CERN's Evolution to Cloud Computing Portends Revolution in Extreme IT Productivity

Transcript of a BriefingsDirect podcast on the move to cloud computing for data-intensive operations, focusing on the work being done by the European Organization for Nuclear Research.

Listen to the podcast. Find it on iTunes/iPod and Podcast.com. Download the transcript. Sponsor: Platform Computing.

Dana Gardner: Hi, this is Dana Gardner, principal analyst at Interarbor Solutions, and you're listening to BriefingsDirect. Today, we present a sponsored podcast discussion on some likely directions for cloud computing based on the exploration of expected cloud benefits at a cutting edge global IT organization.

We are going to explore the thinking on how cloud computing both the private and public varieties might be useful at CERN, the European Organization for Nuclear Research in Geneva.

CERN has long been an influential bellwether on how extreme IT problems can be solved. Indeed, the World Wide Web owes a lot of its usefulness to early work done at CERN. Now the focus is on cloud computing. How real is it, and how might an organization like CERN approach cloud?

In many ways CERN is quite possibly the New York of cloud computing. If cloud can make it there, it can probably make it anywhere. That's because CERN deals with fantastically large datasets, massive throughput requirements, a global workforce, finite budgets, and an emphasis on standards and openness.

So, please join us, as we track the evolution of high-performance computing (HPC) from clusters, to grid, to cloud models through the eyes of CERN, and with analysis and perspective from IDC, as well as technical thought leadership from Platform Computing.

Join me in welcoming our panel today. We are here with Tony Cass, Group Leader for Fabric Infrastructure and Operations at CERN. Welcome, Tony.

Tony Cass: Pleased to meet you.

Gardner: We're also here with Steve Conway, Vice President in the High Performance Computing Group at IDC. Welcome, Steve.

Steve Conway: Thanks. Welcome to everyone.

Gardner: And, we're also here with Randy Clark, Chief Marketing Officer at Platform Computing. Welcome Randy.

Randy Clark: Thank you. Glad to be here.

Gardner: Over the last several years, we've seen cloud computing become quite popular as a concept. It remains largely confined to experimentation, but this notion of private cloud computing is being scoped out by many large and influential enterprises as well as large early adopters like CERN.

Let me go to you Steve Conway. What's the difference between private and public cloud and how far away are any tangible benefits of cloud computing from your perspective?

Already here

Conway: Private cloud computing is already here, and quite a few companies are exploring it. We already have some early adopters. CERN is one of them. Public clouds are coming. We see a lot of activity there, but it's a little bit further out on the horizon than private or enterprise cloud computing.

Just to give you an example, we just did a piece of research for one of the major oil and gas companies, and they're actively looking at moving part of their workload out to cloud computing in the next 6-12 months. So, this is really coming up quickly.

Gardner: So, this notion of having a cohesive approach to computing and blending what you do on premises with these other providers isn't just pie in the sky. This is really something people are serious about.

Conway: Well, CERN is clearly serious about it in their environment. As I said, we're also starting to see activity pick up with cloud computing in the private sector with adoption starting somewhere between six months from now and, for some, more like 12-24 months out.

Gardner: Randy Clark, from your perspective, how many customers of Platform Computing would you consider to be seriously evaluating what we now refer to as public or private cloud?

Clark: We have formally interviewed over 200 customers out of our installed base of 2,000. A significant portion -- I wouldn't put an exact number on that, but it's higher than we initially anticipated -- are looking at private-cloud computing and considering how they can leverage external resources such as Amazon, Rackspace and others. So, it's easily a third and possibly more.

Gardner: Tony Cass, let's go to you at CERN. Tell us first a little bit about CERN for those of our readers who don't know that much or aren't that familiar.

Tell us about the organization and what it does, and then we can start to discuss your perceptions about cloud.

Cass: We're a laboratory that exists to enable, initially Europe's and now the world's, physicists to study fundamental questions. Where does mass come from? Why don't we see anti-matter in large quantities? What's the missing mass in the universe? They're really fundamental questions about where we are and what the universe is.

