Topics

- Glossary and Definitions
- Determinants for network path
- Incoming connections
- Outgoing connections
  - Routing module on NSK
  - Routing module on CLIM
- Source Address Selection
- Sample configurations
- Guidelines
Interfaces:
- the data interfaces (eth1 thru eth5) are marked **e1-e5**
- the interfaces which are not configured are colored grey
- for the sake of a simpler diagram, interfaces shown in the examples are all physical, but an interface could be bonded interface for the configuration/topics discussed in this document

Connections:
- a connection is shown on the interface which is used for the connection
- the **directed block** of the connection diagram contains the **destination address** used and the **rectangular block** of the connection contains the **source address** used
- the **color of the destination box** matches that of the destination Node and the **color of the source box** matches that of the interface whose IP address is being used as the source address
IP Addresses:
- if the application is bound to an IP address, that one is marked within ‘[]’ in the IP address list
- the IP address selected to be the source address for the connection is marked with ‘--’ in front of the IP address

Commands:
- climconfig ip -add eth1 -ipaddress b.20 -netmask 255.255.255.0
- climconfig ip -add eth1 -ipaddress c.21 -netmask 255.255.255.0
Routes:
- routes implicitly added by the kernel (when an IP address is added to the interface) are marked in \textit{white-italics}
- routes added explicitly by the user (using climconfig) are \textcolor{green}{colored green}
- the selected route for this connection is marked with ‘\textbf{--}’
- the kernel route \texttt{b.0/24} is a result of adding the IP address \texttt{b.20/24} to the interface
- the user added route \texttt{c.0/24} shows the route added with options ‘\texttt{-net -target c.0 -netmask 255.255.255.0}’
- the user added route \texttt{d.1/0} is the \texttt{default route} added with ‘\texttt{-default -gateway d.1}’

Commands:
- climconfig route -add eth1 -default -gateway d.1
Connections:

- a connection is shown on the interface connection that it traverses
- the directed block of the connection diagram contains the destination address used and the rectangular block of the connection contains the source address used
- the color of the destination box matches that of the destination Node and the color of the source box matches that of the interface whose IP address is being used as the source address
- the connection shown horizontally-down shifted from the interface is an example of a second connection originated from the Provider
Connections:

- a connection is shown on the interface connection that it uses
  
  - the **directed block** of the connection diagram contains the destination address used and the rectangular block of the connection contains the source address used
  
  - the color of the destination box matches that of the destination Node and the color of the source box matches that of the interface whose IP address is being used as the source address

- the connection shown horizontally shifted from the interface is an example of a second connection sent from the Provider

- the connection shown with a grey destination box is the representation of the connection that is right above that one, originated after the failover
Prefix:

- the netmask (Genmask column in the netstat output) in the route is also referred as prefix in this presentation.

- longest prefixes are listed first. Same colored values are of the same prefix length.

- default has no prefix match (netmask of 0.0.0.0).

Commands:

- climconfig ip -add eth1 -ipaddress 10.10.10.11 -netmask 255.255.255.0
- climconfig ip -add eth4 -ipaddress 10.10.10.44 -netmask 255.255.255.0
- climconfig ip -add eth5 -ipaddress 10.15.50.20 -netmask 255.0.0.0
- climconfig ip -add bond0 -ipaddress 10.11.60.100 -netmask 255.255.0.0
- climconfig ip -add bond0 -ipaddress 10.12.60.200 -netmask 255.255.0.0

- climconfig route -add eth4 -host -target 10.10.10.200
- climconfig route -add eth4 -default -gateway 10.10.10.1
- climconfig route -add bond0 -net -target 20.0.0.0 -netmask 255.0.0.0 -gateway 10.11.1.1
Activation:

- same colored values are of the same prefix length.
- routes with the same prefix-length are listed in the order of activation.
Routes:

- host route has a netmask (prefix) of 32
- climconfig route -add eth4 -host -target 10.10.10.200
Routes:

- host route has a prefix (netmask) of 32
  - climconfig route -add eth4 -host -target 10.10.10.200

