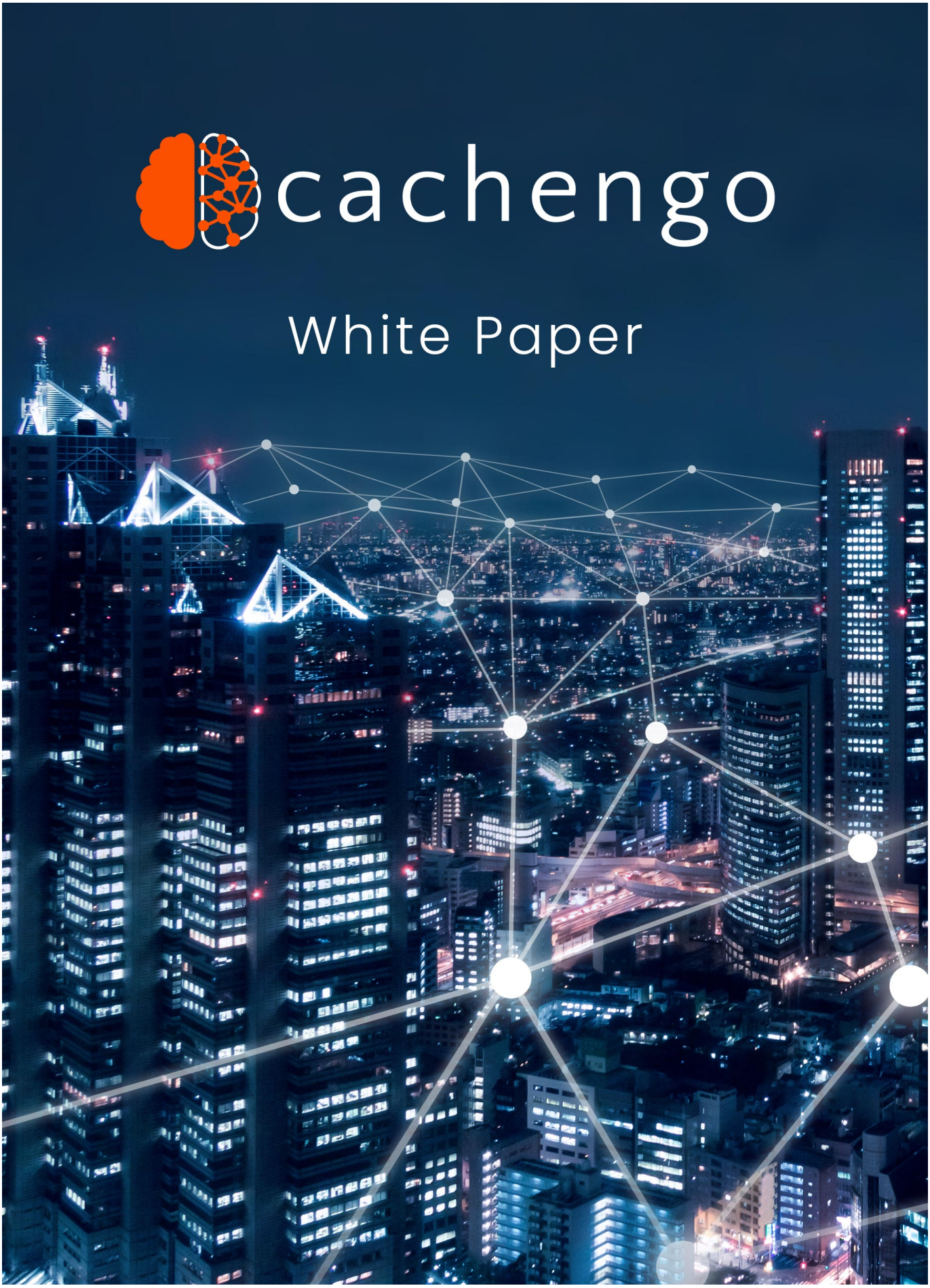
cachengo

White Paper



Advanced Edge AI for Computer Vision

Executive Summary

Most want to take advantage of data at the edge of the enterprise where businesses can make a difference but don't know how. Cachengo™ creates an environment for innovation at the edge known as an Analytical Edge architecture by providing the right combination of security, scale, economics and performance. What makes Cachengo unique is the technology that supports these capabilities.

Security comes from a Cachengo Connect peer to peer network that sits on top of Ethernet that makes it virtually impossible to hack or disintermediate.

Scale comes from the computational storage model of adding processing power with storage so that capability and capacity expand together. Scale doesn't need to take a lot of space, our compact form factor can package 32 independent nodes in a single rack unit.

Favorable economics come from a simplified edge architecture to reduce CAPEX by 5x and OPEX by 4x compared with traditional Intel® architectures and by employing the Cachengo Connect management schema that scales from the edge to the cloud to the data center seamlessly. Additionally, the Cachengo Market provides a portfolio of applications from different parties to enrich the functionality of the edge and control costs of implementation.

Performance comes from not just CPUs added to storage, but also GPUs or TPUs to optimize analytics at the edge. Instead of moving your data at the edge to the cloud or data center for processing, Cachengo's Analytical Edge allows you to move the compute to the edge and relieve your network of unnecessary backhaul traffic and allow your applications to work on data at the origin. You don't need a data center or a cloud to do analytics. You need a Cachengo.