We do that by operating an accelerator, the Large Hadron Collider, which collides protons thousands of times a second. These collisions take place in certain areas around the accelerator, where huge detectors analyze the collisions and take something like a digital photograph of the collision to understand what's happening. These detectors generate huge amounts of data, which have to be stored and processed at CERN and the collaborating institutes around the world.

We have something like 100,000 processors around the world, 50 petabytes of disk, and over 60 petabytes of tape. The tape is in just a small number of the centers, not all of the hundred centers that we have. We call it "computing at the terra-scale," that's terra with two R's. We've developed a worldwide computing grid to coordinate all the resources that we have with the jobs of the many physicists that are working on these detectors.

Gardner: So, to look at the IT problem and unpack it a little bit. You're dealing with such enormous amounts of data. You've been in the distribution of these workloads for quite some time. Maybe you could explain a little bit the evolution of how you've distributed and managed such extreme workload?

No central management

Cass: If you look at the past, in the 1990's, we had people collaborating, but there was no central management. Everybody was based at different institutes and people had to submit the workloads, the analysis, or the Monte Carlo simulations of the experiments they needed.

We realized in 2000-2001 that this wasn't going to work and also that the scale of resources that we needed was so vast that it couldn't all be installed at CERN. It had to be shared between CERN, a small number of very reliable centers we call the Tier One centers and then 100 or so Tier Two centers at the universities. We were developing this thinking around the same time as the grid model was becoming popular. So, this is what we've done.

What a lot of the grid academics have done is in understanding or exploring what could be done with the grid, as an idea. What we've been focusing on is making it work and not pushing the envelope in terms of the technology, but pushing the envelope in terms of the scale to make sure that it works for the users. We connect the sites. We run tens of thousands of jobs a day across this and gradually we've run through a number of exercises to distribute the data at gigabytes a second and tens of thousands of jobs a day.

We've progressively deployed grid technology, not developed it. We've looked at things that are going on elsewhere and made them work in our environment.

Gardner: As I understand it, the interest you have in cloud isn't strictly a matter of ripping and replacing, but augmenting what you're already doing vis-a-vis these grid models.

Cass: Exactly. The grid solves the problem in which we have data distributed around the world and it will send jobs to the data. But, there are two issues around that. One is that if the grid sends my job to site A, it does so because it thinks that a batch slot will become available at site A first. But, maybe a grid slot becomes available at site B and my job is site A. Somebody else who comes along later actually gets to run their job first.

Today, the experiment team submits a skeleton job to all of the sites in order to detect which site becomes available first. Then, they pull down my job to this site. You have lots of schedulers involved in this -- in the experiment, the grid, and the site -- and we're looking at simplifying that.

These skeleton jobs also install software, because they don't really trust the sites to have installed the software correctly. So, there's a lot of inefficiency there. This is symptomatic of a more general problem. Batch workers are good at sharing resources that are relatively static, but not when the demand for resource types changes dynamically.

So, we're looking at virtualizing the batch workers and dynamically reconfiguring them to meet the changing workload. This is essentially what Amazon does with EC2. When they don't need the resources, they reconfigure them and sell the cycles to other people. This is how we want to work in virtualization and cloud with the grid, which knows where the data is.

Gardner: Steve Conway, you've been tracking HPC for some time at IDC. Maybe you have some perceptions on how CERN is a leading adopter of IT over the years, the types of problems they're solving now, or the types of problems other organizations will be facing in the future. Could you tell us about this management issue and do you think that this is going to become a major requirement for cloud computing?

World technology leader

Conway: Starting with CERN, their scientists have earned multiple Nobel prizes over the years for their work in particle physics. As you said before, CERN is where Tim Berners-Lee and his colleagues invented the World Wide Web in the 1980s.

More generally, CERN is a recognized world leader in technology innovation. What's been driving this, as Tony said, are the massive volumes of data that CERN generates along with the need to make the data available to scientists, not only across Europe, but across the world.

For example, CERN has two major particle detectors. They're called CMS and ATLAS. ATLAS alone generates a petabyte of data per second, when it's running. Not all that data needs to be distributed, but it gives you an idea of the scale or the challenge that CERN is working with.

In the case of CERN's and Platform's collaboration, as Tony said, the idea is not just to distribute the data but also the applications and the capability to run the scientific problem.

