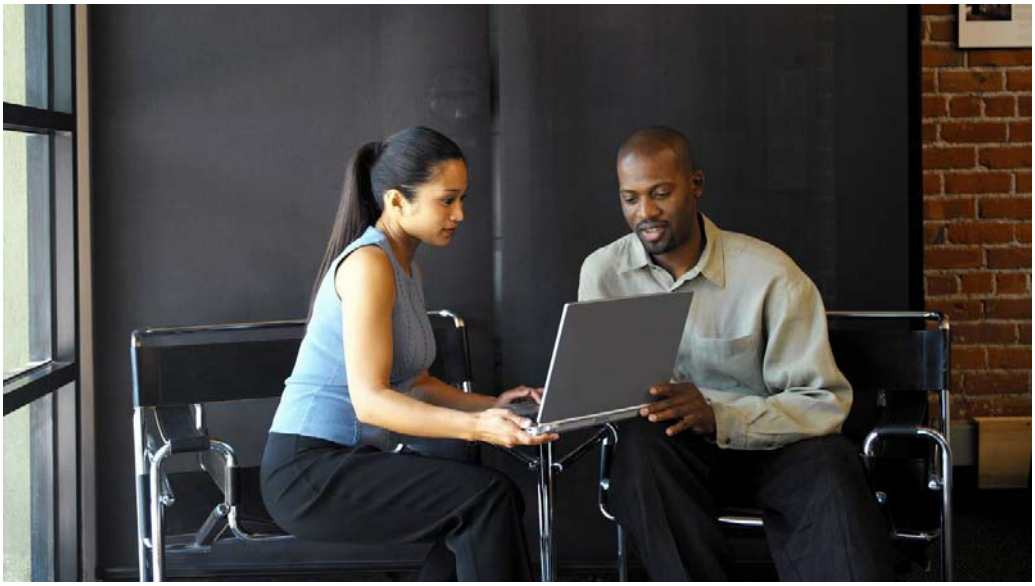
Cisco Public.

Technology Tutorials

UNIFIED COMMUNICATIONS-BEST PRACTICES TRANSCRIPT



Program and Presenter Opening

Ian Pudney: Hello and welcome to the "Cisco on Cisco" Tutorial on Cisco's Unified Communications Best Practices. I'm Ian Pudney, and I'm an IT Manager in the Unified Communications team.

The theme of our show today is an overview of best practices Cisco IT has developed based on its own experiences in deploying, managing, and supporting a Unified Communications environment.

It's my pleasure to introduce the guest of today's show: Kevin O'Healy, a Member of Technical Staff from the Unified Communications Operations team.

Kevin O'Healy: Thank you, Ian. It's a real pleasure to be here today and to be able to present our best practices for Unified Communications. So take it away, Kevin.

Solution Redundancy

Kevin O'Healy: So, in any discussion of IP Telephony, best practices is going to start with an end-to-end solution redundancy, and this is going to extend to both the underlying network infrastructure as well as to all of the individual components in the unified communications deployment, including the Cisco Unified Communications Manager servers, the IP IVR servers, voicemail servers.

Let's go ahead and start with the Cisco Unified Communications

Manager servers. The Call Manager cluster has built-in redundancy capabilities in the form of primary subscribers and secondary subscribers. And the key when it comes to best practices is to place these servers in different locations.

These locations in a campus environment can be in different data centers, for example, or in an office where all of the services are located in a single building. The other option is to place the Call Manager servers on different floors of the building.

Really, the key is to have spatial redundancy throughout the environment. So in the event that there is a network failure, or a power failure, in one part of the environment, the alternate secondary servers can take over those services such as call processing, and still ensure that all of the clients have functionality.

The same holds true for the supplementary services such as TFTP servers. In a Call Manager cluster, all of the phones, all of the gateways, pull their configuration files from a TFTP server. So just as you would with a Call Manager server, place your TFTP servers in different locations so that in the event that one part of the network is impacted, the TFTP server in the alternate part of the network can still provide those services.

And then lastly, all of your supplementary services, conferencing bridges, music on hold, transcoders, IP IVRs, the same holds true for these. The key is to implement redundancy within the system itself and then to place all of those services throughout the overall infrastructure so that you have built-in resiliency.

And just to mention, one other option is the cluster over the WAN design model where, in this case, you actually place your servers for the Cisco Unified Communications Manager in different cities. They all participate in the same Call Manager cluster, but there's spatial redundancy.

And there are some geographic limitations that have to do with bandwidth and latency that are tied to this solution, but it is one solution that works very well, especially when it comes to expanding geographic reach for a particular region of the world.

Ian Pudney: So Kevin, questions on that. What are the latency limitations when it comes to clustering over the WAN?

Kevin O'Healy: The primary limitation is going to be the 20-millisecond response time. So from one Call Manager to another Call Manager needs to be 20 milliseconds one way, or 40 milliseconds roundtrip.

Ian Pudney: Okay.

Kevin O'Healy: But there are some other bandwidth limitations as well. The key is to look at the Cisco Unified Communications Manager SRND where it outlines exactly what those limitations are.

Solution Redundancy (cont…)

So continuing on with end-to-end solution redundancy, this also extends to our voice gateways, so our voice gateways for routing outbound calls and receiving inbound calls. Many organizations that I look at, they standardize on one single vendor for all of their calling. And the downside of that is they are completely dependent on that one carrier. If that carrier is having an issue within their own environment, there's no option to reroute those calls.

The best practices call for multiple paths for routing. So for example, using different carriers, one for local exchange and one for inter-

exchange. In this case, if you attempt to place a long distance call out of your cluster, first utilize your dedicate long distance trunks, but in the event that all circuits are busy, then reroute the call automatically through your local trunks just by pre-fixing the necessary area code.