Business Challenge

The action is outside the enterprise data center. This is where things get made, customer interactions happen, sensor data is generated, and the all-seeing electric eye is looking. In the past this data was perhaps used for limited purposes to solve a single problem. Now people are seeing that this data can be used in a larger context to understand more fully the environment around them. This often requires rapid response to changing events at the edge of the enterprise.

To take advantage of this opportunity at the edge, a new approach is required. No longer can enterprises wait for data to be sent back to the data center for processing and then make changes in the edge environment. New technology enables processing at the edge for action at the edge in real-time that can improve business results and responsiveness to change.

Solution Overview

The Cachengo architecture is the optimal combination of management, networking, and hardware that enable analytics at the edge where the data originates. The management environment is based on

Cachengo Portal addresses the ugly truth that scale-out architectures are difficult to manage. The solution is making ease of use a priority. Cloud-based Cachengo Portal's touch deployment for Cachengo nodes can integrate servers with a single command. Once integrated, devices can be added to Peer Groups to establish P2P connections securely and without the use of VPNs or port forwarding.

Similarly, the Cachengo Portal can deploy services spanning from your Data Center to the Edge, all with just a single click. Support includes a wide-range of products ranging from AI to Container Orchestration and Object Storage through ecosystem partners including Kubernetes, MinIO and Tensorflow, and more as our app list found in the Cachengo Market keeps growing.

Cachengo's unique networking system allows you to communicate with devices without ever opening a port on your Firewall. You gain additional leverage from your existing Ethernet network by adding Cachengo Connect to bring all the capabilities together in one specialized network. This creates a highly-secure software defined WAN that connects the Cachengo nodes and allows large-scale implementations for massively parallel analytics workloads.

Cachengo server nodes are based on the Symbiote building block which includes a CPU, GPU/TPU, Flash memory and networking. Nodes are delivered with up to 8 Symbiotes in a compact Bento Box or up to 32 Symbiotes in our Pizza Box for compute densities of up to (192) 576 CPU cores in a 1U package.

The benefits of the Symbiote include high density, low power creating a much more efficient way of deploying analytics capability over traditional architectures.

Architectural Design

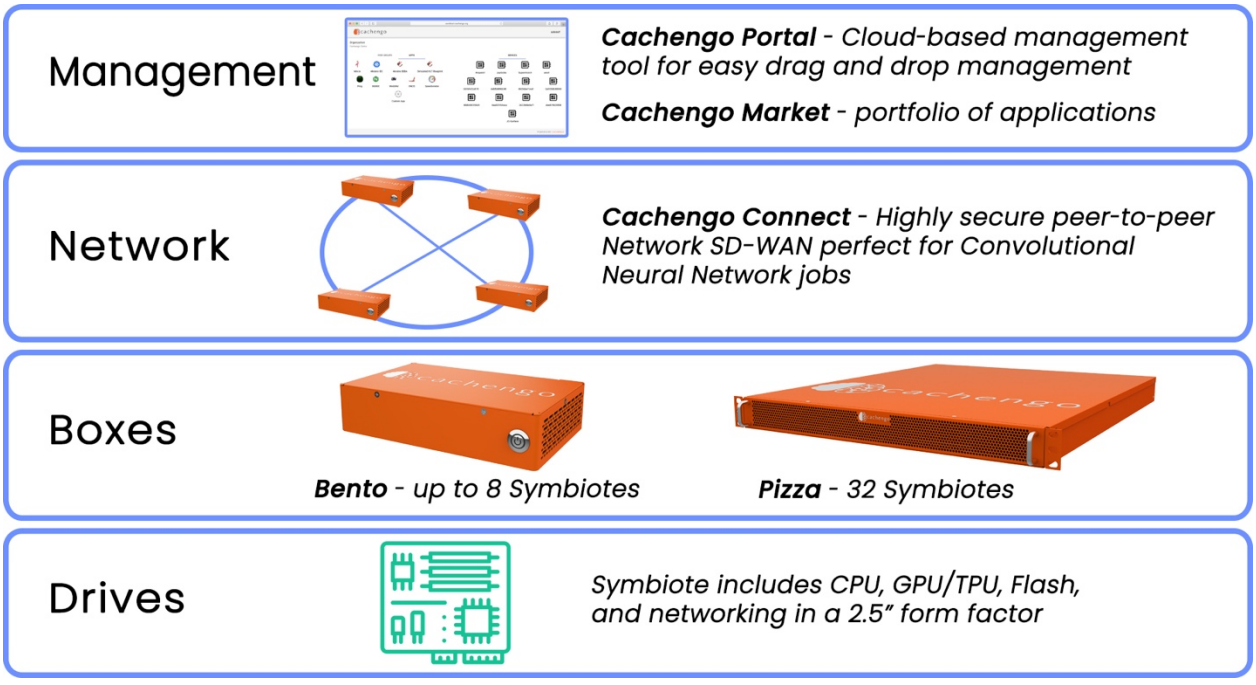


Figure 1: *Cachengo system architecture for the Analytical Edge*

Hardware

The foundational building block of the hardware is the **Symbiote** drive. This is an independently addressable server that combines the components that make analytics at the edge so powerful. There is an Arm processor with six cores to provide compute capability with a minimum of power which is often a requirement in many edge environments. But analytics also requires parallel processing for many jobs, including Computer Vision. A GPU or TPU are incorporated for this purpose in the Symbiote. There is Flash storage for high performance/low power data storage and Ethernet networking that enables a highly secure peer to peer networking that ties it all together. Each Symbiote is an independent device allowing highly parallel workloads to be spread across multiple nodes.

