The Next Wave in Cloud Systems Architecture

Amin Vahdat, Google

Dejan Milojicic, Hewlett Packard Labs

Twenty-five years ago, no one expected services to be seamlessly and consistently delivered across campuses and edge. It is now possible to envision machine learning-powered services capable of scaling to the reliability, performance, and security requirements of billions of users worldwide.

Because of their visionary nature but also their highly volatile success-failure ratio, predictions have always been popular among a wide spectrum of audiences. The COVID-19 pandemic has turned this popularity into a necessity. Time constraints and a lack of thorough experimentation necessitated a dependence on predictions: which vaccines to use, when to close or open countries, when and how to reopen offices, and so on. Suddenly, predictions became a way of life.

After 10 years of making technology predictions for the IEEE Computer Society (see the press releases at https://www.computer.org/press-room), two successful special issues on technology predictions in *Computer* (December 2020 and July 2021), and panels at numerous events (SWITCH 2020; the IEEE Computer Society Computers, Software, and Applications Conference 2020; and so forth), we have decided to initiate the column "Predictions."

We decided to invite a reputable guest for each "Predictions" column. A guest should have a demonstrated vision for the future of computing and a track record of delivering on that vision. Our ideal guest should have a combination of academic rigor, deep technology knowledge, and an understanding of business implications.

Amin Vahdat, an engineering fellow and vice president at Google, is a perfect match for the inaugural "Predictions"

Digital Object Identifier 10.1109/MC.2021.3099955
Date of current version: 24 September 2021

column. He leads systems infrastructure at Google, and prior to that, he was a professor of computer science at the University of California San Diego and Duke University. Vahdat brings a wealth of experience in compute, operating systems, accelerators, storage, and networking. In his professional role, it is essential to predict the workload evolutions, customer demands, and trends in technology.

In this article, Vahdat takes us on a prediction tour of the future of systems infrastructure, addressing topics such as the emergence of new accelerators, impact of the growth of data, disaggregated data center designs, the evolving roles of operating systems and programming languages, application delivery models, the roles of networking, and much more. I hope that you enjoy reading this column as much as I enjoyed working with him to deliver it.

DEJAN MILOJICIC: Interesting technology and business implications are driven by the imminent end of Moore's law, such as a plethora of innovative accelerators, the rise of photonics, and the introduction of new memory technologies. Which one of these is a fundamental disruptor and which is a temporary transition?

AMIN VAHDAT: The computing industry has delivered incredible new business and societal capabilities through sustained, exponential improvements in the scale and performance efficiency of the underlying hardware and software. However, the underlying hardware trends that have powered this march over the past few decades are slowing or stopping with each generation of hardware, including general-purpose CPUs, dynamic RAM (DRAM), storage devices, and network elements, delivering incrementally less benefit relative to the previous generation, alarmingly with lower levels of reliability in many cases. This will mean that software optimizations, including novel algorithms, protocols, and replication strategies, will play an important role in maintaining the cadence of regular system-efficiency improvements.

I believe that all the technologies you point to, and more that are yet to come, are fundamental disruptors we will need to accommodate. For the first time in decades, a range of new technologies is actually a must have—rather than a nice to have—to sustain exponential growth in compute capability. Current capabilities, like real-time language recognition and translation, would simply not have been possible from either a cost or power standpoint without hardware acceleration through accelerators like tensor processing units and GPUs. Similarly, exploding demand for video consumption on a range of devices and networks could not be effectively supported without hardware-supported video coding. These examples are just the beginning of an increasing array of segment-specific hardware architectures that can deliver step-function improvements to domains such as analytics, high-performance computing, and real-time serving.

On the memory and storage front, we are hitting a wall in available improvements in cost per gigabyte, especially as we try to simultaneously meet the needs of different applications that are individually latency-, bandwidth-, and capacity-bound on the same hardware. We now have underlying hardware technologies that can strike tradeoffs between cost, latency, bandwidth, and capacity somewhat independently, meaning that we will deploy servers with a range of memory and storage configurations, enabling higher-level scheduling infrastructure to map applications to available hardware configurations in the most cost- and performance-effective manner.

Photonics have always played a critical role in the scale of available computation in the data center. Today, silicon photonics, or integration of photonics onto integrated circuits, is emerging as a requirement for continued expansion of power-efficient data rates. We are already seeing this play out in traditional transceiver design, with the next question being the role silicon photonics will play in mainboard and ASIC design, with reliability for hundreds of optical engines integrated on a single board or ASIC being one of the most pressing challenges.

