Computing at full capacity

August 2 2015, by Eric Brow, Mit Industrial Liaison Program

Over 12 million servers in 3 million data centers in the U.S. burn about 100 billion kilowatt-hours of electricity every year. Billions of dollars are spent on data center energy every year, with billions more spent on power distribution and cooling infrastructures. Even with the magnitude of these numbers, energy and cooling represent only about 20 percent of the typical total cost of ownership of data centers, which is typically dominated by server hardware (about 40 percent) and software (about 25 percent) costs. Additional costs, including storage, networking, and information technology labor, further swell the price tag.

According to a 2014 study from tech research firm Gartner, Inc., data center systems collectively represent a $143 billion market. With enterprise software adding $320 billion to that and IT services another $963 billion, the overall IT industry represents a whopping $3.8 trillion market.

Companies are increasingly seeking new ways to cut costs and extract the largest possible value from their IT infrastructure. Strategies include placing data centers in cooler climates, switching to more affordable open source software, and virtualizing resources to increase utilization. These solutions just scratch the surface, however.

An MIT-connected startup called Jisto offers businesses a new tool for cutting data center and cloud costs while improving resource utilization. Jisto manages existing enterprise applications by automatically wrapping them in Jisto-managed Docker containers, and intelligently deploying them across all available resources using automated real-time

deployment, monitoring, and analytics algorithms. As the resource utilization profile changes for each server or different parts of the network and storage, Jisto elastically scales its utilization in real-time to compensate.

"We're helping organizations get higher utilization of their data center and cloud resources without worrying about resource contention," says Jisto CEO and co-founder Aleksandr (Sasha) Biberman. So far, the response has been promising. Jisto was a Silver Winner in the 2014 MassChallenge, and early customers include data-intensive companies such as banks, pharmaceutical companies, biotech firms, and research institutions.

"There's pressure on IT departments from two sides: How can they more efficiently reduce data center expenditures, and how can they improve productivity by giving people better access to resources," Biberman says. "In some cases, Jisto can double the productivity with the same resources just by making better use of idle capacity."

Biberman praises the MIT Industrial Liaison Program and Venture Mentoring Service for hosting networking events and providing connections. "The ILP gave us connections to companies that we would have never otherwise have connected to all around the world," he says. "It turned us into a global company."

Putting idle servers back to work

The idea for Jisto came to Biberman while he was a postdoc in electrical engineering at MIT Research Lab of Electronics (RLE), studying silicon photonic communications. While researching how optical technology could improve data center performance and efficiency, he discovered an even larger problem: underutilization of server resources.

"Even with virtualization, companies use only 20 to 50 percent of in-house server capacity," Biberman says. "Collectively, companies are wasting more than $100 billion annually on unused cycles. The public cloud is even worse, where utilization runs at 10 to 40 percent."

In addition to the problem of sheer waste, Biberman also discovered that workload resources are often poorly managed. Even when more than a half of a company's resources are sitting idle, workers often complain they can't get enough access to servers when they need them.

Around the time of Biberman's realization, he and his long-time friend Andrey Turovsky, a Cornell University-educated tech entrepreneur, and now Jisto CTO and co-founder, had been brainstorming some startup ideas. They had just developed a lightweight platform to automatically deploy and manage applications using virtual containers, and they decided to apply it to the utilization and workload management problem.

Underutilization of resources is less a technical issue, than a "corporate risk aversion strategy," Biberman says. Companies tend to err on the side of caution when deploying resources and typically acquire many more servers than they need.

"We started seeing some crazy numbers in data center and cloud provisioning," Biberman explains. "Typically, companies provision for twice as much as they need. One company looks at last year's peak loads, and overprovisions above that by a factor of four for the next year. Companies always plan for a worst-case scenario spike. Nobody wants to be the person who hasn't provisioned enough resources, so critical applications can't run. Nobody gets fired for overprovisioning."

Despite overprovisioning, users in most of the same organizations complain about lack of access to computing resources, says Biberman: "When you ask companies if they have enough resources to run

applications, they typically say they want more even though their resources are sitting there going to waste."