- a net route has a prefix of >0 and <32
  - climconfig route -add bond0 -net -target 10.12.0.0 -netmask 255.255.0.0
Routes:
- host route has a prefix (netmask) of 32
  - climconfig route -add eth4 -host -target 10.10.10.200

- a net route has a prefix of >0 and <32
  - climconfig route -add bond0 -net -target 10.12.0.0 -netmask 255.255.0.0

- a default route has a prefix of 0
- route that matches any destination (matches any destination since the netmask is 0)
  - climconfig route -add eth5 -default -gateway 10.10.10.1
Primary IP address:
- displayed by 'ip addr show' command
- primary IP address is listed as the first on the interface
- non-primary IP addresses are tagged 'secondary'

Source hint:
- shown by the command 'ip route show'
- normally the primary address is used as the source hint
Various components and factors that affect the network path taken by a connection

Destination IP Address:
- the attribute to look up the route
- for a client application, this is specified by the client
- for a server application, it is normally the other end of the connection

Route on the CLIM:
- routes to the destination, activated on the CLIM is looked up first
- subnet routes, added by kernel as part adding an IP address
- routes added by the user, using climconfig

CIP ROUTE object:
- default gateways on multiple CLIMs can be prioritized by ROUTE object
- added in SCF, associated with a CIP provider

Interfaces on the CLIM:
Source IP Address:
- used in connection requests, seen by the network devices on the path
- if set in the socket, picks the CLIM

Configured IP address:
- configuring IP address adds kernel subnet routes
- for server application, this determines incoming interface
- for client applications bound to IP address, this determines the outgoing CLIM
Incoming:
- independent of routing
- the connection comes in at the appropriate interface
Outgoing connections

Two Level routing algorithm
- At the NSK host side (Host Routing - HR)
  - goal is to select the CLIM
  - depends on the local address in the socket
  - depends on the destination address
  - depends on routes from all the CLIMs of the provider
- At the CLIM side (CLIM Routing - CR)
  - goal is to select the interface
  - depend on routes on all the interfaces of the CLIM
  - depends on the activation order of the interfaces

Host side:
  - select the CLIM and send to the CLIM for routing

CLIM side:
  - send to Linux stack for routing
Local IP Address:
- local IP address is already selected for the socket, select that CLIM
- local address is the IP address that an application binds to
- for a connected socket, the local address is already set
- for a server application, it is the IP address of the received connection request

LPM:
- **Longest Prefix Match** (LPM) algorithm is used for picking the route
  - all the routes in the Provider is considered
  - host route has prefix of 32, default route has prefix of 0

Unique LPM:
- if there is a single LPM route, then the CLIM with that route is picked
  - the connection request is sent to that CLIM, for CLIM routing processing

Multiple equal LPMs:
- if there are multiple routes with the same LPM (not default), then the **round-robin** is used to pick the route (and CLIM)
  - if there are two routes with the same LPM on CLIM-A and one route on CLIM-B, then CLIM-A is picked twice as often as CLIM-B
  - starting J06.09/H06.20, the metric attribute(....)
Multiple default routes:
- if there are multiple default routes, the SCF ROUTE object is also considered

ROUTE object:
- used if there are multiple CLIMs on a provider
- used to break tie between default routes on multiple CLIMs
- the CLIM with the highest ROUTE priority is selected
- this object is NOT used for selecting any equal host/net routing.

Round Robin:
- if there is no ROUTE object, then default route is picked using round robin algorithm
- if there are multiple ROUTEs with the same priority, round robin algorithm is used
Unique LPM:
- ‘netstat -rn’ command shows the prefix on the route
- the netstat display has the order of the lookup
- the prefix comes from the netmask of the IP address or the route

Equal LPM:
- if there are multiple LPM routes, the metric on the routes are used
- route with the lowest metric is selected
- metric is a user specified attribute on the route