Just as you would with your Call Manager servers, the key is to place all of your voice gateways throughout the environment so they're not all located in one ISR chassis, for example. But instead, separate those gateways throughout the total infrastructure so that you do have that built-in redundancy.

And then lastly, utilize the PSTN as a backup route for on-net dialing. So if within your organization your preferred route to dial between two different offices is on-net over your WAN, in the event of an out-of-bandwidth condition, utilize the PSTN as an alternate route, Automated Alternate Routing, AAR, is one way of doing this.

You can also do this with the route list and route group constructs within Call Manager so that in the event that the call does not succeed on-net, there is a second route available to route that call, in this case, over the PSTN.

Ian Pudney: Question, Kevin. How does Cisco handle diversity from an inbound call perspective with PSTN?

Kevin O'Healy: Inbound call diversity can be a little more challenging than outbound. The reason being is that we can control outbound, but in the case of inbound, we're reliant upon the carrier to deliver that call to us.

So one of the easy ways of doing inbound diversity is multiple circuits that can terminate inbound calls that are all part of the same trunk route. In this case, any one of those circuits can terminate that call. And if one circuit goes down, another circuit can take over.

Ian Pudney: Okay.

Kevin O'Healy: And then continuing on for end-to-end solution redundancy, we're looking at the overall network. This is the underlying infrastructure that all of our unified communications platforms ride on.

Solution Redundancy (cont 2…)

Of course, we look at the core distribution and access layer model that everyone should be familiar with when it comes to design of the network. But we're looking are the individual devices and the fact that each of those devices have redundant components, redundant power supplies, different power circuits, supervisor modules that should be redundant.

One thing I like to point out is often times I'll walk into an IDF closet and I'll look at a switch, and it is a high availability switch, multiple power supplies, the supervisor modules, redundant line cards. But both of the power supplies will be plugged into the same circuit. That really doesn't benefit you if you have a blown fuse on that circuit.

So in this case, two separate power supplies should be attached to two separate circuits so that in the event one goes down, the switch or the router can still stay online.

And when it comes to power, I also want to mention UPSs, all the devices within the infrastructure, from the data center to the actual closets powering, all of the routers and switches should have a UPS. And our goal is a minimum of two hours of run time. And what that two hours represents is enough time for the local IT staff to troubleshoot that power outage.

But in reality, what we've seen is that in most power outages, if the power isn't restored in about 15 to 20 minutes, the users typically will pack up and work from home or they'll go to a local coffee shop and work remotely rather than stay in the building and work in the dark.

Ian Pudney: Question, Kevin. Mostly if we're involved in IT, we're very familiar with the use of UPS in the main data center. But can you explain a bit more what's the value of having the UPS in all of the floor and wiring closets?

Kevin O'Healy: Absolutely. So if properly designed, IP Telephony infrastructure would utilize inline power, or Power over Ethernet, to power all of the phones.

So in this case, in a data center we have redundant power. In the event of a power outage, the Call Manager servers stay online. All of the network closets where we have our routers and switches are attached to a UPS, so they stay online. And because our phones at the end user stations receive inline power, they also stay online. So even though the building power goes down, we still have voice connectivity.

Ian Pudney: All right, thanks.

Change Management

Kevin O'Healy: Next, let's just take a moment to talk about the change management process. And speaking from experience, there are a lot of little things that we can do with the change management process to make our lives that much easier.

No. 1 is to incorporate all of the necessary teams, from voice and data, business groups and application teams, into a single review board when it comes to looking at change management.

The real goal here is to make sure that all of the necessary eyes are looking at any change that's going to happen to the environment. So in this case, the network team, for example, might be upgrading a WAN circuit or performing an upgrade to a LAN switch. This is going to impact the unified communications teams. So it's important to have the UC teams include on that change management notification so they have visibility to that change.

To facilitate this, what can use are multiple aliases. For example, one alias, which includes all IT personnel interact with the voice and data network, and then a second alias, which includes voice support as well as any of the application and business teams. But probably what's most important at all when it comes to change management is standardized configurations globally throughout the entire infrastructure. If all of the systems utilize the same standard configurations, it makes it very easy to look at a change management request, see what's being proposed, and determine if there any downstream ramifications, any downstream impact. Whereas if each individual system has a different design or configuration, it takes a lot longer to go through and determine what systems will be impacted when an upstream device or router, for example, is upgraded.

Voice Quality

Voice quality. Clearly, this is very important when it comes to IP Telephony best practices. So right off the bat, Voice over IP audit to identify network readiness. Is your network ready to support Voice over IP? In many cases, if you haven't run voice over IP in the past, it probably isn't. It's going to require a QoS redesign. Need to make sure that we have dedicated priority classification for voice traffic.

Not just for voice traffic, however, we also need to prioritize our

signaling traffic. We need to look at our video traffic and then any other important traffic to support either the unified communications environment, but also the business itself. So there could be other business applications that need to have priority that are going to need to compete with call signaling, for example.

And something that comes to mind, a recent example that I've actually seen, is CTIOS, when it comes to our contact centers. It's necessary to go ahead and tag and prioritize the CTIOS traffic as well because it is part of a signaling and not necessarily something that should just be left to best-effort.

Something else to consider is auxiliary VLANs for our voice traffic. The goal is to logically separate our voice traffic from our data traffic. And there are a couple of different benefits. The common ones that you see, as I mentioned, would be separating voice from video. Another one would be to conserve your non-RSC 1918 address space so you can assign private addressing for all of your phones and save your public address for your actual hosts that need to be accessible remotely.

But also, the other advantage in placing your voice traffic in a separate VLAN from data is that it makes your QoS classification very easy because all of your voice traffic is on one segment, or in one VLAN. So those types of classifications become that much easier.