CERN is definitely a leader there, and cloud computing is really confined today to early adopters like CERN. Right now, cloud computing services constitute about $16 billion as a market. That's just about four percent of mainstream IT spending. By 2012, which is not so far away, we project that spending for cloud computing is going to grow nearly threefold to about $42 billion. That would make it about 9 percent of IT spending. So, we predict it's going to move along pretty quickly.

Gardner: How important is this issue that Tony brought up about being able to manage in a dynamic environment and not just more predictable static batch loads?

Conway: It's the single biggest challenge we see for not only cloud computing, but it has affected the whole idea of managing these increasingly complex environments -- first clusters, then grids, and now clouds. Software has been at the center of that.

That's one of the reasons we're here today with Platform and CERN, because that's been Platform's business from the beginning, creating software to manage clusters, then grids, and now clouds, first for very demanding, HPC sites like CERN and, more recently, also for enterprise clients.

Gardner: Randy Clark, as you look at the marketplace and see organizations like CERN changing their requirements, what, in your thinking, is the most important missing part from what you would do in management with HPC and now cloud? What makes cloud different, from a management perspective?

Dynamic resources

Clark: It's what Tony said, which is having the resources be dynamic not static. Historically, clusters and grids have been relatively static, and the workloads have been managed across those. Now, with cloud, we have the ability to have a dynamic set of resources.

The trick is to marry and manage the workloads and the resources in conjunction with each other. Last year, we announced our cloud products -- Platform LSF and Platform ISF Adaptive Cluster -- to address that challenge and to help this evolution.

Gardner: Let's go back to Tony Cass. Tell me what you're doing with cloud in terms of exploration. I know you're not in a position to validate, or you haven't put in place, any large-scale implementation or solutions that would lead the market. But, I'm very curious about what the requirements are. What are the problems that you're trying to solve that you think cloud computing specifically can be useful in?

Cass: The specific problem that we have is to deliver the most physics we can within the fixed budget and the fixed amount of resources. These are limited either by money or by data-center cooling and generally are much less than the experiment wants. The key aim is to deliver the most cycles we can and the most efficient computing we can to the physicists.

I said earlier that we're looking at virtualization to do this. We've been exploring how to make sure that the jobs can work in a virtual environment and that we can instantiate virtual machines (VMs), as necessary, according to the different experiments that are submitting workloads at one time to integrate the instantiation of VMs with the batch system.

Once we got that working, we figured that the real problem was managing the number of VMs. We have something like 4,000 boxes, but if you have a VM per call, plus a few spare, then it can easily get to 60,000, 70,000, or 80,000 VMs. Managing these is the problem that we are trying to explore now, moving away from "can we do it" to "can we do it on a huge scale?"

Gardner: Are you yet at the point where you want to be able to manage the VMs that you have under your own control, and perhaps starting to deploy virtualized environments and workloads in someone else's cloud and make them managed and complimentary.

Cass: There are two aspects to that. The resources in our community are at other sites, and all of the sites are very independent. They are also academic environments. So, they are exploring things in their own way as well. At the moment, we're looking at how you can reliably send a virtual image that's generated at one place to another site.

Amazon does this, but there are tight constraints in the way they manage that cluster, because they built it thinking about this. Universities maybe didn't build their own cluster in a way that separates that out from some of the other computing they're doing. So, there are security and trust implications there that we are looking at. That will be a thing to collaborate on long-term.

More cost effective

Certainly, if we configure things in our own way, when we look in a cloud environment, perhaps it will be more cost effective for us to only purchase the equipment we need for the average workload and they buy resources from Amazon or other providers. But, there are interesting things you have to explore about the fact that the data is not at Amazon, even if they have the cycles.

There are so many things that we're thinking about. The one we're focusing on at the moment is effectively managing the resources that we have here at CERN.

Gardner: Steve Conway, it sounds as if CERN has, with its partnered network, a series of what we might call private-cloud implementations and they're trying to get them to behave in concert at what we might call at a public cloud level. That exercise could, as with the World Wide Web, create some de-facto standards and approaches that might, in fact, help what we call hybrid cloud computing moving forward. Does that fairly surmise where we are?