Equal metric:
- if there are multiple routes with the same metric, the oldest activated interface is selected
- activation of an interface is done by the host module on NSK, in parallel
- activation of the interface depends on the network
- the order of the activation is displayed by ‘netstat -rn’ command
- ‘ip route show’ on the CLIM also shows the list of activated routes
Socket:
- if the socket is already connected, the source address = local address
- if the socket binds to a specific IP address, it is the source address

From route hint:
- if the route has a ‘src’ hint, then that IP address would be used for source address
- command 'ip route show' shows the src hint for the route
- the kernel added routes will always have a src hint

From the interface:
- if there is an IP address in the same IP subnet as the next hop in the route, that would be used
- if there are multiple IP addresses in the next-hop IP subnet, the primary IP address is used
- if a gateway is configured on the route, that is the first hop
- the first activated IP address on the interface will be used
Follow the routing algorithm with examples
The second IP address is NOT used for outgoing connections.

Config: A
- **two IP addresses** are configured, in the **same IP subnet**
- destination is link-local (switch)
  - no user routes added (only kernel added subnet route)

Commands:
- climconfig ip -add eth3 -ipaddress a.11 -netmask 255.255.255.0
- climconfig ip -add eth3 -ipaddress a.12 -netmask 255.255.255.0

Notes:
- ip addr show ➔ lists the IP addresses; non-primary IP addresses are marked ‘**secondary**’ in the output
  - the kernel added route has the source hint with the primary address, as the second IP address did not have to add another route
The second IP address is NOT used for outgoing connections.

Config: A
- **two IP addresses** are configured, in the **same IP subnet**
- destination is link-local (switch)
- no user routes added (only kernel added subnet route)

Notes:
- `ip addr show` lists the IP addresses; non-primary IP addresses are marked ‘**secondary**’ in the output
- the kernel added route has the source hint with the primary address, as the second IP address did not have to add another route
Config: B1
- two IP addresses are configured, in different subnets
- the gateway is in the same subnet as the second configured IP address
- destination is in external network (accessible thru the gateway)
- user added route to the destination network, **thru the gateway**

Commands:
- climconfig route -add eth3 -default -gateway b.1

Behavior:
- user added **default route** will be used
- the source address is **deterministic**, first configured IP address that is on the same network as the gateway

Notes:
- the user added default route does not have a ‘src’ hint
- IP address from the first hop IP subnet is picked for source address
Source address depends on the application behavior.

Config: B2
- two IP addresses are configured, in different subnets
- the gateway is in the same subnet as the second configured IP address
- user added route to the destination network, **thru the gateway**
- socket is bound to the first IP subnet

Behavior:
- user added **default route** will be used
- bound IP address is used for source address

Notes:
- network devices should be configured to handle both the IP addresses
Guidelines - config B: Multiple subnets on the same interface

- destination could see any of the configured IP addresses
- firewall configurations should handle the different IP addresses
- with application knowledge, this configuration can be used
- split it up
Separate the IP subnets to two interfaces..
Config: C1
- **two interfaces** are configured (interface could be physical as shown, or bonded)
- two IP addresses are configured, one on each interface, on **different subnets**
- destinations are **link-local**

Commands:
- `climconfig ip -add eth2 -ipaddress a.11 -netmask 255.255.255.0`
- `climconfig ip -add eth4 -ipaddress b.21 -netmask 255.255.255.0`

Behavior:
- kernel added routes are used
- interface selected is **deterministic**, one with the IP address on the same IP subnet as the destination
  - the source address is **deterministic**, the configured IP address on the interface

Notes:
- the next hop is the destination
- for a.100, route a.0/24 is picked and the IP address in the src hint is used for source address
Config:
- same as previous
- different destination

Notes:
- destination b.200 has the route on e4
- route b.0/24 is used, and source hint from the route is used for source address
Socket bound to different subnet - rare configuration.

Config:
- same as previous
- socket bound to an IP address

Notes:
- e4 has the route to the b.200 destination
- source address from e2 is used, as the application is bound to that IP address
- a.11 is used in connections sent from e4 which has the b.0 IP subnet!
What if two interfaces have the same IP subnet?
Interface and source address for the new connections change after reboot.

