CHAPTER 4

CMX Deployment Models

September 4, 2014

This chapter introduces high-level models for the deployment of infrastructure components necessary for location services and CMX. Considerations around bandwidth utilization and scalability of the MSE are discussed. Finally the high-level models are mapped to campus and branch designs showing physical infrastructure designs for supporting CMX services as well as guest access for CMX Visitor Connect.

Overview

This section of the Cisco CMX CVD is targeted for IT personnel who are looking at deploying CMX services over private Enterprise network infrastructures. The common characteristic of Enterprise networks are that they are deployed for private use by a single business entity.

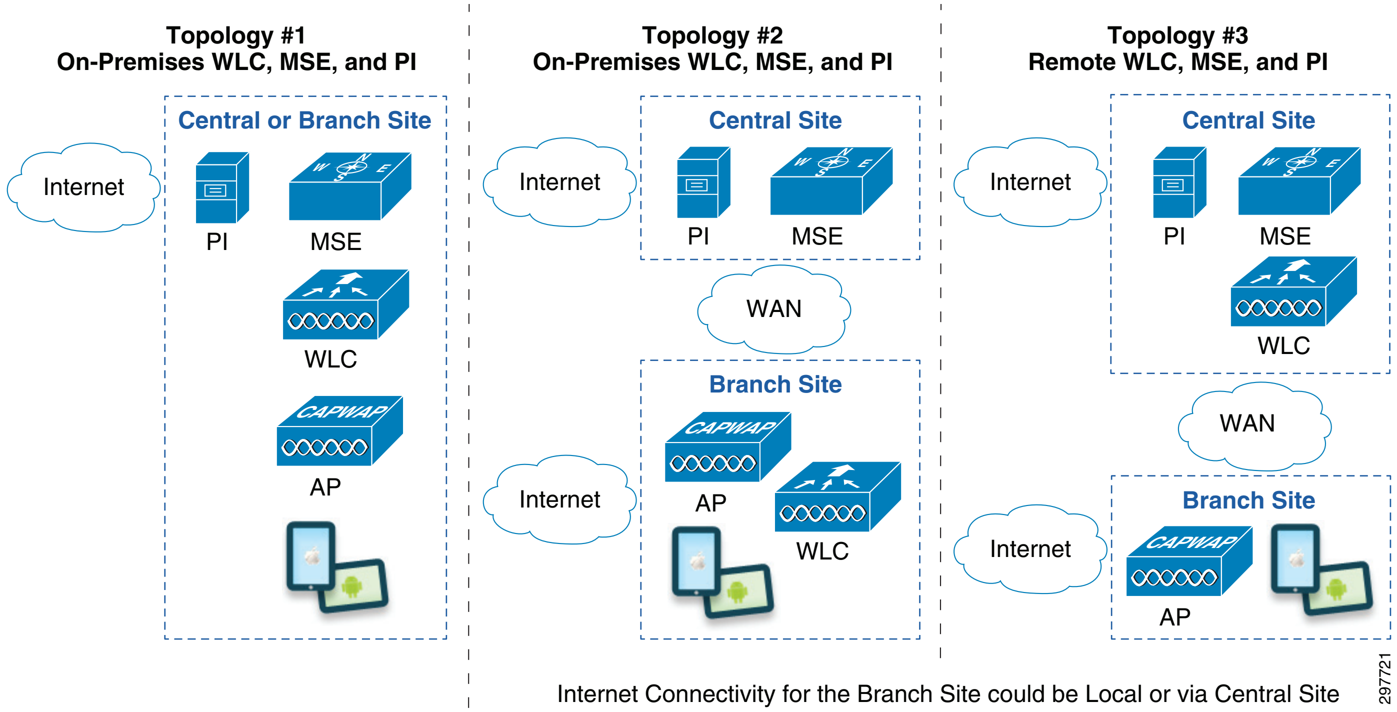
This version of the CMX CVD focuses primarily on meeting the CMX Location Analytics use cases, CMX Presence Analytics use case, and the CMX Visitor Connect use case, which is discussed in Chapter 7, "CMX Use Case Stories." Each of these use cases is deployed over an enterprise wireless network infrastructure which is designed to support location services.

Note This version of the CMX design guide does not discuss deployment of CMX or location services over a Service Provider network infrastructure.

Deployment Topologies

Figure 4-1 shows three basic topologies for deployment of the components discussed in Chapter 3, "CMX Solution Components" in a network infrastructure.

Figure 4-1 High-Level Deployment Topologies



These topologies are based on the physical position of the Mobility Services Engines (MSEs) and Cisco Prime Infrastructure (PI) in relation to the wireless LAN controllers (WLCs) and access points (APs). The location of the APs represents the site where CAS (location services) and CMX services are required.

In Topology #1, all wireless infrastructure components—APs, WLCs, MSEs, and PI—are deployed on-premises at the site. In Topology #2, the WLCs and APs are deployed on-premises at a location such as a branch site, while the MSEs and PI are deployed remotely at a central site. In Topology #3 only the APs are deployed on-premises at a location such as a branch site, while most of the wireless infrastructure—consisting of WLCs, MSEs, and PI—are deployed centrally. Note that the network administrator may not necessarily have a choice in terms of the positioning of the WLCs in relation to the APs. In other words, the infrastructure may be an existing WLAN deployment which now requires CAS (location services) and/or CMX Services.