This paradox emerges from the common practice of splitting access into different resource groups, which have different levels of access to various cluster nodes. "It's tough to fit your work into your slice of the pie," Biberman says. "Say my resource group has access to five servers, and it's agreed that I use them on Monday, and someone else takes Tuesday, and so on. But if I can't get to my project on Monday, those servers are sitting completely idle, and I may have to wait a week. Maybe the person using it on Tuesday only needs one of the five servers, so four will sit idle, and maybe the guy using it the next day realizes he really needs 10 or 20 servers, not just the five he's limited to."

Jisto breaks down the artificial static walls created with ownership profiles and replaces them with a more dynamic environment. "You can still have priority during your server time, but if you don't use it, someone else can," Biberman explains. "That means people can sometimes get access to more servers than were allotted. If there's a mission-critical application that generates a spike we can't predict, we have an elastic method to quickly back off and give it priority."

Financial services companies are using Jisto to free up compute cycles for Monte Carlo simulations that could benefit from many more servers and nodes. Pharma and life science companies, meanwhile, use a similar strategy to do faster DNA sequencing. "The more nodes you have, the more accurately you can run a simulation," Biberman says. "That's a huge advantage."

Docker containers for the enterprise

Jisto is not the only cloud-computing platform that claims to improve resource utilization and reduce costs. The problem with most, however,

is that "if you have a really quick spike in workload, there's not enough time to make intelligent decisions about what to do," Biberman says. "With Jisto, an automatic real-time decision-making process kicks in, enabling true elasticity across the entire data center with granularity as fine as a single core of a CPU."

Jisto not only monitors CPU usage but other parameters such as memory, network bandwidth, and storage. "If there's an important memory transfer happening that requires a lot of bandwidth, Jisto backs off, even if there's plenty of CPU power available," Biberman says. "Jisto can make intelligent decisions about where to send jobs based on all these dynamic factors. As soon as something changes, Jisto decides whether to stop the workload, pause it, or reduce resources. Do you transfer it to another server? Do you add redundancy to reduce the latency tail? People don't have to make and implement those decisions."

The platform also integrates rigorous security provisions, says Biberman. IT directors are understandably cautious about bringing third-party software into their complex data center ecosystems, which are often protected by firewall and regulation settings. Jisto, however, can quickly prove with a beta test how the software can spin its magic without interfering with mission-critical resources, he adds.

Jisto's unobtrusiveness is largely due to its use of Docker containers. "Docker has nice APIs and makes the process much easier, both for us as developers and for Jisto customers," Biberman explains. "Docker is very portable—if you can run it on Linux, you can run it on Docker—and it doesn't care if you're running it on a local data center, a private cloud, or on Amazon. With containers, we don't need to do something complicated like run a VM inside another VM. Docker gives us a lightweight way to let people use the environment that's already set up."

Based in Cambridge, Massachusetts, Jisto was the first, and remains one of few, Docker-based startups in this region.

Moving up to the cloud

Companies are increasingly saving on data center costs by using public cloud resources in a hybrid strategy during peak demand. Jisto can help bridge the gap with better efficiency and flexibility, says Biberman. "If you're a bank, you might have too many regulations on your data to use the public cloud, but most companies can gain efficiencies with public clouds while still keeping their private cloud for confidential, regulated, or mission-critical tasks."

Jisto operates essentially the same whether it's running on-premises, or in a private, public, or hybrid cloud. Companies that exceed the peak level of their private data center can now "burst out" onto the public cloud and take advantage of the elastic nature of services such as Amazon, says Biberman. "Some companies provision hundreds of thousands of nodes on Amazon," he adds. The problem is that Amazon charges by the hour. "If a company only needs five minutes of processing, as many as 100,000 nodes would sit idle for 55 minutes."

Jisto has recently begun to talk to companies that do cloud infrastructure as a service, explaining how Jisto can reprovision wasted resources and let someone else use them. According to Biberman, it's only a matter of time before competitive pressures lead a cloud provider to use something like Jisto.

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology