IoT and the Future of Data Center Architectures
No one denies that the volume of information traversing the Internet is growing at a rate faster than the rabbit population at your local pet store. The anecdotal evidence surrounds us. In almost any public environment, the guy who stands out is the one who isn’t hunched over a glowing screen with his thumbs tapping out a message faster than a woodpecker assaults a tree trunk. A 2014 study conducted by Cisco attempted to quantify the volume this activity and arrived at the following estimates:

- Annual Global IP traffic will pass a zettabyte—that’s a one followed by 21 zeros—by 2016—and surpass 1.6 zettabytes by 2018
- This traffic volume has increased five fold in just the past five years
- By 2018 over half of the traffic will be generated by non-PC devices, the number of which will be double the global population.

The other aspect of this continued escalation in data volume is the immediacy of its availability required by its consumers. The demands of this combination of volume and speed will continue to act as stressors on existing data centers and are forcing providers to examine how to keep data as close to end users as possible to enhance its value and eliminate latency. The identification of the optimal structure to address these requirements is the subject of this white paper.

**The Inefficiencies of the Traditional Data Center Structure**

The nature of the data passing through the data center is evolving. Huge packets (for rich applications such as video, for example) are intermixed with billions of tiny bits of information from billions of devices comprising the continually growing “Internet of Things” (IOT). The common thread that links these seeming disparate data types is their common need for near real term processing on a continuous basis. Due to deficiencies in their design, construction and operation, many of today’s existing facilities cannot ensure that they are up to the challenges posed by these demanding processing requirements.
In effect, many existing data centers and their supporting network structures were not designed and built to effectively process the heterogeneous volumes of data that are increasingly required to simultaneously deliver a video within the window a customer defines as acceptable while also performing the analytics necessary for a manufacturing company to track their inventory status in real time.

The Stratified Structure and New Roles for Data Centers

The underlying diktat in a stratified structure is to keep the data as close to the end user as possible to reduce or eliminate the negative impact of latency. From a structural perspective, a stratified structure features one or more large centralized data centers facilities that work in concert with multiple, smaller “edge” data centers located strategically to be closer in proximity to end users and, in the very near future, micro facilities that will operate in concert with specific edge sites.

Although they have yet to be implemented on a large scale, increasing data volumes and latency “sensitivity” will give rise to micro data centers (<150kW) that will serve as the initial point of interaction with “end users”. In this role they will function as screening and filtering agents for edge facilities to move information to and from a localized area. More specifically they will determine what data or “requests” are passed on to data centers above them in the hierarchy and also deliver “top level” (most common) content directly to end users themselves. The primary impact of this “localization” of delivery will be levels of latency below what is currently achievable. An important consideration for the effective development and implementation of these micro sites is the need for a more sophisticated level of network, compute and storage management than is available today. Due to the current absence of these facilities, all of these functions are currently performed at the edge.

The next level within a stratified architecture is the edge data center. The specific function of edge facilities is to serve as “regional” sites supporting one or more micro locations. As a result, they perform the processing and cache functions within the parameters of what would defined as the lowest acceptable level of latency for their coverage region. As the regional point of aggregation and processing, edge facilities will require a high level of capacity and reliability than their connected micro sites. Tier III certification will be a standard edge data center requirement, with average capacities coalescing around 1MW.

The central facilities within a stratified structure house an organization’s mission applications and serve as the central applications processing points (CAPP) within the architecture with back-ups conducted on an infrequent basis in relation to their connected edge data centers.
Due to their level of criticality, the applications within a centralized data center are “non-divisible”. From a practical perspective this means that it is inefficient in terms of both a cost and overall system overhead to run these functions in multiple facilities. The vast amount of bi-directional information flow prescribed in a stratified structure necessitates that “fat pipe” connectivity be used between them and the edge locations they support.

### Division of Labor

Simply put, the primary efficiency of stratified versus centralized data center network architecture is based upon the historical concept of the division of labor. Each component has a specific role in the structure that promotes efficiency by reducing overhead and negating distance limitations. The multiple advantages of the stratified structure enable it to meet the technical and consumer demands for compute, storage and access and make it the logical next phase in the evolution of the role of the data center:

**Flexibility**

As in most things in life the 80/20 rule also applies to data centers and the operation of mesh architectures. In this case, 80% of the most accessed applications and data take 20% of storage and network capacity, with 20% of the least accessed consuming the other 80%. The important element of this particular equation is that due to factors such as the fickleness of end users or seasonality the content that make up these ratios is continually changing. Due to its distributed structure, with each element serving in a prescribed capacity, stratified structures are uniquely qualified to support these dynamic requirements as opposed to centralized alternatives.

**Security**

Data center security has always been about more than berms and mantraps, but as we are seeing on a far too regular basis, attacks on the content of our data centers can come from virtually anywhere in the world. In a stratified structure security (both physical and cyber) is not only distributed but escalating as well. Housing the organization’s mission critical applications in the CAPP of the configuration makes them both easier to track, and more importantly, defend. Multi-levels of protection—firewalls, access points, etc.—are built into the central facility to immediately identify the initiation of the attack (Denial of Service, for example) and ensure that it is immediately thwarted.

Security is also enhanced in the stratified structure by the existence of the edge sites. Unlike, a conventional architecture in which a disaster recovery site would be called into action in the event of a cyber attack, an action that may require precious time, the existence on multiple edge sites allow traffic to be re-routed or off-loaded from the a site under siege. In short, security in a stratified configuration offers a host of defensive options that are not available in centralized structure.
Mission Critical at all Levels

In stratified architectures, each element (central hub, edge and micro facilities) must be designed to deliver the appropriate level of reliability. Speaking in terms of the Uptime Institute’s Tier system the central or hub locations should be built and certified as either a Tier III or Tier IV facility. In fact, Tier IV would be the preferred standard here since the mission critical applications they house support all of the elements of the configuration. The redundancies at this level also ensure that the system remain operational even during regular periods of hardware refresh.

