Gartner Fellows Interview With Microsoft's Ray Ozzie on Cloud Computing

Neil MacDonald, David Mitchell Smith

Cloud-based computing will fundamentally transform how enterprises procure and consume technology solutions. Vendors such as Microsoft, which rely almost exclusively on revenue from traditional shrink-wrapped software, are threatened by the shift, but must embrace the technological and business model discontinuities. At this critical transition point for Microsoft, and nearly a year after Microsoft publicly disclosed its cloud-computing strategy, Gartner Fellows Neil MacDonald and David Smith met with Microsoft Chief Software Architect Ray Ozzie to talk about how Microsoft and the industry have progressed during the past year, and to look ahead at what's next for cloud computing.

Key Findings

- Ozzie's (and thus, Microsoft's) vision of cloud computing emphasizes hybrid enterprise/cloud computing, where organizations choose when and where computing takes place locally versus in cloud-based infrastructure, and emphasizes that the on-premises and cloud-based solutions work seamlessly together.

- As with early encryption issues, Ozzie believes that security and privacy issues will be addressed as the industry matures in a combination of legislative advances, as well as industry cooperation.

- Ozzie's vision for cloud computing includes system infrastructure, an application platform and finished applications being delivered as a service.

- Microsoft is investing in its own data centers because it must in order to provide its consumer-based services. It believes it is helping to lead the industry in providing innovations in data center architectures.

- Ozzie believes the future of cloud computing is in the experience delivered via a browser across multiple devices — mobile, PC and TV-type screens — which is a vision he refers to as "three screens and a cloud."

Recommendation

- Use this discussion to better understand the vision behind Microsoft's cloud-based strategy and how it differs from other cloud-based providers.

TABLE OF CONTENTS

Gartner

1.0 Ozzie's Vision for Cloud-Based Computing

Gartner: If someone were to ask you, "Ray, what's this cloud computing thing all about," do you have a simple explanation?

Ray Ozzie: I wish I could be that concise. Internally, we have two viewpoints. As an industry, we really need to figure this thing out. We confuse utility computing in many ways with the cloud. One might ask, "What's the difference between efficient use of the computing resources versus cloud computing?" In my mind, I separated these two views by saying that we will all be using utility computing — that is, we will provide ways, and other people will provide ways of using virtualization and good management software to run workloads very efficiently on large sets of relatively horizontal machines in relatively homogeneous hardware environments. We'll all be doing utility computing.

However, cloud computing is the ability to flow workloads into your own infrastructure, as well as outside sets of infrastructure, for whatever reason you want. Whether it's for agility, cost or scale, you would use the same programming models and tools that you would for the utility computing on-premises. I know it gets very confusing in terms of private cloud versus public cloud. I would be very open to the industry having that discussion and standardizing the terminology.

Gartner: What you've described so far focuses on cloud-based system infrastructure and application platform as a service, as well as the ability to move workloads seamlessly between on-premises and the cloud. However, does your vision also include things like applications, information or processes that might be consumed? Is it a "full IT stack" so to speak?

Ozzie: In most conversations, when I'm talking about cloud computing, I'm generally talking about the lower layers. It's just the nature of the conversations that I'm in. When people are at the higher layers, they're using other terms, like "software as a service," to talk about the shift that's happening at those higher layers. We call this "software plus services."

At a macrolevel, the world we're moving into is basically "three screens and a cloud." It's a cloud, in terms of where the back end of the computing happens. That could be a private cloud, a public cloud or a partner cloud. That's the back end of the stuff. And the three screens that I'm talking about are a phone class of screen, a PC class of screen and a TV class of screen. All of them have a browser. The browser is kind of the greatest common factor among all of them. Some people say that the browser is the least common denominator, but I like to pay it more respect as the greatest common factor. In general, the experiences that we consume on these screens, and what we will aspire to do as users, have to do with getting the information we need and connecting with the people we want using the browser. The things that we use a lot will have an experience that's tuned to that class of device — TV experiences, PC experiences, phone experiences and so on.