And then lastly, when it comes to voice quality, what codec selection do you have within your organization? There are several different codecs that Cisco supports, and really, the key is to find a codec that works within your organization based on the amount of bandwidth have in order to support that voice traffic, but also provides a minimum voice quality that you need.

Every organization is different. It's not uncommon to have multiple codecs running within a single organization, but the key is to have standardization. For example, a single codec that runs across all of your WAN links simplifies the call and admission control process because in your Gatekeeper and Call Manager, you can allocate a certain amount of dedicated bandwidth per call and not have to factor in variance where some calls might be one codec and some might be a different codec.

But one thing to consider is you do need to test out your codec before you actually put it into production. Just don't simply rely on a group of engineers to decide what codec is best. Put it in a lab and let your business users run through and test that codec to verify that it's going to meet their requirements.

Ian Pudney: Question, Kevin. What codec has Cisco IT implemented within their environment, and how did we arrive at that decision?

Kevin O'Healy: So a couple of different things within Cisco. Historically, on our low-link WAN sites, small field sales offices, what we'd utilize is G.729. And then in our larger sites where we have additional bandwidth, we would utilize G.711. But one thing within Cisco that we've always done is G.711 for our contact centers. And the reason being is that our contact centers are public facing, and we want to make sure that we have the highest quality voice for our public-facing interactions.

In the future, we'll be moving to G.722, which still provides a very high voice quality. It's a wideband codec, but has a variable bandwidth, so it can actually adjust based on how much bandwidth we have available to us.

Ian Pudney: So we should be getting the same amount of voice quality but for a lower bandwidth overhead.

Kevin O'Healy: Exactly.

Ian Pudney: Okay, good.

Kevin O'Healy: Continuing on with voice quality, a couple of specific things to look at when it comes to QoS. No. 1 is low latency queuing applied at the edge. We want to classify voice traffic as close to the edge as possible and then ensure that that classification is carried all the way through the network end to end.

Voice Quality (cont…)

So one way of doing this is to trust the IP phone traffic, which is tagged by default, and just need to configure the Catalyst switches in order to trust that phone traffic. And then, ideally, rewrite any other traffic that is sourced from a PC, for example, to another classification, perhaps best-effort, ToS value of zero.

One that we found is that consistency is key, standardized QoS policy throughout the entire organization. So if every site has the same basic QoS policies applied, in so far as how you classify traffic for voice, video, signaling, as well as best-effort applications, it makes design, as well as implementation, that much easier.

The advantage is an engineer in the field who's deploying a new site or performing a WAN upgrade, doesn't need to redesign the QoS model. It simply needs to look at a QoS cookbook to determine what actually policies should be applied to that site.

And then also, ensure consistent call and admission control bandwidth configurations. So the important takeaway here is if there's a WAN upgrade performed by the network team, ensure that the voice team is aware of that, and then can update the call and mission bandwidth on Gatekeeper as well as in Call Manager.

So it's important that both teams are walking the same path, aware of what each other are doing, and this ties back into the previous point with change management, having visibility from groups into network infrastructure changes.

One little tricky item I'll mention here, and that is that Call Manager, when it looks at how much bandwidth is utilized for a G.711 call or a G.729 call, for example, is different from how much bandwidth Gatekeeper assigns to that particular call. So in some cases, it is possible – required to go through and do some math in order to truly make sure that your call and mission control policy is in place end-to-end and matches on both the Call Manager as well as in Gatekeeper.

Security

Next, let turn our attention to security. So just as you would with any other application or file server on your network, it's important to have an antivirus software installed as well as have automatic updates for all of the virus definition files. As part of that, however, it's also important to go through and ensure that you have reporting on what the actually virus definition file is. So it's important to know not that you just have antivirus installation, that it's running, but that it has a current definition file that has been updated with the most recent iteration or most recent update of that file.

Install Cisco Security Agent on all of your Call Manager or Communications Manager servers. One thing I'll mention on this point is read the documentation whenever you perform a Call Manager upgrade. It may be required to actually uninstall or disable CSA before

doing that upgrade and then, afterwards, to either re-enable or, in some cases, reinstall CSA after you perform the Call Manager patcher upgrade. So do be sure to read the find print in the documentation.

Then lastly, make your life easy. Streamline the deployment patch in the systems. And this starts with a regular window for patching. Typically, one week every quarter might be designated for patches and upgrades. But also, within your organization, standardized hardware and operating system versions.

And the real advantage is if a new security vulnerability is announced, you don't need to spend time tracking every system you have within your organization to determine what OS is installed on what hardware platform. Instead, you simply need to look at which systems or which platforms are vulnerable and then match up which systems have that vulnerability. And if you have standards in place, it makes it easier. You don't have to touch every system, just those that are truly vulnerable.

Security (cont…)

Continuing on with security, tight control on external network access, but also on the internal network control access. So many companies I look at and work with, they will implement a firewall or an ACL in front of their Call Manager servers. Many organizations will also block any traffic from the data VLAN to the voice VLAN. This works well.

So for example, a user on the data VLAN might not be able to access the web of an IP phone, which can perhaps be a security concern to some organizations. But also, it means that some holes in the firewall or ACL will need to be open to support the technical track, which might be maintaining the individual Call Manager servers so that their stations, based on IP address, for example, are authenticated and can access the voice VLAN.

Of course, strong authentication for remote users and VP end users, but also frequent password resets. So whether that means that you have a standardized password per cluster, per region, or perhaps even one for the globe, the important takeaway is to have the passwords reset on a regular basis. A typical policy would be every 90 days.

Ian Pudney: Question, Kevin. These best practices seem very common to all IT services, right. These are not necessarily specific to IP Telephony.