The benefits and disadvantages of each topology from a CAS and CMX perspective are summarized in Table 4-1.

Table 4-1 Benefits and Disadvantages of Deployment Topologies

Topology	Benefits	Disadvantages
#1: On-Premises WLCs, MSEs, and PI	• No additional WAN bandwidth requirements for transporting client RSSI information from APs to WLCs, and from WLCs to MSEs for processing location information. • Potentially increased scale of the overall CMX deployment since each site has one or more MSEs.	• Potentially increased cost of having to deploy and maintain WLCs and MSEs at each site. • No ability to view location and analytics data across multiple sites from a central set of MSEs.

Table 4-1 Benefits and Disadvantages of Deployment Topologies

#2: On-Premises WLCs, Remote MSEs and PI	• Potentially reduced cost of not having to deploy and maintain MSEs at each site. • The ability to view location and analytics data across multiple sites from a central set of MSEs.	• Potentially increased WAN bandwidth required to transport client RSSI information from the WLCs to a central set of MSEs for processing location information. • Potential scale limitations of the overall CMX deployment since all sites rely on a single set of MSEs.
#3: Remote WLCs, MSEs, and PI	• Potentially reduced cost of not having to deploy and maintain WLCs and MSEs at each site. • The ability to view location and analytics data across multiple sites from a central set of MSEs.	• Potentially increased WAN bandwidth required to transport client RSSI information from the APs to a central set of WLCs for processing location information. • Potential scale limitations of the overall CMX deployment since all sites rely on a single set of MSEs.

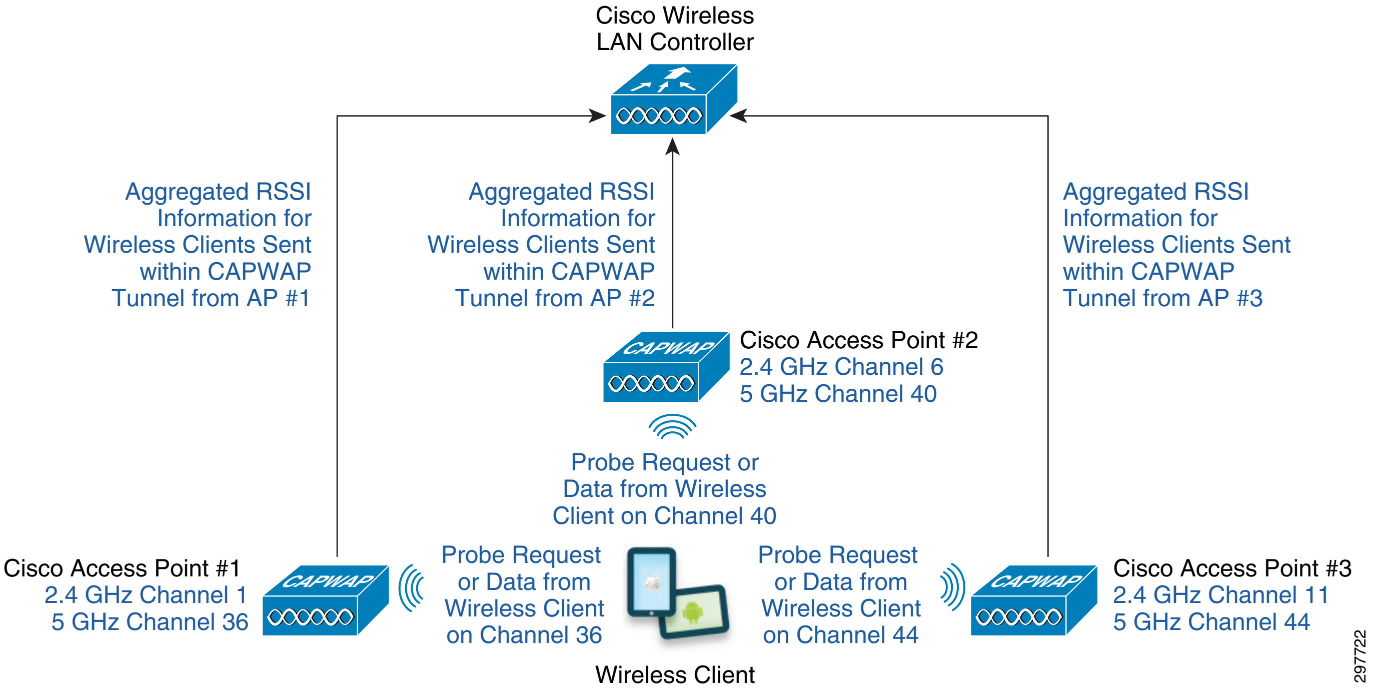
The following sections further discuss some of the considerations resulting from the deployment topologies.