Ultimately, it will be as seamless as clicking on something on the Web and having it deposit and cache that application on the device, or having that data synchronized among the devices or to the cloud and so on. That's more or less the pattern that we're optimizing at Microsoft — whether it's connected to entertainment, productivity or the operating system (OS) itself.

Gartner: Has the cloud computing part of what you described been characterized as a return to centralization, with the industry building these massive data centers all over the world?

Ozzie: It sounds like it's the return of decentralized centralization.

Gartner

Gartner: With the emphasis on centralization and all the focus on building massive data centers, are we possibly going too far in that direction? Are we building the analogy to the water-powered mills that will end up being condos someday because things end up being redistributed back out to personal and enterprise-owned data centers?

Ozzie: I believe in a hybrid model. I fundamentally, deeply believe in a hybrid model at the experience side and at the back-end side.

At the back-end side, it depends on the size of enterprise and the workload, as well as the segment of the enterprise and whether it is highly regulated or whatever. The decisions regarding what to keep on-premises versus what to distribute into the cloud will vary dramatically. Very small businesses will put almost everything into the cloud. Very large businesses will put all their infrastructural systems, such as mail, phone systems and document management, into the cloud. Enterprise applications that have high integration requirements and a lot of legacy issues will stay on-premises. What happens in the middle is a mix.

Gartner: I'm thinking more along the lines of exploiting the local hardware and the ability to use technologies like peer-to-peer networking that will take advantage of the processing power and the storage capacity that people will continue to have in their PCs.

Ozzie: The peer-to-peer capability is not yet well used, but it will be for content delivery networks (CDNs) as one more tier in a CDN hierarchy. Now, we have centralized serving of applications and content. There are CDNs like Akamai and some companies like us and Google that invest in our own CDNs. We float some percentage to commercial CDNs and do some ourselves. But then the third tier will be peer to peer, because the communication costs are high, and it's very efficient to go peer to peer.

Yesterday, all user-generated content was at the edge. The industry digerati paint a picture of all content flowing to the center, but that's unrealistic. It's going to be some hybrid of the two. Even at our scale — or a Facebook scale — it's expensive to put all photographs in the cloud or all videos in the cloud. I'm sure YouTube has this issue. It's problematic to have all that egress being paid for by one party. In that case, peer to peer makes a huge amount of sense. But again, it's a hybrid architecture. If you don't have the center, then you can't rendezvous. You can't find each other. You can't connect in any way, shape or form. However, if you don't have the edge, then you don't have the agility. You pay for ingress and egress when you don't have to.

2.0 The Importance of Microsoft Data Centers to Its Vision

Gartner: Why do you think it is so important for Microsoft to get into owning the physical data centers and spending billions of dollars building its own cloud infrastructure versus just enabling others to do this with your software? It seems like there's a big bet being placed that you have to own this infrastructure.

Ozzie: There are several reasons. The first is we're not just a platform company. We're an application company. So, we have huge infrastructure needs just for the applications we directly serve our consumers — Hotmail on the consumer side, Messenger, Xbox Live and the media service that supports Zune. Just on the consumer side, we have the need for enormous scale.

When I started at Microsoft, each one of these was a stovepiped application. They had their own servers. They had their own people that operated them, even in the common data centers. In essence, each was mirroring its own profit and loss all the way down the technology stack.

I fundamentally believe that all the enterprise applications that we sell as software will also be a service. I know that every time you add a zero to the order of magnitude, you can do it more

Gartner

efficiently. So, if we are serving 100 million Exchange mailboxes, we'll do it better than if we're serving 1 million or 100,000. There is a significant advantage there.

If we're doing it efficiently for ourselves, why not do it for others? One might ask, "Isn't Microsoft a platform company? Doesn't it have a lot of partners and a tremendous ecosystem?" Between 90% and 95% of our revenue comes through partners. It doesn't come directly to us. And we want to keep a healthy worldwide ecosystem of partners alive. Just because we build our own data centers doesn't mean we become the cloud.

