Systems, Networking, and Cybersecurity Ph.D. Qualifier Exam

Spring 2021

The following questions relate to the papers in the reading list on the Spring 2021 qualifier webpage (https://people.cs.vt.edu/vbimal/qual/). For full citations, please see that reading list. Before starting, read and understand the following guide provided by Virginia Tech: Avoiding Plagiarism: A Guide For Graduate Students at Virginia Tech (https://graduateschool.vt.edu/content/dam/graduateschool_vt_edu/graduate-honor-system/avoiding-plagiarism-short-guide.pdf). In your answers, you must avoid unattributed direct quotations and paraphrases and use proper documentation of all sources you use. This requires that you include a bibliography in your response. Failure to follow these guidelines Represents a violation of Virginia Tech's Honor Code and will result in a score of 0.

1. The following questions relate to the paper "Motivation for and Evaluation of the First Tensor Processing Unit".
    a. What is the new problem that the paper looks at?
    b. What are some key insights in the paper?
    c. What are the key sources/causes of performance benefit of TPU over CPU and GPU?
    d. What are the key sources/causes of energy benefit of TPU over CPU and GPU?
    e. Why is there a tradeoff between response time and throughput for DNN inference?
2. The following questions relate to the paper "A Hardware Accelerator for Tracing Garbage Collection".
    a. What is the new problem that the paper looks at?
    b. Describe the high-level concepts of garbage collection (GC). Describe why GC is needed and its high-level steps.
    c. What are the key sources/causes of performance benefit of the hardware accelerated GC in the paper over a CPU-based GC?
    d. What are the key factors that differentiate this work from prior works on hardware support for GC?
    e. What are some major drawbacks of this work?
3. The following questions relate to the paper "MicroScope: Enabling Microarchitectural Replay Attacks".
    a. According to the paper, what are the key benefits that Intel SGX is supposed to provide?
    b. What is the new problem that the paper looks at?
    c. What are some key insights in the paper?
    d. Why does MicroScope require flushing the page table entries of the replay handles?
    e. What is the unique advantage of MicroScope over other attack schemes?

4. The following questions relate to the paper “Fresher content or smoother playback? A brownian-approximation framework for scheduling real-time wireless video streams”.
    a. Summarize the motivations, main objectives, models, approaches, main results, and contributions of this paper.
    b. What is the capacity region for Quality of Experience (QoE)? What does QoE-Optimality mean?
    c. What is the key design tradeoff in the considered scheduling problem for real-time wireless video streams? Does this tradeoff exhibit different behaviors under different traffic loads?
    d. What is the purpose of introducing the Brownian approximation? What is the intuition behind the Brownian approximation? How is the accuracy of the Brownian approximation justified?
5. The following questions relate to the paper “On the Power of Randomization for Scheduling Real-Time Traffic in Wireless Networks”.
    a. Summarize the motivations, main objectives, models, approaches, main results, and contributions of this paper.
    b. What is the definition of efficiency ratio? What does it mean if an algorithm can achieve an efficiency ratio of 0.5?
    c. What is the intuition behind the design of randomized scheduling algorithms?
    d. What is an interference graph? While AMIX-ND (Algorithm 2) is designed for collocated networks, AMIX-MS (Algorithm 3) is designed for general interference graphs. Now, suppose one applies both AMIX-MS and AMIX-ND to collocated networks, which one achieves a better efficiency ratio?
6. The following questions relate to the paper “Rateless Codes for Near-Perfect Load Balancing in Distributed Matrix-Vector Multiplication”.
    a. Summarize the motivations, main objectives, models, approaches, main results, and contributions of this paper.
    b. What is the problem of stragglers in distributed matrix-vector multiplication?
    c. Describe different straggler-mitigation strategies and compare their pros and cons.
    d. The proposed rateless codes exploit the linearity of the computation (matrix-vector multiplication). What if the computation is more sophisticated, e.g., non-linear? Can rateless codes still be applied?
7. The following questions relate to the paper "TESSERACT: Eliminating Experimental Bias in Malware Classification across Space and Time”.
    a. Briefly describe the key ideas in the paper.
    b. Give an example of spatial and temporal bias in a security context different from the one considered in the paper (i.e., other than malware classification).
    c. What is "concept drift"? How is concept drift related to temporal bias in the context of evaluating malware classification?
    d. Explain the intuition behind the proposed "Area Under Time" metric, and why is it useful?
    e. Is TESSERACT applicable in all security settings where ML is used (assume prediction tasks)? What are the challenges with using TESSERACT?

8. The following questions relate to the paper "High Precision Open-World Website Fingerprinting".
    a. What is a website fingerprinting attack? Is it a realistic attack?
    b. Why is the open-world setting more challenging, compared to the closed-world setting?
    c. Explain one website fingerprinting method (any method from the literature is fine). For the chosen method, briefly explain the features used. Are certain features more useful than others? If yes, which features are more useful?
    d. What does the paper mean by base rate fallacy in the open-world setting? Why is it a challenge?
    e. What is the intuition behind r-precision, and why is it useful?
    f. What are the key ideas behind the optimized classifiers proposed in the paper to improve open-world website fingerprinting?
    g. How can we defend against website fingerprinting attacks (in an open world setting)? Explain the intuition behind any defense discussed in this paper (or other prior work), and the challenges with building an effective defense.
9. The following questions relate to the paper "Using GANs for sharing networked time series data: Challenges, Initial Promise, and Open Questions".
    a. Explain a few applications of ML-generated synthetic data in computer networking. Do you think ML-generated synthetic data can help in your research domain? If yes, give an example.
    b. What are the challenges with generating synthetic network data using ML (in the context of this paper)?
    c. What are the advantages of GAN-based data synthesis, compared to using simulations or expert-driven models?
    d. This work proposes modifications to a standard GAN to generate time series data. Explain the intuition behind each modification, and why it is required.
    e. Is the proposed GAN useful in a predictive modeling setting (based on the results in the paper? Explain your answer highlighting benefits and/or limitations.
    f. Consider a network security setting where the task is to detect malicious web bots by analyzing web request traces. Do you think ML-generated synthetic data can help in such a case? If yes, would it be a straightforward application of the GAN described in this paper or are there new challenges?