Kevin O'Healy: It's going to be a combination. So the ones that you mention are – you're absolutely right. These are going to apply to any web or file sharing application, and it's business critical, but focusing on the voice specific security concerns. Things like your IP phone that is an access point into the network. A user can plug their PC into the back of a phone and gain access to the network, potentially pulling an IP address and access the internal internet.

So it's important when you have publicly facing phones. For example, in a lobby, you want to make sure that the PC port is disabled so a user can't come in and plug their PC in and use that as a point of access.

But going one step further, what if that same malicious user were to walk into a lobby where we have a publicly facing phone, unplug the phone, and then plug their PC directly into that network jack. In this case, we'd want to make sure that we have that network jack configured as a device on a dead net, so even if they did pull an IP address, they can't off of that subnet, there's nothing they can access.

And just one more item I'll mention on that topic is, once again, a lobby phone, we don't want somebody to be able to come in off the street,

access the directories on a public phone, and see the phone number for our executives. So we do want to make sure that we change our service URLs and directory URLs on all of our public access phones.

Ian Pudney: Right. Interesting.

Monitoring

Kevin O'Healy: Next, we have a discussion of monitoring. In the past, monitoring really consisted of device monitoring where we would simply look at an individual device and is it up or is it down. And one way of doing that is simply, for example, doing an ICM ping. Is the box up? If yes, then it's assumed that the device is available. If there's no ping response, then it's assumed that the device is down. The downside of this approach is that it doesn't look at the actual service itself. For example, just because a Call Manager server is responding to pings does not mean that it's doing call processing. The Call Manager service may be stopped, for example. So now, when we look at monitoring, it's important to look at an end-to-end service availability, not just individual device availability.

A couple of different things to consider, so for example, we can utilize SNMP to pull out key statistics on the Call Manager servers and voice gateways. In the case of a server CPU and memory utilization, the key advantage of many SNMP management agents is that it can provide historical tracking or trending. So we can go back in time and see if we are, for example, utilizing more virtual memory today than we were a month ago in our system.

When it comes to our voice gateways, our voicemail system, our IVRs, all of these have designated ports and we also need to look at what our utilization is. So we're not monitoring just the Call Manager servers, but also of the individual services that are tied to Call Manager.

One key thing is to establish thresholds. And for every system, what that the threshold is is going to be different, but the key is to have consistent monitoring variables, or always monitoring the same key components.

So in the case of a Call Manager server, for example, I would suggest monitoring the Cisco Call Manager heartbeat, an indication of whether or not the Call Manager service is running. I would also choose to monitor how many phones we have registered to a particular Call Manager server and how many phones you have registered a secondary Call Manager server.

And to just make sure that we have standardized monitoring in place for all of our services on all of your servers.

Ian Pudney: So Kevin, question. What are some of the instances that we've had Cisco where we've used our monitoring setup?

Kevin O'Healy: We actually use it on a regular basis. So for example, one type of thing that our monitoring keys us into is if we start to see phones that register on our secondary servers, it indicates we might have a problem on our primary servers. So for example, we will use is either a real-time monitor or a Microsoft Performance Monitor on some of the previous versions of Call Manager and will track the number of registered devices, whether it's a phone or a gateway, on it's of our servers. So in this case, on a primary server, we know how many devices we should have registered. And if we drop below that number, we will generate an alert.

And then on our secondary server, we would expect to have no devices registered. And if we see devices start to register to the secondary, we

generate an alert ‘cause in this case we know they are losing devices from the primary.

Support

Next, let’s talk about support as it relates to unified communications best practices. So when it comes to our operating system, what I want to talk about very briefly is easy ways today to upgrade or patch our system.

I remember back several years ago, four or five years, for example, where all of our patches and upgrades had to be performed from the console in the data center. And in many cases, an upgrade might take us several hours, the better part of a day. Whereas today, with the faster CPUs, the better hardware, as well as the remote administration tools, whether that’s VNC or the RILO, to be honest, I perform most of my upgrades from home rather than in the cold data center at the actual location where we have the servers.

And then when it comes to our application patches and upgrades, in the past, we’ve always pulled the redundant hard drives on our Call Managers prior to doing an actual upgrade. And the logic was in the event that we started an upgrade and for whatever reason we had to back out of that upgrade, we didn’t want to have to restore that system from scratch by rebuilding the operating system, reinstalling, in this case, the Call Manager application, and then re-synchronizing the databases.

So instead, what we would do on previous versions of Call Manager is actually prior to the upgrade, pull one of the redundant drives, insert a new drive and allow mirroring to complete, and then we would perform the upgrade.

In the event that we did have to back out, we can simply shut down the server and reinsert the redundant drives that we had pulled prior to the upgrade process.

Then the last thing I’ll mention on this particular slide is minimizing impact during the upgrade. And there are lots of little things that you can learn over the years about how to do this.

No. 1 is to minimize the number of reboots associated with what any upgrade cycle. So for example, when we upgrade a Call Manager system, we would typically upgrade the publisher first, which will have some impact to the clients, but minimal. For example, see if any changes might be impacted, Extension Mobility, login and logout, during that reboot might be impacted. But no call processing impact the clients.

Next, we would upgrade the TFTP servers. Ideally, we would have redundant TFTP and, again, minimal impact by – minimal impact to the cluster when the TFTP servers are rebooted.

And then going back to the fact that we would have all of our phones and devices registered to the primary subscribers, we would actually upgrade the secondary subscribers next; bring them up to whatever the new version is.

And then when it comes time to upgrade the primaries, we would go in and stop the Call Manager service on the primary servers. This would force all of the phones and gateways to register to the secondary servers, and now we can reboot the primary subscriber as many times as we need to and the phones and gateways won’t attempt to reregister back to the primary each time cause an impact to the client.

And then when we're complete with the upgrade, we simply re-enable the Call Manager service on the primaries. All of the devices sail back from the secondary to the primary with a minimal number of resets for all of the phones and gateways in the process.

One thing I'll mention here is disable alerting during the upgrades. And the reason being is that we have correct monitoring in place, which means all of the phones and gateways are monitored. We're tracking how many are registered and to which server. That will definitely generate alerts during an upgrade process with all of the resets. So be sure to disable your alerting during the upgrade process.

Support (cont…)

When it comes to Day 2 support of your unified communications, there are a couple key things that we've learned over the years. No. 1 is user privileges. Not every user who's going to be supporting the unified communications infrastructure needs to have administration rights. And in the early days with Call Manager, that unfortunately is what we had to do was grant all of our users who were doing phone max, for example, administration privileges on the actual Call Manager application. Today, however, we use MLA, multilevel access, so that we can grant specific privileges to users based on what their actual job function is. And this extends beyond just the individual Call Manager servers, but also into the network infrastructure. So for example, using user mode versus privilege mode on our servers and routers.

One of our engineers, for example, might need to determine if a voice circuit is up or down, or if a WAN circuit is taking errors. They don't need privilege level to do this. Instead, they simply need to have user access. And so best practice is to restrict who has full administration access to all of your devices within the environment.

Next, documentation. Hopefully, you have standardized deployments as I've mentioned earlier in the discussion. But really, the key here is to document and keep that information in a centralized store where all of the support teams, design teams, and implementation teams can access that documentation.

Standardization is critical as one of the most important things that I can call out. And being able to reference the actual design is very helpful when troubleshooting an issue. And so generating a centralized storage place for all of that documentation, perhaps sorted by technology, is a good way to go.

And then tied to that is a list of frequently asked questions. And the reason being is that many of our users will need to open a case to do something relatively simple such as reset a password or identify what a dial-in number might be for voicemail or for a meeting place call.

Having a list of frequently asked questions published on an internal website can make it easier for our users to actually solve their own questions without needing to open up a case, which ties up their own time as well as our support team's time as well.

And then lastly, what I'll mention is the support case management. Often times, what we've seen is users will try and bypass the actual process of escalating cases. In this case, perhaps going from a Tier 1 to a Tier 3 team bypassing Tier 2.

And while this may be effective in resolving the their underlying issue, the downside is we lose the visibility into actual trends, and that's why it's important for us to be able to enforce all of the escalation procedures, typically a Tier 1, Tier 2, Tier 3 support model. So we're gonna start aggregate the case that are similar and look for trends

throughout the process.

Unified Communications Deployment at Cisco: Best Practices in Action

So next, what we're going to do is take a look at a real world example of unified communications best practices as we've deployed them when in Cisco by looking at our global Cisco deployment, as well as one of our specific clusters, in this case, in San Jose.

Best Practices Deployment at Cisco

To provide a little bit of context, the global deployment of unified communications to support Cisco is, today, over 100,000 IP Telephony endpoints. These include both our Cisco hard phones, typically going to be the 796X and the 797X series, but also includes about 30 to 35,000 Cisco IP communicators, or soft phones, deployed on laptops.

Of those 100,000 phones, about 40,000 of those endpoints are deployed on a single Call Manager cluster in San Jose. So what I'm going to do in this slide, and the next couple of slides, is talk through exactly how we build our clusters using that cluster as the example. But everything I talk about is identical for all of our other clusters that we have throughout the entire Cisco organization.

So the first thing is device redundancy. In the case of San Jose, we have a 19-server cluster, which is a single publisher, two TFTPs, eight primary subscribers, and eight secondary subscribers.

All of our publisher and primary subscriber servers are located in one data center, and all of our secondary servers are located in another data center about a mile away. So going back to the start of discussion, we have spatial redundancy.

It goes a little further than that. We're also operating on two separate power gates. Our voice gateways for the environment are located in eight different building. We have over a hundred PRIs in this particular environment. As I mentioned, in voice gateways in eight different buildings, and the key here is that we're tied to different central offices depending on where we are in campus.

So even if we have an outage in one of our central offices, we can still route outbound calls, for example, through an alternate central office on the other side of campus.

One thing I'll also mention is that we do have multiple paths in and out of all of our clusters. For this particular cluster, for example, all of our calling between Cisco sites is done on net. So in this case, our first preference for routing calls from on Cisco office to another is going to be over the Cisco WAN utilizing Gatekeeper to perform a dial plan resolution.

In the event that Gatekeeper cannot complete the call or get a reject on the mission request, then what we have in place with our route list and route group constructs is to reroute that call through the PSTN, in this case, using long distance circuits.

Best Practices Deployment at Cisco: Cisco UCM Configuration

When we actually go through the process of building out a new Call Manager server, one thing we do is we utilize standardized naming. And the advantage here is by looking at a server's name, we should be able to determine what type of server is that. Is that a Cisco Call Manager server? Is it an IVR server? What location is it in? For example, SJC represents San Jose. RTP, Raleigh, North Carolina.

And then also, that same naming construct enables us to determine what cluster and potentially what function that server performs within that cluster. And so by simply looking at a server name, we can very easily determine the type of server, the location, and its function.

When we build out a server, one thing we do is remap the drive letters according to our own internal standards. The advantage here is all of our operating system files and folders are going to be on one partition, where all of our trace files are going to be on another partition. We'll even go one step further, and that is dedicate other driver arrays specifically for Call Manager tracers.

And the advantage here is our trace file write procedures are not competing with the actual operating system write procedures when it comes to the hard disk. We're operating on different driver arrays, different physical drives as well.

When it comes to the antivirus software typical installation, I've already mentioned how that works with the automated virus definition files. And then I'll add to that is we do report on what those virus definition files are and if a system falls out of date.

One thing that we've noticed internally that is very important is the network configuration. We do hard set all of our Call Manager server, Unity server, IVR server, speed and duplex for the network interface cards, and we also do that the switches. And while the servers and the switches usually sync out correctly on speed, sometimes we do see a mismatch on the duplex. So by going through a hard set in both sides, documenting what each port is, we know exactly what it is. We know exactly that it's hard set and that they are matching. We don't have speed or duplex mismatches.

Ian Pudney: Kevin, a question. What are some of the symptoms that you might see in your network if you do have an issue with your network configuration, a duplex or a speed mismatch?

Kevin O'Healy: Sure. Typical places where you might see that is if you're doing a new installation and you have a speed or duplex mismatch and you're trying to replicate data from one server to another where the entire database has to be replicated. What you would see in that case is just a very long time for actually completing that database replication. In many cases, can actually set you back hours on the upgrade or the installation process.

But also, when it comes to copying out trace file servers, everything that touches that system is going to be operating at a slower response time simply because many packets are being dropped simply because of collisions.

Ian Pudney: Right. So what about a normal operation of phones?

Kevin O'Healy: Mmm hmm. The normal operation of a phone is we do allow those to do an auto-negotiate, in this case, where they communicate directly with the switch and auto-negotiate speed and duplex. It's really going to be on the servers where we have a need for high throughput where we really need to focus on it.

Ian Pudney: I see. I see. Thanks.

Kevin O'Healy: One final point is time configuration. It's important to have a time reference installed, and whether that's Network Time Protocol or a Windows time service, one thing I will mention is I've seen many servers where we actually end up with both services running simultaneously and that can actually skew the time on the box.

So it is important to verify that you have one or the other configured, but not both configured at the same time. Also, make sure that all of the servers within your cluster are referencing the same time source.

Best Practices Deployment at Cisco: Remote Offices

So Cisco not only operates a couple of large campus sites such as San Jose with 40,000 phones, but we also have voice and data services to approximately 300 other offices worldwide. We have approximately 20 Cisco Unified Communication Manager clusters located throughout the globe, and we place these in 12 regional hubs.

And really, what I want to point out here is consolidation. What we're doing is placing all of our critical components, whether it's a Call Manager server, a Unity, an IVR, active directory, even exchange, we're placing those in the same location. And the benefit is we're now staffing our IT resources in those locations instead of having to disperse those same IT resources throughout the globe.

One thing to consider is a cluster over the WAN solution, and this does provide spatial redundancy. For example, the ability to place your Call Manager servers in different cities throughout the globe and have all of those servers participate in the same Call Manager cluster. And the real advantage to this solution is that you can spread your geographic reach of a cluster. So for example, it's possible to span multiple countries depending on the physical proximity of those servers.

And the takeaway for the end client is that the services are going to work throughout that entire region of the world. So for example, Extension Mobility, anywhere that the user attempts to log in with Extension Mobility, that it's covered by that particular cluster, that service is going to function for them.

Implementing Cisco SRST for the remote sites provides centralized dial plan administration at the central headquarters, for example, the regional hub where you might have the Call Manager cluster, but still provides a fallback mechanism in the event of a Call Manager outage or a WAN outage.

But there are a couple of key things when working with SRST that need to be tested. And it's important to test the dial plan for all of the remote sites using both normal operation where the devices are registered to Call Manager, but also in SRST mode, where all of the phones are registered to the local ISR Gateway. Verify the dial plan, both inbound and outbound. This also needs to include voicemail rerouting.

For example, in the event of a WAN outage, if all of your voicemail servers are located in a regional hub, do you have rerouting through the PSTN for that voicemail to function.

And then don't forget about your emergency services dialing. That should be tested under both normal operation as well as SRST.

A couple of additional items when looking at the dial plan. Depending on what region of the world you're working in, you may have a fixed link dial plan, or even a variable link dial plan. Even in the states where we have a fixed link dial plan, what we've seen is that many of our sites require seven-digit dialing for local calls, whereas other sites, also within the United States, would require ten-digit dialing for local calls.

So again, it's important to actually test all of the different call patterns to verify that calls are completing according to the way that the local staff expects them to complete.

Ian Pudney: So Kevin, tell me about how the AAR feature is used within Cisco.

Kevin O'Healy: AAR is automated alternate routing, and it works on a centralized call processing cluster to reroute calls through the PSTN in the event that there is enough bandwidth throughout that call line net.

And so a couple of key things to consider when it comes to configuring AAR. No. 1 is that it's based on the E.164 mask of the destination device. And so what you do need to make sure of is all of the devices have DIDs in the event to reroute that call through the PSTN. So unfortunately, AAR is not going to work if you are utilizing non-DIDs.

Second piece of that is make sure that the E.164 mask actually does match the true E.164 number for that particular end station.

And then from there, just make sure that all of the AAR group configurations are correct as well as number prefixing, especially if you're dialing, for example, from one country to another country using AAR.

So it's really an advantage. In the event of an out-of-bandwidth condition, gives the user one more shot at being able to complete that call over the PSTN without that user needed to redial using a bypass number or a true E.164 number.

Ian Pudney: Right. So it's an advantage to the user to provide a seamless calling experience.

Kevin O'Healy: And the nice thing is that the user rarely knows. They may see a number display change on their phone, but there's no actual function that the user needs to perform. It happens in the background automatically.

Ian Pudney: Right, okay. I understand.

Best Practices Deployment at Cisco: Cisco UCM Configuration (cont…)

Kevin O'Healy: Continuing on, we talk about best practices, specifically when we're looking at our Call Manager servers, the trace file configurations, and then also how we monitor our Call Manager servers.