For this to be successful to enterprises, we must have this capability in every country — literally in every jurisdiction, all of which have different laws about information handling and privacy. It is a very complex environment. Although we can touch major markets, we cannot get everywhere. Therefore, we need ecosystem partners.

Gartner: How can Microsoft differentiate itself in data centers?

Ozzie: In terms of data center evolution, I don't think people really understand the importance of the advancement of data-center-level technology:

- In a first-generation data center, you'd get screwdrivers out and boot CDs. You'd install the machines in the racks, lay down the OS, configure the networks and so on. When I first came Microsoft just four years ago, that's where we were — and I think that's where most people were. Let's just stay at the hardware level. I won't even touch on the software level yet.

- A second-generation data center is what we referred to as a rack level of deployment. You might buy up to 100 machines or a couple of racks. You'd have them prebuilt by some third party. You'd bring it in and configure it all at once. It's not like you're assembling them — you aren't building things yourself.

- The third generation is what we're deploying now, which is bringing modularity to the level of containers. We build the shell of the building and deploy these containers into it. The first, second and third generation required that you determine how big the property you're going to acquire is, as well as the availability of power. You build massive generators out to cover when the power fails. You build a huge shell of a building where all these racks will go, and then you start filling them until they are done. That filling takes a while, so you've got this big empty shell and big generators sitting there for a while, while you are filling them. The third-generation data center still has that shell, but it's now containers — 2,000 or so computers in each— not a rack wall at a time.

- The fourth generation, which is in trial now — and I'll say we'll move to that as a primary model within a year or so — is fully modular where you're bringing in machines and power backup, and cooling on a modular basis as needed. This generation includes the whole supply chain from creation of the computers all the way up, and we're trying to optimize that to weeks of lead time.

This is all important because for a company like ours, which is building a lot of stuff, we don't want to spend hundreds of millions of dollars that's going to be sitting idle. Every enterprise will have some level of prebuild that they need to do. Every telecom partner and every government that builds data centers have the issue of how much they plan for and how much power they are wasting.

That's just the hardware foundation. Then you overlay the software on top of that so that the systems we're deploying are no longer deployed as stovepipes. There is a management fabric monitoring these, and dynamically moving workloads and powering off machines when they're not in use and so on.

Gartner

If we didn't need the scale that we're doing, we wouldn't be driving the innovation down this curve. I never realized until we started to engage with these major telecom partners how much they need R&D, because they're mostly still at the first generation, or between the first and second generation.

Gartner: Do you think the typical enterprise data center will follow this same path of evolution?

Ozzie: Most enterprises don't have the dynamic expansion and contraction that we need at our scale. Most still operate with stovepiped workload machine combinations. Most of them are still built with heterogeneous hardware configurations — not homogeneous. I think over time enterprises will absolutely benefit — even on-premises — from the fact that we will drive the hardware ecosystem into producing modular containers and providing more choice.

Gartner: Do you see the evolution of the enterprise data center and the evolution of cloud computing infrastructure as being tightly coupled? Do you think that they converge?

Ozzie: I believe that, generally, you take what you have and you incrementally get it to where it's going. Historically, we've been building scale-up architectures in the data centers. We've been building systems with storage area networks, and we've been compensating for reliability by buying expensive hardware that's reliable, as opposed to doing horizontal application models where you can kill a node and the system keeps operating. We then took the scale-up architectures and layered virtualization onto them to have more-flexible consolidation of workload management on this heterogeneous scale-up hardware. One customer's cluster of machines that it runs virtualization on may be completely different from another's. Even in a data center, you might have different classes of clustered machines. Some have a lot of memory, and some have reliable disks. It's a patchwork.

The difference between where we're going is that we're very rigorous about saying, "No. It's all horizontal." We're going for homogeneity. And you don't get the choice of having 10 machines with a ton of memory and 10 machines with bigger disks. We just say there are limited footprints. And maybe there are three: a big disk footprint, a big memory footprint and a big input/output footprint. However, there is limited choice, and you must program in an environment that lets you move in this direction. I think that this model will come into the data center, and you'll end up with a split data center. You'll have the current heterogeneous model and sitting side-by-side with it, literally, will be the homogeneous one. And you'll begin to move workloads from one to the other as the vendors provide rewritten software to fit this new model.

