1

How We Got Here: History of Data Centers and Current Choices

COPYRIGHTED MATERIAL

Data center processing capabilities were designed to do multiple and complex equations, transactions, executions, and storage. The limitations of the mainframe are often the abilities and brain trust of the information technology (IT) director/operator in using the mainframe, and the bandwidth to the mainframe often limits its use. For example, one mainframe properly utilized can collapse 10,000 to 20,000 square feet of legacy servers and white space. This is not only one big single point of potential failure but one remarkably efficient use of space and environmentals.

The utilization of the mainframe is often 50% or less, which is not a good return on investment (ROI). The protocols of the mainframe functions are not as nimble as those of enterprise systems, unless the programmers are confident and fluent. Older mainframes were nine feet by five feet wide, broken into modules (5 to 11 modular components). A minimum of three feet had to be left for service accessibility on all sides. Mainframes had fixed power from whips and were fixed to plumbing for cooling; they did not easily move once they were set. In the 20-year total cost of ownership model, the 20-year environmentals would service three to four IT life cycles of equipment with nonsevere (low-velocity) increases of power distributions or cooling.

Mainframes were expensive, and they did a myriad of complex functions simultaneously. They cost millions of dollars, old and new, and it took an act of Congress or multiple senior executive signatures to authorize the purchase and installation and growth of mainframe installations. The downside of mainframes were:

- Very expensive IT spend
- Very expensive and exacting environmentals (installation) to operate 24/7

- Difficult to move into a live data center (monolithic)
- Expensive to maintain vendor-driven costs, vendors on-site (licensed engineer required)
- Migration or upgrade challenges
- Tape storage a challenge
- Bandwidth of telecom connectivity (blend of copper and fiber inside plant and outside plant)

Plenty of the mainframe computers were placed in vertical assets in urban environments in the 1980s. They stayed urban because companies wanted them to be near or close to fiber optics central offices and the IT personnel, who in the old days stayed at headquarters for political and public transportation reasons.

As we will discuss later, in the 1990s, data centers became populated with just-in-time solutions precipitated by a flurry in overspending. Financing of kit was easy. If a piece of kit could provide marginal or a few added features, it was sold and provisioned. Tactical solutions eclipsed strategic solutions. Money was abundant for IT spending. The leasing terms for the new, most relevant equipment were almost silly. To please shareholders, vendors were making it too easy to get the equipment out the door. Users were unsure of the solutions they were buying. The total cost of ownership models were thrown out the window. The question at that time was: What is a data center?

My introduction to the world of data centers came on the heels of the futuristic high-rise development working for a rock star, high-rise developer and a few prewar renovations developers in the cast-iron and SoHo areas of lower Manhattan, for assets meant to lease, not to last.

My boot camp (following the above work), for the world of acronyms or mission-critical facilities was an eight-story concrete grain warehouse on a pier. The LeFrak Organization (Sam and Richard LeFrak) added eight stories of steel to turn it into a 15-story office building (no thirteenth floor), named Newport Center I. (It was renamed Newport Financial Center [NFC], a name I suggested in a naming contest).

A Japanese publishing company (Recruit, USA) came to New York in the mid- to late 1980s. The chairman went up in a helicopter, pointed to the building, and told the broker to buy it. When they found that 111 Pavonia Avenue in Jersey City, New Jersey, was not for sale, they leased it.

The NFC was one of the largest, most dynamic and forward-thinking developments in America's history. Recruit deserves all the credit. The company missed the market by 18 months, but it was brilliant nonetheless. Newport rivals the turnarounds of Boston's Faneuil Hall, Baltimore's Inner Harbor, and London's Canary Wharf. Sam LeFrak was a visionary who created the LeFrak City housing development in New York. He saw the value of the views from New Jersey to New York and realized that Manhattan as an island had limits. New Jersey's PATH train, which was cleaner and less expensive than the New York City subway and connected to the New York City subway system, enhanced the sale of NFC. Sam's son Richard would carefully craft and successfully implement the multiuse and mature development that we now see. Richard's visionary qualities are eclipsed only by his humility. Although the LeFraks were best known for the development of the masses and not the classes, their greatest assets are not real estate but the relentless efforts and loyal human infrastructure that keep the machine running. (Mike Sabet, the on-site development manager, is a good example of this.) Like Sam, the Japanese saw an opportunity to leverage the close proximity to the city and decided to develop the land as they did in Japan: developing mission-critical assets on the fringe of an urban environment. Businesses in New York were moving in the same direction. There was a movement to get back-office operations out of Manhattan to the boroughs of New York and New Jersey to support the boom of the mid-1980s (remember the age of excess, "bright lights, big city"?). Development that characterizes this time is the Staten Island "Teleport" (satellite farm with office building on low-cost real estate), Paine Webber in Weehawken, New Jersey, and Newport in Jersey City. Manhattan was moving back-office operations, call centers, and noncritical functions to Connecticut, New Jersey, Queens, Brooklyn, and warm-weather points in the South.

Recruit wanted to take a business plan developed in Japan and export it to the United States, to the undisputed financial capital of the world: New York City. Jersey City would do. The concept was "shared infrastructure."