MILOJICIC: Data center design has been motivated by cost and performance for quite some time. Given the rate of evolution of individual technologies, is there an opportunity for rethinking interfaces to optimize performance? For example, accelerating access to data and/or interconnections by deploying computations closer to it. The so-called in- or near-memory computations and similar for interconnects with smart network interface controllers (NICs), data processing units, and so on?

VAHDAT: Yes, absolutely. The rate of growth of data and the increasing computation required to summarize and transform data into suitable formats is growing well beyond what our CPU budgets will allow for power- and cost-efficient processing. Today, we strive to treat data as a large, location-independent "blob" with intermediate processing steps composed together in a manner not far from classic UNIX-pipe composition. While convenient, this means that data must often move back and forth across kilometers of compute within a large campus, compressed, encrypted, decompressed, decrypted, transformed, processed, and replicated dozens of times. Processor architects optimize pJ/bit, and it is likely that we will need similar measures in considering the cost of end-to-end processing of every ingested byte of data.

The resulting metrics will lead to a number of necessary innovations in system design. First, some amount of computation power will be needed closer to data stores with closely coupled data ideally placed closer to one another in a logical topology. In the extreme, new hardware will colocate server-class processing power or even hardware accelerators in the same server or rack enclosure as the data. However, more generally, such sophisticated processing will require deep understanding of the logical data processing pipeline and an orchestration and runtime layer able to place and migrate data computation according to overall data lifetimes. Understanding the provenance of derived data will open opportunities to create efficient, intermediate formats, perhaps allowing different processing elements to handle a range of native formats and summaries, ideally with a "data compiler" capable of placing the right preprocessing at the right point for a specified data flow. Smart NICs and their generalization will form the nervous system in the above runtime, likely managing both the movement of the data from location to location but also its intermediate processing.

While there is substantial complexity required to realize this, the next wave of efficiency improvements will come from end-to-end performance considerations rather than lower-level measures such as million instructions per second and storage capacity/input–output per second. How much processing, energy, and storage is required to deliver an insight or result in the end-to-end composition? There are integer factors of performance, cost, and energy efficiency available in our infrastructure when viewed through this lens.

MILOJICIC: Do we need new operating systems, programming languages, programming environments, and middleware to make all this seamlessly work, or would those from the previous era suffice?

VAHDAT: Software will drive the success of this next wave in infrastructure evolution. What we will see is that the shape of compute containers can change dynamically at runtime, perhaps adding and removing memory, storage, or accelerators based on the needs of the services currently scheduled on the server. Real-time performance monitoring will support isolation among tenants while accounting for antagonists at multiple levels of the system hierarchy; for example, L3 or DRAM capacity versus bandwidth requirements, and behaviors of individual applications. The level of malleability in the composition of servers will need to fundamentally change from the operating systems' current view of a fixed "hardware" environment from boot to shutdown.

At the distributed systems level, services will require much more predictability and determinism in accessing remote storage and compute resources. This will, in turn, require a runtime capable of delivering isolation among millions of concurrent communication channels, all while ensuring efficiency and the mapping of requests to the underlying replica or resource best capable of fulfilling them.

MILOJICIC: How can we most effectively accomplish hardware–software co-design to account for optimizations adequately at different levels of the stack?

VAHDAT: The level of visibility we have into the dynamic nature of real data center computation is incredibly limited. We understand what industry benchmarks like SPEC and TPC-H look like in isolation and know how to optimize them for hardware. However, actual data center workloads are increasingly heterogeneous, multitenant, and distributed, substantially reducing the predictive power of existing benchmarks.

Moving forward, we will need to start with much deeper hardware measurement infrastructure to characterize workloads in the wild. This can, in turn, support new benchmarks that account for wider variability in computational structure, potential hardware offload capability, and the fundamental distributed and multitenant nature of modern computation. Single-server, single-tenant benchmarks cannot be the basis for projecting the value of future infrastructure. An open question is the required scale, heterogeneity, and variability for sufficient predictive power.

MILOJICIC: To optimize the data center cost, disaggregated design has been taking off, with accelerators being separated from compute—deployed in racks-size units—disaggregated memory is emerging and storage has been deployed (for example, storage area network and network attached storage) for quite some time. Is this trend going to continue in the future?

VAHDAT: The interesting thing about disaggregation is that, while it holds fundamental technical appeal, including for me personally, its progression has been rather limited since we deployed the first disaggregated storage solutions. Systems like GFS (the Google File System) showed how to treat thousands of hard drives spread across the data center as the underlying storage for a distributed and disaggregated storage system. However, the bandwidth and latency of hard drives are modest compared to the speeds of data center networks and, as importantly, the software layers responsible for managing I/O requests.

