IP Instances – Network Isolation meets Zones

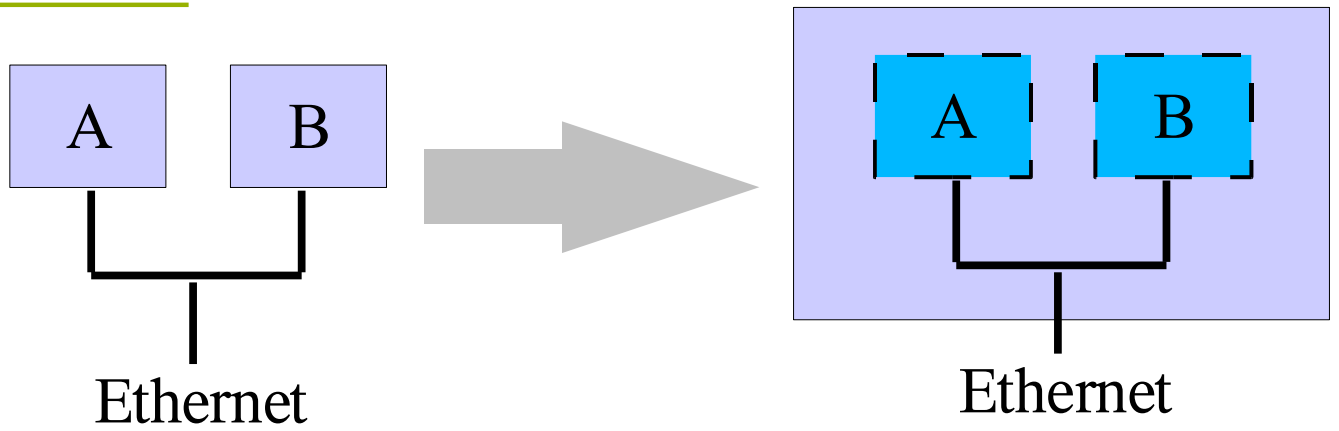
Erik Nordmark
Distinguished Engineer
Solaris Networking
October 2006

IP Instances is a part of CrossBow

- CrossBow provides various pieces of Network Virtualization and Resource Control
  - Sunay has talked about that at a previous SUG meeting
- CrossBow can be downloaded from
  - http://opensolaris.org/os/project/crossbow
- And you can participate
  - Lively discussions, design docs, FAQs, source code drops, preliminary binary releases, etc...

Zones – Application Containers

- Each zone is securely separated from other zones
  - A process belongs to one zone
- Some set of resources assigned to zone
  - CPU, file systems
- Zone has one or more IP addresses
- Allows consolidation of servers that are on the same network



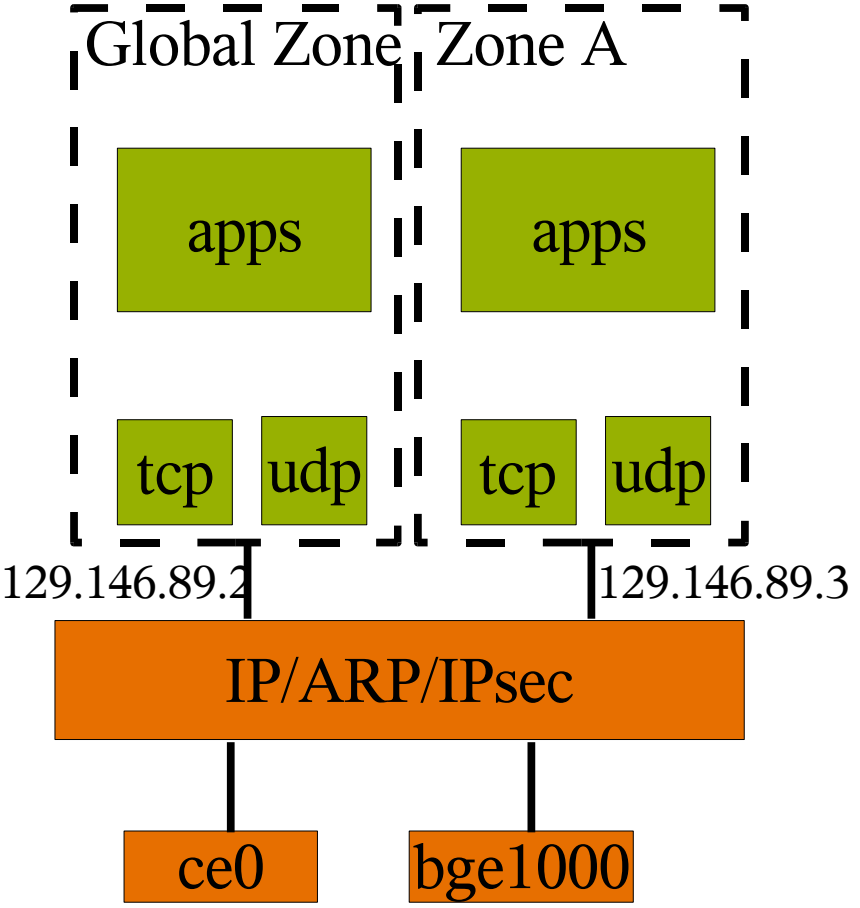
Zone Network Configuration and Deployment – Static IP

```
bilen# zonecfg -z foo
zonecfg:foo> set zonepath=/export2/foo
zonecfg:foo> add net
zonecfg:foo:net> set physical=bge0
zonecfg:foo:net> set address=10.0.0.1
zonecfg:foo:net> end
zonecfg:foo> commit
bilen# zoneadm -z foo install # or clone
bilen# zoneadm -z foo boot
bilen# zlogin -C foo
```