3.0 Changes Needed for Developers and Operations to Embrace Cloud Computing

Gartner: How do people get from one model to the other — particularly developers. What kind of tools will be available? Will this require a change in mind-set? Or can they just rely on you for what's necessary?

Ozzie: Cloud computing won't be successful if organizations and developers have to reinvent everything. That's not what customers want. They want a smooth transition. This notion of the two approaches sitting side by side is very important. That's why, at least from Microsoft's perspective, the System Center management portal will let you manage workloads in both. The technologies in Azure are the technologies that are being brought into the enterprise for the more homogeneous side of that equation.

From a developer perspective, I can only model what I have been through at Microsoft with our own development groups. That's the best pattern that I have in my mind. When I first started talking about this vision in 2006, when I really started getting traction, talking to internal development groups about "you have to shift," the developers went through seven stages of

Gartner

denial. People who have spent their whole life building software the same way really don't get that they can't get to the next level without revisiting something.

The thing that benefited Microsoft internally was that we have had MSN since the mid-1990s growing and serving half a billion users. I created a small internal conference called "SoftServe" that meets annually and brings software development people together with service people. At these events, the service people get up and tell horror stories, such as:

> It's 4 a.m. and the phone rings. It's this guy who works for me:
>
>> "I have bad news and really bad news."
>>
>> "What's the bad news?"
>>
>> "The bad news is we're down and we've lost *n* customers worth of data."
>>
>> "What's the really bad news?"
>>
>> "I think we might not be able to get it up again for two weeks. If we get it up, then we'll permanently lose all the customers' data. There's no way to recover."

There are many horror stories, such as the inability to scale, having to debug complex interdependencies among live services and how you have to develop systems differently to change the engines in flight, which I don't think people appreciate enough. We as an industry have come to assume that the IT guy can bring the server down over the weekend, do the migration or upgrade, and then bring it up. It's just how we've operated. You just can't do that anymore. This means having a very rigorous model view separation, preprovisioning the database in a phase on a live system, and adding the columns that you need so you can start to bring the service live for subsets of users. It means teaching people about behavioral analytics, using A/B testing and watching a subset of users as they start to use the new service, and getting the bugs fixed with them before they start deploying further and further. There's a lot of learning. And then there are things like horizontal programming. The programming in roles are deployed as front ends, as midtier and as back ends. How do you do system integration creating a VLAN or equivalent between the on-premises and off-premises systems that need to integrate to deliver a real solution to a customer? These are tremendous issues.

The opportunity for Microsoft is to lay down design patterns for applications, and to provide sample applications and sample solutions that people can look at and then just delete the insides and start putting their own code in it. This gets people started into the best practices of this new world, because I don't think people will just slip into that mentality overnight.

4.0 Addressing the Security Concerns of Cloud Computing

Gartner: How do you see the security concerns of moving to cloud computing being addressed?

Ozzie: There's no perfect solution. Security is inherently risk management. If it's described as a "black-and-white issue," we'll never get there. Whether on-premises or off-premises, everything is vulnerable. So, we just basically invest at different layers of the architecture. There are different aspects of that investment. Oddly enough, it starts with the lawyers and with our policy folks. We have to understand the regulatory environment in every single jurisdiction that we or our customers want to serve. The analogy that I'll make that might resonate is that we are with cloud computing right now where we were with encryption with mass market products and exports controls in the early 1990s — which is that everybody had their own export and import restrictions, which prevented a software developer from writing something with crypto in it and getting it shipped.

