CITRIX

Deployment Modes

Contents

Customizing the Ethernet ports 3

Port Parameters 3

Accelerated Bridges (apA and apB) 4

Motherboard Ports 6

VLAN Support 6

Inline Mode 7

Ethernet Bypass and Link-Down Propagation 7

Accelerating an Entire Site 8

Partial-Site Acceleration 8

WCCP Mode 9

WCCP Mode (Non-Clustered) 13

WCCP Clustering 13

Deployment Topology 15

Deployment Worksheet and Cluster Limitations 16

Configuring WCCP Clustering 18

Testing and Troubleshooting 19

Virtual Inline Mode 19

Configuring Packet Forwarding on the Appliance 20

Router Configuration 21

Monitoring and Troubleshooting 22

Group Mode 22

When to Use Group Mode 23

How Group Mode Works 24

Enabling Group Mode **24**

Forwarding Rules **26**

Monitoring and Troubleshooting Group Mode **27**

High-Availability Mode **27**

Cabling Requirements **28**

Other Requirements **28**

Updating Software on a High-Availability Pair **29**

Customizing the Ethernet ports

August 17, 2018

A typical appliance has four Ethernet ports: two accelerated bridged ports, called *accelerated pair A* (apA.1 and apA.2), with a bypass (fail-to-wire) relay, and two unaccelerated motherboard ports, called Primary and Aux1. The bridged ports provide acceleration, while the motherboard ports are sometimes used for secondary purposes. Most installations use only the bridged ports.

Some SD-WAN units have only the motherboard ports. In this case, the two motherboard ports are bridged.

The appliance's user interface can be accessed by a VLAN or non-VLAN network. You can assign a VLAN to any of the appliance's bridged ports or motherboard ports for management purposes.

Figure 1. Ethernet Ports

Port List

Table 1. Ethernet Port Names

The ports are named as follows:

Motherboard port 1	Primary (or apA.1 if no bypass card is present)
Motherboard port 2	Auxiliary1 or Aux1 (or apA.2 if no bypass card is present)
Bridge #1	Accelerated Pair A (apA, with ports apA.1 and apA.2)
Bridge #2	Accelerated Pair B (apB, with ports apB.1 and apB.2)

Port Parameters

August 20, 2018

Each bridge and motherboard port can be:

- Enabled or disabled
- Assigned an IP address and subnet mask
- Assigned a default gateway
- Assigned to a VLAN
- Set to 1000 Mbps, 100 Mbps, or 10 Mbps
- Set to full duplex, half-duplex, or auto (on SD-WAN 4000/5000 appliances, some ports can be set to 10 Gbps)

All of these parameters except the speed/duplex setting are set on the Configuration: IP Address page. The speed/duplex settings are set on the Configuration: Interface page.

> **Note**s about parameters:

- Disabled ports do not respond to any traffic.
- The browser-based UI can be enabled or disabled independently on all ports.
- To secure the UI on ports with IP addresses, select HTTPS instead of HTTP on the Configuration: Administrator Interface: Web Access page.
- Inline mode works even if a bridge has no IP address. All other modes require that an IP address be assigned to the port.
- Traffic is not routed between interfaces. For example, a connection on bridge apA does not cross over to the Primary or Aux1 ports, but remains on bridge apA. All routing issues are left to your routers.

Accelerated Bridges (apA and apB)

August 17, 2018

Every appliance has at least one pair of Ethernet ports that function as an accelerated bridge, called *apA* (for *accelerated pair A*). A bridge can act in inline mode, functioning as a transparent bridge, as if it were an Ethernet switch. Packets flow in one port and out the other. Bridges can also act in one arm mode, in which packets flow in one port and back out the same port.

An appliance that has a bypass card maintains network continuity if a bridge or appliance malfunctions.

Some units have more than one accelerated pair, and these additional accelerated pairs are named apB, apC, and so on.

Bypass Card

If the appliance loses power or fails in some other way, an internal relay closes and the two bridged ports are electrically connected. This connection maintains network continuity but makes the bridge

ports inaccessible. Therefore you might want to use one of the motherboard ports for management access.

Caution: Do not enable the Primary port if it is not connected to your network. Otherwise, you cannot access the appliance, as explained in
Ethernet Bypass and Link-Down Propagation

Bypass cards are standard on some models and optional on others. Citrix recommends that you purchase appliances with bypass cards for all inline deployments.

The bypass feature is wired as if a cross-over cable connected the two ports, which is the correct behavior in properly wired installations.

Important: Bypass installations must be tested - Improper cabling might work in normal operation but not in bypass mode. The Ethernet ports are tolerant of improper cabling and often silently adjust to it. Bypass mode is hard-wired and has no such adaptability. Test inline installations with the appliance turned off to verify that the cabling is correct for bypass mode.

Using Multiple Bridges

If the appliance is equipped with two accelerated bridges, they can be used to accelerate two different links. These links can either be fully independent or they can be redundant links connecting to the same site. Redundant links can be either load-balanced or used as a main link and a failover link.

Figure 1. Using dual bridges

When it is time for the appliance to send a packet for a given connection, the packet is sent over the same bridge from which the appliance received the most recent input packet for that connection. Thus, the appliance honors whatever link decisions are made by the router, and automatically tracks the prevailing load-balancing or main-link/failover-link algorithm in real time. For non-load-balanced links, the latter algorithm also ensures that packets always use the correct bridge.

WCCP and Virtual Inline Modes

Multiple bridges are supported in both WCCP mode and virtual inline mode. Usage is the same as in the single-bridge case, except that WCCP has the additional limitation that all traffic for a given WCCP service group must arrive on the same bridge.

High Availability with Multiple Bridges

Two units with multiple bridges can be used in a high-availability pair. Simply match up the bridges so that all links pass through both appliances.

Motherboard Ports

August 17, 2018

Although the Ethernet ports on a bypass card are inaccessible when the bypass relay is closed, the motherboard ports remain active. You can sometimes access a failed appliance through the motherboard ports if the bridged ports are inaccessible.

