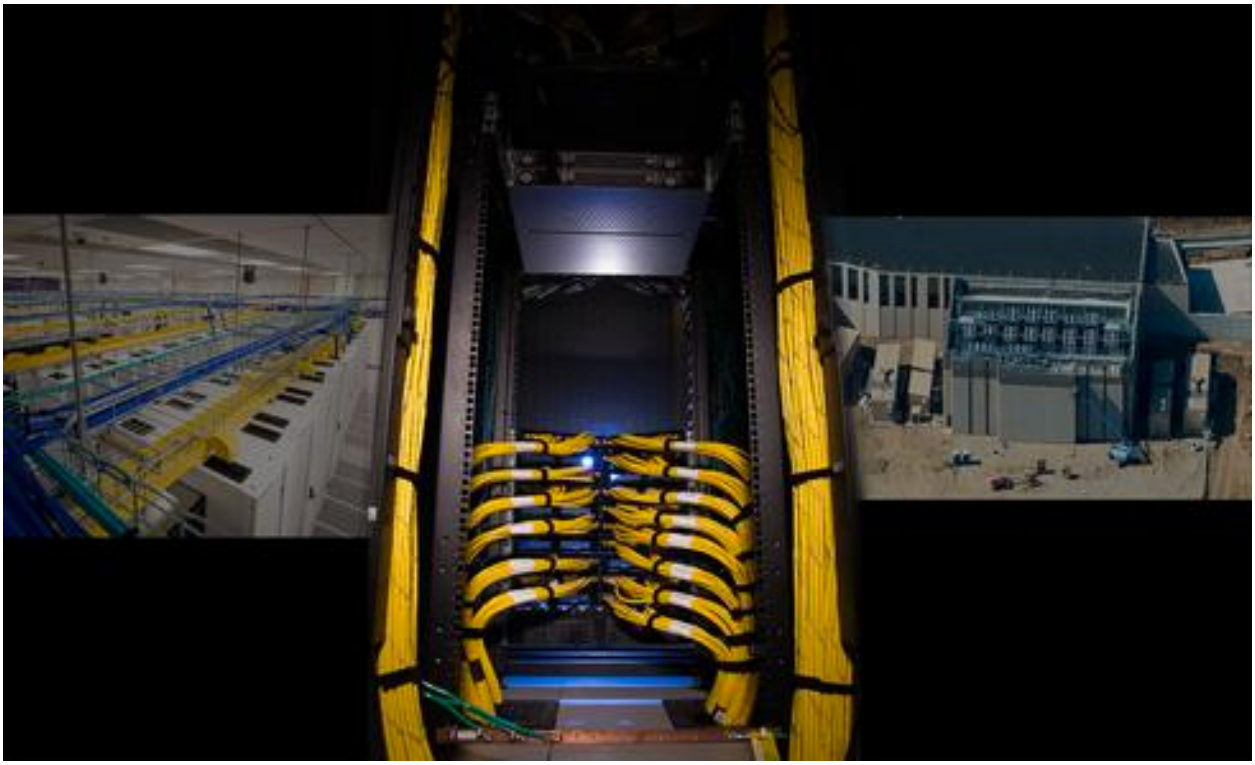
6 Models Of The Modern Data Center

InformationWeek, Charles Babcock
http://www.informationweek.com/big-data/hardware-architectures/6-models-of-the-modern-data-center/d/d-id/1269326

Our exclusive look inside the new data centers of Fidelity, GM, Capital One, Equinix, ServiceNow, and Bank Of America shows the future of computing.



The cloud hasn't killed the company-owned data center.

Companies from General Motors to Fidelity are taking the best infrastructure ideas from Web giants such as Facebook and Google and adapting them to their very different business needs in order to build a new generation of data centers.

To highlight the different approaches and strategies these companies are taking, we looked at six companies beyond the Web giants that have made or are planning major data center investments. Our point isn't that data center construction is booming; Gartner forecasts only 2.3% growth in data center spending in 2014, to $143 billion. Companies will spend more than twice that on enterprise software, by comparison, at a growth rate of nearly 7%.