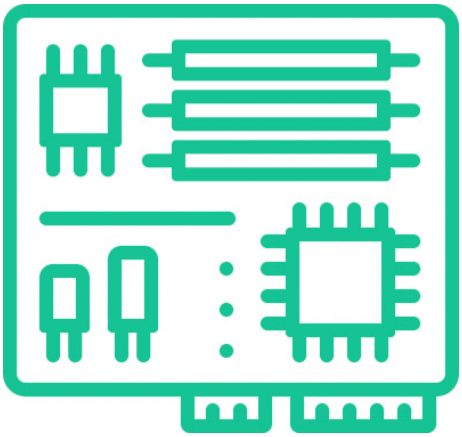


Figure 2 : *Symbiote*

The Symbiote drives are packaged into server nodes as Bento or Pizza boxes. The **Bento** box houses 8 Symbiotes and is designed for deployment in a wide variety of environments and is not a rack device. The **Pizza** box, however, is designed as a 1RU rack device, suitable for deployment in a standard equipment rack and provides a high density of 32 Symbiotes.

Network

To support a highly parallel workload like Computer Vision, a highly efficient network is required. To keep costs low and still provide a high performance, high security and high scalability we found that 10Gb Ethernet is the right answer. **Cachengo Connect** is a software defined WAN architecture over Ethernet to connect the Cachengo nodes. For security the peer to peer network does not publish addresses in a directory to virtually eliminate "man in the middle" attacks and creates a very secure network at a reasonable cost. Drives are configured at setup and remain private addresses. As workloads change the drives can be reconfigured and repurposed.

Most SD-WAN technologies simply focus on connecting or extending networks. At Cachengo, we needed to do much more than that. We needed to provide a way to connect to potentially millions of **Symbiotes**, which are almost exclusively deployed behind secure firewalls and can be configured as object storage devices (OSDs), or even as application servers.

We wanted to handle all of our communications without a dependency upon Secure Shell (SSH). It is with this idea in mind that we created Cachengo Connect. Cachengo Connect is a messenger and a protocol at the same time, making it a unique product that allows us to seamlessly connect to our devices, no matter where they sit.

We can use Cachengo Connect in combination with Cachengo Portal to quickly deploy applications. Anything we can do via SSH can be done via Cachengo, but in a fully-trackable, secure, and auditable fashion.

The worst thing you can do for your edge computing latency is to connect everything up to a proxy. We efficiently and securely connect all of our managed devices, regardless of whether they sit behind different firewalls. Proxies are inherently bad for scaling for numerous reasons— the biggest is that they ultimately create bottlenecks.

Imagine wanting to go from point A to point B, but having to go through point Z to do so, only to find that there is major construction going on at point Z. Why go through this experience if you don't have to?

Connecting devices up via VPN tunnels is equivalent to throwing a proxy in the middle of all of your traffic. You can do it, but you don't have to.

In addition to the bottleneck potential, proxies are inherently bad because they also provide a risk for man in the middle attacks.