The Primary Port

If the Primary port is enabled and has an IP address assigned to it, the appliance uses that IP address to identify itself to other acceleration units. This address is used internally for a variety of purposes, and is most visible to users as the Partner Unit field on the Monitoring: Optimization: Connections page. If no motherboard port is enabled, the appliance uses the IP address of Accelerated Pair A.

The Primary port is used for:

- Administration through the web based UI
- A back channel for group mode
- A back channel for high-availability mode

The Aux1 Port

The Aux1 port is identical to the Primary port. If the Aux1 port is enabled and the Primary port is not, the appliance takes its identity from the Aux1 port's IP address. If both are enabled, the Primary port's IP address is the unit's identity

VLAN Support

July 19, 2018

A virtual local area network (VLAN) uses part of the Ethernet header to indicate which virtual network a given Ethernet frame belongs to. SD-WAN appliances support VLAN trunking in all forwarding modes (inline, WCCP, virtual inline, and group mode). Traffic with any combination of VLAN tags is handled and accelerated correctly.

For example, if one traffic stream passing through the accelerated bridge is addressed to 10.0.0.1, VLAN 100, and another is addressed to 10.0.0.1, VLAN 111, the appliance knows that these are two distinct destinations, even though the two VLANs have the same IP address.

You can assign a VLAN to all, some, or none of the appliance's Ethernet ports. If a VLAN is assigned to a port, the management interfaces (GUI and CLI) listen only to traffic on that VLAN. If no VLAN is assigned, the management interfaces listen only to traffic without a VLAN. This selection is made on the Configuration: Appliance Settings: Network Adapters: IP Addresses tab.

Inline Mode

August 20, 2018

In inline mode, traffic passes into one of the appliance's Ethernet ports and out of the other. When two sites with inline appliances communicate, every TCP connection passing between them is accelerated. All other traffic is passed through transparently, as if the appliance were not there.

Figure 1. Inline mode, Accelerating All Traffic on a WAN

> **Note**: Any TCP-based traffic passing through both units is accelerated. No address translation, proxying, or per-site setup is required. Inline mode is auto-detecting and auto-configuring.

Configuration is minimized with inline mode, because your WAN router need not be aware of the appliance's existence.

Depending on your configuration, inline mode's link-down propagation can affect management access to the appliance if a link goes down.

Inline mode is most effective when applied to all traffic flowing into and out of a site, but it can be used for only some of the site's traffic.

Ethernet Bypass and Link-Down Propagation

October 27, 2018

Note: Link-Down propagation was added to the SD-WAN 2000, 3000, 4000, and 5000 appliances with the 7.2.1 release.

Most appliance models include a "fail-to-wire" (Ethernet bypass) feature for inline mode. If power fails, a relay closes and the input and output ports become electrically connected, allowing the Ethernet signal to pass through from one port to the other as if the appliance were not there. In fail-to-wire mode, the appliance looks like a cross-over cable connecting the two ports.

Any failure of the appliance hardware or software also closes the relay. When the appliance is restarted, the bypass relay remains closed until the appliance is fully initialized, maintaining network continuity at all times. This feature is automatic and requires no user configuration.

When the bypass relay is closed, the appliance's bridge ports are inaccessible.

If carrier is lost on one of the bridge ports, the carrier is dropped on the other bridge port to ensure that the link-down condition is propagated to the device on the other side of the appliance. Units that monitor link state (such as routers) are thus notified of conditions on the other side of the bridge.

Link-down propagation has two operating modes:

- If the Primary port is not enabled, the link-down state on one bridge port is mirrored briefly on the other bridge port, and then the port is re-enabled. This allows the appliance to be reached through the still-connected port for management, high availability heartbeat, and other tasks.
- If the Primary port is enabled, the appliance assumes (without checking) that the Primary port is used for management, high availability heartbeat, and other tasks. The link-down condition on one bridge port is mirrored persistently on the other port, until carrier is restored or the unit is rebooted. This is true even if the Primary port is enabled in the GUI but not connected to a network, so the Primary port should be disabled (the default) when not in use.

Accelerating an Entire Site

July 19, 2018

Inline mode, Accelerating All Traffic on a WAN shows a typical configuration for inline mode. For both sites, the appliances are placed between the LAN and the WAN, so all WAN traffic that can be accelerated is accelerated. This is the simplest method for implementing acceleration, and it should be used when practical.

Because all the link traffic is flowing through the appliances, the benefits of fair queuing and flow control prevent the link from being overrun.

In IP networks, the bottleneck gateway determines the queuing behavior for the entire link. By becoming the bottleneck gateway, the appliance gains control of the link and can manage it intelligently. This is done by setting the bandwidth limit slightly lower than the link speed. When this is done, link performance is ideal, with minimal latency and loss even at full link utilization.

Partial-Site Acceleration

August 17, 2018

To reserve the appliance's accelerated bandwidth for a particular group of systems, such as remote backup servers, you can install the appliance on a branch network that includes only those systems. This is shown in the following figure.

Figure 1. Inline Mode, Accelerating Selected Systems Only

SD-WAN traffic shaping relies on controlling the entire link, so traffic shaping is not effective with this topology, because the appliance sees only a portion of link traffic. Latency control is up to the bottleneck gateway, and interactive responsiveness can suffer.

WCCP Mode

November 14, 2018

Web Cache Communication Protocol (WCCP) is a dynamic routing protocol introduced by Cisco. Originally intended only for web caching, WCCP version 2 became a more general-purpose protocol, suitable for use by accelerators such as Citrix SD-WAN appliances.

WCCP mode is the simplest way of installing a SD-WAN appliance when inline operation is impractical. It is also useful where asymmetric routing occurs, that is, when packets from the same connection arrive over different WAN links. In WCCP mode, the routers use the WCCP 2.0 protocol to divert traffic through the appliance. Once received by the appliance, the traffic is treated by the acceleration engine and traffic shaper as if it were received in inline mode.