But we are seeing a boom in data center innovation, and it's coming not just from Web and cloud service companies, but from conventional companies that still see running world-class data centers as part of their competitive advantage.

Here's an overview of the innovations that we'll explore in this slideshow:

Fidelity is opening a new data center in Nebraska this fall. The investment giant is one of the biggest advocates outside Silicon Valley for open source hardware, and the new building itself is a modular, just-in-time construction design.

General Motors spent $130 million on a new data center in Michigan, with a second one coming online this summer. It's a private-cloud-meets-mainframe operation, as the company supports both Web-ready apps and long-running legacy software.

Capital One christened its new $150 million Virginia facility in March, replacing capacity by third-party operators as the company shifts from an outsourcing strategy to insourcing most of its IT. The goal is to make sure infrastructure doesn't slow down its new Agile development initiatives.

Bank of America is entirely reimagining its data center infrastructure, with a private cloud architecture and commodity hardware taking center stage. And it's looking like modular, container-based capacity will play a major new role.

Equinix is tapping into companies' need for faster transactions and processing by building telecom-centric data centers. These facilities can connect cloud computing and storage resources, for example, to an array of telecom options -- 130 different carriers from the hub that sits near Amazon Web Service's Virginia data center complex.

ServiceNow is a fast-growing software-as-a-service business that has doubled its data center capacity over the past two years, opting to lease space from a co-location provider rather than build. It trusts in its resilient architecture design and management software to keep the services up and running -- along with staff working around-the-clock at the co-lo's facility.

Fidelity keeps its options open

Fidelity Investments is set to open a state-of-the-art data center in Nebraska based on a design it has been working on for five years. Fidelity aims to use as much open source code and standardized Open Compute hardware as possible in its data center, along with its own proprietary "Click to Compute" server orchestration and management system.

As an early member of Facebook's Open Compute Project, Fidelity is hoping to see competing suppliers produce servers for a networked, rack-based hardware platform that encourages rapid cycles of innovation. As part of a highly regulated industry, it also wants a data center that it owns and manages and in which it retains company data.

The data center, slated to open in September, implements the Centercore design Fidelity has been working on since 2009, as it settled on the right blend of elements for a leading financial services company. Its intention is to capture the elasticity of "hyperscale" data centers built by the likes of Google and Amazon, says Eric Wells, Fidelity's VP of data center services. "It's a very open design that can evolve as we decide to add capacity in the future."

What's different

In the previous generation of Fidelity data centers, Wells says, "We found a lot of stranded power and IT capacity, where the infrastructure couldn't take full advantage of the resources available to it because of a crowding together of the wrong mix of elements." Fidelity adds units to its Centercore design in 500-kilowatt or 1-megawatt units, with all the power capable of being consumed by the equipment in the unit. A 500-kilowatt CoreUnit might typically represent 2,200 square feet of data center space.

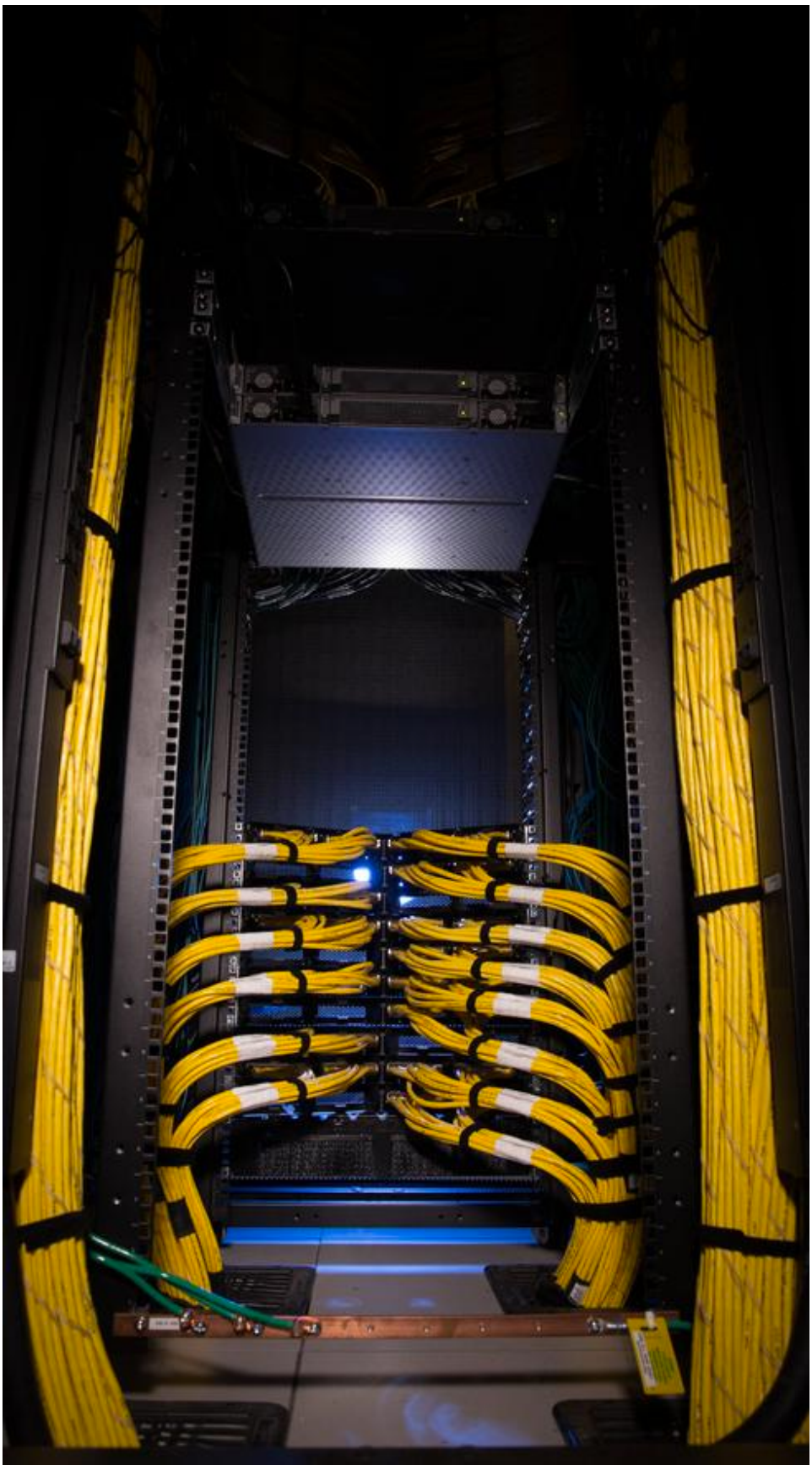
CoreUnits are steel-frame, one-story rooms that can be assembled together like Lego pieces coming together at the site, Wells says. CoreUnits have sliding panels in the walls that allow a new unit added to the data center to open up and provide contiguous space to another CoreUnit. The units are built off-site to Fidelity's specs by an independent fabricator, Environmental Air Systems, and then trucked to Fidelity's data center construction site. The units, unlike earlier modular designs based on shipping containers, may be stacked into a multistory building, which can be particularly useful in an urban location.

Within days of arrival, they've been equipped with the power connections and cabling they need to take up their station. An entire data center can be constructed in this way in six months, and expanded as needed. Fidelity calls it "just-in-time data center construction" and builds no more than it needs at any one time.

Each CoreUnit has its own cooling system and power distribution system. They're designed to run at a warm 90 degrees, collecting hot air off the equipment and either cooling it or venting it to the outside. CoreUnits can withstand F3-force winds, which can occur in the large tornadoes that strike the Midwest.

Fidelity's Nebraska data center is expected to use 40% less energy than the company's previous data centers.

It will contain thousands of x86 servers, but unlike Google's and Amazon's facilities, it will also contain some RISC/Unix servers, Wells says. The servers and switches are based on Open Compute standards.



