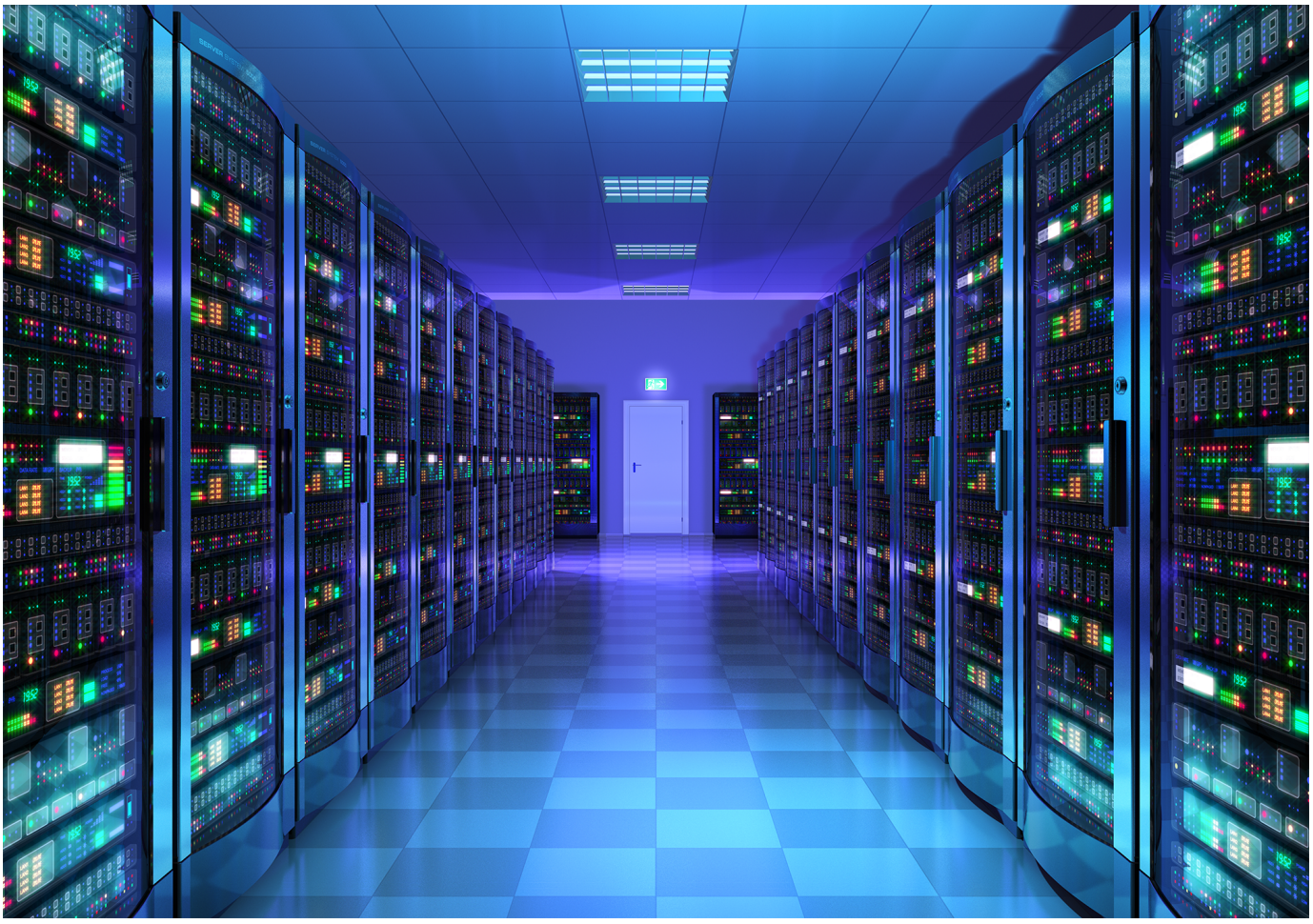
Predictable Latency



Introduction to Latency

Latency, quite simply, is a delay given in units of time, which in the computing world could also be defined

as the response time. As servers have scaled up, they must respond to millions of requests a second. Latency has become critical to make sure the user experience – for example, when shopping online at Amazon – is smooth and uninterrupted. Outages are a very serious concern but even more minor interruptions, for example those caused by a Distributed Denial of Service (DDoS) attack, can be costly as clients are driven to other services.

Historically, bandwidth has often been a concern for end users, for example when moving from 56K modems to DSL and eventually cable or faster fiber. Files were commonly large enough that bandwidth was a limitation and, further, streaming music, video, and games has since moved that bar upwards. Content Delivery Networks (CDNs) certainly have their hands full in that regard but there are many transactions that are more latency-sensitive. For example, trading and financial markets require precision where microsecond (µs) differences can be significant.

Predictability

While New York has often been called the "City That Never Sleeps," the Internet must literally be operational at all times. When discussing predictable latency, it's important to realize that uptime – that is, the amount of time that servers are actually available – is the first priority. Storage inevitably fails and there must be a way to swap in replacements transparently, that is without the client ever noticing. If a larger failure does occur there needs to be a graceful fallback, hopefully also with minimal client impact. These events can add latency if the system is not robustly designed.