> **Note**:
>
> - For the purposes of this discussion, WCCP version 1 is considered obsolete and only WCCP version 2 is presented.
> - The standard WCCP documentation calls WCCP clients “caches.” To avoid confusion with actual caches, Citrix generally avoids calling a WCCP client a “cache.” Instead, WCCP clients are typically called “appliances.”
> - This discussion uses the term “router” to indicate WCCP-capable routers and WCCP-capable switches. Though the term “router” is used here, some high-end switches also support WCCP, and can be used with SD-WAN WANOP appliances.

The SD-WAN WANOP appliances support two WCCP modes:

- WCCP is the original SD-WAN WANOP WCCP offering supported since release 3.x. It supports a single appliance service group (no clustering).
- WCCP clustering, introduced in release 7.2, allows your router to load-balance traffic between multiple appliances.

How WCCP Mode Works

The physical mode for WCCP deployment of a SD-WAN WANOP appliance is one-arm mode in which the SD-WAN appliance is connected directly to a dedicated port on the WAN router. The WCCP stan-

dard includes a protocol negotiation in which the appliance registers itself with the router, and the two negotiate the use of features they support in common. Once this negotiation is successful, traffic is routed between the router and the appliance according to the WCCP router and redirection rules defined on the router.

A WCCP-mode appliance requires only a single Ethernet port. The appliance should either be deployed on a dedicated router port (or WCCP-capable switch port) or isolated from other traffic through a VLAN. Do not mix inline and WCCP modes.

The following figure shows how a router is configured to intercept traffic on selected interfaces and forward it to the WCCP-enabled appliance. Whenever the WCCP-enabled appliance is not available, the traffic is not intercepted, and is forwarded normally.

Figure 1. WCCP Traffic Flow

Traffic Encapsulation

WCCP allows traffic to be forwarded between the router and the appliance in either of the following modes:

- L2 Mode—Requires that the router and appliance be on the same L2 segment (typically an Ethernet segment). The IP packet is unmodified, and only the L2 addressing is altered to forward the packet. In many devices, L2 forwarding is performed at the hardware layer, giving it the maximum performance. Because of its performance advantage, L2 forwarding is the preferred mode, but not all WCCP-capable devices support it.
- GRE Mode—Generic Routing Encapsulation (GRE) is a routed protocol and the appliance can in theory be placed anywhere, but for performance it should be placed close to the router, on a fast, uncongested path that traverses as few switches and routers as possible. GRE is the original WCCP mode. A GRE header is created and the data packet is appended to it. The receiving device removes the GRE header. With encapsulation, the appliance can be on a subnet that is not directly attached to the router. However, both the encapsulation process and the subsequent routing add CPU overhead to the router, and the addition of the 28-byte GRE header can lead to packet fragmentation, which adds additional overhead.

WCCP mode supports multiple routers and both GRE vs. L2 forwarding. Each router can have multiple WAN links. Each link can have its own WCCP service group.

Traffic shaping is not effective unless the appliance manages UDP traffic as well as TCP traffic. A second service group, with a UDP service group for each WAN link, is recommended if traffic shaping is desired.

Registration and Status Updates

A WCCP client (an appliance) uses UDP port 2048 to register itself with the router and to negotiate which traffic should be sent to it, and also which WCCP features should be used for this traffic. The appliance operates on this traffic and forwards the resulting traffic to the original endpoint. The status of an appliance is tracked through the WCCP registration process and a heartbeat protocol. The appliance first contacts the router over the WCCP control channel (UDP port 2048), and the appliance and router exchange information with packets named "Here_I_Am" and "I_See_You," respectively. By default, this process is repeated every ten seconds. If the router fails to receive a message from the appliance for three of these intervals, it considers the appliance to have failed and stops forwarding traffic to it until contact is reestablished.

Services and Service Groups

Different appliances using the same router can provide different services. To keep track of which services are assigned to which appliances, the WCCP protocol uses a service group identifier, a one-byte integer. When an appliance registers itself with a router, it includes service group numbers as well.

- A single appliance can support more than one service group.
- A single router can support more than one service group.
- A single appliance can use the same service group with more than one router.
- A single router can use the same service group with more than one appliance. For SD-WAN appliances, multiple appliances are supported in WCCP cluster mode, and a single appliance is supported in WCCP mode.
- Each appliance specifies a "return type" (L2 or GRE) independently for each direction and each service group. SD-WAN WANOP 4000/5000 appliances always specify the same return type for both directions. Other SD-WAN appliances allow the return type to be different.

Figure 2. Using different WCCP service groups for different services

Multiple service groups can be used with WCCP on the same appliance. For example, the appliance can receive service-group 51 traffic from one WAN link and service-group 62 traffic from another WAN link. The appliance also supports multiple routers. It is indifferent to whether all the routers use the same service group or different routers use different service groups.

Service Group Tracking. If a packet arrives on one service group, output packets for the same connection are sent on the same service group. If packets arrive for the same connection on multiple service groups, output packets track the most recently seen service group for that connection.

High Availability Behavior

When WCCP is used with high-availability mode, the primary appliance sends its own apA or apB management IP address, not the virtual address of the high availability pair, when it contacts the router. If failover occurs, the new primary appliance contacts the router automatically, reestablishing the WCCP channel. In most cases the WCCP timeout period and the high availability failover time overlap. As a result, the network outage is less than the sum of the two delays.

Standard WCCP allows only a single appliance in a WCCP service group. If a new appliance attempts to contact the router, it discovers that the other appliance is handling the service group, and the new appliance sets an Alert. It periodically checks to determine whether the service group is still active with the other appliance, and the new appliance handles the service group when the other appliance becomes inactive. WCCP clustering allows multiple appliances per service group.

Deployment Topology

The following figure shows a simple WCCP deployment, suitable for either L2 or GRE. The traffic port (1/1) is connected directly to a dedicated router port (Gig 4/12).

Figure 3. Simple WCCP deployment

In this example, the SD-WAN WANOP 4000/5000 is deployed in one-arm mode, with the traffic port (1/1) and the management port (0/1) each connecting to its own dedicated router port.

On the router, WCCP is configured with identical ip wccp redirect in statements on the WAN and LAN ports. Two service groups are used, 71 and 72. Service group 71 is used for TCP traffic and service group 72 is used for UDP traffic. The SD-WAN appliance does not accelerate UDP traffic, but can apply traffic shaping policies to it.

> **Note**: The WCCP specification does not allow protocols other than TCP and UDP to be forwarded, so protocols such as ICMP and GRE always bypass the appliance.

WCCP Clustering

SD-WAN release 7.2 or later supports WCCP clustering, which enables your router to load-balance your traffic between multiple appliances. For more information about deploying SD-WAN appliances as a cluster, see WCCP Clustering.

WCCP Specification

For more information about WCCP, see Web Cache Communication Protocol V2, Revision 1, http://tools.ietf.org/html/draft-mclaggan-wccp-v2rev1-00.

WCCP Mode (Non-Clustered)

November 16, 2018

WCCP mode allows only a single appliance in a WCCP service group. If a new appliance attempts to contact the router, it discovers that the other appliance is handling the service group, and the new appliance sets an Alert. It periodically checks to determine whether the service group is still active with the other appliance, and the new appliance handles the service group when the other appliance becomes inactive.

> Note: WCCP clustering allows multiple appliances per service group.

Limitations and best practices

Following are limitations and best practices for (non-clustered) WCCP mode:

- On appliances with more than one accelerated pair, all the traffic for a given WCCP service group must arrive on the same accelerated pair.
- Do not mix inline and WCCP traffic on the same appliance. The appliance does not enforce this guideline, but violating it can cause difficulties with acceleration. (WCCP and virtual inline modes can be mixed, but only if the WCCP and virtual inline traffic are coming from different routers).
- For sites with a single WAN router, use WCCP whenever inline mode is not practical.
- Only one appliance is supported per service group. If more than one appliance attempts to connect to the same router with the same service group, the negotiation succeeds only for the first appliance.
- For sites with multiple WAN routers serviced by the same appliance, WCCP can be used to support one, some, or all of your WAN routers. Other routers can use virtual inline mode.

WCCP Clustering

November 16, 2018

The WCCP clustering feature enables you to multiply your acceleration capacity by assigning more than one SD-WAN appliance to the same links. You can cluster up to 32 identical appliances, for up to 32 times the capacity. Because it uses the WCCP 2.0 standard, WCCP clustering works on most routers and some smart switches, most likely including those you are already using.

Because it uses a decentralized protocol, WCCP clustering allows SD-WAN appliances to be added or removed at will. If an appliance fails, its traffic is rerouted to the surviving appliances.

Unlike SD-WAN high-availability, an active/passive pair that uses two appliances to provide the performance of a single appliance, the same appliances deployed as a WCCP cluster has twice the performance of a single appliance, delivering both redundancy and improved performance.

In addition to adding more appliances as your site's needs increase, you can use Citrix's "Pay as You Grow" feature to increase your appliances' capabilities through license upgrades.

Citrix Command Center is recommended for managing WCCP clusters. The following figure shows a basic network of a cluster of SD-WAN appliances in WCCP mode, administered by using Citrix Command Center.

Load-Balanced WCCP Clusters

The WCCP protocol supports up to 32 appliances in a fault-tolerant, load balanced array called a cluster. In the example below, three identical appliances (same model, same software version) are cabled identically and configured identically except for their IP addresses. Appliances using the same service groups with the same router can become a load balanced WCCP cluster. When a new appliance registers itself with the router, it can join the existing pool of appliances and receive its share of traffic. If an appliance leaves the network (as indicated by the absence of heartbeat signals), the cluster is rebalanced so that only the remaining appliances are used.

Figure 1. A load-balanced WCCP cluster with three appliances

One appliance in the cluster is selected as the designated cache, and controls the load-balancing behavior of the appliances in the cluster. The designated cache is the appliance with the lowest IP address. Because the appliances have identical configurations, it doesn't matter which one is the designated cache. If the current designated cache goes offline, a different appliance becomes the designated cache.

The designated cache determines how the load-balanced traffic is allocated and informs the router of these decisions. The router shares information with all members of the cluster, so the cluster can operate even if the designated cache goes offline.

> Note: As normally configured, an SD-WAN WANOP 4000/5000 appliance appears as two WCCP caches to the router.

- Load-Balancing Algorithm

  Load balancing in WCCP is static, except when an appliance enters or leaves the cluster, which causes the cluster to be rebalanced among its current members.

  The WCCP standard supports load balancing based on a mask or a hash. For example, SD-WAN WANOP WCCP clustering uses the mask method only, using a mask of 1-6 bits of the 32-bit IP address. These address bits can be non-consecutive. All addresses yielding the same result when masked are sent to the same appliance. Load balancing effectiveness depends on choosing an

appropriate mask value: a poor mask choice can result in poor load-balancing or even none, with all traffic sent to a single appliance.

Deployment Topology

November 16, 2018

Depending on your network topology, you can deploy WCCP cluster either with a single router or with multiple routers. Whether connected to a single router or multiple routers, each appliance in the cluster must be connected identically to all routers in use.

Deploying appliances in a WCCP cluster requires more planning than does deploying a single appliance. Read the following sections carefully before proceeding.

Single router deployment

In the following diagram, three SD-WAN appliances accelerate the datacenter's 200 Mbps WAN. The site supports 750 XenApp users.

As shown on the NetScaler SD-WAN Datasheet, an SD-WAN 3000-100 can support 100 Mbps and 400 users, so a pair of these appliances supports 200 Mbps and 800 users, which satisfies the datacenter's requirements of a 200 Mbps link and 750 users.

For fault tolerance, however, the WCCP cluster should continue to operate without becoming overloaded if one appliance fails. That can be accomplished by using three appliances when the calculations call for two. This is called the N+1 rule.

Failure is an unusual event, so usually all three appliances are in operation. In this case, each appliance is supporting only 67 Mbps and 250 users, leaving plenty of headroom, and making good use of the fact that the cluster has three times the CPU power and three times the compression history of a single appliance.

Without WCCP clustering, as much capacity and fault-tolerance would require a pair of SD-WAN WANOP 4000-500 appliances in high availability mode. Only one of these appliances is active at a time.

Multiple router deployment

Using multiple WAN routers is similar to using a single WAN router. If the previous example is changed to include two 100 Mbps links instead of one 200 Mbps link, the topology changes, but the calculations do not.

Deployment Worksheet and Cluster Limitations

February 28, 2019

On the following worksheet, you can calculate the number of appliances needed for your installation and the recommended mask field size. The recommended mask size is 1–2 bits larger than the minimum mask size for your installation.

Parameter	Value	Notes
Appliance Model Used		—
Supported XenApp and XenDesktop Users Per Appliance	Uspec =	From data sheet
XenApp and XenDesktop Users on WAN Link	Uwan =	—
User overload Factor	Uoverload = Uwan/Uspec =	—
Supported BW Per Appliance	BWspec =	From data sheet
WAN Link BW	BWwan =	—
BW Overload Factor	BWoverload = BWwan/BWspec =	—
Number of appliances required	N = max(Uoverload, BWoverload) +1 =	Includes one spare
		—
Min number of buckets	Bmin = N, rounded up a power of 2 =	—
If SD-WAN 4000 or 5000,	Bmin = 2N, rounded up to a power of 2 =	—
Recommended value	B = 4 X Bmin if Bmin <= 16, else 2 X Bmin =	—
Number of “one” bits in address mask	M = log2(B)	If B=16, M=4.

Mask value: The mask value is a 32-bit address mask with several one bits equal to M in the worksheet provided earlier. Often these bits can be the least-significant bits in the WAN subnet mask used by

your remote sites. If the masks at your remote sites vary, use the median mask. (Example: With /24 subnets, the least significant bits of the subnet are 0x00 00 nn 00. The number of bits to set to one is log2(mask size): if mask size is 16, set 4 bits to one. So with a mask size of 16 and a /24 subnet, set the mask value to 0x00 00 0f 00.): _____

The above guidelines work only if the selected subnet field is evenly distributed in your traffic, that is, that each address bit selected by the mask is a one for half the remote hosts, and a zero for the other half. Otherwise, load-balancing is impaired. This even distribution might be true for only a few bits in the network field (only 2 bits). If so with your network, instead of masking bits in the offending area of the subnet field, displace those bits to a portion of the host address field that has the 50/50 property. For example, if only three subnet bits in a /24 subnet have the 50/50 property, and you are using four mask bits, a mask of 0x00 00 07 10 avoids the offending bit at 0x00 00 0800 and displaces it to 0x00 00 00 10, a portion of the address field that is likely to have the 50/50 property if your remote subnets generally use at least 32 IP addresses each.

Parameter	Value	Notes
Final Mask Value		—
Accelerated Bridge		Usually apA
WAN Service Group		A service group not already in use on your router (51-255)
LAN Service Group		Another unused service group
Router IP address		IP address of router interface on port facing the appliance
WCCP Protocol (usually “Auto”)		—
DC Algorithm		Use “Deterministic” if you have only two appliances or are using dynamic load balancing like HSRP or GSLB. Otherwise, use “Least Disruptive.”

Configuring appliances in a WCCP cluster has the following limitations:

- All appliances within a cluster must be the same model and use the same software release.
- Parameter synchronization between appliances within the cluster is not automatic. Use Command Center to manage the appliances as a group.

- SD-WAN traffic shaping is not effective, because it relies on controlling the entire link as a unit, and none of the appliances are in a position to do this. Router QoS can be used instead.
- The WCCP-based load-balancing algorithms do not vary dynamically with load, so achieving a good load balance can require some tuning.
- The hash method of cache assignment is not supported. Mask assignment is the supported method.
- While the WCCP standard allows mask lengths of 1-7 bits, the appliance supports masks of 1-6 bits.
- Multicast service groups are not supported; only unicast service groups are supported.
- All routers using the same service group pair must support the same forwarding method (GRE or L2).
- The forwarding and return method negotiated with the router must match: both must be GRE or both must be L2. Some routers do not support L2 in both directions, resulting in an error of "Router's forward or return or assignment capability mismatch." In this case, the service group must be configured as GRE.
- SD-WAN VPX does not support WCCP clustering.
- The appliance supports (and negotiates) only unweighted (equal) cache assignments. Weighted assignments are not supported.
- Some older appliances, such as the SD-WAN 700, do not support WCCP clustering.
- (SD-WAN WANOP 4000/5000 only) Two accelerator instances are required per interface in L2 mode. No more than three interfaces are supported per appliance (and then on appliances with six or more accelerator instances.)
- (SD-WAN 4000/5000 only) WCCP control packets from the router must match one of the router IP addresses configured on the appliance for the service group. In practice, the router's IP address for the interface that connects it to the appliance should be used. The router's loopback IP cannot be used.

Configuring WCCP Clustering

November 16, 2018

After you have finalized the deployment topology, considered all limitations, and filled in the deployment worksheet, you are ready to deploy your appliances in a WCCP cluster. To configure the WCCP cluster, you need to perform the following tasks:

- Configuring the Router
- Configuring the Appliance

Testing and Troubleshooting

November 16, 2018

The **Monitoring** > **Appliance** > **Application Performance** > **WCCP** page shows the current state of not only the local appliance but of all other appliances that have joined the cluster. Select a WCCP cache and click **Get Info**.

The Cache Status tab shows the local appliance's status. When all is well, the status is "25: has assignment." You must refresh the page manually to monitor changes in status. If the appliance does not reach the status of "25: has assignment" within a timeout period, other informative status messages are displayed.

Additional information is displayed when you click on the **Service Group or the Routers** tabs.

The Cluster Summary tab displays information about the WCCP cluster as a whole. As a side effect of the WCCP protocol, each member of the cluster has information about all the others, so this information can be monitored from any appliance in the cluster.