GM puts computing near power users

The people who soak up the most computing power at General Motors are engineers doing CAD drawings of new vehicles and those simulating crashes. It's no coincidence that the automaker built its two new data centers alongside two existing GM locations in Michigan where most of those teams work.

Limiting the distance all that data must travel saves on networking costs and improves responsiveness, notes Jeff Liedel, who leads the data centers as executive director of global IT operations. GM modeled its two new data centers on the cutting-edge practices of Web giants such as Facebook and Google -- for example, it uses in-row cooling and x86 commodity servers. But it also needed to accommodate legacy apps not built for a Web architecture.

Name-brand, highly virtualized x86 servers provide most of the automaker's computer power, running a standard software stack that includes Linux OS, VMware, Oracle database, and WebLogic Java application server. But this isn't a private cloud for purists. It isn't one single, general-purpose pool of compute, storage, and networking, because many of GM's applications run best on dedicated hardware. "Like any other 50- or 100-year-old company, we have a lot of other stuff," Liedel says. That other stuff includes Solaris servers, as well as mainframes that run 300 different applications, including systems that process tens of billions of dollars in material and parts purchases. GM's Outlook and Exchange apps run on Windows servers.

Liedel distinguishes between "cloud-ready apps" that can run on the shared private cloud part of GM's environment and older apps that don't fit that model. CAD/CAM probably will never be a cloud app, he says, because it requires so much graphics-intensive local computing. But other apps, such as expense reporting, run on a private cloud environment.

GM's $130 million Warren, Mich., data center opened last summer, and its Milford, Mich., center is due to open this summer. GM is closing 23 data centers worldwide (some of them operated by outsourcers) and moving most of that capacity to these two, which are built to be identical and provide failover for each other.

Here are some other features of the data centers:

Energy efficiency: GM's Warren data center is running at about 1.5 PUE; PUE is the standard ratio that measures how much energy the facility uses divided by the energy the computing equipment itself uses (as opposed to cooling or lighting). The closer to 1, the better, and the most efficient enterprise data centers today run about 1.2 or 1.3. "We'll get there," Liedel says.

In-row coolers: Like most modern data centers, GM's has cooling systems that react to heat sensors in a particular server or rack and can cool only that area rather than trying to cool an entire room. It has Plexiglas rooms of servers in which the temperature can run regularly at about 90 degrees, compared with about 70 degrees for the rest of the room. Those "hot aisle containments" max out at 130 degrees -- at which point the roof pops open, releasing the steaming air into the rest of the building. GM also uses evaporative cooling, using water chilled by the cool Michigan air, for much of the year instead of standard air conditioning.

Flywheels over batteries: In the event of a power loss, GM's data centers use a flywheel system, which Liedel describes as a "mechanical UPS." If the electricity goes out, the flywheels are released to run the facility until conventional diesel generators can kick in. The flywheel system consumes about as much energy as a more conventional uninterruptible power supply system, Liedel says, but it takes far less maintenance and "replaces a roomful of lead acid batteries." The Warren center is LEED Gold-certified, and GM will pursue the same standard for Milford.

Power use per square foot: Liedel keeps an eye on a data center metric you don't hear much about -- power use per square foot -- to determine if GM is at risk of running out of capacity. Companies usually build a data center with physical space to expand, and GM is no different. The Warren center is only about 60% to 70% filled. But data centers also can run out of power, from either the utility or its own backup generators. Today, the Warren data center is using less than 50% of its power capacity. As the electronics get smaller, they'll draw more power per square foot, which means Liedel is watching whether he has to add juice well before he needs to pour more concrete.

Capital One goes for fast and simple

Capital One has trained 5,000 of its employees on Agile development in the past three years, with the goal of constantly spinning out new software features to customers. The future of financial services, Capital One believes, is digital. But what's the point of knocking out apps in six weeks if it takes another two months to get them deployed and actually in customers' hands because of infrastructure shortfalls?

"The objective we set is that we never want the infrastructure to slow down the pace at which our Agile teams can operate," CIO Rob Alexander says. That need for speed in getting new digital features out to customers is a primary reason Capital One opened a $150 million data center outside Richmond, Va., in March.

