Statement on Research, and Impact

SERHAT ARSLAN, Stanford University, USA

Ubiquitous Internet connectivity has become an essential utility like electricity and water with nearly 5 billion regular users [12]. It enables an immense amount of computing power, i.e., the cloud, for critical applications such as autonomous vehicles, online banking, and telehealth. Such applications make more than 330 billion Gigabytes of data travel every day in different kinds of networks, i.e., broadband or cellular [9]. With such a large amount of traffic, it is impossible to monitor or debug the entire system manually as it was done in the past. If we were to generate a logline (metadata) every time a packet is processed by the networking infrastructure, we would need a whole new and bigger Internet to communicate these logs, making it excessively expensive. Hence, the literature is limited in precisely measuring/understanding the state of the network and how much each user can be served by it. On the other hand, the emergence of network-heavy technologies such as distributed ML, IoT, edge computing, and virtual reality expeditiously increases the need to automatize the infrastructure management for optimal utilization of the capacity.

My mission in this landscape is to build systems, with strong grounding in theory, that can improve the performance of such emerging applications over the Internet with greater visibility into what is happening in the communication infrastructure. These systems leverage state-of-the-art *network programmability* and optimize the monitoring mechanisms. They extract only the essential bit of information with minimal overhead for maximum performance [4, 5, 10, 11]. Then, I design *closed-loop control* algorithms which exploit the collected metadata to allocate network resources with higher efficiency [7] and enable new applications that were not even feasible before [1, 2]. With such an end-to-end mission and a background in electrical engineering, I study all the layers of networking systems down from the hardware [3, 8] up to the application layer [2] in different kinds of networks. All of this work is available in the public domain to facilitate community efforts for collaboration and sharing. This is fundamental for tangible impact in the industry.

1 PROGRAMMABLE HARDWARE FOR ULTRA LOW LATENCY

Modern distributed applications send huge numbers of Remote Procedure Calls (RPCs) between large groups of servers. Therefore, reducing the processing times of these RPCs plays a significant role in the overall performance of the applications. With a novel architecture called NanoPU, we implemented the message reassembly, transport layer, thread scheduling, and core selection functions on hardware, and opened up a direct path between the network and the application code [8]. By delivering the incoming data straight into the register arrays of the processor cores without any software overheads, NanoPU accelerated the RPC processing times by an order of magnitude compared to the fastest commercially available Network Interface Cards (NICs).

An accelerated NIC-CPU co-architecture is not the end of the story for the lowest latency in data centers. The minimum RPC latencies are achieved when both the end-host *and* the network are optimized for this purpose. The end-host processing dictates how fast an incoming RPC is delivered to the application thread on the CPU whereas the network processing dictates how fast this RPC can be delivered to the remote end-host in the first place. The design of the transport layer and congestion control algorithm have big consequences on the network latency. Indeed, new protocols are constantly being proposed to reduce network latency for RPCs, but it is too expensive to implement these protocols on fixed-function hardware due to the costs of designing and deploying a new chip after each protocol.

Bringing programmability to accelerator architectures such as the NanoPU is crucial for enabling cloud service providers to design and deploy their favorite transport protocol and congestion control algorithm as the state-of-the-art makes progress. With this goal in mind, I designed NanoTransport – a programmable transport layer abstraction on hardware. It exposes interfaces to packet processing pipelines so that developers can implement new networking protocols and algorithms on the NIC with 3 orders of magnitude lower transport layer processing latency [3]. The ideas proposed in this work are being evaluated by Intel's smart-NIC team to include in future products.

2 UTMOST PRECISION FOR CONGESTION CONTROL

To continue lowering latencies in networks, I explored the transport layer protocols as well as congestion control algorithms. In [11] we identified how to best utilize the network links with minimal queuing. While working on traffic patterns that create congestion, I made the observation that data center workloads are evolving towards highly parallel, and mostly bursty applications, i.e., distributed ML training. Therefore even a single incorrect or slow congestion control decision ends up creating tens of microseconds of tail queuing or under-utilization which prolongs the flow completion times.

To avoid incorrect or slow congestion control reactions, it is vital to precisely measure the congestion that takes place at the switches. In [5], I showed that collecting the instantaneous queue occupancy from switches reveals high-resolution insights about the state of congestion which can not be obtained through popular proxy signals in the industry, i.e., packet-loss, delay (RTT), and ECN. Hence it was inevitable to use this information for congestion control purposes.

