large scales? Do the characteristics of IoT devices (homogeneity, dependence on physical environment, resource constraints, thin or no OS) bring out new reliability challenges?

3. What are the energy and power issues in designing large-scale IoT systems? How can techniques such as energy harvesting be used to make IoT devices self-powered? How can protocols be made more energy efficient? How can protocols be developed to account for intermittent execution?

4. How should IoT devices leverage emerging wireless networking technologies such as LoRa and 5G?

5. How do specific smart city application domains such as connected healthcare, renewable energy, buildings, transportation impact system design? For example, how do their requirements impact whether IoT data processing is done on-device, at the edge, or in the cloud?

Cloud, Serverless Computing and HPC

Kirk Cameron, Xipeng Shen

Location: Mount Vernon

Cloud Computing and High Performance Computing (HPC) have many differences, but share many common concerns, such as computing efficiency, scalability, throughput, cost, energy efficiency. This session aims to examine the differences and potential connections between Cloud Computing and HPC, in the backdrop of recent advances in cloud execution models (e.g., serverless computing), workload evolvement (e.g., machine learning-based applications), new trends in scientific computing, and the emergence of new hardware (memory, accelerators, etc.). Example questions include what are the grand challenges in each of them; what common challenges they share; what are the new opportunities; what synergy can be built between Cloud Computing and HPC; what can be provided to cultivate the synergy.

Network and systems security

Ang Chen

Location: Salon 6

Computer networks and systems are foundational to modern applications. Despite significant progress in performance, reliability, and many other dimensions, we have not seen commensurate growth in security. Retrofitting security in these systems leads to temporary fixes at best, as the resulting solutions tend to be very brittle. We need a more fundamental approach where security is explicitly designed into the system as a "first-class" goal. This could be achieved, for instance, using principled approaches to system design and development, or using formal methods to provide strong guarantees. We will discuss recent trends in networking and systems, such as network programmability, resource disaggregation, etc., and examine their security implications. Can we leverage these trends to make the future more secure? Would these trends lead to new, unforeseen security problems? We hope to achieve, through community effort, a systematic approach to security in network and systems design.

Reproducability in Systems/Networking Research

Nick Feamster, Alex Snoeren

Location: Lee

In this breakout, we plan to discuss the following questions:

1. What are current best practices regarding (a) The reproducibility of published research in the CSR/NeTS community? (b) The release of source code and data to facilitate such reproducibility?

2. What are the current impediments to fostering more reproducibility in published research in the CSR/NeTS community?

3. How can the NSF encourage researchers to release artifacts that assist in the reproducibility of published research in the CSR/NeTS community?

4. What types of activities do other research communities do to foster reproducibility? How do researchers engage in those processes? Are they successful?

5. What is the appropriate way to address reproducibility for CSR/NeTS-funded research that is (completely or partially) based on proprietary datasets, systems, or code that might make the research difficult to reproduce?

6. Should the NSF alter its framing of proposal review or reporting processes (e.g., articulation of broader impacts, additional information requested at reporting time) to help implement any of the recommendations discussed in the above questions?

Emerging Hardware

Hadi Esmaeilzadeh, Don Porter

Location: Rosslyn 2

We are at a point where traditional hardware scaling techniques have plateaued, both in CPUs and in storage technologies, and designers are exploring a range of more restricted or special-purpose designs. For instance, hard-drive vendors are scaling capacity with shingled and interlaced magnetic recording, which introduce considerable, non-backward compatible restrictions on I/O patterns. Of course, this can be somewhat hidden behind a more complex layer of firmware, but at a significant hardware cost and significantly perturbing application performance.

This creates new opportunities and challenges for researchers:

1) How to get the most from an increasingly constrained hardware budget?

2) How to balance general-purpose versus special-purpose designs?

3) How to redefine hardware/software abstractions that enable updating specialized hardware without redoing its software stack?

4) How to write future-proof software that will not only work, but perform well, on hardware that did not exist when the software was written?

5) How to move from domain-specific hardware to domain-specific computational stack?

6) How to do performance analysis and tuning when performance-opaque firmware has a first-order impact on performance?

7) How to facilitate developers designing their own special-purpose hardware?

Integrated sensing, computation and feedback in wearable devices, from e-tattoos to near-zero computation to digital therapy

Roozbeh Jafari, Hassan Ghasemzadeh, Tinoosh Mohsenin

Location: Fairfax

No abstract available yet.