Your router can also provide status information. See your router documentation.

Virtual Inline Mode

August 20, 2018

> **Note**: Use virtual inline mode only when both inline mode and WCCP mode are impractical. Do not mix inline and virtual inline modes within the same appliance. However, you can mix virtual inline and WCCP modes within the same appliance. Citrix does not recommend virtual inline mode with routers that do not support health monitoring.

In virtual inline mode, the router uses policy based routing (PBR) rules to redirect incoming and outgoing WAN traffic to the appliance for acceleration, and the appliance forwards the processed packets back to the router. Almost all of the configuration tasks are performed on the router. The only thing to be configured on the appliance is the forwarding method, and the default method is recommended.

Like WCCP, Virtual inline deployment requires no rewiring and no downtime, and it provides a solution for asymmetric routing issues faced in a deployment with two or more WAN links. Unlike WCCP, it contains no built-in status monitoring or health checking, making troubleshooting difficult. WCCP is thus the recommended mode, and virtual inline is recommended only when inline and WCCP modes are both impractical.

Example

The following figure shows a simple network in which all traffic destined for or received from the remote site is redirected to the appliance. In this example, both the local site and remote site use virtual inline mode.

Figure 1. Virtual Inline Example

Following are some configuration details for the network in this example:

- Endpoint systems have their gateways set to the local router (which is not unique to virtual inline mode).
- Each router is configured to redirect both incoming and outgoing WAN traffic to the local appliance.
- Each appliance processes the traffic received from its local router and forwards it back to the router.
- PBR rules configured on the router prevent routing loops by allowing packets to make only one trip to and from the appliance. The packets that the appliance forwards back to the router are sent to their original (local or remote) destination.
- Each appliance has its default gateway set to the address of the local router, as usual (on the **Configuration: Network Adapters** page). The options for forwarding packets back to the router are Return to Ethernet Sender and Send to Gateway.

Configuring Packet Forwarding on the Appliance

July 19, 2018

Virtual inline mode offers two packet-forwarding options:

Return to Ethernet Sender (default)—This mode allows multiple routers to share an appliance. The appliance forwards virtual inline output packets back to where they came from, as indicated by the Ethernet address of the incoming packet. If two routers share a single appliance, each gets its own traffic back, but not the traffic from the other router. This mode also works with a single router.

Send to Gateway (not recommended)—In this mode, virtual inline output packets are forwarded to the default gateway for delivery, even if they are destined for hosts on the local subnet. This option is usually less desirable than the Return to Ethernet Sender option, because it adds an easily forgotten element of complexity to the routing structure.

To specify the packet-forwarding option—On the Configuration: Optimization Rules: Tuning page, next to Virtual Inline, select Return to Ethernet Sender or Send to Gateway.

Router Configuration

August 17, 2018

The router has three tasks when supporting virtual inline mode:

1. It must forward both incoming and outgoing WAN traffic to the SD-WAN appliance.
2. It must forward traffic to its destination (WAN or LAN).
3. It must monitor the health of the SD-WAN appliance so that the appliance can be bypassed if it fails.

Policy-Based Rules

In virtual inline mode, the packet forwarding methods can create routing loops if the routing rules do not distinguish between a packet that has been forwarded by the appliance and one that has not. You can use any method that makes that distinction.

A typical method involves dedicating one of the router's Ethernet ports to the appliance and creating routing rules that are based on the Ethernet port on which packets arrive. Packets that arrive on the interface dedicated to the appliance are never forwarded back to the appliance, but packets arriving on any other interface can be.

The basic routing algorithm is:

- Do not forward packets from the appliance back to the appliance.
- If the packet arrives from the WAN, forward it to the appliance.
- If packet is destined for the WAN, forward to the appliance.
- Do not forward LAN-to-LAN traffic to the appliance.
- Traffic shaping is not effective unless all WAN traffic passes through the appliance.

Note: When considering routing options, keep in mind that returning data, not just outgoing data, must flow through the appliance. For example, placing the appliance on the local subnet and designating it as the default router for local systems does not work in a virtual inline deployment. Outgoing data would flow through the appliance, but incoming data would bypass it. To force data through the appliance without router reconfiguration, use inline mode.

Health Monitoring

If the appliance fails, data should not be routed to it. By default, Cisco policy based routing does no health monitoring. To enable health monitoring, define a rule to monitor the appliance's availability, and specify the "verify-availability" option for the "set ip next-hop" command. With this configuration, if the appliance is not available, the route is not applied, and the appliance is bypassed.

Important: Citrix recommends virtual inline mode only when used with health monitoring. Many routers that support policy-based routing do not support health-checking. The health-monitoring feature is relatively new. It became available in Cisco IOS release 12.3(4)T.

Following is an example of a rule for monitoring the availability of the appliance:

```
pre codeblock !- Use a ping (ICMP echo)to see if appliance is connected
 track 123 rtr 1 reachabilit y ! rtr 1 type echo protocol IpIcmpecho
192.168.1.200 schedule 1 life forever start-time now
```

This rule pings the appliance at 192.168.1.200 periodically. You can test against 123 to see if the unit is up.

Monitoring and Troubleshooting

July 19, 2018

In virtual inline mode, unlike WCCP mode, the appliance provides no virtual inline-specific monitoring. To troubleshoot a virtual inline deployment, log into the appliance and use the Dashboard page to verify that traffic is flowing into and out of the appliance. Traffic forwarding failures are typically caused by errors in router configuration.

If the Monitoring: Usage or Monitoring: Connections pages show that traffic is being forwarded but no acceleration is taking place (assuming that an appliance is already installed on the other end of the WAN link), check to make sure that both incoming WAN traffic and outgoing WAN traffic are being forwarded to the appliance. If only one direction is forwarded, acceleration cannot take place.

To test health-checking, power down the appliance. The router should stop forwarding traffic after the health-checking algorithm times out.