Later on, I collaborated with Google to design a novel congestion control algorithm, Bolt [4], that reduces the tail latency in Google's infrastructure by 80% without a noticeable overhead compared to the existing production algorithm. To achieve such a drastic improvement, Bolt leverages precise queue occupancy or utilization information from programmable switches with Sub-RTT feedback delay which is the shortest feedback loop any congestion control algorithm can ever have. With such timely and precise feedback, Bolt senders avoid making implicit estimations about the severity and exact location of the congestion or the number of competing flows, freeing them from manually tuned hard-coded parameters and inaccurate reactions. As a result, up to 3× faster flow completions are achieved with Bolt compared to the state-of-the-art.

3 DIRECTIONS FOR FUTURE RESEARCH

My existing work answers the question: *what level of visibility and programmability is required and/or achievable to optimize communication latency in data center networks?* Moving forward, I intend to continue building **interpretable, high-performance, software-defined networks and systems** with a focus on 3 rising concerns about how networks serve people today.

3.1 Protocols and Algorithms for Energy-Aware Networking

The Internet consumes 400-800 Terawatt-Hours of electricity each year accounting for approximately 3% of the global electricity usage, a figure projected to rise significantly in the future with the rise of large generative AI models such as ChatGPT. Therefore, any improvement in the energy efficiency of communication networks will be beneficial—both environmentally, and financially for network owners. Hence I plan to investigate: (a) *What are the energy consequences of the existing networking algorithms and protocols, i.e., hardware offloading, routing, load balancing, and congestion control?* (b) *Is there a trade-off between energy efficiency and performance?* (c) *Given the energy footprint of networking protocols and algorithms, how can we optimize them with existing trade-offs?*

There is a large body of energy-aware networking research in the wireless/mobile setting where the radio communication is the main factor of energy consumption. Yet, the questions I asked above

are mostly unanswered in the wired settings, i.e., data centers and wide area networks, making it a large potential area of new research. In this regard, I started exploring how congestion control in particular affects the energy bill of data centers and showed that energy savings of around $10 Million/year can be obtained when congestion control algorithms approximate the Shortest Remaining Processing Time First paradigm as opposed to fair resource allocation per-flow [6]. This result suggests that we as a community should rethink our current approach to congestion control and potentially more for substantial savings and environmental impact. Hence, I plan to collaborate with large network owners (e.g., Google, Meta, Microsoft, AT&T, Verizon) through grants, and student internships to obtain such savings while building a greener future.

3.2 On-Demand Performance for Trust-Free Cellular Networks

Most countries are serviced by only a handful of Mobile Network Operators (MNOs) offering users little or no choice of provider, and little incentive for MNOs to invest in new infrastructure or services. Such uninvested infrastructure mostly serves users on a best-effort basis, constraining the quality of service. Lowering the barrier to entry for new providers promises to be the solution for increasing competition and service quality. With this in mind, I proposed a preliminary decentralized wireless (DeWi) architecture that aligns the incentives for providers and users to participate in a more competitive cellular connectivity market [1]. The key is to allow users to dynamically choose the network to join, creating competition among operators, large and small.

For this to work, the user and the DeWi operator need to authenticate each other, the operator needs to offer a service to the user without a legal agreement in advance, and the user needs to decide if the operator is providing the service as promised while appropriately paying for the provided service. All of this requires an element of trust between users and providers, as well as a means to pay for service, leading to the question: *Without any prior legal contract between users and providers, how can users know they are receiving the promised service, and how can providers know they will be paid for the service they provide?*

I am developing a framework, called d-Cellular, that can form the foundations of the decentralized environment described above [2]. It leverages network programmability for extracting telemetry from base stations and user equipment to reach a consensus on the quality of innovative services. It also introduces a novel negotiation protocol for users and providers to settle on the terms and payments for the service. This work received the **Best Paper** award at the IEEE Future Networks World Forum 2023. However, more aspects of cellular connectivity need to be studied for this environment to become commercially available. For example, (a) *how can we develop the right incentive mechanism for inter-provider handovers without prior legal agreements?* (b) *How can privacy and security of users be protected in such a decentralized environment?* (c) *What kind of spectrum-sharing mechanisms or protocols are needed to enable this decentralized environment?* Given that citizen welfare is the ultimate utility for this work, I seek to work with government agencies (e.g., NSF, DARPA, or ERC), and standards bodies (e.g., IETF) to help fund this stream of research.