Answer questions about timezone, root passwd, NIS/DNS configuration

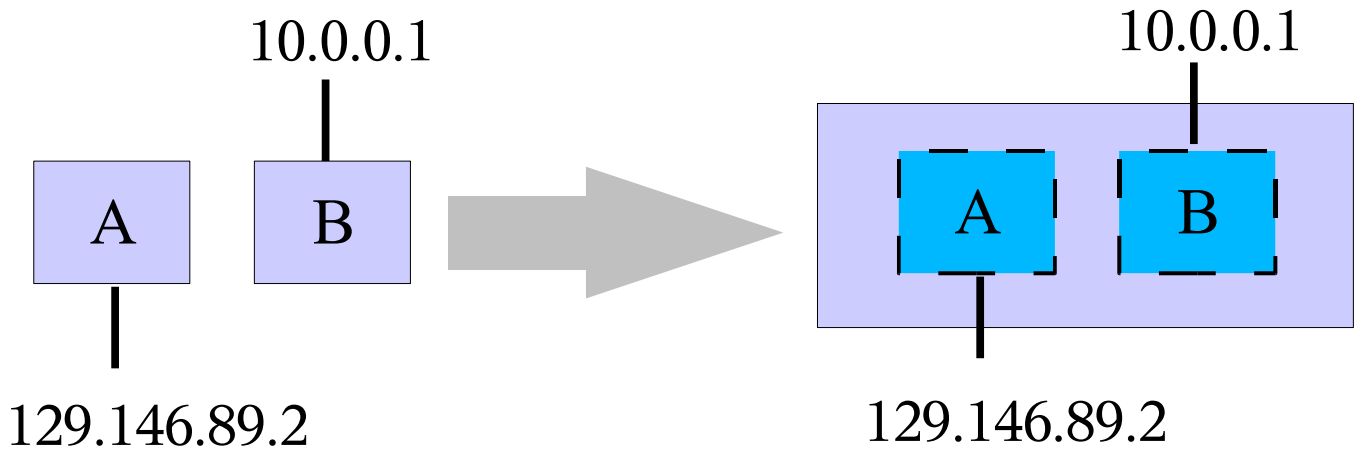
Networking in Zones

- User level is separated
- Zone constrained to use its IP addresses
- Conceptually IP/ARP/IPsec common for all zones
  - > Shared routing, ARP, configuration
- Conceptually TCP, UDP, SCTP separate for each zone
  - > Implementation has part of transports in IP

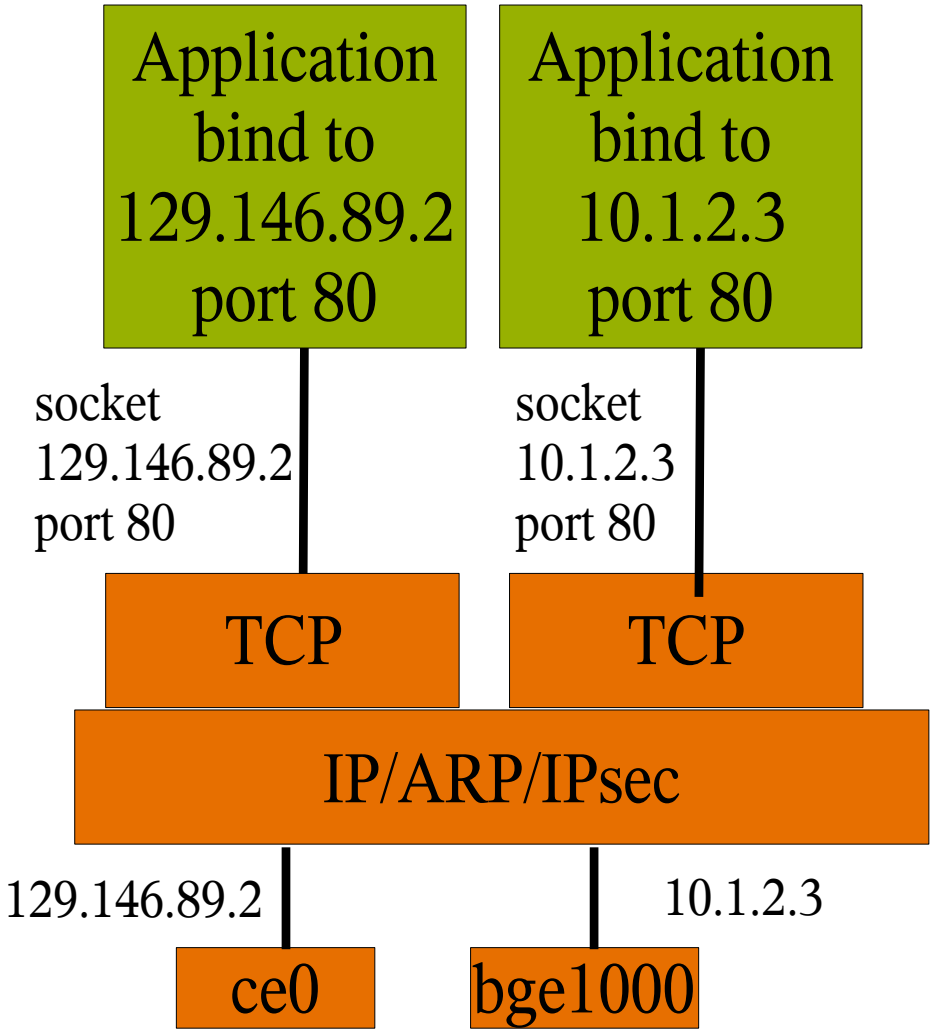


Zones + separate networks?

- How can we attach separate networks to separate zones?
  - > Separate LANs or separate VLANs
- The networks do not “connect” inside the box
- Seems to be common in data centers
- Could even be separate net 10 networks – overlapping IP addresses

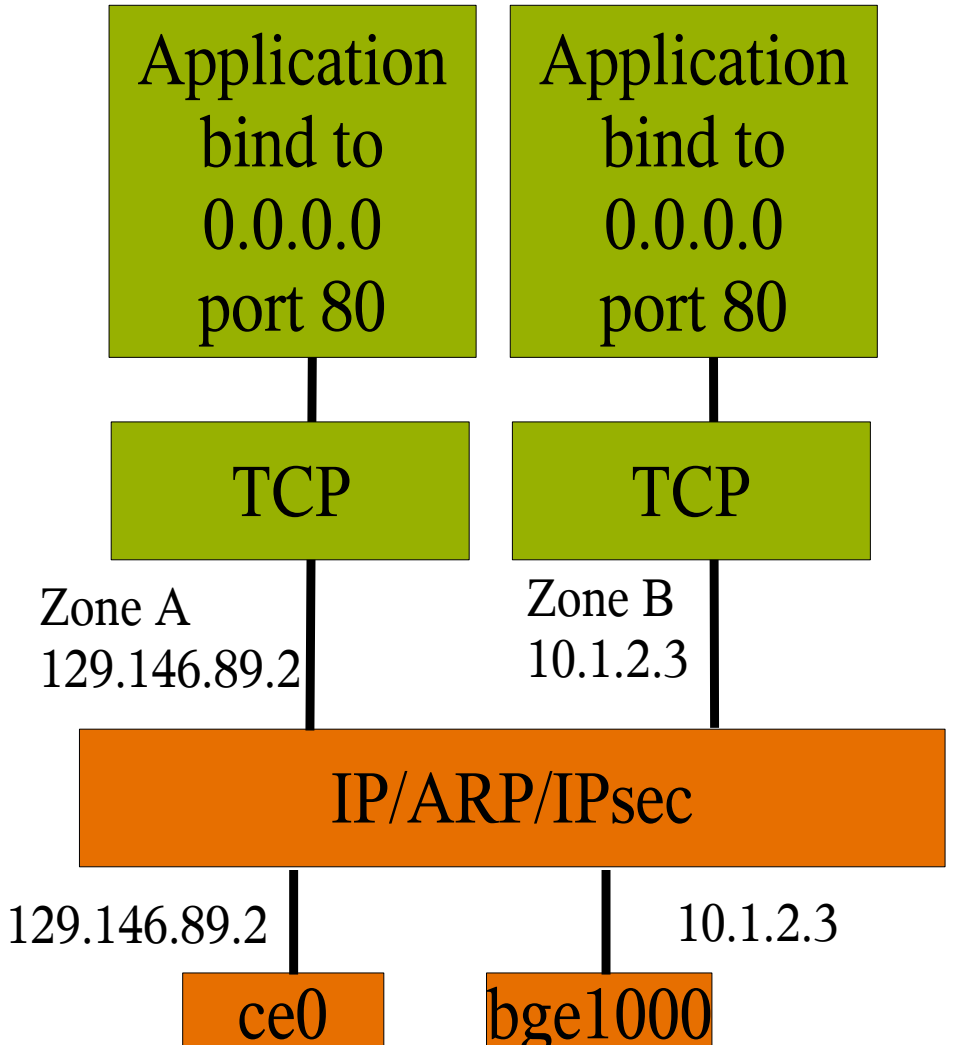


Network model prior to Zones – multihomed host support



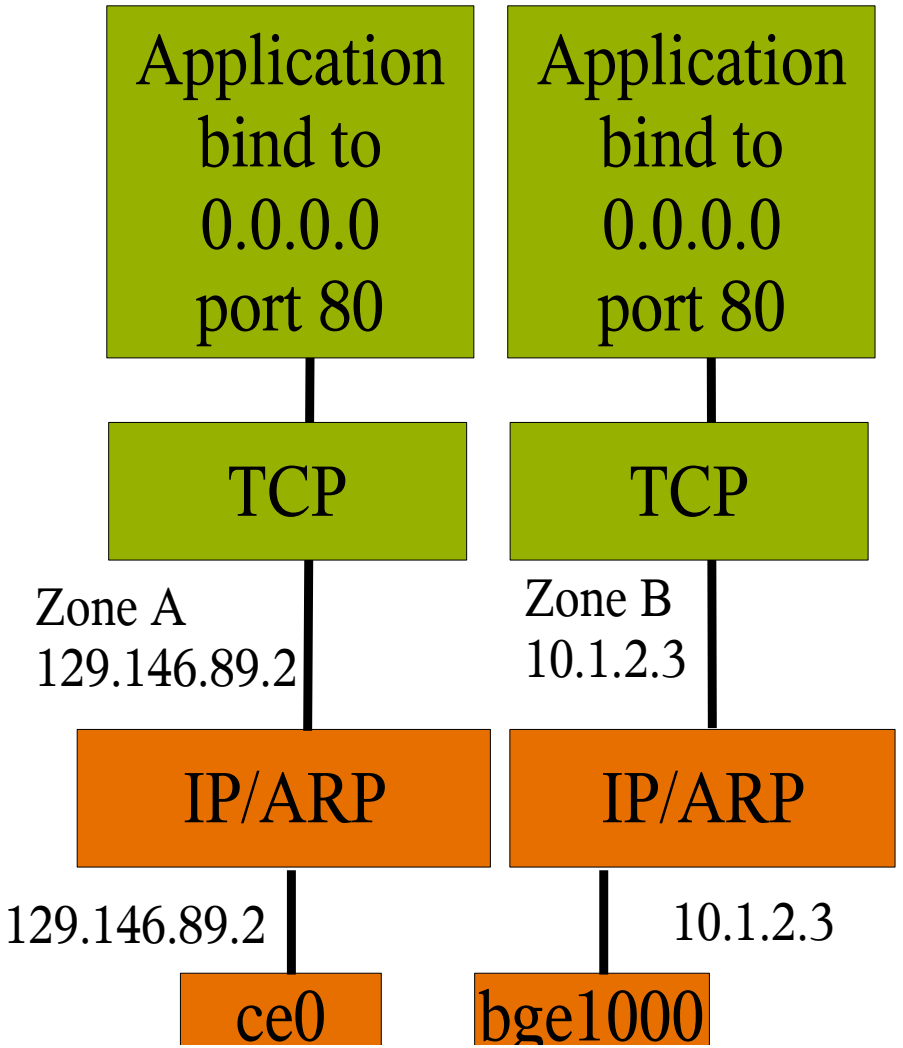
- Stack can have multiple network interfaces and multiple IP addresses
- An application can explicitly bind a socket to a particular address
  - Assumes application code has this capability