Edge facilities will be built to be Tier III constructed certified to provide a consistent level of durability at the points closet to their users. The reliability requirements of edge data center will also place a premium of strict adherence to documented standards and codes. For example, the potential for system failure, and personal injury, due to the conduction of “hot work” on energized equipment would be eliminated via the following of workplace electrical safety guidelines like those specified in NFPA 70E.

Adaptability

Due to the investments made in stratified structures, the ability for sites to be able to adapt technological changes will be an essential requirement. This will be especially true for edge facilities due to function as the primary point of interface of customers. The raised floor environments in these facilities should feature to primary capabilities to deliver the level of adaptability these sites will require:

1) Columnless construction: Layout flexibility is a key element of a data center’s ability to adapt to new requirements. A data center that is populated with columns naturally imposes limitations on the floor layout options of the user. Only a column free raised floor area is able to offer the user the maximum number of possible design alternatives.

The flexibility provided via the absence of columns is a necessary design element when we consider the dynamic nature of both current, and certainly, data center environments.

Rack layouts will have a direct impact on a site’s capacity for example. OpenStack configurations call for 21” racks as opposed to the historical 19” standard. Although seemingly small, the impact of modified technical specifications—even those as seemingly small as 2”—will have substantial ripple effects that will require the full area of a raised floor as opposed to capacity limitations imposed within a column-based configuration.

2) Ability to support high and low density applications: While the jury remains out on the ultimate direction for the average amount of power that will be required per rack to support future applications, a data center must be able to support both if it is to be truly adaptable. Locations that do not possess this ability are in danger of becoming the future poster children for stranded capacity.
The need for both mixed density support and column free design within edge facilities combine to simplify and minimize the impact of the more frequent moves, adds and changes that can be expected within a mesh structure. The attributes of these two elements will contribute to edge facilities being able to support more racks (effectively increasing capacity) versus a column centric raised floor. As a result, new racks and servers can be added, tested and cut over without the need to remove existing hardware.

**Geography**

In a mesh structure the ability to locate an edge data center wherever it is required is essential. Almost by definition, the majority of edge facilities will be located in markets outside the top five to ten major metropolitan areas (typically these will be the domain of central hubs). Secondary and tertiary markets will find themselves as the optimal locations for edge facilities. While modular solutions may initially be seen as likely edge solutions, their special and physical (the lack of a shell hardened withstand sever weather and seismic events) limitations make their ability to serve as flexible and reliable edge locations dubious at best. In an environment where equipment at the edge moves from telco gear to sophisticated processing equipment housed in $3 million racks, pre-fab solutions are the metaphorical version of a house made of sticks when one of brick is required.

**Network**

Many of today’s mega facilities boast of the cornucopia of network options available within their POP rooms or even localized rings that connect their facilities within a region. While these attributes may be attractive to an end user seeking one or two facilities, they are insufficient when viewed from the perspective mesh architectures where the hub must support three specific network capabilities:

**The ability to scale**- Localized rings and even the most robust POP room are not designed to provide the ability to support multiple edge facilities. Traditional data center architectures have been with their focus on the access they can provide from a single location. A mesh hub, however, must be designed and built with the ability to support multiple satellites as one of its fundamental responsibilities.

**Multiple providers**- Even the most robust carrier hotel/data center the network offerings are constrained by the number of providers who service that area. Mesh designs dictate that the hub provide supra-regional network support at a minimum and preferably nationwide connectivity support. The vast majority of today’s mega facilities are not equipped to adequate meet this challenge.

**Cost and Control**- Network costs typically constitute a large share of the operating expense for even a single data center.
Cost consideration becomes an even more important element in a mesh environment, thereby creating the need for hub locations to provide a multitude of connectivity options, including those outside of traditional providers, and concentration that can be leveraged to deliver network costs far below those that can be delivered in the current site by site environment that currently characterizes the data center industry.

Planning a Stratified Structure

Successful implementation of mesh structures must be part of a long-term strategic plan. For many organizations this will impose a new discipline on their data center planning process. In contrast to the quasi-reactive—“we’re almost out of capacity at data center “X” so we need to put together an RFP for new site”—mode of planning that characterizes most organizations today, planning for a mesh network architecture will require end users to conduct their planning activities using an expanded time horizon, in which the questions center around how many data centers will we need in the next 5-10 years?

As part of this extended planning process, businesses will begin to think in multiples. This will be of particular importance on the edge where increased importance will be placed on:

- Replicable design for ease of operations and longevity
- Reliability
- Flexibility
- Pay as You Go Expansion to enable end users to add capacity based on their own schedule and budget requirements
- Management
- Network

The geographic placement of facilities (both for edge locations and one or more hubs) will also be a leading factor in planning as both existing and future “coverage” must be taken into consideration due to the interrelationship requirements of all of the eventual locations within the structure.

Summary

The massive volumes of information and the continually declining definition of “acceptable latency” are placing new demands not just on the individual data center but their aggregated network architectures as well.
In a world where even the youngest consumer of data complains about load times, building bigger data centers isn’t the sole—or best solution—to these requirements. The need to keep data as close to the end user as possible is rapidly becoming the most important factor in data center planning. Relying on the time tested principle of the “division of labor” stratified architectures will come to dominate the data center landscape in the coming years.