Gartner

The No. 1 priority is understanding the environment. The No. 2 priority is making sure that, at the infrastructure level, we understand the roles of the various human beings involved who might touch the hardware, and understand what is on that hardware that is in the clear and encrypted. Understand what is on the wire that is in the clear and encrypted. Moving up the stack to the keys that we manage, we need to understand those keys and their flow. You have to understand the threat models. We have threat models for certain applications that we have to guard against, such as someone coming in and physically taking the computer or gaining physical access. The customer is less concerned with an oscilloscope on the circuit board on a live system, but we are.

Ultimately, what's going to make customers trust the cloud is if we put their most valuable data in there, and others put their most valuable data in there — and, thus, prove over time that we are trustworthy. The same will be true of our competitors. Thank goodness we went through HailStorm and Passport, because we learned a lot, and I would argue that we are the most trustworthy — and the Federal Trade Commission has made sure that we understand privacy practices and that we understand how to handle personally identifiable information within our infrastructure. The European Union (EU) is making sure that we understand how to deal with privacy in a way that reflects its laws and its regulatory environment, which is different from the U.S.

Gartner: With the encryption export controls, ultimately we relaxed the export controls. For cloud computing, do you think that the answer is that Canada, the U.S. and the EU should relax some of these constraints?

Ozzie: I think things will ultimately get relaxed. I'll give you an example of something that probably should be relaxed. I was talking to a customer who was looking at cloud computing and who deals with health information in the U.K. A citizen in Ireland and a citizen in England cannot store their records in the other country, even within one health service that services both. They can't even have replica copies for redundant backup. Therefore, you have to spend the money to have all the redundancy within one jurisdiction. I think there are things where we just haven't been educated yet to understand what Larry Lessig [professor at Stanford Law School] means by "the code is the law." The architectural possibilities have to come together with the law in some way, shape or form, and that will take time. It isn't that all these things will be repealed. Encryption keys are more important than the physical location of the data.

I don't see legislation that says that the keys must be in this or that country. They say the data must be in this or that country. There are a lot of nuances that we have going on between developers/architects and lawyers trying to get around the laws based on architecture. Cooler heads must prevail, and that only happens with time. It happens when it's driven by customers. The customers will want to do things that they can't do. The vendors can spend money on lawyers and lobbyists to help. But, ultimately, it's going to be the customers who come back, apply pressure and make these things happen.

Gartner: Do you think that this is a decade off?

Ozzie: I'm more optimistic because the economic value of what we are talking about is so huge and because there are environmental issues.

We'll begin to see progress locally and jurisdictionally in the three- to five-year time frame. Ten years from now, we'll look back and say, "I remember when… ." Universally, things will be pretty clear in 10 or more years.

Gartner: Is this a scenario where you, Google and Amazon collectively work to get the regulations changed?

Gartner

Ozzie: Consider the crypto analogy. We had the BSA [Business Software Alliance] and all the competitors, such as Microsoft and Lotus, give a common voice. We went on a roadshow to all the senators to force change.

Gartner: Is there a similar initiative for cloud computing?

Ozzie: I don't think it is as formal as the BSA right now, but I'm sure these things will happen. There are places where competitors get together and make these changes happen. A lot of it is just education. Everybody needs to know what this cloud thing is. For reasons that you said, the industry is not helping.

5.0 Beyond Cloud Computing

Gartner: What's next beyond the cloud?

Ozzie: I believe we are moving to a world with three screens and the cloud. Our experience model is repivoting to be cloud-based as the hub, and delivered across the phone, PC and TV. We've only begun to conceptualize that. If there are no programming models for the back end and the cloud stuff, then patterns and practices will emerge. We don't have things to copy on the front end.

We've talked about how we can use the parallel resources on the client to make a better client. But we really haven't talked about the design patterns for solutions. To deliver your photos, it's a canonical one. What is the right architecture if you wanted to write a photo service? How do you build a project and write it so that there is a service component that delivers stuff through a browser, and has the right back-end pieces? Can you build a separate but related piece of code that relates to the service and the phone that does very efficient notifications, so that it doesn't drain the battery? Is there code for the TV so that it understands how you deal with remotes, so that when you come into a living room and you happen to have a phone and the phone pairs with the TV, you can flip photos from the phone onto the TV? There are scenarios that we will want to develop, and we don't just want to build a vertical solution. We want developers to have the capabilities of figuring out how to build solutions and stand by these things. Beyond the cloud is related to the cloud, but it's more at the experience level.