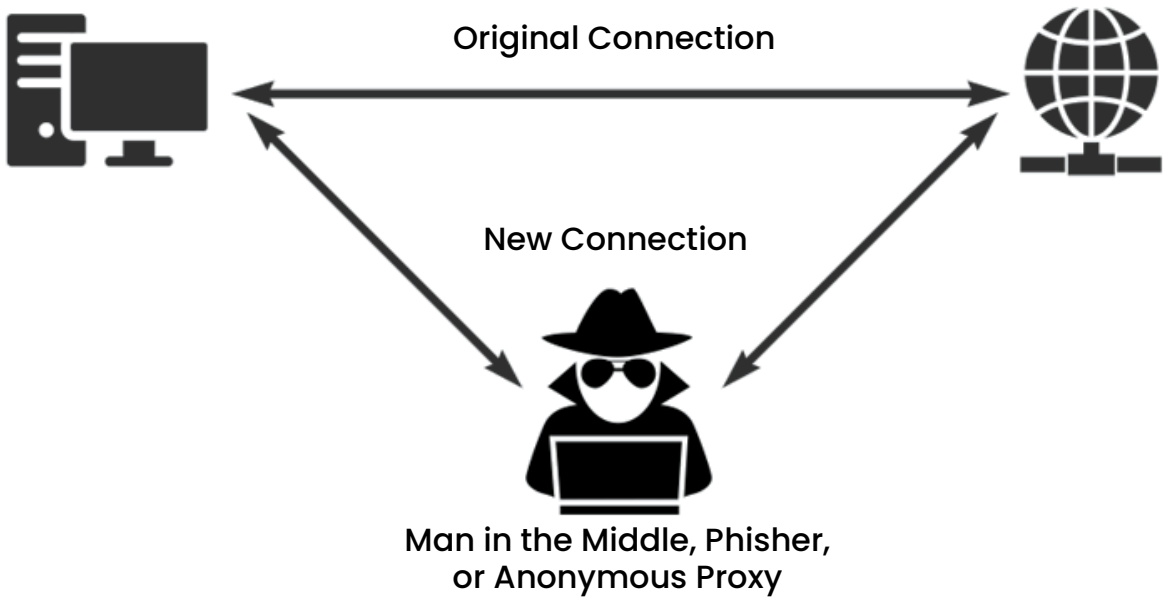


Figure 3 : *Example of a proxy server MITM exploit*

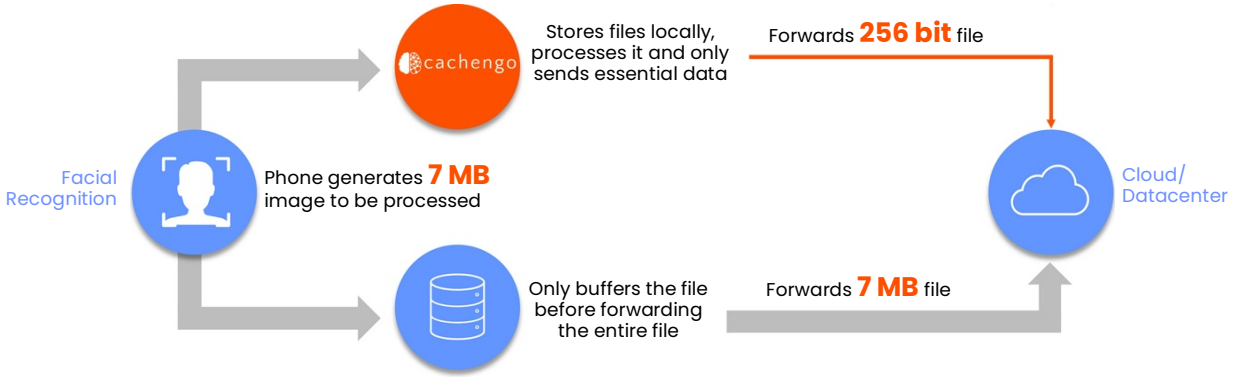
With a man in the middle attack on a proxy, any vulnerability can be exploited to create an opportunity for data to be hijacked, manipulated, or otherwise compromised. Furthermore, any connected devices or services can also be attacked through these types of exploits.

Besides the proxy concerns, there is also a concern for exposing endpoints to the public. Ever hear of DoS or DDoS? DoS stands for denial of service. When you expose a service on the internet it becomes extremely vulnerable to attacks.

This is because the service must broadcast itself to whatever is trying to find it and utilize it. People with malicious intentions can discover these exposed endpoints and disrupt such services by overloading them with requests until the underlying resources crash.

With our Secure Routes we can connect many endpoints without the use of VPN tunnels and without exposing resource endpoints. Imagine being in a building with many offices. Now, remove all doors, windows, and even hallways, but you want to go from one office to another.

While such a scenario might wreak havoc to someone suffering from claustrophobia, it is actually ideal when it comes to network security. This is how we connect devices efficiently and securely. Of course, you are free to continue to utilize your VPNs if you like…but we wouldn't recommend it.



In this instance, Cachengo achieves a 29,000:1 reduction in file transfer size

Figure 4 : Backhaul traffic reduction diagram

Software

The management software known as **Cachengo Portal** is a cloud-based tool to manage the nodes. It is designed as an intuitive low-touch tool where many functions are automated to enable efficient management of large-scale deployments. For large deployments management can be a challenge. Cachengo meets that challenge by making it easy to add peer groups and install apps. Customers like how easy it is to add peer groups and install apps, with a simple "point and click" interface.

The Cachengo operating system allows third party apps, and one critical feature is app installation. A company can create their own apps to deploy into their devices or they can deploy third-party apps via the Cachengo marketplace. Creating an internal or marketplace app can be as simple as writing a couple of lines in bash and making them executable or as complex as you want to make it.
App installation manages both the marketplace or a company can create their own apps to deploy into their devices. Under the hood, these two actions are the same for Cachengo (except where third parties may charge for a marketplace app). Creating an app for internal or marketplace use can be as simple

as writing a couple of lines in bash and making them executable (you could also make pretty complex installers).