Group Mode

August 20, 2018

In group mode, two or more appliances become a single virtual appliance. This mode is one solution to the problem of asymmetric routing, which is defined as any case in which some packets in a given connection pass through a given appliance but others do not. A limitation of the appliance architecture is that acceleration cannot take place unless all packets in a given connection pass through the same two appliances. Group mode overcomes this limitation.

Group mode can be used with multiple or redundant links without reconfiguring your routers.

> **Note**: Group mode is not supported on the SD-WAN 4000 or 5000 appliances.

Group mode applies only to the appliances on one side of the WAN link; the local appliances neither know nor care whether the remote appliances are using group mode.

Group mode uses a heartbeat mechanism to verify that other members of the group are active. Packets are forwarded to active group members only.

Avoiding asymmetric routing is the main reason to use group mode, but group mode is not the only method available for that purpose. If you decide that it is the best method for your environment, you can enable it by setting a few parameters. If the default mechanism for determining which appliance is responsible for a particular connection does not provide optimal acceleration, you can change the forwarding rules.

Figure 1. Group Mode With Redundant Links

Figure 2. Group Mode With Non-Redundant Links with Possible Asymmetric Routing

Figure 3. Group Mode On Nearby Campuses

When to Use Group Mode

July 19, 2018

Use group mode in the following set of circumstances:

- You have multiple WAN links.
- There is a chance of asymmetric routing (a packet on a given connection might travel over either link).
- Group mode seems simpler and more practical than alternatives that use a single appliance.

The alternatives are:

- WCCP mode, in which traffic from two or more links is sent to the same appliance by WAN routers, by means of the WCCP protocol.
- Virtual inline mode, in which your routers send traffic from two or more links through the same appliance (or high-availability pair).
- Multiple bridges, where each link passes through a different accelerated bridge in the same appliance.
- LAN-level aggregation, which places an appliance (or high-availability pair) closer to the LAN, before the point where WAN traffic is split into two or more paths.

How Group Mode Works

August 17, 2018

In group mode, the appliances that are part of the group each take ownership for a portion of the group's connections. If a given appliance is the owner of a connection, it makes all the acceleration decisions about that connection and is responsible for compression, flow control, packet retransmission, and so on.

If an appliance receives a packet for a connection for which it is not the owner, it forwards the packet to the appliance that is the owner. The owner examines the packet, makes the appropriate acceleration decisions, and forwards any output packets back to the non-owning appliance. This process preserves the link selection made by the router, while allowing all packets in the connection to be managed by the owning appliance. For the routers, the introduction of the appliances has no consequences. The routers do not need to be reconfigured in any way, and the appliances do not need to understand the routing mechanism. They simply accept the routers' forwarding decisions.

Figure 1. Sending-side Traffic in Group Mode

Figure 2. Receiving-side traffic flow in group mode

Group mode has two, user-selectable failure modes, which control how the group members interact with each other if one of them fails. The failure mode also determines whether the failed appliance's bypass card opens (blocking traffic through the appliance) or remains closed (allowing traffic to pass through). The failure modes are:

Continue to accelerate- If a group member fails, its bypass card is opened and no traffic passes through the failed appliance. The result is presumably a fail-over if redundant links are used. Otherwise, the link is simply inaccessible. The other appliances in the group continue to accelerate. The usual hashing algorithm handles the changed conditions. (That is, the old hashing algorithm is used, and if the failed unit is indicated as the owner, a hashing algorithm based on the new, smaller group is applied. This preserves as many older connections as possible.)

Do not accelerate- If a group member fails, its bypass card closes, allowing traffic to pass through without acceleration. Because an unaccelerated path introduces asymmetric routing, the other members of the group also go into pass-through mode when they detect the failure.

Enabling Group Mode

October 27, 2018

To enable group mode, create a group of two or more appliances. An appliance can be a member of only one group. Group members are identified by IP address and the SSL common name in the appliance license.

All group mode parameters are on the Settings: Group Mode page, in the Configure Settings: Group Mode table.

Figure 1. Group Mode Page

To enable group mode

1. Select the address to use for group communication. At the top of the Group Mode Configuration table on the Configuration: Advanced Deployments: Group Mode tab, the table cell under Member VIP contains the management address of the port used to communicate with other group members. Use the (unlabeled) drop-down menu to select the correct address (for example, to use the Aux1 port, select the IP address you assigned to the Aux1 port). Then, click Change VIP.
2. Add at least one more group member to the list. (Groups of three or more are supported but are rarely used.) In the next cell of the Member VIP column, type the IP address of the port used by the other appliance for group-mode communication.
3. Type the other group member's SSL common name in the SSL Common Name column. The SSL common name is listed on the other appliance's Configure: Advanced Deployments: High Availability tab. If the other group member is a high-availability pair, the name listed is the SSL common name of the primary appliance.
   Note: If the local appliance is not part of a high-availability pair, the first cell in the high availability Secondary SSL Common Name is blank. If the other group member is a high-availability pair, specify the SSL Common Name of the high availability secondary appliance in the high availability Secondary SSL Common Name column.
4. Click Add.
5. Repeat steps 2-4 for any additional appliances or high-availability pairs in the group.
6. The three buttons under the list of group members are toggles, so each is labeled as the opposite of its current setting:
   a) The top button reads either, **Do not accelerate when member failure is detected** or **Continue to accelerate when member failure is detected**. The "Do not accelerate…" setting always works and does not block traffic, but if any member fails, the other group members go into bypass mode, which causes a complete loss of acceleration. With the "Continue to accelerate" option, the failing appliance's bridge becomes an open circuit, and the link fails. This option is appropriate if the WAN router responds by causing a failover. New connections, and open connections belonging to the surviving appliances, are accelerated.
   b) The bottom button should now be labeled Disable Group Mode. If it is not, enable group mode by clicking the button.
7. Refresh the screen. The top of the page should list the group mode partners, but display warnings about their status, because they haven't been configured for group mode yet. For example,

it might indicate that the partner cannot be found or is running a different software release.

8. Repeat this procedure with the other members of the group. Within 20 seconds after enabling the last member of the group, the Group Mode Status line should show NORMAL, and the other group mode members should be listed with Status: On-Line and Configuration: OK.