Network model for Zones – IP address assigned to a zone



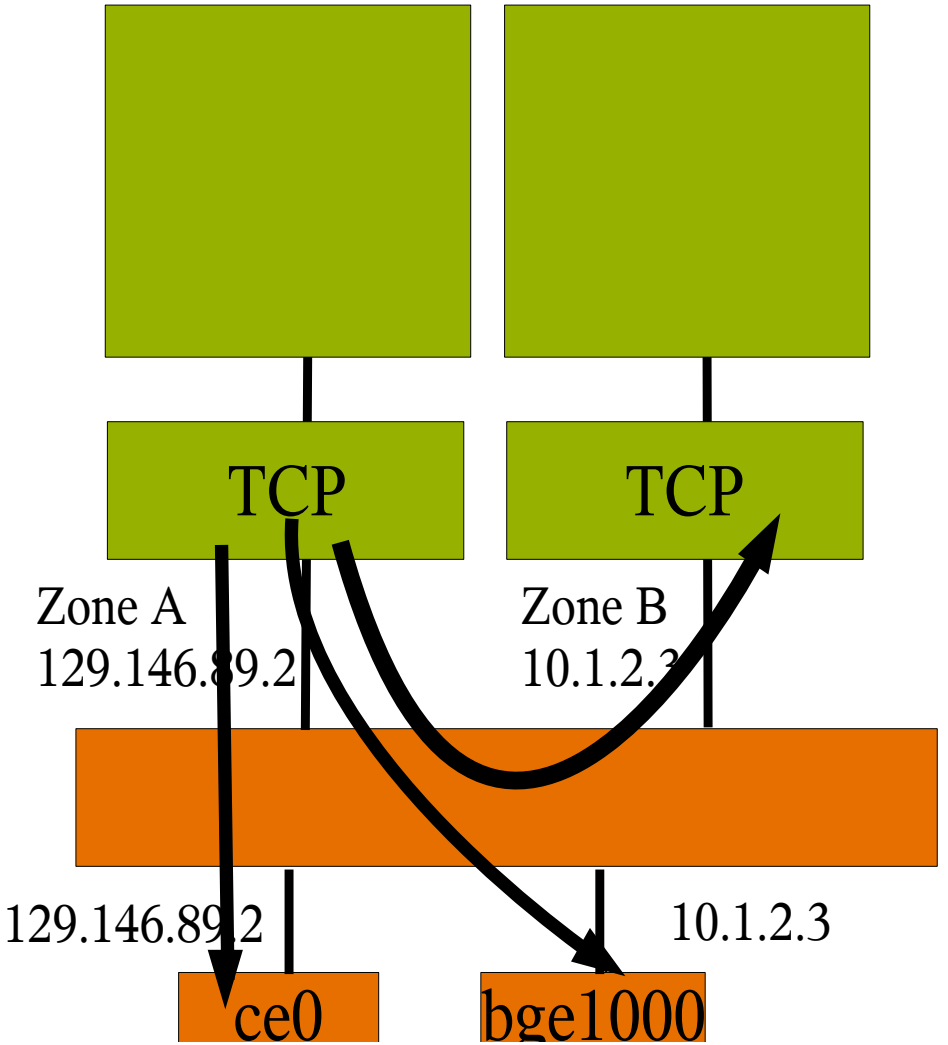
- Application in zone (including uid 0) is constrained
  - > Can't mess with other zones – change routing etc.
  - > Can't change its IP address
  - > Can't snoop