What we've learned over the past few years when it comes to Call Manager trace files is that they're invaluable from an operational perspective, looking at a potential outage or a fault, being able to provided those Call Manager tracers to the business unit or to TAC. So one thing we do on all of our internal Call Manager servers is we configure the tracers to be written to a dedicated drive space.

Again, the benefit here is a dedicated physical disk as well as a dedicated logical disk, separate from the operating system, is reading and writing from.

By default, the Call Manager servers will record traces, but the reality is they're usually not at a detailed enough level to really be useful. So often times, what would happen is we would need to pull Call Manager tracers to examine a particular incident, but the tracers wouldn't be detailed enough to actually identify the true cause of the problem. So we have to change the level of tracing and then wait for a reoccurrence of that particular issue.

So instead, best practices actually to set your Call Manager trace level to either detailed or arbitrary, so you're capturing much more data. So it does require more space, but in this case, you don't need to wait for a reoccurrence of the issue to actually pull traces and see what it is.

Ian Pudney: So Kevin, is there an impact to the Call Manager system by running

these high levels of tracing?

Kevin O'Healy: There absolutely is an impact, and if you read the documentation, it's going to mention that as a particular caveat that you need to be aware of.

Best Practices Deployment at Cisco: Service Monitoring

So one thing to consider is monitoring the system, once again, ties back into SNP monitoring, memory, CPU, utilization, but also your disk IO, what's the actual write-time back and forth between the disk, are you doing paging files, for example, where it's actually writing the memory.

Ian Pudney: So if you have a really busy system, you maybe don't want to run detailed tracing that might push it over the edge.

Kevin O'Healy: You know, it's entirely possible. For what it's worth, I haven't seen that occur. Even on our large cluster of 40,000 phones, I haven't seen the situation where detailed tracing has actually impacted our system.

But I will point out what we're doing today, especially with a new appliance model for our Call Manager, is we're designating specific servers as trace servers. These are going to be dedicated Windows operating system servers, and all they do is collect Call Manager traces from all of the Call Manager servers. We utilize a real-time monitor to grab all of those Call Manager traces, and now they're on one dedicated server. We're not storing large amounts of trace file data on our actual Call Manager servers.

Ian Pudney: Right. So basically, as long as you're following all of the best practices, you're stepping out here as far as monitoring the system, making sure that you're monitoring the thresholds. Then it's okay to run the detailed tracing because you should have your system running at basically an optimal level of performance.

Kevin O'Healy: Exactly.

Ian Pudney: Okay.

Kevin O'Healy: Really, our goal is to be able to restore service as quickly as possible in the event of an outage and not have to wait for a reoccurrence of a particular issue.

Ian Pudney: Yeah, okay. I see.

Kevin O'Healy: So one final point I will mention just to make our lives that much easier, is that by default, all of the Call Manager servers utilize the same naming convention for trace files. So if you're looking at a trace file from three or four servers in the same cluster, it's very difficult just by looking at the trace file name to determine which server wrote that trace file. So going into the Call Manager trace configuration, you can prefix an actual value, for example, the server name that's going to be written as part of the Call Manager trace file name.

And now, when looking at the same three or four files from different systems, you can very easily determine which server wrote that particular trace without having to open the file and look at the details.

Ian Pudney: Right. That's a great tip. I didn't know that.

Kevin O'Healy: It works out really well. And then the second piece is configuring monitoring alerts. One way in the past was to use the Microsoft performance monitor. The Cisco solution is the real-time monitor. And this works with the previous versions of Cisco Call Manager as well as with the new versions of Cisco Unified Communications Monitor.

There are several different variables that you can look at. I've highlighted a few, for example, drive space, registered devices, as well as the heartbeat. Heartbeat, as I mentioned previously, is an indication of whether or not that service, Call Manager, for example, is running.

One thing I would recommend is separating your alerts into two different classifications: minor alerts versus a major alert. A minor alert is something that's informational. You need to be aware of it, but it's not something that you need to respond to right now. In fact, if it were to occur after hours, you probably don't want to have a page on that type of alert. And this might be a situation where a particular gateway has unregistered. As long as you have enough other gateways to cover that loss, it's okay to wait 'til the following morning to look into that issue.

Major alerts are gonna be just the opposite. This is going to be something where it's critical. You need to be notified right now that there's an issue within your system. So for example, if you were to lose perhaps 10% of your gateways, maybe that is the threshold that you set, at which point you should start generating pages, even overnight, to all of your support staff personnel.

The key takeaway on this is to have minor alerts and major alerts, but to have consistent monitoring as well. All of the systems should have the same objects and counters monitored, and all that changes is the individual threshold based on how heavily utilized that particular server is. But all of the base objects and counters should be the same throughout all of the servers in the infrastructure.

Continuing on with service monitoring, one of the goals is end user experience. What is it truly like to be an end user in your environment when utilizing the unified communications applications? One way of replicating – emulating a user is automated synthetic testing.

Lots of different things play into this. For example, the delay to dial tone. In automated systems test that generates an off-hook message to Call Manager and record with the delay to dial tone is, how long does it take to receive the message back from Call Manager that that device should play an off-hook dial tone message. That can be an indication of how busy your Call Manager server is.

If your Call Manager server is overloaded, there will be a delay to processing that dial tone request. It can also be an indication of congestion of the network. Hopefully, you have prioritization for of your Call Manager and phone signaling. So hopefully, it's not getting delayed in transit from one device to that other.

This is required for all of the unified communications services within the infrastructure, and so this can also extend to, example, voicemail. The same automated test can dial into your voicemail system and verify that you receive a "Welcome to Cisco Unity" message. It can be used to dial into an IVR that might be used for a meeting place just to verify that a user does a LAN on that IVR and then perhaps even do some digit input to see if we can translate or route that call according to the logic on that IVR.