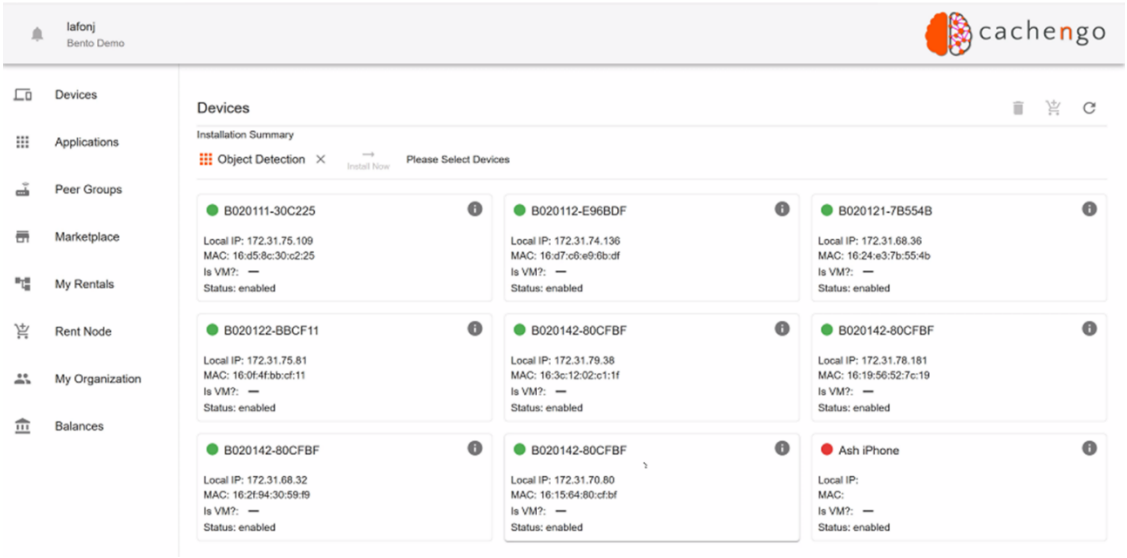


Figure 5 : Cachengo Portal management screen

The Cachengo API is a great tool for ecosystem partners. We are proud to say that 100% of our management features are exposed through our APIs. We document 100% of our endpoints so that we can make them accessible to our clients.

Ecosystem Environment

Analytics is a world of many possibilities and many partners to enable those possibilities. The **Cachengo Market** is a collection of tools that are available in Cachengo Portal to enable different functions. Watch us grow as we add new partners.

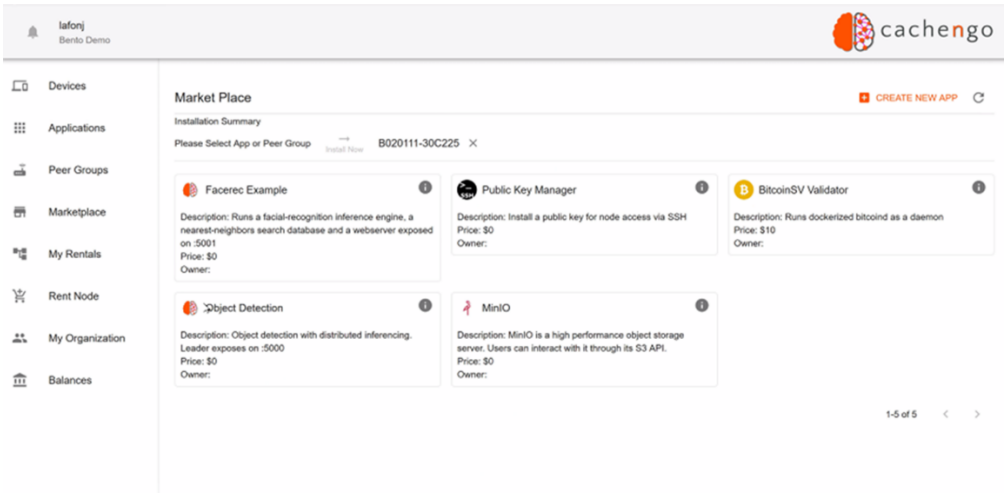


Figure 6 : Example of the Cachengo Market ecosystem

Computer Vision

A popular workload for the Cachengo is Computer Vision.

Today's complex analytics require data - and lots of it. If we are talking about machine learning (ML), artificial intelligence (AI), or just some type of augmented reality (AR), the need for fast storage is substantial. Latency and throughput become much more critical in order to deliver the data to where the analytics are processed.

The whole genre of 'Computational Storage' was created to address the needs of these new workloads. But analytics processing (also referred to as neural networks) is very different from simple offload. While other solutions incorporate compute offload capabilities, they cannot implement complex neural networks. This is because they do not possess graphics processing units (GPUs/TPUs).

Our Symbiotes feature embedded GPUs/TPUs and can perform sophisticated neural networking functions locally, right where the data is stored. What this means is you can ingest the data where it will permanently reside at the lowest CAPEX and OPEX possible, perform your analytics locally, and then merely send the results to wherever your central indexing resides.

This approach is not only cheaper, but is faster, less complex, and also simply smarter. As we continue to usher in 5G and Edge capabilities, we have a responsibility to be ecologically conscious. Processing data at the drive significantly reduces this network backhaul!