Zone model vs. separated LANs/VLANs



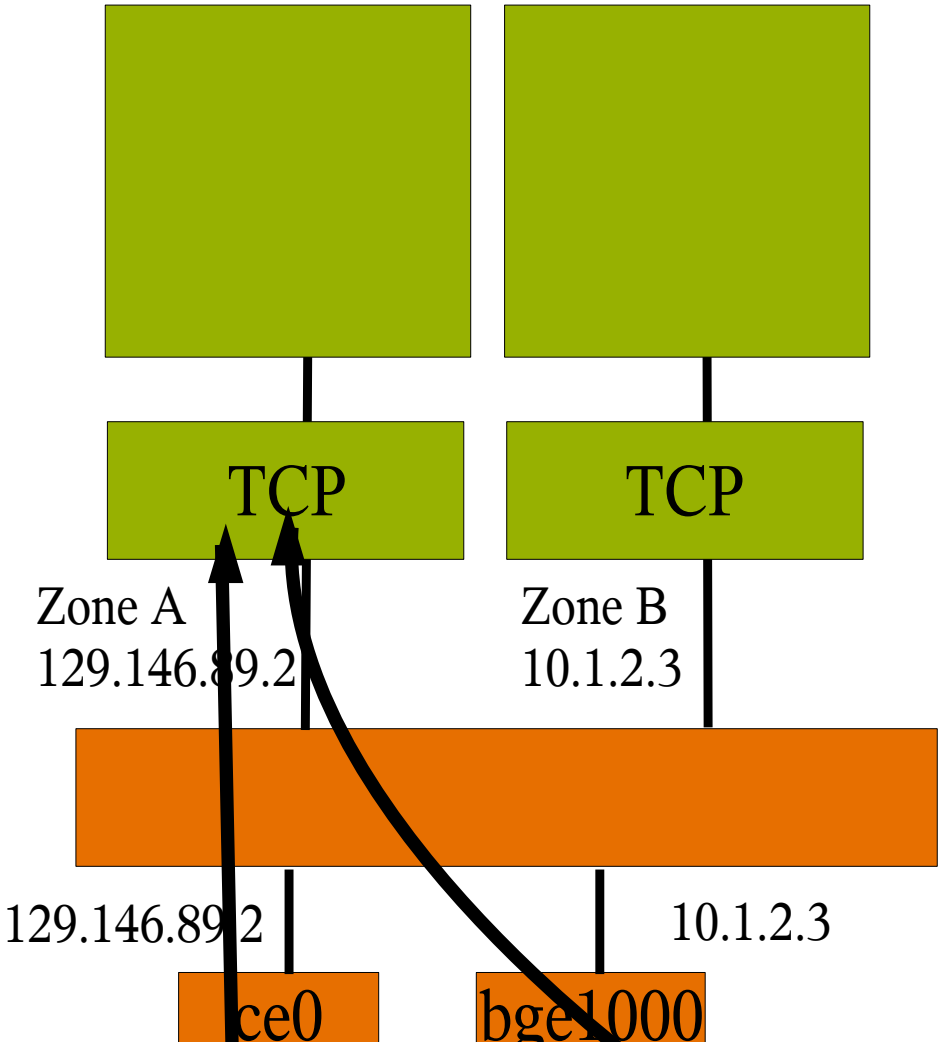
- From what we can tell, when there are separate LANs or VLANs associated with different zones, users expect IP level separation between them.
- But this isn't how the system was designed.
- Several observed differences between desired and observed behavior

Implications of Network model - outbound



- Packets sent from zone A take the shortest route to the destination
  - Out any network interface
  - Including internal loopback on another zone
  - Just as in the multihomed host case
- Various attempted fixes in S10 and escalated bugs
- How about broadcast and multicast?

Implications of Network model - inbound



- Packet destination address must match zone (or broadcast/multicast address)
- But packets can arrive from any interface
- Can potentially address for unicast with ndd's ip_strict_dst_multihoming, or IP Filter rules
- IPMP configurations?
- Broadcast, multicast?