The problem with data center infrastructure, Alexander says, is that "so often, it's the impediment to getting things moving." Capital One used to get much of its data center capacity through facilities run by third-party operators. The new data center, paired with its existing facility on the other side of Richmond, will allow it to deliver most of that capacity in-house.

In terms of the data center architecture, the Richmond center is a bit over 70% virtualized. It includes sections of high-density "blocks" of servers, storage, and

networking located in the same racks, with name-brand hardware configured to a design that Capital One customized to its needs. To deliver deployment speed, architects are working to standardize software and hardware stacks for certain platforms -- so all Java software, for example, will run on the same infrastructure block.

Standardized platforms and DevOps tactics for deploying software more efficiently are helping to meet rising demand: In March, developers were doing 1,200 builds a day on that Java platform, compared with just 300 six months earlier, before the new data center went online.

Here are some other features of Capital One's new data center:

Hot-aisle containment for high density: Much of the data center uses a typical alternating hot-aisle/cold-aisle approach. But for its new high-density stacks, architects put a Plexiglas room around the equipment. That lets them run those racks much hotter than in the rest of the data center and cool a smaller space. There are only a few of those contained double aisles today, but the data center is built so that as other rows are converted to a high-density private cloud infrastructure, they can be enclosed in the same way.

Cool Virginia air? If you've been to Virginia in July, "cool outside air" isn't what comes to mind -- it's more like muggy and hot. But the Richmond data center can use outside air for cooling about 150 days of the year. The center is LEED Gold-certified.

Active-active data centers: Capital One has its headquarters in the Richmond area, and it has another data center on the other side of the city from its new one. The goal is to run both data centers for active workloads, but also use them to back each other up during periods of maintenance or brief outages, much like public cloud providers do with their facilities. Data can travel a double round trip between the data centers in 1 millisecond over fiber. The centers aren't yet running a true active-active setup, but that's the goal.

Hardware inventory: When new hardware lands in the data center, it gets an RFID tag so that with one scan, an employee can look up the machine's specs, when it was purchased and delivered, and the like. The team built a tall, slim cart covered with RFID readers that an employee can slowly push past a stack of equipment and know exactly what's there. It's used primarily in the receiving area

today, but data center managers plan to use the system to do a monthly inventory of all the equipment deployed on the data center floors, with no manual effort beyond walking a cart past the racks. The collected data will help IT assess when it needs to do tech refreshes of aging gear.

Redundancy: Capital One displays, in 3-foot-tall letters along a main hallway of its data center, the word "Simplicity." Here's an example of that ethos at work: Since you need power from two independent sources coming into every server rack for redundancy, Capital One color-coded those A and B sources -- one red, one blue -- so that anyone can tell at a glance if a piece of equipment has a backup power source. And if it's a type of equipment that doesn't support two power inputs, it gets a yellow cord as a warning that pulling it will turn it off.

Bank of America thinks modular

Three boxcar-sized shipping containers are sitting right now on a cement floor in a Virginia data center, each packed full of servers, storage, networking, and cooling. What they hold just might be a big part of the future for Bank of America's global computing infrastructure.

Bank of America is in the midst of an initiative to reimagine its computing infrastructure. It's testing two private cloud management stacks, one based on OpenStack and one from a proprietary vendor. David Reilly, BofA's global technology infrastructure executive, thinks 80% of the bank's workloads will be able to run on this software-defined data center stack, using "ruthlessly standardized" commodity hardware.

In the past few months, Reilly and the bank's infrastructure team have become increasingly convinced that modular data center containers will play a big part in the company's computing needs. BofA will still have some large-scale data center capacity. But whereas today it has large data centers supported by third-party colocation capacity, modular containers could provide much of that flexible capacity, helping the bank move quickly to seize new business opportunities and also letting it cut back capacity if a business shrinks.

As Reilly and team explore these ideas of modular computing, commodity hardware, and software-defined infrastructure, Reilly says, "One of the pitfalls is if you make the new data center look like the old data center."