Once availability is secured the next challenges are "wobble" and tail latency. Unpredictable latency could be defined as a wobbling line when measuring latency for the same operation over time; ideally, that is, you want a straight line. More importantly, it's crucial that latency is consistent even when the hardware is pushed to its peak with the maximum Input/Output Operations Per Second (IOPS). This is known as the "tail" or "long tail" and is most commonly benchmarked to determine the predictability of latency under extreme but probable conditions.

Hardware

There are two primary approaches to removing long tail latency bottlenecks in a server array: by improving hardware, and by reducing software overhead. Which side is more important depends on the implementation as well as the workload. For example, with High Performance Computing (HPC) the software side is often the larger source of latency because HPC operates from a dedicated platform. However, improvements can and should be made for both, particularly with an eye on maintaining congruency – that is, it's important to locate and alleviate specific bottlenecks rather than develop more generally.

Throwing more hardware at the issue isn't always conducive because latency limitations may not be a compute problem. For example, complex server configurations are more often limited by the network implementation. Networking hardware does tend to have far lower latency than on the software or Operating System (OS) end, so it's often more useful to improve or replace existing hardware rather than add to hard line complexity. One possibility is to introduce complex Quality of Service (QoS) rules to improve network traffic flow.

One example of a hardware bottleneck would be a lack of server resources – CPU and memory – which can be overcome with offloading I/O management to Network Interface Controllers/Cards (NICs). I/O and compute are now being offloaded to storage devices, also, for example with the Persistent Memory Region (PMR) and Controller Memory Buffer (CMB) functions from the Non-Volatile Media Express (NVMe™) specification. NVMe-oF™ allows this remotely through networking with, for example, Remote Direct Memory Access (RDMA). These features allow stronger flexibility and for a superior distribution of resource loads. SSDs are also moving to Zoned Namespaces (ZNS) and Key-Value (KV) implementations to reduce memory load.

Software

Predictable latency has become particularly important as servers have moved away from direct service to containers and virtualized environments which have complex, on-demand interactions. This type of setup, as utilized by Google Cloud, Microsoft Azure, Amazon Web Services, and more, is often more limited by software implementation. That is because their services operate with a Software-Defined Networking (SDN) configuration where, for example, there are software switches with shared memory allocation. This has a number of possible areas for bottlenecks, such as with scheduling – due in part to context switching delays and mismatches between software and hardware – and packet processing speed.

Improving the SDN stack has therefore become a priority of larger cloud hosts and services. The need to get "closer to the metal," or closer to native hardware performance, has long been a goal. This is especially true with hypervisors which must be as efficient as possible with the same amount of raw hardware resources. One way to do this is to bypass the networking stack with direct communication. Superior planning for Remote Procedure Calls (RPCs) can also make the latency tail more manageable. This goes hand-in-hand with many of the improvements we are seeing in storage management with regard to host-controller cooperation.

Summary

The increasingly virtualized landscape of the Cloud, coupled with unprecedented growth in the need for services, has made evident that scalable resources pushed to the maximum are often in a fragile state. It's not uncommon for game launches to arrive with broken and slow servers, for example. Cyberattacks can also often bring interruptions to social media, online banking, and more. Maintenance and redundancy can only go so far; the underlying server resources, including storage, must be able to respond to stochastic demand. The Cloud must be elastic and able to smooth out spikes by delivering predictable response times.

The vast increase in memory usage, and this includes solid state drives with NAND flash, has brought unique challenges. The incredible amount of possible IOPS can strain server hardware resources, but this overhead can be offloaded in various ways. More drastically, the move to containers and Virtual Machines

(VMs) has increased the software load and exacerbated the possibility for hardware-software mismatch. Bringing this into lockstep through new standards is one approach, but overall improvements to the efficiency of hardware and the implementation of software networking remain the best remedies.

One new player on the scene here is, of course Artificial Intelligence (AI), often with the assistance of Machine Learning (ML). Recognizing and developing patterns is the name of the game – modeling with feedback loops can create a dynamic response to seemingly unexpected latency spikes, for example. Certainly, the resources utilized for AI must be weighed against the possible improvements, so this is an area of continued development. Nevertheless, the idea of Skynet may not forever be considered far-fetched, but for today we are still eliminating conventional bottlenecks.

Our SSD Solutions

PCIe™ - Our ED1 Series is a powerful, high performance SSD made for edge storage applications. It comes in M.2 and U.2 form factors.

SATA - Our ER2 SSD Series delivers affordability and performance with superior random read/write speeds of up to 90,000/45,000 IOPS. It comes in M.2 and 2.5" form factors.

Please contact our Solid State Storage Technology Corp. expert for more information.

*Specifications and features are subject to change without prior notice. Images are samples only, not actual products.

Request Full Specs Sheets

ABOUT US

A subsidiary of KIOXIA Corporation, **Solid State Storage Technology Corporation** is a global leader in the design, development, and manufacturing of digital storage solutions. We offer a comprehensive lineup of high-performance customizable SSDs for the Enterprise, Industrial, and Business Client markets. With various form factors and interfaces, our SSD solutions help businesses simplify their storage infrastructures accelerating variable workloads, improving efficiency, and reducing total cost of ownership.

Learn more at www.ssstc.com