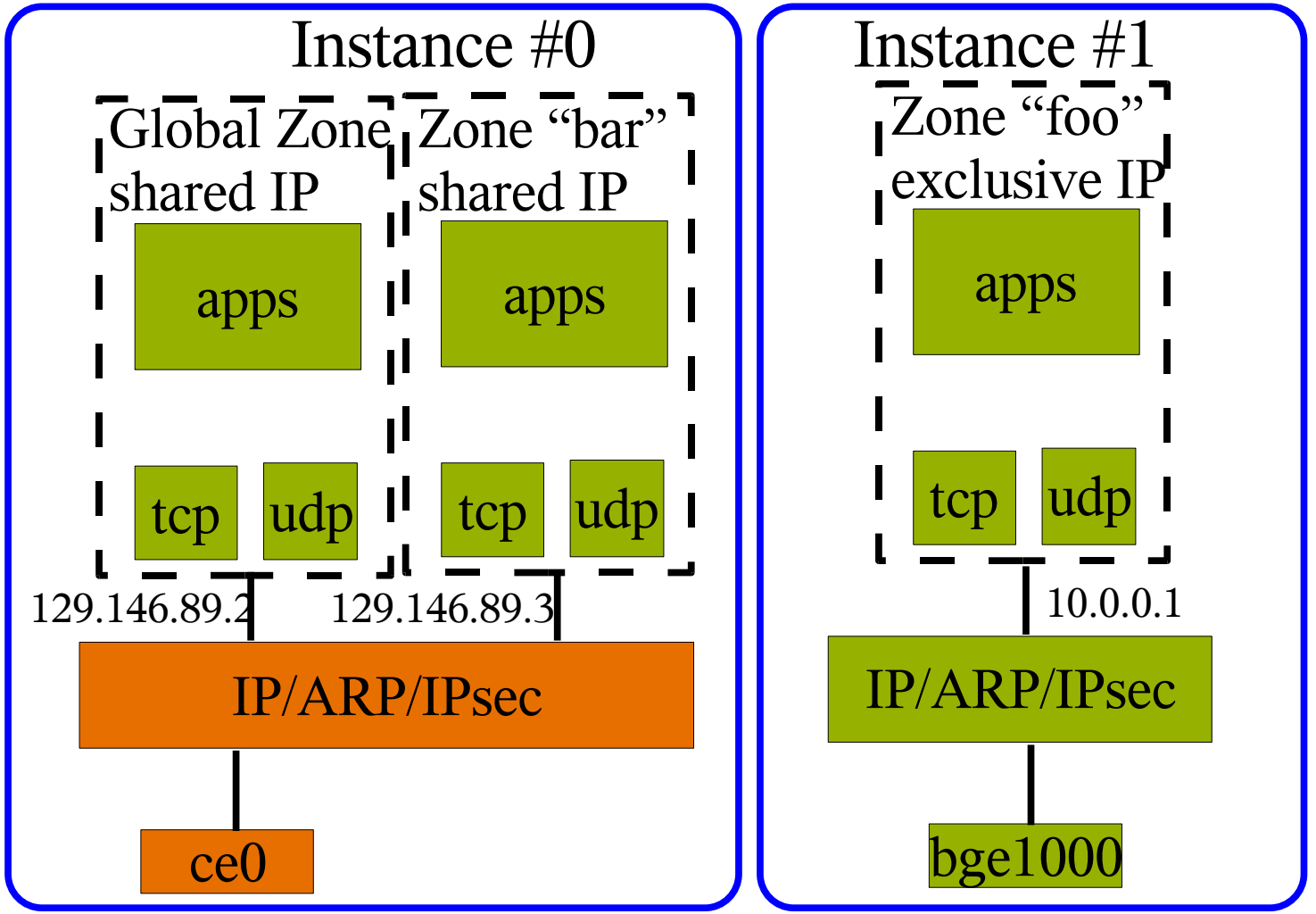
Networking in Zones

- IP addresses assigned to zones
  - A zone will only receive packets sent to its IP addresses
  - A zone can not sent packets using somebody elses IP address
  - Separate port number space per zone
  - Zone prohibited from ARP spoofing, snooping the network
- But a shared IP instance
  - Shared routing and forwarding
  - Shared IPsec and IP Filter configuration
  - Can't configure zones IP addresses using DHCP
- IP instances introduces ability for multiple IP instances – one for each zone

IP Instances for Zones - The Benefits of Separation

- Separate LANs or VLANs can be attached to different zones with no IP leakage between them
  - For instance, possible to have a management network separate from the data network
- Enables the use of IP-level features for zones
  - DHCP
  - IPsec
  - IP Filter
- Per instance network configuration (routing tables, transport tunable, etc)

IP Separation: Multiple IP Instances



Configuration of exclusive-IP zones

- zonecfg just specifies the physical datalink name
  - No IP address
- Inside the zone IP is configured the same as for global zone
  - sysidtools today
  - Including the ability to put a sysidcfg file in the zone's /etc directory before the first boot
  - Including the ability to use DHCP and IPv6 stateless address autoconfiguration for the zone
- No change to how shared-IP zones are configured

Network security and exclusive-IP zones

- The zone is given a datalink name (e.g., bge1000)
  - Compared to an IP address for a shared-IP zone
- The zone can do whatever it wants with that datalink
  - Includes snooping, stealing IP addresses, etc
  - But can not access other datalinks on the machine
- Very similar to having separate servers on separate LANs or VLANs
- Larger issue of how we prevent network threats with virtualization
  - For Xen and Zones
  - Prevent ICMP redirects, RIP, etc, etc
  - Possible future: GLD-level filtering as the general mechanism

Configuration – Static IP

```
bilen# zonecfg -z foo
zonecfg:foo> set zonepath=/export2/foo
zonecfg:foo> set ip-type=exclusive
zonecfg:foo> add net
zonecfg:foo:net> set physical=bge1000
zonecfg:foo:net> end
```

In /export2/foo/root/etc/sysidcfg put:

```
network_interface=bge1000
 {hostname=host_name
  default_route=10.0.0.1
  ip_address=10.1.2.3
  netmask=255.0.0.0
  protocol_ipv6=no}
```

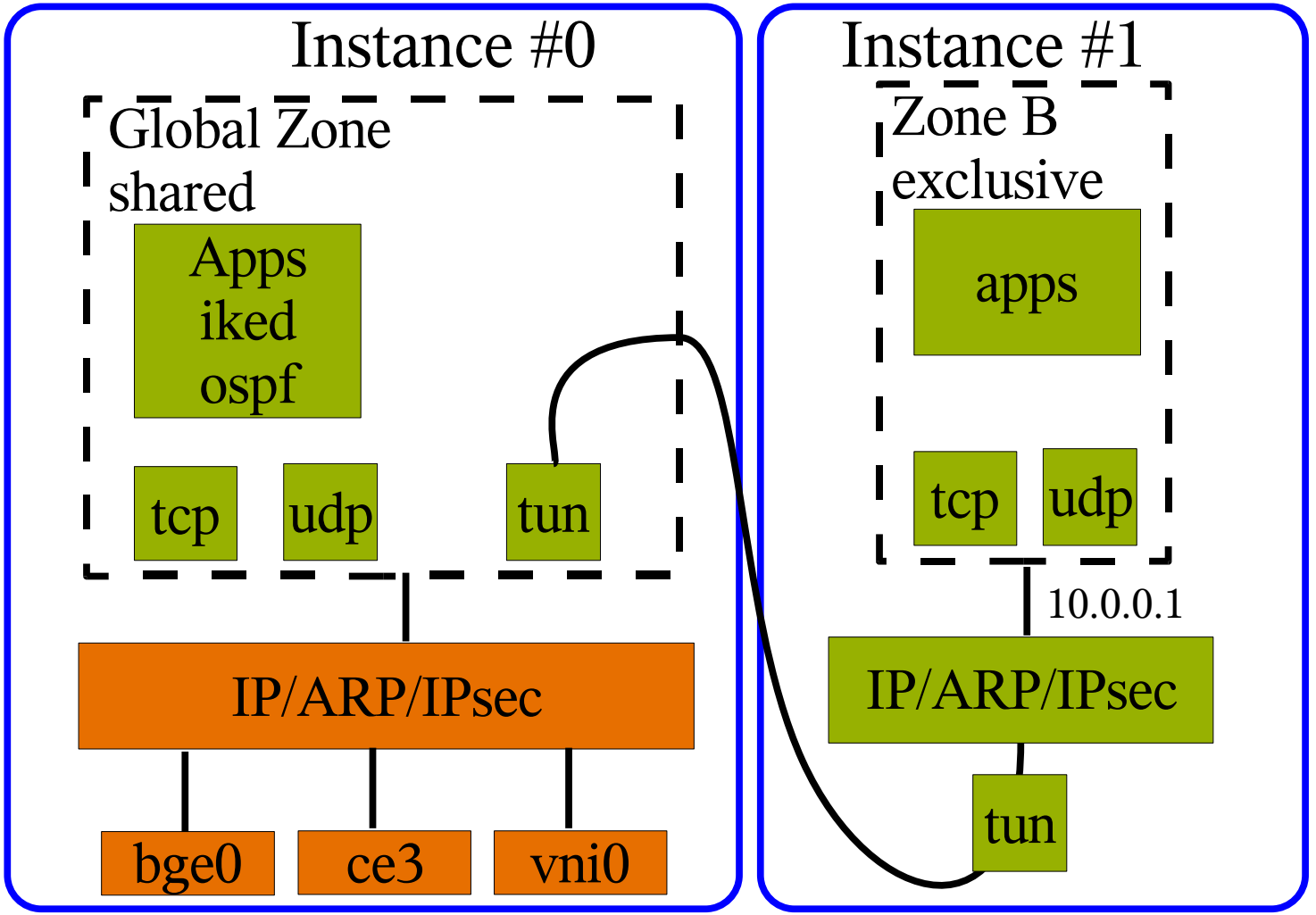
Configuration – DHCP + IPv6

```
bilen# zonecfg -z foo
zonecfg:foo> set zonepath=/export2/foo
zonecfg:foo> set ip-type=exclusive
zonecfg:foo> add net
zonecfg:foo:net> set physical=bge1000
zonecfg:foo:net> end
```