Another thing to consider, especially when it comes to voice quality, is what is the MOS that you had running throughout your organization? What level of quality do you expect within your organization? And are you achieving that quality?

Within Cisco today, we are looking at MOS scores. What we utilize is the Cisco Unified Operations Manager and Cisco Unified Service Manager, and the 1040 sensors that we place in front of our Unity Systems as well as in front of our meeting place IVRs, and we're

capturing real traffic from our end users' phones back to those systems, voicemail and meeting place. And specifically, what we're looking at is voice quality. What is the MOS score associated with that call, and does it meet a minimum threshold? Today, that minimum threshold within in Cisco is 3.0. Anything below that is something that we look into. In the future, we'll likely increase that threshold to a higher number so that we're looking at a greater number of calls. Once again, our goal here is to provide the highest quality voice conversation.

Prioritized Recommendations

And finally, what I'm going to do is conclude with a list of five specific prioritized recommendations, and these are based on what I've seen over the past several years working in an enterprise unified communications deployment as the top items I see that are required in order to be successful.

Prioritized Recommendations (cont…)

No. 1 is change management. Having a single change management review board that has visibility from the voice teams, the data teams, as well as the all of the affected application groups. And within Cisco, for example, those affected application groups might be a Cisco TAC. It might be our internal contact center teams. All of the different groups within Cisco have their own projects and plans in place, and we need to be aware when one organization's plan might be stuck in another organization's plan.

QoS. This arguably could be No. 1 on my list. It's important to have standardized QoS configurations, to audit these QoS configurations to make sure that they're in place end to end. Periodic testing with a packet capture utility is critical to verify QoS is working properly.

Management metrics. How do you know if you are successful, and as you said, a threshold or a benchmark to achieve? For example, that could be a MAS score. That could be system availability based on ICMP, SNMP, or also service availability.

Standardized configurations. All of the Call Manager servers, all of the WAN gateways, all of the catalyst switches should all utilize standardized configurations. The goal is to make troubleshooting as well as deployment that much easier by simply having all of the components that are interchangeable. For example, the same hardware platform, but also the same configurations on all devices.

And then lastly is documentation. Create that internal storage point for all of your documentation, the designs, how to operate and support the infrastructure, once again, to make the job of the support team that much easier, but also as a single point of reference for our users with a list of FAQs.

Q&A

Ian Pudney: Thanks, Kevin. Great. Great presentation. Lots of great information. One of the topics I'd just like to pick up a couple of questions on – I know there's a lot of interest around the clustering of the WAN solution. What is Cisco IT's use of clustering of the WAN?

Kevin O'Healy: Mmm hmm. So today, what we have for clustering of the WAN is a single cluster with servers located in London, Amsterdam, and Brussels. And the geographic reach of that cluster is most of Africa as well as all of Western Europe, and it extends into Russia as well.

The cluster has about 10,000 phones registered to, at any given time, one of those three locations. And the real advantage is if we were to

have a network failure that impacted one site, say London, for example, all of the devices and gateways that register to the London Call Manager servers would reregister automatically to other Amsterdam and Brussels, still providing that same look of unified communications. In this case, the primary servers are unavailable. They're utilizing a secondary server for call processing.

Ian Pudney: Right. So you get the benefit of additional resiliency as well as extending those features seamlessly across all those countries.

Kevin O'Healy: Absolutely. It works out really well. We have looked at clustering of the WAN in other parts of the world as well. It's not something we would rule out. I think as we do more large acquisitions, we're going to do more and more of cluster of the WAN deployments where we simply bring in one of our own servers and place that into, for example, one of our acquired site's networks.

Ian Pudney: Right. Another question. A lot of these best practices have been focused around the Unified Communications Manager platform. Are there similar practices across the other components of Cisco unified communications?

Kevin O'Healy: I think you'll find that most things we talked today will translate directly over to other unified communications technologies. So whether that's voicemail with Cisco Unity, MeetingPlace, all of the other additional features, we still want to monitor the individual devices.

We still want to look at end-to-end service components, service availability. But also, we want to do our monitoring just as we would with the Call Manager. In the case of voicemail, with Unity, generating alerts and classifying them either as a major alert or a minor alert. We look at things such as disk space. Just like with a Call Manager, we can find that it sometimes will exceed the amount of disk space where the trace the files need to go in and clean that up.

Further Resources

Ian Pudney: Thank you, Kevin, I'm afraid that's all the time we have for questions today.

For more information about technologies and solutions deployed at Cisco, you can go to the Cisco on Cisco site where you can find Case Studies with information about: what we deploy, what benefits we've gained, what lessons we've learned, and some operational practices and presentations to help you learn more.

Below that, you'll see a toll-free number you can call for more information or to place an order; and you can order Cisco resources on the web from the URL at the bottom of this page.

I'd like to thank you watching, and for being interested in what the Cisco on Cisco Technology Tutorials. We hope that you've enjoyed this seminar and that it has helped answer some of your questions about best practices for deploying, managing, and supporting Unified Communications solutions.

And thank you, Kevin, for spending your time today, and for your enthusiasm and expertise in unified communications.

Kevin O'Healy: It was a pleasure to be here, Ian.

Ian Pudney: Thank you again for watching, and good-bye.

Americas Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 527-0883

Asia Pacific Headquarters
Cisco Systems, Inc.
168 Robinson Road
#28-01 Capital Tower
Singapore 068912
www.cisco.com
Tel: +65 6317 7777
Fax: +65 6317 7799

Europe Headquarters
Cisco Systems International BV
Haarlerbergpark
Haarlerbergweg 13-19
1101 CH Amsterdam
The Netherlands
www-europe.cisco.com
Tel: +31 0 800 020 0791
Fax: +31 0 20 357 1100

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices.**

Printed in USA C78-331727-01 10/06