Config: D1
- two interfaces are configured (interface could be physical as shown, or bonded)
- two IP addresses are configured, one on each interface, on the same subnet
- connections from **two different instances** are shown (second instance after a CLIM reboot)

Behavior:
- kernel added routes are used
- the kernel route from the **first activated interface** is used
- the order interfaces are activated is **not deterministic** (and could change after a CLIM reboot/failure, or SCF clim abort)
- the interface (and thus source address) used is not deterministic across CLIM reboots or CLIM starts

Notes:
- e2 was activated first, thus only the route from that interface is used
Config:

- same as previous
- same destination
- e4 interface was activated first

Notes:

- e4 is used for originating connections to the same destination
How about adding a host route for a better LPM?
Host route to control the interface and source address

Config: D2
- same as previous
- specific **host route** to the destination added to an interface

Commands:
- climconfig ip -add eth2 -ipaddress c.31 -netmask 255.255.255.0
- climconfig route -add eth2 -host -target c.130
- climconfig ip -add eth4 -ipaddress c.32 -netmask 255.255.255.0

Notes:
- the LPM picks the interface with the host route
- the source address belongs to the interface with the user route
- independent of the activation order, the host route is picked due to the LPM algorithm
- e4 is never used for c.130 destination, unless e2 interface is failed
What if the socket is bound to an IP address?
A client application bound to an IP address.

Config: D3
- destination is external, accessible thru the gateway
- same default route is added on both the interfaces

Commands:
- climconfig route -add eth2 -default -gateway c.1
- climconfig route -add eth4 -default -gateway c.1

Notes:
- the first activated interface used
- source address is the IP address that is bound to the socket
- there are two default routes for the destination, either could be used
- destination would see the same source address (that the application bound to, [c.31])
- connection uses IP address that is not configured on the interface (c.31 v/s e4)
Incoming server connection is similar to binding!
Server application uses different outgoing interface than incoming interface.

Config: D3
- destination is external, accessible thru the gateway
- same default route is added on both the interfaces

Behavior:
- user added default route is used
- the first activated interface used
- source address is the IP address that is bound to the socket

Notes:
- network traces will only show traffic in one direction!

Guidelines:
- combine interfaces on the same subnet to bonded interface (say e2+e4 = bond0)
- metric attribute could be used to make the default route deterministic
Guidelines - config D: Multiple interfaces have the same IP subnet

- bonded interface
- host/net routes help LPM independent of the order of activation
- metric defines further preference for same LPM routes
- sniff all the interfaces for trouble shooting
- expect different IP addresses at the other end of the communication
- configure firewall/gateway to expect/allow both IP addresses (and MAC addresses)
Okay, so different subnets on multiple interfaces
Two default gateways on the same CLIM.

**Config:** E
- destination is external, accessible thru two different gateways
- **two different default routes** are added on the interfaces

**Notes:**
- there are two default routes for the destination, either could be used
- the first IP address that belongs to that of the gateway subnet on the interface is used as source address
  - source address belong to the interface
  - two connections to the same destination could have two different source address and two different paths
Config:
- same as previous
- same destination
- e5 activated after e1

Notes:
- a different gateway (b.1) is used for the same destination (d.140)
- both the gateways must be configured to allow connections to the destination
Guidelines - config E: Multiple interfaces have the same IP subnet

- host/net routes help LPM independent of the order of activation
- metric defines further preference for same LPM routes
- expect different IP addresses at the other end of the communication
- configure firewall/gateway to expect/allow both IP addresses (and MAC addresses)
How about multiple CLIMs on the Provider?
Basic multiple CLIM configuration.

Config: F1
- two CLIMs under a Provider  - each interface has **one IP address in the same subnet**
- one interface has failover configured (CY.e2 to CX.e3)
- no user added routes, destination is link-local

Notes:
- routes are selected using round-robin algorithm (round-robin of all available equivalent routes)
- each CLIM has one route each to the same destination
- the ROUTE object does not affect this config (it only affects default route selection)
Second outgoing connection, uses round-robin algorithm; picks the other CLIM.