I will take a circuitous route in defining just what a data center is via the shared infrastructure concept because the brain trust and intellectual capital was great, among multiple users, and because the importance of human infrastructure, documentation, and proactive preventive maintenance was highlighted with a co-op of users and their inherent best practices and unique requirements.

The shared concept delivered by the Japanese was based on the challenges of cooling mainframe computers. IBM was dominant in the early 1980s. Its "big iron," or mainframe, computers were being spit out of Ulster, Kingston, and Poughkeepsie in New York State and elsewhere as fast as they could be made.

In Japan, due to the challenges of expensive or unavailable real estate in urban environments (specifically Tokyo and Roppongi, respectively), businesses realized that these data centers were going to have to be vertical assets and that a shared “heartbeat,” or critical cooling system, could be far more efficient and economical. By more economical, I mean that users would not have to acquire a relatively small data center space and the associated relatively huge expense and delay associated with land acquisition, and the protracted and expensive design and construction required for a 15- to 20-year asset. For small users, this may not be the best use of valuable resources of money, time, and talented in-house staff.

Cooling the mainframe was the challenge. The mainframe is fixed (but modular) for a long time; even in the 1980s, this meant for a period of five to seven years. As we know, “necessity is the mother of invention.” The decision to build data centers vertically was then, and is still, a concern because of possible water flow from floors above to a data center, and could force an outage. Mixing water and electricity is traditionally a bad idea. Also, another tenant or user could force an interruption within a shared asset and could create a security breach or evacuation. Mainframes had water on the floor and in the equipment.

The compromise was *shared infrastructure*. The benefits were:

- Lower capital expense-entry barrier
- Faster to market
- Maintenance performed by experts in their core business
- Clustered IT and facilities vendors led to economies of “buy,” improving service-level agreements
- Modular moves, adds, and changes

Traditional data center components often were an afterthought for corporate facilities. Many times they were placed in odd and nonstrategic locations:

- Basement locations were problematic due to their proximity to piping (water, steam, etc.) and flooding.
- Top-floor locations were liable to roof flooding, and expensive long power and fiber runs.
- Locations under or contiguous to cafeterias on second floors led to smoke and water flow issues.
- Locations with center core assets were problematic due to building services running through the data center.

In the 1980s, data centers were underfunded and poorly maintained with a host of single points of failure. At the time, we were still living in the paper age of transactional work and batch work. Storage was requested but rarely required, and enforcement was archaic. Consequences of noncompliance of storage had rare enforcement consequences. There was no "hammer" to speak of.

In vertical assets, data centers were often placed in odd or unusual locations. Cooling and power upgrades were brought to the white space under the most challenging and circuitous routing and conditions.

Outside plant upgrades or power had to be worked out with the utility. That meant the utility dictated how much and how long to make proposed improvements. Then easements with the landlord du jour had to be satisfied, often requiring additional real estate for a substation or step-down or step-up transformers of power in a basement. Then a vertical right-of-way had to be mapped out with paper drawings. (There were no "as-builts," so a time-consuming and dirty process of climbing into shaftways, elevator risers, and duct banks needed to be verified before a landlord would allow the work; then charge by the linear foot to run and maintain the distribution and usage of new risers, decommissioned elevators, etc.)

The same challenges were waiting for the cooling component of the data center for the placement of cooling towers, or DX units, the conduits to serve the source of the air-handling units and computer room air-conditioning units.

The raised floor to underside of dropped ceiling was challenging but not a showstopper due to the lower density of heat dissipation and respective cooling needed. Getting the mainframe—a piece of equipment the size of a small car—into the building, up the freight car, down the hall, in the tight turns was always an interesting challenge. Although the mainframe was built in modular units, and most were designed to make it through a 30-inch doorway, moving the mainframe from the truck to its designated footprint was often an arduous process. Face it, servers are cheaper and faster to place in cabinets.

Today these data center design and implementation challenges for vertical assets within urban environments seem humorous. Yet this was the situation until the crash of 1987 and the economic slowdown that lasted until 1991. IT spending was dramatically reduced and the moves, adds, and changes during these years were generally required, not merely requested.

So why were data centers not given the reverence that we give them today? What has changed?

- Data centers were tied at the hip to large office installations due to the human infrastructure; both IT and facilities management are needed to make these things happen. Large office installations were driven to public transportation services and prestigious addresses.
- The only diverse and redundant telecom infrastructure in the regulated world of telecommunications that served with large bandwidth

solutions based on multiple business and large human populations to serve. Follow-the-money infrastructure is built around fast and economical penetration of markets. At the time, most businesses and dense human populations were urban. Meaningful telecom facilities serviced meaningful billing opportunities. While by law they had to bring services to rural or suburban areas, diverse, scalable, burstable, and synchronous optical networks (SONET) did not exist or were prohibitively expensive and challenging to build, operate, and maintain. (My first T1 was $1,500 a month in Jersey City.)

- Access to large power substations or near large telecom infrastructure (central offices) was in urban environments, not just in New York City but in most parts of the country.