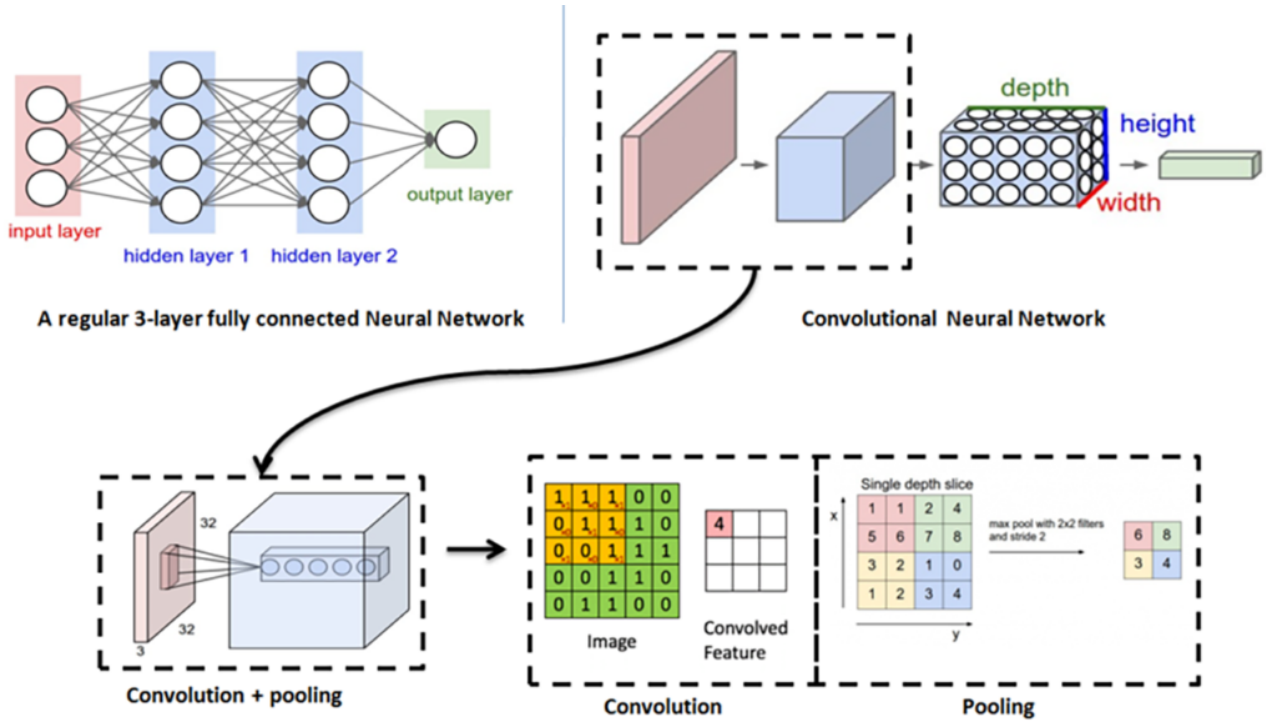


Figure 7 : Example diagram of a convolutional neural network

Convolutional neural networks (CNNs) take machine learning to the next level, especially for image and video data found in Computer Vision applications. While it is conceivable that implementing them onto others' smart SSDs is possible, doing so would require FPGAs with far too many gates and far too high cost to be effective, power efficient or scalable.

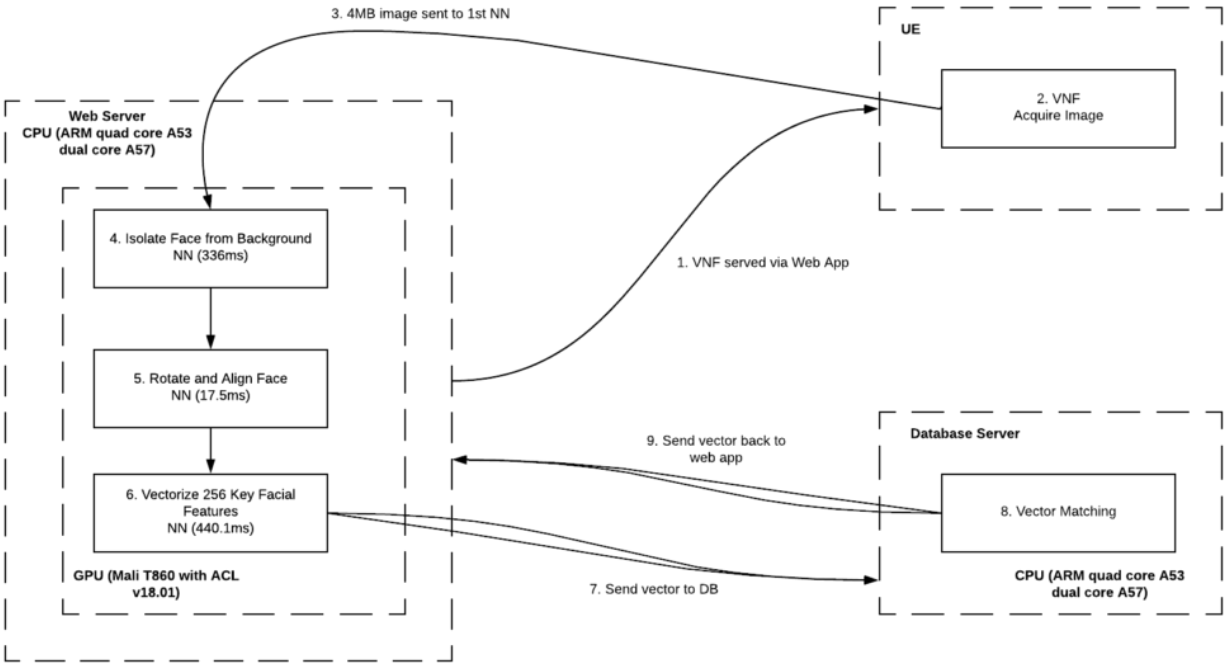


Figure 8 : Computer Vision, as a specific implementation of neural networking, demonstrated by earliest Symbiote prototype

Contrary to the current crop of smart SSDs, **Symbiotes** are fully capable of running CNNs. The above diagram shows a series of three CNNs that were used to demonstrate 10ms facial recognition as a VNF (Virtual Network Function) we deployed at the Edge as part of a telecom M-CORD (Mobile Central Office Re-architected as a Data center) deployment at Mobile World Congress (MWC) 2019 in Barcelona, Spain.