WAN Bandwidth Utilization

For deployments with limited amounts of bandwidth between the different sites where the MSEs, WLCs, and APs are deployed, the network administrator must ensure that sufficient bandwidth is provisioned to meet the requirements for existing applications at the site as well as the requirements for CAS (location services) and CMX services.

For CAS (location services), bandwidth is required to transmit aggregated RSSI data for each mobile device within the CAPWAP tunnel from each AP to the WLC, as shown in Figure 4-2.

Figure 4-2 RSSI Information Sent within the CAPWAP Tunnel from APs to the WLC


Cisco Wireless LAN Controller
Aggregated RSSI Information for Wireless Clients Sent within CAPWAP Tunnel from AP #1
Aggregated RSSI Information for Wireless Clients Sent within CAPWAP Tunnel from AP #2
Aggregated RSSI Information for Wireless Clients Sent within CAPWAP Tunnel from AP #3
CAPWAP
Cisco Access Point #2
2.4 GHz Channel 6
5 GHz Channel 40
Probe Request or Data from Wireless Client on Channel 40
Cisco Access Point #1
2.4 GHz Channel 1
5 GHz Channel 36
CAPWAP
Probe Request or Data from Wireless Client on Channel 36
Probe Request or Data from Wireless Client on Channel 44
CAPWAP
Cisco Access Point #3
2.4 GHz Channel 11
5 GHz Channel 44
Wireless Client
297722

The amount of bandwidth required for transmitting RSSI data between the APs and WLC depends upon multiple factors, including:

- The number or mobile devices at the site. Each mobile device either periodically generate Probe Requests or generate traffic which the FastLocate feature uses to calculate RSSI information. Hence the more mobile devices at the site, generally the more Probe Requests or traffic seen by FastLocate for a given time interval and the more bandwidth needed to accommodate the RSSI information within the CAPWAP tunnel.
- The frequency of packets generated by each mobile device which are used for location determination:
  - If Probe Request RSSI is implemented for CAS, then only Probe Requests from the wireless device are used by the AP to collect and send RSSI information to the WLC within the CAPWAP tunnel. Although the frequency of Probe Requests can vary from under a second to five minutes, a realistic number often used for the frequency of Probe Requests from mobile devices is 30 seconds.
  - If the FastLocate feature is implemented for CAS, then all packets from the wireless device aare used by the WSM module within the AP to collect and send RSSI information to the WLC within the CAPWAP tunnel. With the FastLocate feature enabled, RSSI data can be collected from wireless clients as frequently as approximately every 4-6 seconds, depending upon whether CleanAir is implemented, the channels which the WSM scans, and whether a particular wireless client is transmitting when the WCM scans the particular channel on which the wireless client is operating. This could potentially increase the amount of RSSI traffic by 5-7 times over Probe Request RSSI. Alternatively, if wireless clients only periodically transmit packets such that FastLocate relies mostly on the BAR feature to update unresponsive clients, then there may be little to no increase in the amount of RSSI traffic over Probe Request RSSI.

- The number of APs deployed at the site which hear and report information on a given mobile device. If multiple APs hear either a Probe Request or data packets when using FastLocate, each AP collects RSSI information for the mobile device and send that information to the WLC. Therefore the more APs which hear the mobile device, the more bandwidth may be required to accommodate the RSSI information to the WLC within each of the CAPWAP tunnels.

The network administrator should note that Cisco APs automatically aggregate Probe Request information from mobile devices and forward them within CAPWAP messages to the WLC, regardless of whether or not an MSE is deployed within the infrastructure for CAS (location services) or CMX services. The aggregation timer for transmission of aggregated Probe Request information is, by default, set to 500 milliseconds.

Note Probe Requests could come in at a rate high enough to fill the buffer used to hold the aggregated data before the aggregation timer expires. This could be due to a WLAN deployment with lots of mobile devices, each probing infrequently, or a WLAN deployment with fewer mobile devices, each probing frequently. In either case when the buffer has filled, the aggregated Probe Request information is sent immediately and the aggregation timer reset.

With the FastLocate feature enabled, APs with WSM modules collect RSSI values based on all data packets for all mobile devices heard while the radio within the WSM dwells on a particular channel. The Data RSSI values are again aggregated and forwarded to the WLC within CAPWAP messages. The aggregation timer is, by default, set to 500 milliseconds.

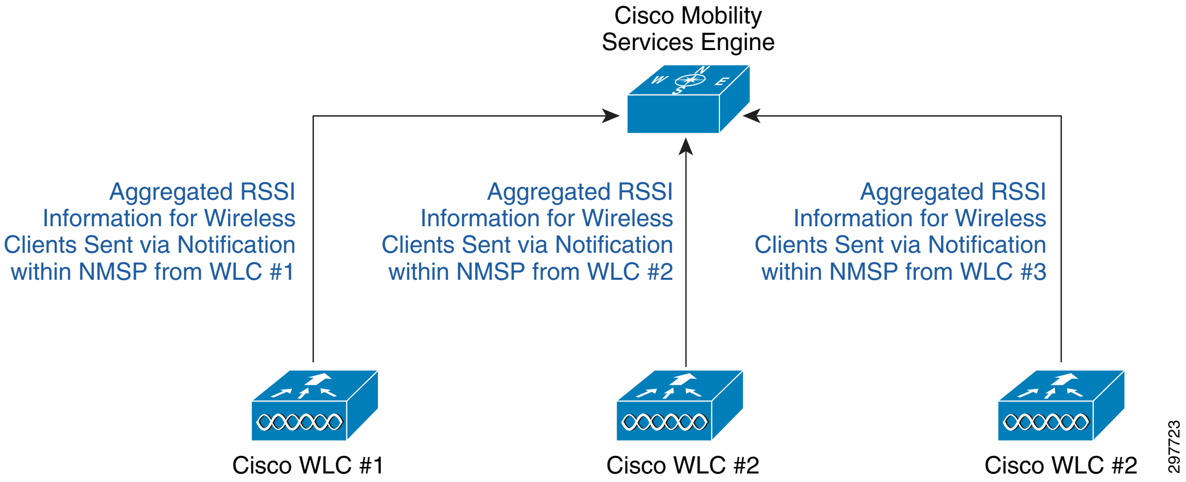
For Topologies #1 and #2 shown in Figure 4-1, since the WLCs and APs are both on-premises, it is typically assumed there is sufficient bandwidth on the LAN to accommodate the RSSI information. Often the APs are deployed in an overlay network (Local Mode) in which all wireless traffic is backhauled to the WLC before being terminated on the LAN. Hence the amount of traffic due to RSSI information within the CAPWAP tunnel is also typically small compared to the actual WLAN traffic encapsulated within the CAPWAP tunnel.

For Topology #3 however, since the WLCs are remote from the APs, the network administrator must ensure sufficient bandwidth on the WAN to accommodate the RSSI information. Topology #3 is indicative of a typical FlexConnect WLAN deployment within a small branch in which wireless traffic is often terminated locally at the AP.

Note The CAPWAP control channel between the APs and WLCs has multiple functions besides transmitting aggregated RSSI information. Hence additional bandwidth overhead is incurred when deploying a FlexConnect WLAN design. These details are not covered within this design guide.

For CAS (location services), bandwidth is also required to transmit aggregated RSSI information, within NMSP Measurement Notification messages, from each WLC to the MSE, as shown in Figure 4-3.

Figure 4-3 Aggregated RSSI Information Sent within the CAPWAP Tunnel from APs to the WLC



The amount of bandwidth required for transmitting RSSI data between the WLCs and the MSE depends on the same factors as discussed in the previous section since the RSSI information from the APs is aggregated by the WLCs before being sent to the MSE. The amount of required bandwidth is also dependent on the following additional factors:

- The number of WLCs deployed. The overall deployment may consist of multiple sites, each with its own WLC, or a single site with multiple WLCs. Each WLC transmits all NMSP messages (containing information for all of its APs) to all MSEs associated with it. RSSI Notification Messages sent within NMSP contain the aggregated RSSI information. These are sent every two seconds by default.
- The number of MSEs deployed. Again, each WLC sends all NMSP messages to all MSEs associated with it (regardless of maps). Hence if a given WLC has two MSEs associated with it, the amount of NMSP traffic is effectively doubled.

For Topologies #1 and #3 shown in Figure 4-1, since the WLCs and MSEs are both on-premises, it is typically assumed there is sufficient bandwidth on the LAN to accommodate the RSSI information.

For Topology #2 however, since the MSE is remote from the WLCs and APs, the network administrator must ensure sufficient bandwidth is provisioned within the WAN to accommodate the RSSI information.

Note The NMSP protocol between the WLC and MSE has multiple functions besides transmitting aggregated RSSI information, such as transmitting wIPS information. Hence additional WAN bandwidth overhead may be incurred when deploying a design in which the WLCs are remote from the MSEs. These details are not covered within this design guide.

For CMX services, additional WAN bandwidth may be required to support guest traffic if CMX Visitor Connect is deployed at the site and if the guest traffic is sent back to a central location (auto-anchored) before being sent to the Internet. The amount of WAN bandwidth required depends on the number of guests simultaneously supported at the site and also varies highly based on whether the guest is simply browsing a web site or downloading streaming video. Per user rate-limiting can be applied to the B2C guest WLAN at the WLC to limit the amount of bandwidth that each guest can consume to partially alleviate this issue.

Note The guest WLAN traffic can be sent to the Internet directly from the site. Future versions of the CMX design guide may address these designs.

Additional WAN bandwidth may be required to support push notifications and/or application specific information to the mobile device. This is required if the mobile device is again associated to the guest WLAN at the site and the client is configured with a CMX mobile application and if the guest traffic is auto-anchored back to the central location before being sent to the Internet. Future versions of the CMX design guide will discuss the Cisco CMX Mobile Application Server & CMX SDK.

It is assumed that end-users may also access location and presence analytics data both from within branches and from within campus locations. Hence bandwidth may also be utilized when end-users access analytics data from within a branch and the MSE is located within a centralized campus location.