To sum up, data centers of the 1980s and the early 1990s often were underfunded afterthoughts of the corporate world. Chief financial officers (CFOs) often were in charge of the go/no go decisions for large capital spending, and the IT executive or chief information officer did not have a place at the table to speak for or on the behalf of the data center. Neither the CFO nor the chief executive officer knew the difference between a mainframe or main gate, kilovolt ampere or British thermal unit. These important disciplines were far down the food chain with poor corporate visibility and were largely undercompensated. They were viewed as taskmasters at best, rarely acknowledged or appreciated for their long hours or years of dedicated service. These were the data center warriors.

Besides having low power, data centers had low cooling standards and were monolithic and not nimble in design. Big box in, add a box, big box out (usually in pieces), storage tapes in, storage tapes out. There was no regulation other than peers or best practices. Power moves were limited to the distance of the whips or seal-tight conduit from the power distribution unit or junction box.

Users would move when the lease expired, and the same madness would continue over and over until 1996, and the era of telecom deregulation. There was no real, compelling reason to change the model. To state the obvious:

- There were no extraordinary events outside of the periodic hurricanes or tornadoes. (Data centers in these regions had no excuse.)
- Earthquakes with meaningful destruction were limited to California. (Most financial institutions and exchanges are in the Northeast.)
- Floods, although national and often underreported, had not had catastrophic consequences in urban environments. Suburban employees had been inconvenienced, but there were no marked events of extensive

duration and great financial loss to power, telecom, the exchanges, national economy, or security.

Because no meaningful consequences had interfered with the U.S. economy, there were few or no drivers to improve the design, installation, or maintenance of the data centers of the 1980s or early 1990s.

These events were noteworthy largely because of the loss of lives; they had no impact on data center siting (locations), design, implementation, or maintenance:

- 1938: The Great Hurricane (i.e., the "Long Island Express") killed 50 people.
- 1985: Hurricane Gloria touched down on the Atlantic coast as a category 2 storm, with winds under 65 miles an hour.
- 1992: Hurricane Andrew killed 23 people and inflicted $43.7 billion in damages. The losses to homes and some businesses were so great that 11 insurance companies went bankrupt.
- 1995: A heat wave killed over 750 people in Chicago.
- 2003: During a heat wave, 35,000 Europeans died. The United Nations predicts urban deaths in the United States to double by 2020 thanks to global warming.
- 2005: In a 52-day period, hurricanes Katrina, Rita, and Wilma caused approximately $20 billion in damages and approximately 1,500 lives were lost. These are 100% new. What is noteworthy is that seven of the top ten most powerful hurricanes in 154 years of record keeping happened between 2004 and 2005.

It is ironic that heat waves traditionally kill more Americans than any other national disaster.

Now that we have identified the general reason or causes to spend the money and time to design, build, and maintain a data center based on traditional acts of God, we can articulate the choices available since the early 1990s.

Currently there are three choices for data centers:

1. **Stand-alone data center.** Such centers are built as greenfield single- or multistory single-purpose assets. Some are found in multitenanted and multistory buildings, but they are not preferred. Some stand-alone data centers can also be 100% new or augment existing centers (warehouse, manufacturing, etc.). Most users start out by trying to save time or money by improving an antiquated asset or partially improved asset with some of the improvements in place for a data center.

2. **Shared infrastructure, multitenanted asset.** This center is a single-purpose and often multistory asset with inside plant improvements that often include emergency power service generator only, uninterruptible power supply, battery, and rectifiers for clean, computer-grade power, air-cooled, DX, dry coolers, security, monitoring, and maintenance. Tenants take largely unimproved space by service provider, inclusive of maintenance, et cetera, all at a premium and expense. Users can buy 100 KVA of uninterruptible power supply and 200 kilowatts of generators. Tenants can buy 40 tons of heating, ventilation, and air conditioning. They also can buy "hot hands," or modified services, for fixed or "cross-connect fees" in a protracted menu. Services are à la carte. Users need to make sure services will be in place during the full term of the lease or licensing agreement to ensure that the asset does not run out of infrastructure capacity. This is a buyer-beware program. The devil is in the details of the service-level agreements (SLAs). No one should expect compensation for lost revenue or brand damage due to outage. Compensation comes in the form of future free rent or minimal setoffs. In fact, SLAs are really little more than facility descriptions; more often than not, the sales staff does not know the difference between the two.
3. **Collocation-caged environments.** These centers often are multi-tenant or telecom assets in which there is shared improvements of generators, uninterrupted power supply, air conditioning, and often cabinets and IT equipment. The services are sold by the cabinet, power circuits, cross-connects, IT circuits, and pretty much everything you look at you pay for. The "hot hands" for all IT work is priced by task or duration of tasks per month (e.g., five hours per month). Like the shared infrastructure model, this is also a buyer-beware program.

These choices can satisfy mission-critical needs in urban environments, where cross-connectivity or layer-one, layer-two, and layer-three solutions can have choice and various architecture. The TELCO assets have the confluence of facilities- and nonfacilities-based fiber optics with a plethora of optic multiplexing and manipulating to suit users' needs with competitive prices based on multiple vendors and, therefore, competition.