The use of GPUs/TPUs has become synonymous with ML/AI. The reason is quite simple – complex neural processing is computationally intensive. GPUs are capable of massively parallelizing these calculations in a way that general purpose CPUs are not. However, in order for these GPUs/TPUs to be able to work their magic, the data must be accessible.

With Edge Computing, the source for data is also massively parallel since data comes from millions of subscribers. Funneling data from these sources and getting them into the GPU/TPU results in queueing– the exact opposite of what we want to implement. It is essentially a funnel that creates a bottleneck for your application!

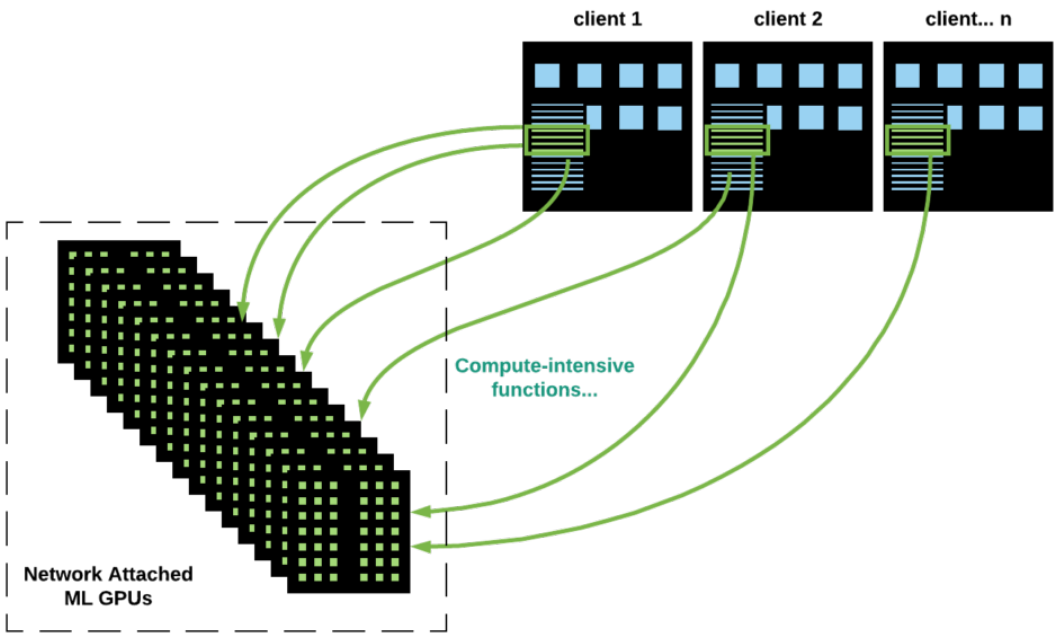


Figure 9 : Edge Analytics with integrated arrays of GPUs to scale with demand

By distributing massive arrays of low-cost GPUs/TPUs, Local Edge Analytics can better accommodate the types of ML/AI/AR requests that are associated with such a vast distribution of data sources. Unlike with single, large GPUs/TPUs, these GPU/TPU arrays do not create large choke points for data to be processed. While it is possible to create massive arrays with very powerful GPUs/TPUs, the costs and power consumption would be counterproductive with the economic scale associated with Local Edge Analytics. Cachengo architecture is purpose-built for Local Edge Analytics.