Notes:
- the destination will see two different IP addresses from the same client, for two independent connections
CY.e2 fails and the IP address gets activated on CX.e3. The route (being the same IP subnet) is already activated on CX.e3.

Config:
- same as previous
- CX.e3 is configured as the failover interface to CY.e2

Commands:
- (on CX) climconfig failover -add eth3 -dest CY.eth2
- (on CY) climconfig failover -add eth2 -dest CX.eth3

Notes:
- prior to the failure the connection from CY.e2 had a.61 as the source address
- after failover, the source address for the equivalent connection would be a.11
How does host route affect LPM algorithm?
Config: F2
- similar to config F1
- one interface have **a user added host route**

Commands:
- (on CY) climconfig route -add eth2 -host -target a.100

Notes:
- the user added host route (and thus the CLIM) is selected for all connections, LPM algorithm
- the interface is **deterministic**, always the one with the user added host route
- the source address is deterministic (the first IP address on the interface with LPM route)
- the host route (prefix=32) outweighs the route on CX (prefix=24)
All outgoing connections (without bound socket) goes thru the same CLIM.

Notes:

- the round robin of the routes is only used for routes with equal LPM
- the host route on CY.e2 to a.100 has the best LPM
- all the connections to a.100 are sent thru CY.e2 with source address a.61
Now, what if the socket is bound to IP address?
Source address is set, the CLIM with that IP address is picked (no LPM used)

Notes:
- host route is not considered
- incoming connections responded thru the same CLIM
- outgoing connections always pick the CLIM with the bound IP address
Source address is set, the CLIM with that IP address is picked (no LPM used)

Config: F3
- similar to config F2
- application socket is bound an IP address
- response to an incoming connection

Behavior:
- host route is not considered
- incoming connections responded thru the same CLIM
- outgoing connections always pick the CLIM with the bound IP address
But I need gateways to external network!
For the first connection to d.140, route thru a.1 one is picked (CLIM CX)

Config: G1
- each interface has a **different default route** (gateway) configured
- destination is reachable thru all of the gateways

Notes:
- round-robin on all the available default routes are used for routing
- the interface on CX is selected twice (has two default routes) as often as the interface on CY (only one default route)
- there are two routes on CX, only one gateway is used (which got activated first)
- the metric attribute (-mt param in climconfig) could be used to control the default route to be used
- one interface (CX.e5) is not used for any outgoing traffic!
Since all the default routes are with equal LPM (prefix=0), host routing module picks the CLIM with the next route (b.1), CLIM CX.

The CLIM routing module picks the first activated route, which is a.1!

Notes:
- different default routes functional almost like same IP subnets configured on different interfaces
- only the first activated default route is used on a CLIM
- the second interface on CX is not used for outgoing traffic!
The host module picks the route on CY.e2 for the next connection (using round robin) and selects CLIM CY.
Why not prioritize by ROUTE object?
ROUTE object is specified with different priorities (CX=60, higher than CY=40)

Config: G2
- similar to config MD1
- ROUTE object, CLIM **CX has higher priority**

Commands:
- scf add route $zzcip.ztc1.route1, clim CX, pri 60
- scf add route $zzcip.ztc1.route2, clim CY, pri 20

Notes:
- Only the CLIM with the highest ROUTE priority is used, for all destinations that will use the default route
Notes:
- the ROUTE object priority takes precedence for default routes
- CX has a better priority and CY is never used for the destination d.140
Guideline:
- different interfaces with same IP subnet on a CLIM $\rightarrow$ dependency on interface activation order
- multiple subnets on the interface $\rightarrow$ prepare for different source addresses at the destination
- net and host routes have a better LPM than default routes
- for a multiple CLIM provider, both the CLIMs are used for default routing $\rightarrow$ ROUTE can limit that to a CLIM
- within a CLIM, the preference of equal LPM depends on activation order $\rightarrow$ metric can make it deterministic
- when a provider has multiple gateways configured, connection destined for a destination can go thru any of those $\rightarrow$ all gateways should have similar firewall configurations
Technology for better business outcomes