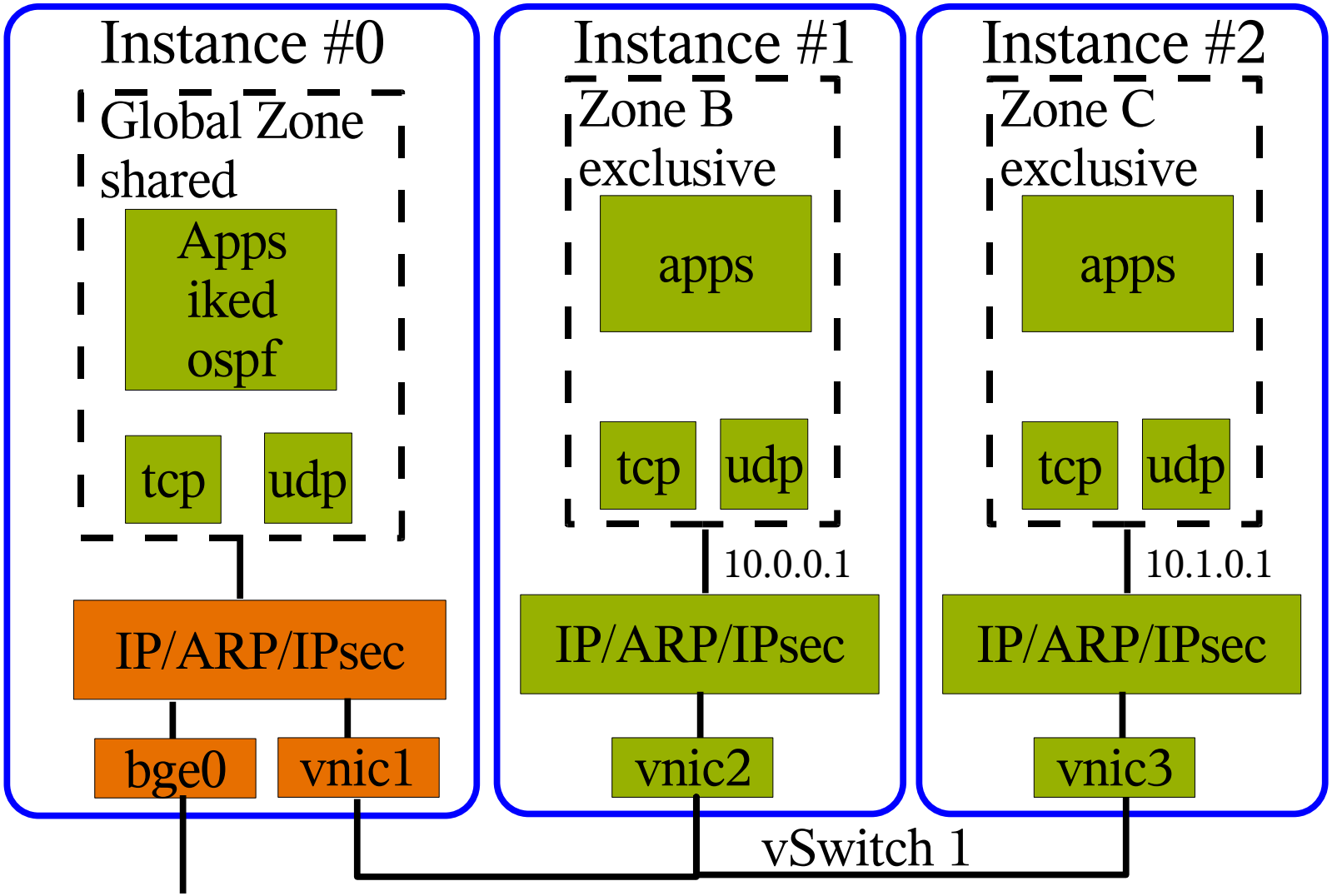
In /export2/foo/root/etc/sysidcfg put:

```
network_interface=bge1000
 {hostname=host_name
  dhcp protocol_ipv6=yes}
```

IPsec “Virtual LANs” - a possible future



Internal networking - vSwitch/VNICs



Ensuring Separation? How did we implement this?