3.3 Systems for Distributed Machine Learning

As the capability and use cases of large language models grow, the pressure on the systems that run/train those models also increases in terms of optimal resource allocations. The communication pattern and the user behavior for such machine-learning applications are unique and require special protocols and algorithms to perform well in constrained environments. In this regard, I am curious to ask: (a) *How differently would distributed ML models train in parallel if they had a complete view of the demand for networking resources from other contending ML tasks?* and (b) *what abstractions are required for ML models to interact with the underlying physical systems they are running on?*

REFERENCES

[1] SVR Anand, Serhat Arslan, Rajat Chopra, Sachin Katti, Milind Kumar Vaddiraju, Ranvir Rana, Peiyao Sheng, Himanshu Tyagi, and Pramod Viswanath. 2022. Trust-Free Service Measurement and Payments for Decentralized Cellular Networks. In *Proceedings of the 21st ACM Workshop on Hot Topics in Networks* (Austin, Texas) *(HotNets '22)*. Association for Computing Machinery, New York, NY, USA, 68–75. https://doi.org/10.1145/3563766.3564093

[2] Serhat Arslan, Ali Abedi, and Sachin Katti. 2023. d-Cellular Trust-Free Connectivity in Decentralized Cellular Networks. In *2023 IEEE Future Networks World Forum (FNWF)*. IEEE.

[3] Serhat Arslan, Stephen Ibanez, Alex Mallery, Changhoon Kim, and Nick McKeown. 2021. NanoTransport: A Low-Latency, Programmable Transport Layer for NICs. In *Proceedings of the ACM SIGCOMM Symposium on SDN Research (SOSR)* (Virtual Event, USA) *(SOSR '21)*. Association for Computing Machinery, New York, NY, USA, 13–26. https://doi.org/10.1145/3482898.3483365

[4] Serhat Arslan, Yuliang Li, Gautam Kumar, and Nandita Dukkipati. 2023. Bolt: Sub-RTT Congestion Control for Ultra-Low Latency. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. USENIX Association, Boston, MA, 219–236. https://www.usenix.org/conference/nsdi23/presentation/arslan

[5] Serhat Arslan and Nick McKeown. 2020. Switches Know the Exact Amount of Congestion. In *Proceedings of the 2019 Workshop on Buffer Sizing* (Palo Alto, CA, USA) *(BS '19)*. Association for Computing Machinery, New York, NY, USA, Article 10, 6 pages. https://doi.org/10.1145/3375235.3375245

[6] Serhat Arslan, Sundararajan Renganathan, and Bruce Spang. 2023. Green With Envy: Unfair Congestion Control Algorithms Can Be More Energy Efficient. In *Proceedings of the 22st ACM Workshop on Hot Topics in Networks* (Cambridge, MA) *(HotNets '23)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3626111.3628200

[7] Denise Barton. 2021 [Online]. Project Pronto: Celebrating Our One-Year Anniversary. Open Network Foundation. https://opennetworking.org/news-and-events/blog/project-pronto-celebrating-our-one-year-anniversary/

[8] Stephen Ibanez, Alex Mallery, Serhat Arslan, Theo Jepsen, Muhammad Shahbaz, Changhoon Kim, and Nick McKeown. 2021. The nanoPU: A Nanosecond Network Stack for Datacenters. In *15th USENIX Symposium on Operating Systems Design and Implementation (OSDI 21)*. USENIX Association, 239–256. https://www.usenix.org/conference/osdi21/presentation/ibanez

[9] IDC and Statista. 2021 [Online]. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025. Statista Inc. https://www.statista.com/statistics/871513/worldwide-data-created/

[10] Yanfang Le, Jeongkeun Lee, Jeremias Blendin, Jiayi Chen, Georgios Nikolaidis, Rong Pan, Robert Soule, Aditya Akella, Pedro Yebenes Segura, Arjun singhvi, Yuliang Li, Qingkai Meng, Changhoon Kim, and Serhat Arslan. 2023. SFC: Near-Source Congestion Signaling and Flow Control. arXiv:2305.00538 [cs.NI]

[11] Bruce Spang, Serhat Arslan, and Nick McKeown. 2021. Updating the theory of buffer sizing. *Performance Evaluation* 151 (2021), 102232. https://doi.org/10.1016/j.peva.2021.102232

[12] Song Bac Toh. 2020 [Online]. The Argument For The Internet As A Utility: Is It Time To Change How It's Delivered? Forbes Technology Council. https://www.forbes.com/sites/forbestechcouncil/2020/06/17/the-argument-for-the-internet-as-a-utility-is-it-time-to-change-how-its-delivered/?sh=4efd0fee7729