MSE Scalability

The scalability of the Cisco MSE, in terms of licensing, is shown in Table 4-2. As of MSE software release 7.4 and above, licensing is based the number of APs supported.

Table 4-2 Mobility Services Engine Scalability

Platform	Location Services Licensing	Advanced Location Services Licensing	wIPS Licensing (Monitor Mode or Enhanced Local Mode)
Cisco 3355 MSE Appliance	Up to 2,500 APs	Up to 2,500 APs	Up to 5,000 APs
Cisco MSE Virtual Appliance (High-end Server)	Up to 5,000 APs	Up to 5,000 APs	Up to 10,000 APs
Cisco MSE Virtual Appliance (Standard Server	Up to 2,500 APs	Up to 2,500 APs	Up to 5,000 APs
Cisco MSE Virtual Appliance (Basic Server)	Up to 200 APs	Does not support Advanced Location (CMX) Services	Up to 2,000 APs

As of MSE release 7.5 and higher, there is no enforcement regarding the number of end-devices on the MSE. However there is a hard limit of 25,000 tracked end-devices on the Cisco 3355 MSE appliance and 50,000 tracked end-devices on the on the high-end virtual MSE server.

Since location calculations are CPU intensive, the MSE also scales based on the number of location calculations it has to perform. A high-end virtual MSE server can handle approximately 90,000 calculations per minute (approximately 1,500 calculations per second). A new location calculation is performed by the MSE for a given wireless client when the following conditions occur:

- New RSSI data is seen by the MSE for the wireless client, which can occur for the following reasons:
  - With Probe Request RSSI—The wireless client sent Probe Requests.

  - With the FastLocate feature enabled —The wireless client sent packets which were seen by WSM modules within the APs during the dwell time on the particular channel the wireless client was transmitting on.

- The signal strength is more than ~5 dBm from the last time RSSI information was received from the wireless client (i.e., the last time the wireless client was seen). Note that this is a configurable parameter. Information about setting this parameter is provided in Chapter 25, "Configuring the Mobility Services Engine for CMX."

In dense AP deployments, if more than five APs see the wireless client, the MSE only uses the top five (the strongest RSSI values) in calculating the X,Y coordinates of the wireless client. However the MSE expends additional CPU cycles sorting and discarding the excess data points.

A very conservative estimate of the number of devices which can be supported by a high-end virtual MSE server would be to assume that updated RSSI information is received from every wireless client every 5-6 seconds (~10 updates per minute). This corresponds to the example of using FastLocate with CleanAir discussed in Probe Request RSSI versus FastLocate in Chapter 3, "CMX Solution Components." Using a maximum 90,000 calculations per minute for a high-end virtual MSE, yields a maximum number of clients for the MSE as follows:

90,000 updates per minute/10 updates per wireless client per minute = 9,000 wireless clients

Hence a very conservative estimate of the number of devices which can be supported by a high-end virtual MSE server is 9,000 wireless devices seen by WLC and tracked by MSE.

As mentioned in Probe Request RSSI versus FastLocate in Chapter 3, "CMX Solution Components," this would require each wireless client to be transmitting packets during the time the WCMs dwell upon the channel which each wireless client is operating on during every scan cycle. Assuming the RSSI information from the 9,000 wireless clients is averaged out—meaning that for every second, approximately 1/6th of the 9,000 wireless clients are heard from—results in a movement percentage of approximately 16.67%.

A more realistic estimate of the number of devices which can be supported by a high-end virtual MSE server would be to assume that updated RSSI information is received from every wireless client every 10-30 seconds (~2-6 updates per minute). Again, using a maximum 90,000 calculations per minute for a high-end virtual MSE yields a maximum of clients for the MSE as follows:

90,000 updates per minute / 6 updates per wireless client per minute = 15,000 wireless clients

90,000 updates per minute / 2 updates per wireless client per minute = 45,000 wireless clients

Hence a more realistic estimate of the number of devices which can be supported by a high-end virtual MSE server is from 15,000-45,000 wireless devices seen by WLC and tracked by MSE. Again, assuming the movement of the wireless clients is averaged out, results in a movement percentage from 3.33% - 10%.

Note The example above highlights an alternative method of estimating scale of the MSE based on movement percentage of wireless devices. For example, for a given a movement percentage of 3%, an estimate of the number of wireless clients a large MSE virtual server can handle is approximately 50,000 wireless devices.

An additional technique which can help scale is to not overprovision cores on the virtual machine which hosts the MSE. In other words the number of virtual CPUs (vCPUs) should equal the number of physical cores. This gives the MSE more CPU time to calculate locations at high scale. Also, a hard disk which supports 1,600 input/output operations per second (IOPS) is recommended. The minimum of 250 IOPS can be used, but may result in slow updates to client location using the MSE API and sluggish response when accessing the MSE graphical user interface (GUI).

Further scaling of the MSE can be accomplished by splitting out services, which may be necessary for large scale WLCs which see more than 50,000 devices and when CMX Analytics and CAS are deployed. The following table shows possible ways of additional MSE scaling by splitting out services.