- Solaris has separate kernel modules for the TCP/IP pieces
  - Due to old STREAMS philosophy, there is no data sharing outside of these modules
- The symbol table tells what data objects we have
  - nm /kernel/drv/arp | grep OBJT | grep -v undef
- Some of that data is read-only
  - Have to read the code to find out
- Take all the other data and convert it from global data to fields in an instance data structure

Example: arp module symbols

- ar_ce_hash_tbl
- ar_ce_mask_entries
- ar_cmd_tbl
- ar_g_head
- ar_g_nd
- ar_m_tbl
- ar_snmp_hash_size
- arl_g_head
- arp_param_arr
- arpinfo
- cb_inet_devops
- fsw
- inet_dev_info
- inet_devops
- info
- modldrv
- modlinkage
- modlstrmod
- netdev_privs
- rinit
- winit

Example: arp instance data structure

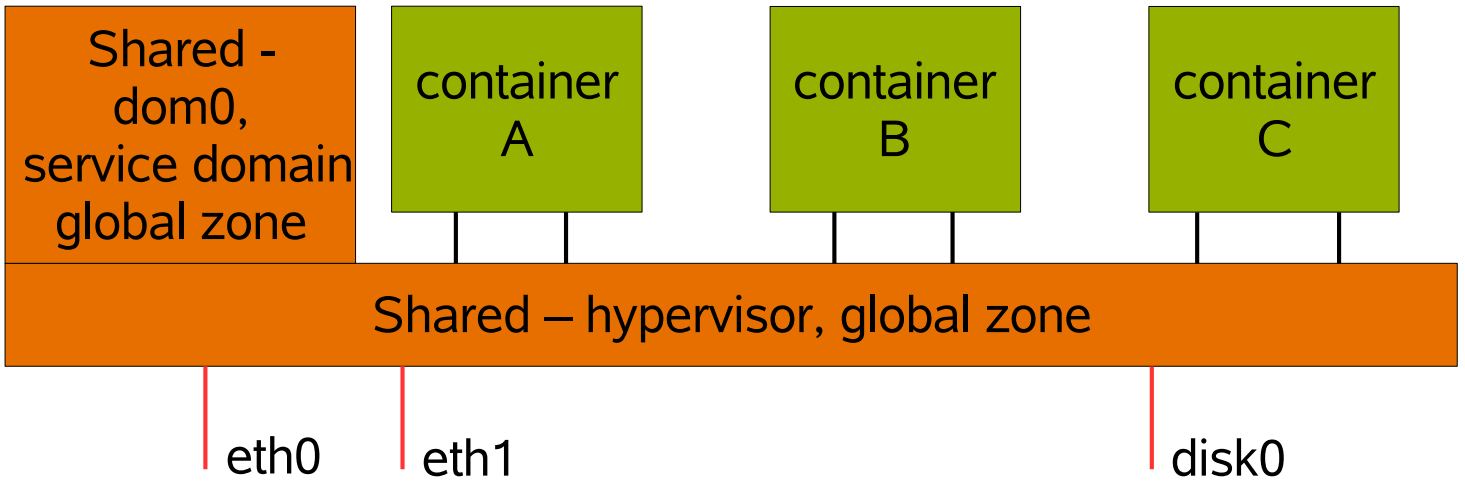
```
struct arp_stack {
  netstack_t    as_netstack;/ Common netstack */
  void          *as_head;
  caddr_t       as_nd;
  struct arl_s  *as_arl_head;
  arpparam_t    *as_param_arr;
  /* ARP Cache Entry Hash Table */
  ace_t         *as_ce_hash_tbl[ARP_HASH_SIZE];
  ace_t         *as_ce_mask_entries;
  int           as_snmp_hash_size;
};
```

Different types of Host Virtualization

- OS virtualization – zones (Solaris Containers)
  - A single OS kernel; separate user-space environments
  - Efficient – shared memory, single-level CPU scheduler, shared read-only file systems
  - Apply OS patches once
- Virtual Machines
  - LDOMS for Sparc, Xen for X64
  - Separate complete OS instances
  - Need separate physical memory
  - More flexibility – can run completely different Oses
- Common approach to Network Virtualization for both

Commonality across the types

- Applications run unmodified
- Single external “wire” to networking and storage
  - (appearance of) virtual devices in each container
- Some software controls the devices/wires



What services can dom0/global zone provide?

- Firewall filtering. For different reasons:
  - Prevent the container from sending bad things out on the network (attack the network)
  - Protect the container from network attacks - enforce security standards without touching the OS (e.g., disallow telnet, but allow ssh)
- Network resource control
  - Assign 10 Mbit/s network bandwith to container A.
- Network multipathing
  - Run link aggregation or OSPF-MP so that the container sees a functioning network across some network failures.
- NAT
  - For desktop/laptop usage where multiple containers are used

Summary

- IP Instances provide for network isolation when separate zones need to be connected to different (V)LANs
- Get IP feature set (DHCP, Ipsec, IP Filter) as a side effect
- With CrossBow VNICs and vSwitches we can build interesting network topologies inside the box
- You can join us an opensolaris.org

Questions?