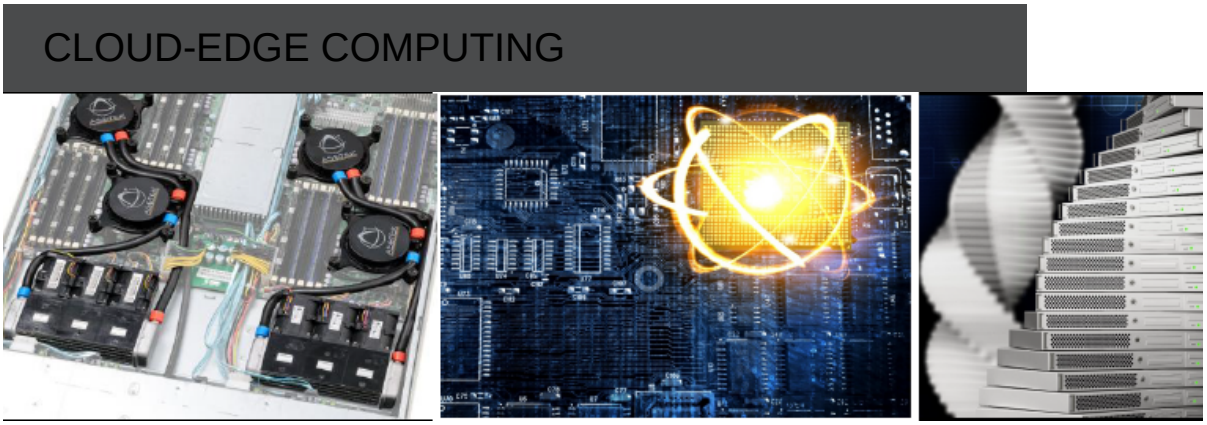
Cloud-Edge Computing



2020

2021-2022

2023-2024

2025+

The rise of application workloads requiring direct liquid cooling techniques is expected to grow; machine learning systems accelerated by GPUs or TPUs, are just one example. Google is one company incorporating direct-to-chip liquid cooling designs into its data centers .AI processing at the edge is also expected to bring direct liquid cooling to edge based data centers.

Edge computing is growing in popularity with hybrid models working in tandem with the cloud. The edge-to-cloud architecture will consist of “distributed clouds” that process most of the data at the edge while sending and receiving some data from centralized data centers, operated either by the end-user companies or by third-party cloud providers like Amazon Web Services and Microsoft Azure. Streaming video, machine learning, and other latency intensive applications will all benefit from this shift.

Virtualization of networks is a key theme that enables greater flexibility, choice, and potential operational savings. Network virtualization will help support the deployment of computing resources to the edge and 5G network slicing. For Communications Service Providers, network virtualization is expected to enable more customized service levels (bandwidth tiers), as well as, a move to vCPEs.

With a proliferation of micro data, traditional network management methods need to evolve in a way that supports DevOps and rapid application change. One way to reduce complexity is to adopt a common policy framework that can automate provisioning and managing of resources. Cisco’s Application Centric Infrastructure approach is an example of this new approach to managing networks in a sensor-driven world.

Data Localization – moving data and processing closer to the source -- reduces bandwidth utilization and latency while improving security. Data and compute localization allow devices to act locally on the data they generate while offloading management and analytics to centralized functions that only act on the data that is filtered and relevant. AWS IoT Greengrass is one service which supports the localization of compute and data and enables the deployment of optimized IoT edge-based services.

With improvements in hardware and software technology, it will become possible to process complex AI models at the source of the data, or the edge of the network rather than in the cloud. Issues like data privacy, connectivity, security will dictate the distribution of compute in the edge vs cloud AI. Improvement in ASICs like the use of novel graph-based AI architectures, combined with increasing ability to compress complex AI models in software will lead to a proliferation of Edge AI. In areas like autonomous driving, robotics, IoT, industrial and medical vision, mobile and personal computing we are likely to see AI models be increasingly deployed on the edge, allowing for both inference as well as real-time on-going training.

Communication Service Providers will be able to leverage the physical presence of their Radio Access Networks (RANs) for new "Wireless Edge" based computing and storage opportunities. This RAN edge will be able to support new low-latency, high bandwidth capabilities and location-based services, crucial for application and micro-service developers. SDN/NFV infrastructures will be a key requirement enabling WEC deployments.

Programmable Data Planes in the datacenter will migrate to the edge and customer premise networks; this will further enable field programmability of the packet forwarding functions of the network data plane to unleash new functionality and modify existing behaviors while maintaining real time processing of the traffic. Data plane edge programmability opens an array of new possibilities by treating packet processing as software that can adapt to events on the network as they occur. P4, and open source language, are examples of solutions which will be extended and used to configure programmable edge data planes.

In a converged infrastructure, the compute, storage, and network components are discrete unit that can be used “individually” for their intended purposes. A Hyper-converged Infrastructure combines compute, storage, networking, and a software defined management layer, including virtualization, into a single component which enables ease of use, management, and scaling.

The proliferation of sensors (sometimes referred to as “the swarm”) prompts rethinking the environment when it comes to networking and data sharing – building resiliency and redundancy into our architectures to enable capture and application of data in new ways. UC Berkeley is experimenting with new always-on infrastructure in their SWARM technology lab.

Microsoft is buying ten million strands of DNA from biology startup Twist Bioscience to investigate the use of genetic material to store data. The data density of DNA is orders of magnitude higher than conventional storage systems, with 1 gram of DNA able to represent close to 1 billion terabytes (1 zettabyte) of data. DNA is also remarkably robust; DNA fragments thousands of years old have been successfully sequenced…commercial viability of synthetic DNA storage is still some way off, but the technology itself works. Microsoft says that Twist, in initial trials, has shown that the process allowed full retrieval of the encoded data from the DNA. If the technology can be made cheap enough, it means that one day long-term data archiving could use the same technology as life itself.

Beyond IoT, we will see autonomous devices performing many tasks which today require human intervention. These devices may broadly range from infrastructure maintenance equipment to assistive technologies in the home. To achieve low cost and high efficiency, these devices will wirelessly offload much of their computing to near edge computing. High bandwidth and low latency networks will be a requirement.

AI powers the self-healing infrastructure or “self-driving IT”. Autonomous Data Centers and automation will be accelerated with the rise of repair robotics and self-healing infrastructures . This could take many shapes and forms such as software that can write or rewrite itself to prevent, amend or stop any faults, or hardware that can be replaced without humanlabor. Nanotechnology and the development of materials that can repair themselves will be key.

Intent Based Networks (IBNS) are networks which can automatically respond and optimize themselves in real-time using network orchestration software that implements a set of pre-defined policies. Advances in Machine Learning are helping to bring IBNS to life.

Classical information is encoded with binary “bits” of 1s and 0s. In contrast, Quantum Computers operate with “qubits” which enable them to process multiple operations simultaneously. Quantum computing promises to dramatically speed up processing for crypto-security applications which factor very large numbers, as well as, for other hard optimization problems. IBM currently provides a 16-qubit quantum computer, in the cloud, for use in experiments. Affordable and commercially available quantum computers will be slow to emerge because quantum computers need to operate at absolute zero temperatures.