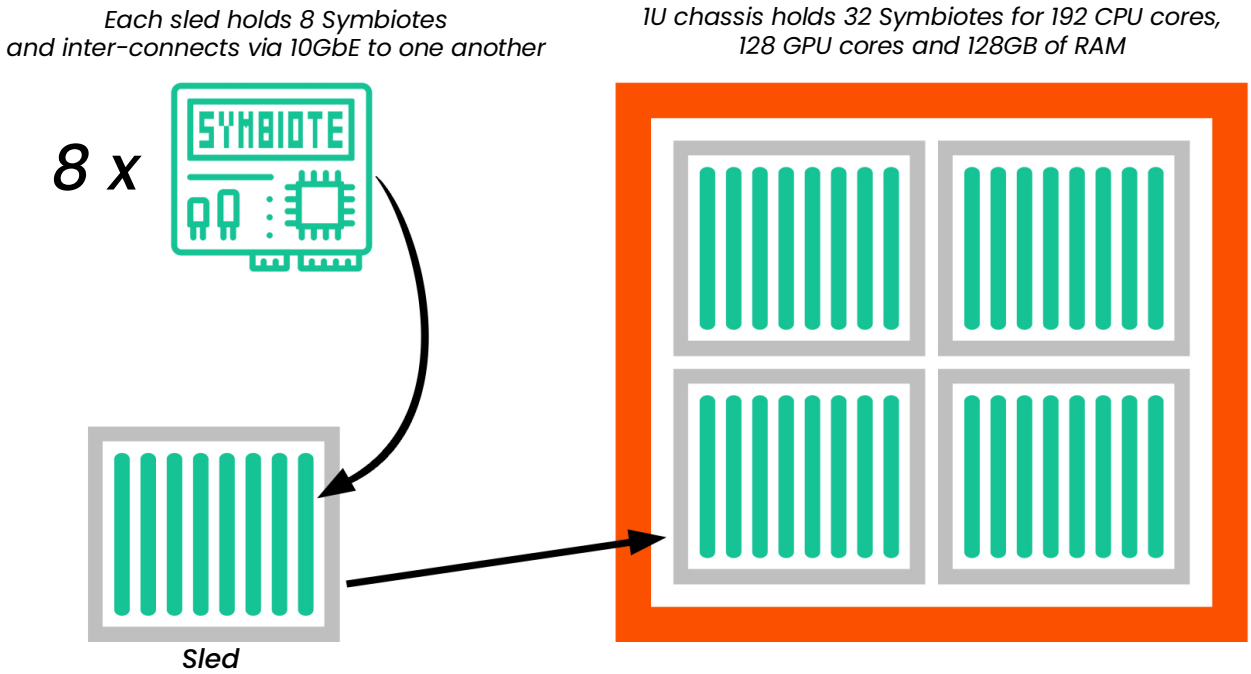


Figure 10 : Edge Analytics with a modular-based system with Symbiote Drives

Many have completely miscalculated what it takes to scale from the Data Center to the edge. This is reflected by their choice of processors to deploy. The de facto choice for deployments is based on an x86-64 architecture, a safe choice which has been the standard for several years. The issue is not so much the type of processor architecture, but rather the number of cores and the amount of memory, which impacts costs and power consumption.

When we started our original Cachengo concept, we began with comparing architectures for 24-drive storage enclosures. The reason for this was simple, as traditional network storage systems are represented by 12, 16, 24 and 48 drive configurations. At the time it was not deemed very feasible to be able to fit 24 server boards AND 24 drives within a 1U chassis, which became our target form-factor for our solutions.

What we soon discovered shocked us…we found our reference design was 1/10th the cost and 1/15th the power consumption compared to another system using high-volume, standard component parts, based on the status quo architecture. This resulted in an enormous reduction in CAPEX and OPEX!

The end result is a scalable, building-blocks approach to large systems. Rather than base a system architecture on a single large CPU, we now have the flexibility to build really large, fast systems with several smaller components. This ability to extract the capabilities of the many cores within a single CPU is known as coherency.

Achieving this coherent ability is oftentimes assumed, but seldom achieved in this era of very large processors. It is much easier to take a smaller processor, completely saturate it and its I/O, then aggregate many to fill much larger pipes. This is the basis for our design. It also results in the creation of large, scalable GPU/TPU arrays as depicted in Figure 10.

Results

The ability to move analytics out of the data center will provide several benefits:

- Greatly reduce backhaul traffic freeing up network capacity and saving infrastructure costs
- Improve the timeliness of analytics at the edge for a more responsive action and potential competitive advantage
- Inexpensive independent nodes will enable new multi-thread applications such as computer vision that can improve operations, manufacturing, autonomous vehicles, customer interaction, merchandizing and more.
- Offload work from the data center and cloud and putting capability where the data is being generated
- Highly secure networking provides a more secure environment while still based on the highly scalable and economic platform of Ethernet
- The Cachengo Market makes the ecosystem easy to implement with proven solutions for different application needs such as object storage or containers
- Cloud-based management found in Cachengo Portal makes management at scale efficient and effective

Conclusions

If you are wondering how to implement an analytics strategy outside the data center, consider a system approach with Cachengo. It's not about the hardware. It's not about the software. You need a purpose-built system approach to get the most from your local edge data, and the Cachengo approach provides the security, scale, economics and performance you need to get edge right.

To find out more or schedule a demo go to **www.cachengo.com** or give us a call at **+01.731.418.3238**

About Cachengo

The early storage solutions were expensive, large and consisted of multiple rack-mountable parts. Our founder, Ash Young, realized this and felt things could become more mainstream if the appliances were all-in-one, utilized a more open-source approach to the OS, and could leverage commodity components.

This quickly became the recipe for modern Enterprise and Cloud-based storage systems. The current practice of always throwing the latest and greatest CPU into each storage appliance, then increasing the number of appliances in the cluster has failed to reduce either CAPEX or OPEX.

Many would argue this methodology hasn't impacted latency or performance in a meaningful way. This was the turning point where we decided to focus our attention.

What we did was relatively simple, but it was so obvious that no one else really bothered to do it. Just as the economic climate for components had shifted 20 years earlier, this same climate now allowed us to migrate from a single CPU complex to placing a CPU and analytics engine via GPU/TPU onto each drive.

By doing this we realized that not only could we reduce latencies and increase performance, but also significantly reduce both CAPEX and OPEX along the way. When we say "significantly" we mean by an order of 10. This breakthrough discovery was simply too important to ignore.