True disaggregation of SSDs has been slower to progress in part because local, dedicated devices have high bandwidth and low latency relative to data center network speeds but also because there can be little software on the path between a client request and device access while maintaining the illusion of "local" access to a disaggregated device.

While accelerators are deployed in rack-scale units, they are also typically deployed with dedicated servers and often dedicated secondary networks. There is very little virtualization support for accelerators today.

I do believe that the next level of end-to-end system efficiency and flexibility will require support for disaggregation. However, it will also require some breakthroughs in hardware/software organization because the assumptions built up over decades about local device access over a dedicated PCIe link will be hard to break.

MILOJICIC: Can you tell us how networking is evolving in the data center and how it is affected by the trends and needs of data and compute?

VAHDAT: The network today constitutes the smallest portion of our data center spend, but it offers some of the biggest challenges and opportunities. For example, the network is often the largest source of large-scale failures and extended downtime. Because it fundamentally connects computation and storage together at scale, a failure in the network most easily cascades resulting in large-scale outages. Similarly, managing the lifecycle of the network from turnup to upgrades to turndown is often the most complex and toilsome for the operations team. Tying the two together, many network outages are correlated with network operations.

Given the increasing societal reliance on compute infrastructure, and the thousands of seemingly independent cloud services, we multiplex onto shared underlying network infrastructure, the reliability of the network must fundamentally improve. Since, in the end, individual network components and systems are unavoidable at cloud's scale of deployment and its rapid rate of evolution, solutions likely lie in designs that ensure hard levels of network isolation and multiplexing of services onto multiple independently operated network infrastructure.

Network performance and performance predictability will also be critical, as we discussed earlier. The disaggregated and data-centric data center will require not just higher performance and lower latency but predictability and isolation under a range of highly variable and bursty communication patterns. This will require us to continue the evolution of software-defined networking to enable visibility all the way to the end applications and their composite, multinode communication patterns, with microsecond-granularity actuation loops allocating bandwidth and rate limits to meet the real-time application SLOs, focusing on remote procedure calls and coflows rather than lower-level metrics focused on packets.

MILOJICIC: Data and compute delivery models have evolved from on premise to public cloud, to hybrid cloud to edge. What is the next step in delivery models?

VAHDAT: One important note is that, while there is a lot of enthusiasm around emerging compute delivery models from public to hybrid to edge cloud, we are still in the very early stages of the migration to these emerging hosting approaches. The vast majority of computing and storage still runs in enterprise data centers with many challenges that must be overcome before we can get to baseline modernization of digital infrastructure.

This is one reason why hybrid connectivity and the ability to seamlessly integrate on-premises infrastructure with cloud hosted and managed infrastructure is so critical to enabling more rapid migration. Here, there are many practical and research challenges to providing a "one-network" administrative view across multiple sites and cloud providers with individual services ideally, transparently, and incrementally migrating from one side to another.

As such, the next step in the delivery model must take on the software tooling, monitoring, and management necessary to make hybrid, multicloud, and edge-deployment scenarios seamless from the perspective of individual businesses. Similarly, application developers will need the runtime support and consistent APIs to enable them to write once and run anywhere.

MILOJICIC: What is the role of virtualization, and how is it evolving? From bare metal to virtual machines to containers and most recently to functions as a service (serverless), virtualization has been increasingly getting more hardware support. What is the next step in evolution?

VAHDAT: Yes, this is a great observation. We are seeing the continuing evolution of virtualization to the point where individual bare-metal servers can be securely configured for individual customers, potentially allocated and reallocated at fine time scales. The hardware architecture, from the BMC, network connectivity, root of trust, and more to allow this in a multitenant data center has been understudied in academia and underdeveloped industry wide.

On the other extreme, we will need to move beyond "static slice of hardware" virtualization capability to more flexible and dynamic container shapes in support of higher-level application structure that benefit from dynamic scale out under failures and bursty access conditions. These containers will need to deliver requisite levels of isolation and security among colocated tenants, all while managing tail latency. As discussed earlier, this will mean that the compute, accelerator, memory, storage, and network resources available to a container will come and go with both the application and the container OS, evolving to understand and manage this new dynamism in line with the trends toward disaggregation of the underlying hardware components across the data center.