Table 4-3 MSE Scaling by Splitting Out Services

Number of MSEs	Services Running on Each MSE	Scale	Devices	Caveats
1	CAS Analytics wIPS Connect/Engage Mobile Concierge	Demo purposes—up to 1,000 devices	1,000	Recommended to run aWIPS on a separate MSE, since aWIPS requires UTC time zone. Otherwise Analytics information will have the incorrect time zone.
1	CAS Analytics Connect/Engage Mobile Concierge	Demo purposes—up to 1,000 devices without wIPS	1,000	No aWIPS. Good for small deployments or demonstrations.
2	1 CAS 1 Analytics	Highest scale for CAS to run on a separate box. Scale to 50,000 devices with 3% movement rate	10,000–50,000 per WLC/MSE	No aWIPS, CMX SDK, or guest services such as CMX Visitor Connect. MSE REST API and Analytics only.
4 + 1 CMX Mobile App Server	1 CAS 1 Analytics 1 wIPS 1 Connect/Engage & Mobile Concierge 1 CMX Mobile App Server	Highest overall scale by separating all services into individual chassis. Use this architecture when there are over 10,000 clients and you need SDK + Analytics + Visitor Connect	10,000–50,000 per WLC/MSE	Splitting out Analytics save ~ 15% CPU when extracting analytics data from raw db. Splitting out Connect/Engage save ~15% CPU for rendering Guest Login web pages under load. Adding CMX Mobile App Server decreases latency for location API calls.

Campus and Branch Designs

Applying the basic high-level topologies (as shown in Figure 4-1) to traditional enterprise campus and branch designs which may require location and CMX services results in the customer deployment models shown in Table 4-4.

T

Table 4-4 Campus and Branch Deployment Models

Deployment Model	Infrastructure	Environment
Single Campus (or Large Branch) Deployment Model	On-Premises APs, WLCs, MSEs, and PI	This type of deployment applies to single-site campus networks, such as a standalone hospital or a single-campus college/university. This type of deployment also applies to single-site large branch network, such as a large retail store.
Multiple Campuses (or Large Branches) Deployment Model	On-Premises APs and WLCs; Remote MSEs and PI	This type of deployment applies to networks that support multiple large campus sites, such as regional hospitals and universities with multiple campuses. This type of deployment also applies to networks that support multiple large-sized branches, such as large retail stores connected to a campus site.
Small Branch Deployment Model	On-premises APs; Remote WLCs, MSEs, and PI	This type of deployment applies to networks that support multiple small-sized remote branches, such as small retail stores or branches of financial institutions which connect to a campus site.

This version of the CMX design guide highlights two of the deployment models—the Single Campus (or Large Branch) Deployment Model and the Small Branch deployment model—in relation to the CMX Location Analytics, CMX Presence Analytics, and CMX Visitor Connect use cases.

Single Campus (or Large Branch) Deployment Model

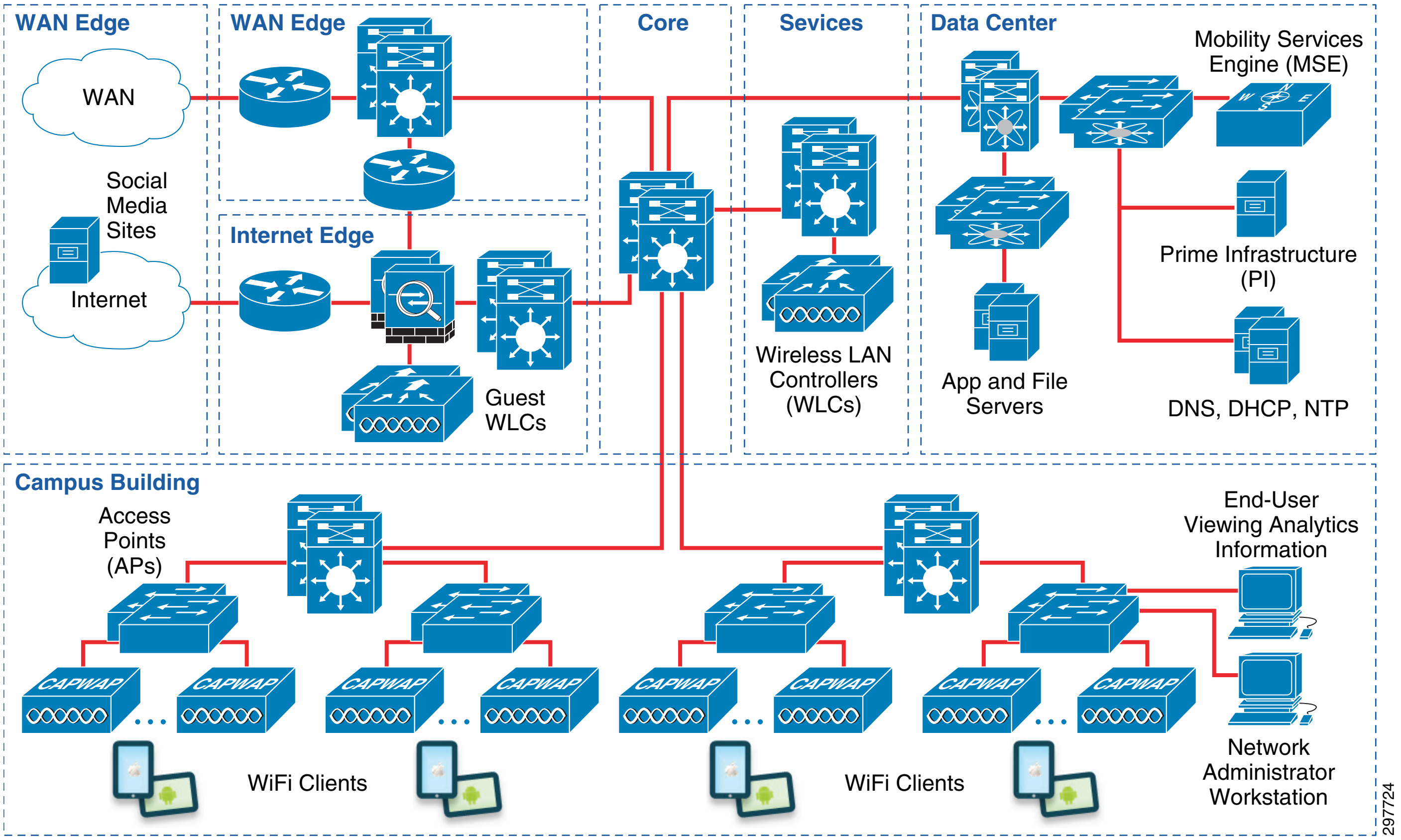
From the perspective of CAS and CMX services, the difference between a single campus and a single large branch is the really just the size of network infrastructure within the site. Campus networks tend to have multiple buildings, requiring a traditional three-tiered (core, distribution, access) switch infrastructure and a more modular design (Internet Edge Module, Services Module, Data Center Module, etc.) for separation of function and scalability. Large branch networks tend to be much simpler, requiring a two-tiered (distribution and access) switch infrastructure and often no modularity within the design. Both campus and large branch network have the APs, WLCs, MSEs, and PI in the same physical location relative to each other. Hence these have been combined into a Single Campus Deployment Model in the following discussion for simplicity.