What is the experience transformation? If you want something to reference when you are writing it, then take a look at the office lab site. We've been playing "the future of productivity" knowledge navigator era videos that lead you through the life of someone who is traveling. Yes, they are kind of out there, but they ask the questions: Where is that thing running? How did it move from that thing in their hand to that thing on the wall? How did that thing get on the desktop that was just sitting in the room? It's not all science fiction. There are real system issues that we need to develop to get to that world, and it's very relevant.

6.0 Thoughts on Following Bill Gates

Gartner: It's been several years now that you have been taking on the role of chief software architect and somehow filling the shoes of the founder of the company, Bill Gates. How is this evolving for you?

Ozzie: The reason I'm pausing is … Craig [Mundie] and I *aren't* able to fill Bill's shoes. Bill will be present forever, even though he's not present. If we had approached it like "we're filling the founder's shoes," then we would have failed — absolutely and miserably — because we are not Bill. The transition of doing the things that we do is going well. There are challenges, and there are things that have worked out way better than I ever could have imagined. It varies person by person by person.

Gartner

You've heard the axiom, "leadership takes followership." For those who want to embrace things moving forward, things have gone amazingly well. Some people were in fear of what would happen and in fear that there would be paralysis, and that didn't happen. Everyone kind of realized that we are moving on, and it's a new thing. My engagement style is far different from Bill's. For a number of groups, that has worked out really well. With others, there are challenges. Some people have a different style or a different view of how they want to take it.

There's some empirical evidence, though, when you go back to that memo I wrote back in 2005. In those days, I had conversations with Bill and he'd say, "Well that's pretty dramatic or radical in terms of what you are trying to accomplish. It's the right thing to do and if you do it, that will be great." And I said, "How?" And he'd say, "I don't know. It starts with a memo, and I don't know what happens after that." I didn't have a path to figure out how it was going to pan out. When I look back and I read the memo, so many of the things that I had written have come to pass, not because I drove them to make it happen, but because the organization made it happen. It may have happened a little differently here or there, but it happened. So, I'm very pleased about that.

RECOMMENDED READING

"Vendor Rating: Microsoft"

"Vendor Focus for Microsoft: cloud Computing Is the Next Big Thing"

"Windows Azure Completes the Microsoft Platform Triangle"

"Application Infrastructure for Cloud Computing: An Emerging Market"

"Microsoft's Cloud Vision Reaches for the Stars but Is Grounded in Reality"

"Live Mesh Is a Step Toward Microsoft's Cloud and Web Platform"

"Web Platforms: Web-Centric Delivery Models for Infrastructure, Applications, Data and Business Services"

"Exchange Market Vulnerability Starts With Outlook"

"Microsoft Collaboration Offering Ushers in New Era for SaaS"

"SharePoint Online: Microsoft's Collaboration Platform in the Sky"

Gartner

REGIONAL HEADQUARTERS

Corporate Headquarters
56 Top Gallant Road
Stamford, CT 06902-7700
U.S.A.
+1 203 964 0096

European Headquarters
Tamesis
The Glanty
Egham
Surrey, TW20 9AW
UNITED KINGDOM
+44 1784 431611

Asia/Pacific Headquarters
Gartner Australasia Pty. Ltd.
Level 9, 141 Walker Street
North Sydney
New South Wales 2060
AUSTRALIA
+61 2 9459 4600

Japan Headquarters
Gartner Japan Ltd.
Aobadai Hills, 6F
7-7, Aobadai, 4-chome
Meguro-ku, Tokyo 153-0042
JAPAN
+81 3 3481 3670

Latin America Headquarters
Gartner do Brazil
Av. das Nações Unidas, 12551
9° andar—World Trade Center
04578-903—São Paulo SP
BRAZIL
+55 11 3443 1509

Gartner