Forwarding Rules

August 17, 2018

By default, the *owner* of a group-mode connection is set by a hash of the source and destination IP addresses. Each appliance in the group uses the same algorithm to determine which group member owns a given connection. This method requires no configuration. The owner can optionally be specified through user-settable rules.

Because the group-mode hash is not identical to that used by load balancers, about half of the traffic tends to be forwarded to the owning appliance in a two-Appliance group. In the worst case, forwarding causes the load on the LAN-side interface to be doubled, which halves the appliance's peak forwarding rate for actual WAN traffic.

This speed penalty can be reduced if the Primary or Aux1 Ethernet ports are used for traffic between group members. For example, if you have a group of two appliances, you can use an Ethernet cable to connect the two units' Primary ports, then specify the Primary port on the Group Mode page on each unit. However, maximum performance is achieved if the amount of traffic forwarded between the group-mode members is minimized.

The owner can optionally be set according to specific IP/port-based rules. These rules must be identical on all appliances in the group. Each member of the group verifies that its group-mode configuration is identical to the others. If not all of the configurations are identical, none of the member appliances enter group mode.

If traffic arrives first at the appliance that owns the connection, it is accelerated and forwarded normally. If it arrives first at a different appliance in the group, it is forwarded to its owner over a GRE tunnel, which accelerates it and returns it to the original appliance for forwarding. Thus, group mode leaves the router's link selection unchanged.

Using explicit IP-based forwarding rules can reduce the amount of group-mode forwarding. This is especially useful in primary-link/backup-link scenarios, where each link handles a particular range of IP addresses, but can act as a backup when the other link is down.

Figure 1. IP-Based Owner Selection

Forwarding rules can ensure that group members handle only their "natural" traffic. In many installations, where traffic is usually routed over its normal link and only rarely crosses the other one, these

rules can reduce overhead substantially.

Rules are evaluated in order, from top to bottom, and the first matching rule is used. Rules are matched against an optional IP address/mask pair (which is compared against both source and destination addresses), and against an optional port range.

Regardless of the ordering of rules, if the partner appliance is not available, traffic is not forwarded to it, whether a rule matches or not.

For example, in the figure below, member 172.16.1.102 is the owner of all traffic to or from its own subnet (172.16.1.0/24), while member 172.16.0.184 is the owner of all other traffic.

If a packet arrives at unit 172.16.1.102, and it is not addressed to/from net 172.16.1.0/24, it is forwarded to 172.16.0.184.

If unit 172.16.0.184 fails, however, unit 172.16.1.102 no longer forwards packets. It attempts to handle the traffic itself. This behavior can be inhibited by clicking **Do NOT Accelerate When Member Failure Detected** on the Group Mode tab.

In a setup with a primary WAN link and a backup WAN link, write the forwarding rules to send all traffic to the appliance on the primary link. If the primary WAN link fails, but the primary appliance does not, the WAN router fails over and sends traffic over the secondary link. The appliance on the secondary link forwards traffic to the primary-link appliance, and acceleration continues undisturbed. This configuration maintains accelerated connections after the link failover.

Figure 2. Forwarding Rules

Monitoring and Troubleshooting Group Mode

July 19, 2018

Two things should be checked in a group-mode installation:

- That the two appliances have entered group mode, which can be determined on either appliance's Configuration: Advanced Deployments: Group Mode page.
- That the behavior of the group-mode pair is as desired when the other member fails, and when one of the links fail, as determined by disabling the other appliance and temporarily disconnecting one of the links, respectively.

High-Availability Mode

July 19, 2018

Two identical appliances on the same subnet can be combined as a *high-availability pair.* The appliances each monitor the other's status by using the standard *Virtual Router Redundancy Protocol (VRRP)* heartbeat mechanism. The pair has a common virtual IP address for management, in addition to each appliance's management IP address. If the primary appliance fails, the secondary appliance takes over. Failover takes approximately five seconds.

High availability mode is a standard feature.

Cabling Requirements

August 17, 2018

The two appliances in the high availability pair are installed onto the same subnet in either a parallel arrangement or a one-arm arrangement, both of which are shown in the following figure. In a one-arm arrangement, use the apA.2 port (and, optionally, the apB.2 port), not the apA.1 port. Some models require a separate management LAN, whether deployed in inline or one-armed mode. This is depicted only in the middle diagram.

Figure 1. Cabling for High-Availability Pairs

Do not break the above topology with additional switches. Random switch arrangements are not supported. Each of the switches must be either a single, monolithic switch, a single logical switch, or part of the same chassis.

If the spanning-tree protocol (STP) is enabled on the router or switch ports attached to the appliances, failover will work, but the failover time may increase to roughly thirty seconds. Without STP, failover time is roughly five seconds. Thus, to achieve the briefest possible failover interval, disable STP on the ports connecting to the appliances.

Figure 2. Ethernet Port Locations (Older Models)

Other Requirements

October 27, 2018

Both appliances in an high availability pair must meet the following criteria:

- Have identical hardware, as shown by on the System Hardware entry on the Dashboard page.
- Run exactly the same software release.

- Be equipped with Ethernet bypass cards. To determine what is installed in your appliances, see the Dashboard page.

Appliances that do not support high availability display a warning on the Configuration: High Availability page.

Updating Software on a High-Availability Pair

October 27, 2018

Updating the SD-WAN software on an high availability pair causes a failover at one point during the update.

Note: Clicking the Update button terminates all open TCP connections.

To update the software on an high availability pair

1. Log on to both appliances.
2. On the secondary appliance, update the software and reboot. After the reboot, the appliance is still the secondary. Verify that the installation succeeded. The primary appliance should show that the secondary appliance exists but that automatic parameter synchronization is not working, due to a version mismatch.
3. On the primary appliance, update the software, and then reboot. The reboot causes a failover, and the secondary appliance becomes the primary. When the reboot is completed, high availability should become fully established, because both appliances are running the same software.