The Single Campus Deployment Model is the topology used for the CMX Location Analytics use cases provided in Chapter 7, "CMX Use Case Stories."

Note A campus deployment is shown as the example in this section for simplicity.

Figure 4-4 shows the placement of the wireless infrastructure components (APs, WLCs, MSE, and PI) within the campus for the Single Campus Deployment Model.

Figure 4-4 Single Campus Deployment Model



This deployment model is the classic centralized (Local Mode) design applied to a campus site which has sufficient infrastructure to support a three-tiered network design. Cisco second generation 3700, 3600, 2700, or 2600 APs deployed within a campus are controlled by Cisco 5500 Series WLCs deployed locally within a separate services module of the campus.

Wireless data traffic is terminated centrally (Local Mode) on the wireless controllers, with the exception of the B2C guest WLAN/SSID which is discussed in B2C Guest Access for CMX Visitor Connect. One or more MSEs (either 3355 standalone appliances or virtual machines running on Cisco UCS infrastructure) are also deployed within the campus data center to support the Context Aware Service (location), CMX Analytics, and CMX Visitor Connect. Cisco Prime Infrastructure running on a virtual machine also within the data center is used to import, configure, and synchronize floor maps to the MSE as well as enable the MSE services.

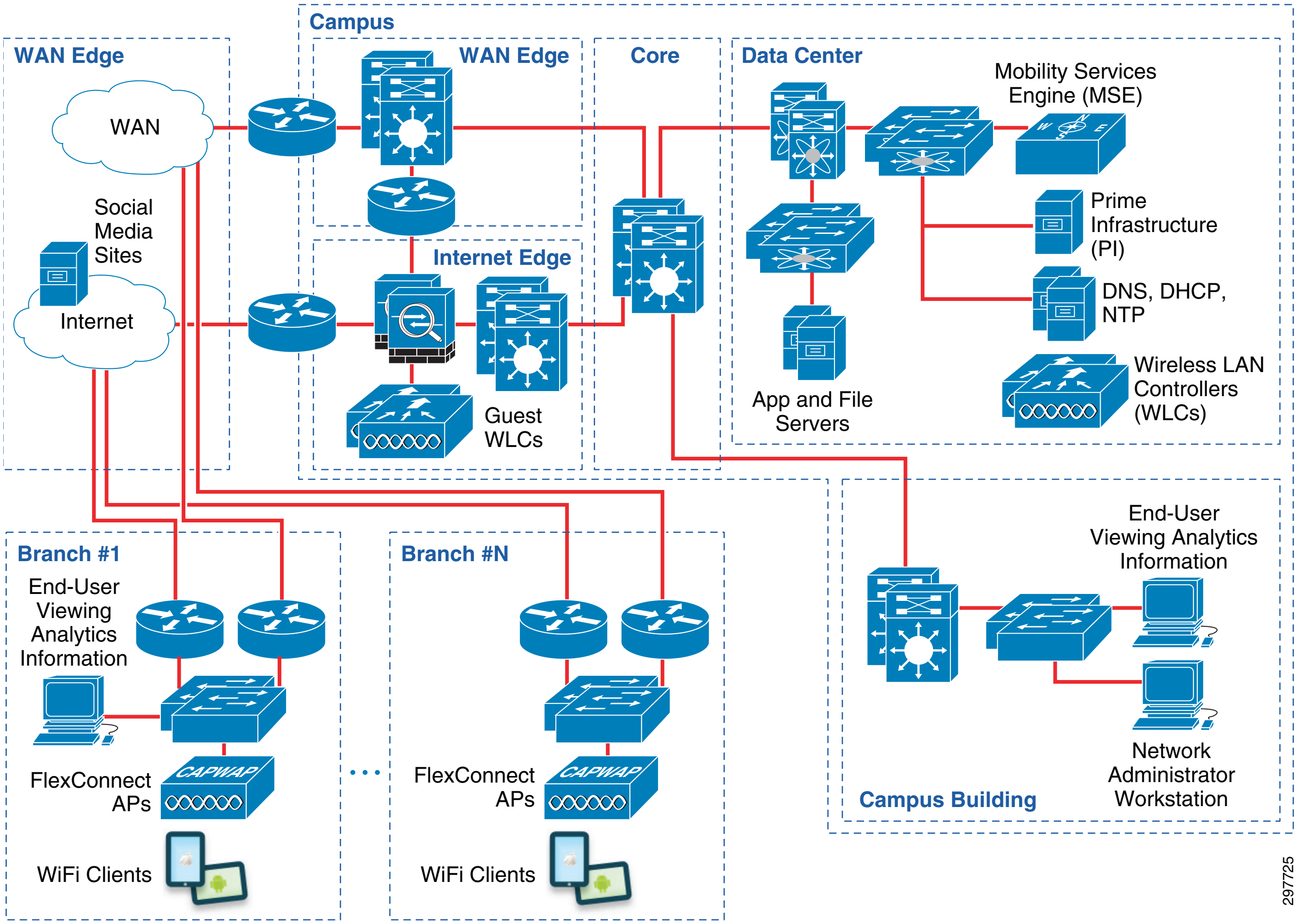
Since the APs, WLCs, and MSE are all on-premises, no WAN bandwidth is utilized in transporting RSSI information from APs to the WLCs and from the WLCs to the MSE. This is one of the advantages of this design. However a disadvantage of this design is that a system-wide view as it relates to Cisco CleanAir, wIPS, Context Aware Services, or CMX cannot be achieved if each Campus location runs its own set of MSEs.

Small Branch Deployment Model

The Small Branch deployment model is the topology used for the CMX Presence Analytics use case and the CMX Visitor Connect use case that is provided in Chapter 7, "CMX Use Case Stories."

Figure 4-5 shows the placement of the wireless infrastructure components (APs, WLCs, MSE, and PI) both within the branch and the campus, for the Small Branch Deployment Model.

Figure 4-5 Small Branch Deployment Model