MILOJICIC: Will 5G and follow-on technologies affect data center designs? For example, will 5G make edge much more real time, putting more pressure on data centers? How much will edge impact data center design and evolution of computing technologies?

VAHDAT: 5G promises to bring affordable, lower latency, and predictable bandwidth connectivity to a whole new range of computing devices. Data rates will continue to explode, accelerating from the already-impressive place we are at today. One of the key questions will then be the tradeoff between: 1) very low-latency reaction times to data generation/the sensing to actuation loop, 2) data provenance and sovereignty requirements for the data, 3) the costs of wide area network transport, and 4) the inherent efficiencies associated with large-scale, centralized computing campuses built from the ground up for power and cooling efficiency relative to more expensive, smaller-scale, and less-optimized colocation facilities located in the heart of metropolitan areas. For example, can we afford to pay the tens of milliseconds in speed of light latency required to transport newly generated data to the nearest large-scale campus? Does the data need to be stored and processed within the jurisdiction of a particular company or country?

The reality is that the tradeoff will be application and scenario specific, which makes the emergence of 5G particularly interesting from the perspective of dynamically configuring machine learning inference, general-purpose computing, storage, and communication pipelines across local, metropolitan, and wide area network infrastructures.

MILOJICIC: Historically, there was a dichotomy between large-scale distributed systems and large parallel systems. It appears that the former have won, except for the needs of high-performance computing (HPC) and high-end analytics systems. Going forward, for economic reasons, there appears to be a convergence of general-purpose compute, artificial intelligence, HPC, and data analytics systems. Will a unified system design be possible to meet the requirements of traditional HPC and analytics applications?

VAHDAT: Yes, this is another really nice observation. Historically, parallel applications have required predictable, isolated, homogeneous, single-tenant execution environments to achieve reasonable performance given the regular nature of the underlying computation and communication loops. Distributed data center applications have, on the other hand, been loosely coupled, fault tolerant, multitenant, and capable of running on multiple generations of hardware and low-level software simultaneously.

While both evolutionary paths have been pragmatic, we are seeing a convergence between the two, also for pragmatic reasons. High-performance computing and machine learning are increasingly migrating to cloud environments where the availability of large-scale batch resources and efficient virtualization to achieve uniformity is making it attractive to run in more heterogeneous and unpredictable environments, with software and hardware providing the illusion of a uniform, dedicated execution environment. At the same time, the availability of efficient, microsecond-scale communication stacks with hardware support for isolated multitenancy is exposing a number of opportunities for traditional distributed applications to achieve much higher levels of performance and efficiency through reduction of expensive caching and data denormalization.

MILOJICIC: Any closing thoughts that you would like to leave our readers with?

VAHDAT: It's a really exciting time to be working in computing infrastructure. It was about 20–25 years that a combination of academic research and industrial breakthroughs laid the foundation of modern Internet service delivery. Back then, no one even dared dream that services would be seamlessly and consistently delivered across campuses and edge consisting of tens of thousands of commodity servers running open source operating systems all interconnected by a dedicated wide area and data center networks comparable in scale to the public Internet, with software-managed computing and storage providing the illusion of a single system image across these same exascale computing platforms.

And yet, this seemingly impossible vision is now considered standard practice. It is now correspondingly simple, or at least possible, to dream of new planetary-scale, interactive services that can seamlessly scale to the reliability, performance, and security requirements of billions of users across the globe for enterprises, large and small.

At the same time, the underlying forces and design patterns that have powered this last revolution in computing are slowing or stopping. Without the winds of exponential growth in compute and storage capacity at fixed cost at our back with the simultaneous headwind of increasing exponential growth of data rates and processing needs, we as a community have to develop fundamentally new design points focused on end-to-end distributed systems efficiency metrics, powered by new segment-specific hardware and increasingly sophisticated dynamic runtimes.

My personal excitement is fueled by the belief that this shift in thinking will enable a whole new set of capabilities that perhaps no one dares dream today. One higher-level prediction I feel comfortable with is that the new capabilities will afford a shift from planetary-scale interactive services to the real-time generation of proactive insights driven by secure, privacy-preserving data analytics frameworks capable of shifting through an ever-increasing explosion of available data, affording fundamental benefits from health care to the sciences to manufacturing and more. C

AMIN VAHDAT is an engineering fellow and vice president for systems infrastructure at Google, Mountain View, California, 94043, USA. Contact him at vahdat@google.com.

DEJAN MILOJICIC is a distinguished technologist at Hewlett Packard Labs, Palo Alto, 94306, California, USA. Contact him at dejan.milojicic@hpe.com.