Conway: That's right. There are going to have to be more rigorous open standards for the clouds. What Tony was talking about at CERN is something that we see elsewhere. People are turning to public clouds today -- "turning to" just meaning exploring at this point for a way to handle overload work and search workloads.

The Internet itself is a pretty high latency network, if you think of it that way. People are looking to send portions of the workload that doesn't have a lot of communication dependencies particularly inter-processor communication dependencies, because the latency doesn't support that.

But, we're seeing some smaller and medium-size businesses looking to public clouds as a way to avoid having to purchase their own internal resources, clusters for example, and also as a way of avoiding having to hire experts who know how to operate them. For example, engineering services firms don't have those experts in house today.

Gardner: Back to you Tony Cass, I know this is still a bit hypothetical, but if there were the standards in place, and you were able to go to a third-party cloud provider for some of these spikes or occasionally dynamically generated workloads that perhaps exceed your current on-premise's capabilities, would this be a financial boon to you, where you could protect your pricing and you could decide the right supply and demand fit when it comes to these extreme computing problems?

Cass: It would certainly be a boon. The possibility is being demonstrated by experiments that are actually based at Brookhaven to do simulations that are CPU-intensive, where they don't need much data transfer or data access. They have been able to run simulations cost-effectively with EC2.

Although their cycles, compared to some of the things we're doing, are more expensive, if we don't have to buy all of the resources, we could certainly save money. Another aspect is that it is beyond money in some sense. If you need to get something fixed for a conference, and you are desperately trying to decide whether or not you've discovered the Higgs then it's not a case of "money's no object," but you can get the resources from a cloud much more quickly than you can install capacity at CERN. So both aspects are definitely of interest.

Gardner: Randy Clark, this makes a great deal of sense from the perspective of a large research organization. But, we're not just talking about specific workloads. We're talking about workloads that will be common across many other vertical industries or computing environments. Can you name a few, or mention some from your experience, where we should expect the same sorts of economic benefits to play out.

Different use cases

Clark: What we're seeing is across industries. Financial services is certainly taking a leadership role. There's a lot going on in the semiconductor or electronic industry. Business intelligence (BI) is across industries and government. So, across industries, we see different use cases.

To your point, these use cases are enterprise applications to run the business, and we're seeing that in Java applications, test and development environments, and traditional HPC environments.

That's something driven by the top of the organization. Tony and Steve laid it out well. They look at the public/private cloud economically, and say, "Architecturally, what does this mean for our business?" Without any particular application in mind they're asking how to evolve to this new model. So, we're seeing it very horizontally and, to your point, in enterprise and HPC applications.

Gardner: Steve Conway, thinking about these large datasets, Randy brought up BI, and that, of course, means warehousing, data analytics, and advanced analytics. A lot of organizations are creating datasets at a scale never anticipated, never mind seen before, things from sensors, mobile devices, network computing, or social networking.

How do we bring together these compute resources, the raw power with these large datasets. I think this is an issue that CERN might also be a bellwether on, in somehow managing these large datasets and the compute power, bringing them architecturally into alignment.

Conway: BI is one of those markets that, in its attributes, straddles the world of HPC and enterprise computing just as financial services does, in the sense that they have workloads that don't have a whole lot of communications dependencies. They don't need networks with very high latency for the most part.

You see organizations like the University of Phoenix, which has 280,000 online students, that have already made this evolution -- in this case, with Platform helping them out -- from clusters to grid computing today. Now, they're looking toward cloud computing as a way to take them further.

You also see that not just in the private sector side. One of the other active customers that's really looking in that same direction is the Centers for Disease Control (CDC), which has moved to from clusters to grid computing.

What you're seeing here is people who have already stepped through the earlier stages of this evolution. They've gone from clusters to grid computing for the most part and now are contemplating the next move to cloud computing. It's an evolutionary move. It could have some revolutionary implications, but, from a technological standpoint, sometimes evolutionary is much safer and better than revolutionary.

Gardner: Tell us about some of the solutions that you now need to bring to market or are bringing to market around management and other issues? Where have you found that the rubber hits the road, in terms of where people can take this in real time? What's the current state of the art? Rather than talking about hypothetical, what's now possible, when it comes to moving from cluster and grid to the revolution of cloud?