Bank of America has a few hundred workloads on the new architecture today, with plans to get up to about 7,000 by year's end, mostly for development and testing. If successful, the bank could push for major adoption next year.

BofA can't rely entirely on mega-scale, global data centers the way a Google or Facebook can, because the bank faces tougher rules on where it can house customers' data. If regulations call for it to keep certain data from a country's customers in that country, BofA wants capacity there that's best in class for speed and performance, but at a cost that's competitive with large-scale centers. That's why a modular option looks so intriguing.

BofA also wants a data center that can scale up and down as business needs change. Down is the hard part. "No one's really been able to crack that yet," Reilly says. With containers, BofA can lease smaller increments of space, and it wouldn't need a classic raised-floor space with lots of built-in cooling systems. Containers need only reliable power and networking, and they must be in physically secure space, he says.

Here are some other interesting elements of BofA's emerging data center strategy:

Open and proprietary stacks: The bank has found both the OpenStack and proprietary management stacks to be as flexible as they're touted to be. Even once it launches those 7,000 workloads toward the end of this year, Reilly hopes to hold off on deciding which stack to use. He's keeping both options open so the bank is as minimally locked in to one software stack as possible.

The hardware matters less than before: BofA is rewriting its applications for a private cloud architecture. "Perhaps the biggest impact the software-defined move will deliver is the way we build systems and software," Reilly says. Each application will tell the infrastructure what resources, such as backup and recovery, it needs, for example. "The infrastructure becomes less central," he says. "It's just less important."

Public cloud not ready: Reilly doesn't think public cloud services are a viable near-term option for much of the bank's computing needs. The security isn't sufficient, he says, and the economics "are not compelling yet."

Don't go too early: Reilly sees a competitive advantage in getting on to this next-generation of computing infrastructure before rivals do, but a global bank can't have a learn-on-the-job IT backbone. "We have a business to run," he says, "and we won't take undue risk in the adoption of this technology."

Equinix focuses on communications

Data center operator Equinix bases its decision on where to build a new facility on four main considerations: communications capabilities and location, location, location.

Equinix is one of the premier builders of communications hubs and co-location facilities, with a specialty of locating just outside some of the world's largest trading and retail markets -- New York City, Los Angeles, Atlanta, and Chicago, among others. One of its newest is an eight-story building in Seattle.

Equinix caters to data center customers who put a premium on speed, reliability, and multiple providers of their communications -- thus Equinix's special attention to telecom links and physical location. For example, a software-as-a-service vendor may use Amazon Web Services or Microsoft Azure for cloud computing and storage but connect that computing power via a high-speed private-line access to a nearby Equinix communications hub. In that hub, Equinix can let the SaaS vendor deliver data to end customers using a carrier that isn't available inside an Amazon or Microsoft cloud facility. AWS's Direct Connect service, Microsoft's ExpressRoute, and Verizon's Secure Cloud Interconnect services all use Equinix for this kind of routing to other telecom carriers or services.

Carrier-neutral communications hubs help businesses looking to extend into territories that their primary telecom suppliers don't serve well. For a global presence, a business will almost certainly need a number of carriers to make sure its website and services have fast response times in different parts of the world.

Equinix concentrates many communications carriers in a "network neutral" facility -- any carrier can be connected to any other carrier to let a company optimize speed, cost, or reliability. Today Equinix builds those connections manually, running a fiber jumper from carrier A to carrier B, says Equinix CIO Brian Lillie. In the near future, it will create new connections via a software function, implemented over a virtual network taking advantage of established physical connections, Lillie says. That will let customers switch carriers more readily than they do today.

The data center as communications hub has become more important as businesses become more digital, based on e-commerce and online customer interaction. It's hard for companies building their own data centers to replicate a

telecom hub, since generally a company only gets one or two carriers as primary providers. By comparison, Equinix's Ashburn, Va., center -- near Amazon's US East data center complex -- houses 130 carriers; a typical Equinix center has 50 to 60.