This deployment model is suited for customer deployments in which a large number of branches, each of which only requires one or two APs, need to be supported. Cisco second generation 3700, 2700, 3600, or 2600 Series APs within each branch are configured to operate in FlexConnect mode. APs deployed within the multiple small branches are controlled by Cisco Flex 7500 Series WLCs located within the data center of a remote campus.

Wireless data traffic is terminated directly on the APs, with the exception of the B2C guest WLAN/ SSID, which is discussed in B2C Guest Access for CMX Visitor Connect. One or more MSEs (either 3355 standalone appliances or virtual machines running on Cisco UCS infrastructure) are also deployed within the campus data center to support the Context Aware Service (location), CMX Analytics, and

CMX Visitor Connect. Cisco Prime Infrastructure running on a virtual machine also within the data center is used to import, configure, and synchronize floor maps to the MSE as well as enable the MSE services.

The bandwidth utilization of transporting RSSI information from each AP to the WLC is one of the disadvantages of this design. However the advantage of this design is the lower cost of supporting one or more centralized MSEs within a campus location which service all branches versus deploying one or more MSEs at each branch. Further, a system-wide view as it relates to Cisco CleanAir, wIPS, Context Aware Services, or CMX can be achieved if the MSE is servicing WLCs throughout the network.

B2C Guest Access for CMX Visitor Connect

For this version of the Cisco CMX design guide, guest wireless Internet access is provided by auto-anchoring the traffic from a B2C guest WLAN/SSID back through either the Flex 7500 Series WLCs (for the Small Branch Deployment model) or the 5500 Series WLCs (for the Single Campus Deployment model) and terminating the traffic on a DMZ segment off of the dedicated guest 5508 WLC within the Internet Edge of the campus. The B2C guest WLAN/SSID, used for CMX Visitor Connect, is configured for open authentication with Web Passthrough on both the internal (foreign) WLC and the guest (anchor) WLC. The MSE is configured as the external web portal within the WLCs for web redirection. Details for configuring the WLCs are provided in Chapter 23, “Configuring Cisco Wireless LAN Controllers.”