Interaction of technologies

Clark: What Platform sees is the interaction of distributed computing and new technologies like virtualization requiring management. What I mean by that is the ability, in a large farm or shared environment, to share resources and then make those resources dynamic. It's the ability to add virtualization into those on the resource side, and then, on the server side, to make it Internet accessible, have a service catalog, and move from providing IT support to truly IT as a competitive service.

The state of the art is that you can get the best of Amazon, ease of use, cost, accessibility with the enterprise configuration, scale, and dependability of the enterprise grid environment.

There isn't one particular technology or implementation that I would point to, to say "That is state of the art," but if you look across the installations we see in our installed base, you can see best practices in different dimensions with each of those customers.

Gardner: Randy, what are some typical ways that you're seeing people getting started, when they want to make these leaps from evolutionary progression to revolutionary paybacks? Where do they start making that sort of catalytic difference?

Clark: The evolution is the technology, as Steve said. The revolution is in the approach architecturally to how to get to that new spot.

Taking a step back, we see customers thinking about architecturally how do they want to have that management layer? What is that management layer going to mean to them going forward? And, can they quickly identify a set of applications and resources and get started.

So, there is an architecture piece to it, thinking about what the future will hold, but then there is a very pragmatic piece -- let's get going, let's engage, let's build something and be able to scale that out over time. We saw that approach in grid computing. We're encouraging folks to think, but then also to get started.

Gardner: Tony Cass at CERN, what are your next steps? Where would you expect to be heading next as you explore the benefits and possible real-world opportunities?

Cass: We're definitely concentrating for the moment on how we exploit effective resources here. The wider benefits we'll have to discuss with our community.

Gardner: What would you like to see happen next?

Focusing on delivery

Cass: What I would like to see happen next is a definite cloud environment at CERN, where we move from something that we're thinking about to something that is in operation, where we have the ability to use resources that aren't primarily dedicated for physics computing to deliver cycles to experiment. I'd like to see a cloud, a dynamically evolving environment in our computer center. We're convinced it's possible, but delivering that is what we're focusing on.

Gardner: Steve Conway, where do you see things headed next? What are the next steps that we should look for, as we move from that evolutionary progression to more of a revolutionary productivity?

Conway: It's along a couple of dimensions. One is the dimension of people actually working in these environments. In that sense, the CERN-Platform collaboration is going to help drive the whole state of the art forward over the next period of time.

The other one, as Randy mentioned before, it that the evolution of standards is going to be important. For example, right now, one of the barriers to public-cloud computing is vendor lock-in, where the cloud, the Amazons, the Yahoos, and so forth are not necessarily interoperable. People are a little bit concerned about testing their data there. The evolution of standards is going to accelerate this trend.

Gardner: Why don't I give the last word today to Randy? Tell us about some information that's available out there for folks who are looking to explore and take some first steps towards this more revolutionary benefit.

Clark: I'd encourage everybody to visit our website. There are a number of white papers, webinars, and webcasts that we've done with other customers to highlight some other use cases within development, test, and production environments. I'd point people to the resource page on our website www.platform.com.

Gardner: I want to thank our guests. This has been a very interesting discussion, and I certainly look forward to following what CERN does, because I do think that they're going to be a leader in terms of what many others will be end up doing in B2B cloud computing.

Thank you to Tony Cass, Group Leader for Fabric Infrastructure and Operations at CERN. Thank you, sir.

Cass: Thank you.

Gardner: And also a good, big thank you to Steve Conway, Vice President in the High Performance Computing Group at IDC. Thank you, Steve.

Conway: Thanks.

Gardner: And also, of course, thank you to Randy Clark, Chief Marketing Officer at Platform Computing.

Clark: Thank you for the opportunity.

Gardner: This is Dana Gardner, principal analyst at Interarbor Solutions. You've been listening to a sponsored BriefingsDirect podcast on what likely outcomes we can expect from cloud computing and architecture, on the progression from grid to cloud computing, and moving into a more revolutionary set of benefits. Thanks for listening and come back next time.

Listen to the podcast. Find it on iTunes/iPod and Podcast.com. Download the transcript. Sponsor: Platform Computing.