Equinix operates 100 data centers in 33 urban centers around the globe, which means building an Equinix data center with the latest switching equipment and 10-Gbps, 40-Gbps, and soon, 100-Gbps fiber optic lines. A total of 975 carriers can be found in one or more of its data centers. Equinix operates seven such centers in the New York City area, including two in Manhattan, catering to financial services companies for functions such as high-speed trading.

Other notable features include:

Testing for trouble: Lillie says it's difficult to get two electricity sources in some dense urban locations, so Equinix has to build even more uninterruptible power supply and backup generator infrastructure into those than most data centers. It also does some interesting testing. When those huge diesel backup generators kick in during a power loss, they can produce vibrations that cause other equipment to vibrate and potentially overheat. So when testing the generators, technicians scan the data center with infrared cameras, looking for an equipment heat signature that might indicate trouble. "Just because you build for reliability doesn't mean that you have it operationally," Lillie warns.

Cloud Exchange: A new Equinix service lets companies move enterprise data and workloads with less exposure on the public Internet. A customer may then need to make only limited use of the Internet to get to an Equinix center, or it may establish a private connection, making the route all private.

ServiceNow adapts co-lo to its needs

ServiceNow, a supplier of IT service desk and other business service automation in the form of software-as-a-service, has doubled its data center capacity over the last two years, but it's taking a different approach from these other companies that are building their own facilities. ServiceNow has no single physical data center blueprint and is using colocation centers in CenturyLink Savvis, Verizon Terremark, and Equinix facilities rather than building its own. But inside each location is ServiceNow's own resilient architecture and specialized software infrastructure.

A cage of ServiceNow servers at a co-lo site is matched by another cage running an identical batch of applications. While both are active, if a server in one fails, all its customers and their live data are moved to the other within 90 seconds because live data is continuously being replicated between the two. Furthermore, each co-lo site is paired with another nearby so that if a disaster threatens an entire facility all customers can be transferred to another.

Disaster struck ServiceNow's Switzerland location earlier this year when fire took out an electrical substation supplying a colocation facility it uses outside of Zurich. The co-lo operator warned customers that they had four hours before backup generators would run out of fuel, and that they should make plans

accordingly. ServiceNow transferred all its customers to the backup center in Geneva, well before the threatened generator stopped, the company says.

"We apologized to some customers for having had to move them," says Allan Leinwand, ServiceNow's VP and CTO of cloud platform and infrastructure. "They said, 'We never knew you did it.' "

In addition to providing resiliency, Leinwand says he needs to ensure customer privacy and data integrity. Enterprise cloud users are debating whether virtual machines provide sufficient isolation, or whether a mechanism such as Docker containers can do the job better. Leinwand uses neither. Instead, ServiceNow runs each customer in its own Java Virtual Machine on a shared host and on its own isolated database. That way each customer has its own implementation of the database system in its own JVM, providing process and data isolation, as opposed to virtual machine or container isolation with shared database resources.

In contrast, many Salesforce.com customers are sharing one implementation of an Oracle database system, relying on it to maintain the borders between different customers' data. Leinwand thinks a deeper form of isolation comes from providing individualized instances of the database in a JVM, with logical borders around a customer's data as well as its application. Each customer then has a separate process through which it's accessing the database and manipulating data.

ServiceNow has added so much data center capacity in the last two years to provide greater assurance of continuous availability, Leinwand says. ServiceNow has been fortunate not to suffer any major outage in its 11-year history, he says, "but we don't think we're immune. The unexpected happens, and we want to be prepared for it." He and his staff continue to work on eliminating possible causes of failure and reducing the time to recovery if one occurs.

Other notable ServiceNow features include:

Guarantees: The vendor's service-level agreements guarantee uptime of 99.8%, after which it starts paying credits for lost service. Amazon Web Services and Microsoft offer SLAs of 99.5%.

Staffing: It keeps round-the-clock data center staff on hand in colocation facilities to perform repairs and maintenance.

Multi-instance, not multitenant: Under this SaaS architecture, ServiceNow can do an upgrade or maintenance for a single customer without disturbing any other operations on the server. With a multitenant architecture, all customers on a given host must upgrade at the same time.