

# Improvements to OpenVMS clustering and IP Cluster Interconnect (IPCI) - aka OpenVMS Cluster over IP

Nilakantan Mahadevan  
OpenVMS Engineering, HP



# Agenda

- IP Cluster Interconnect (IPCI) - aka OpenVMS cluster over IP
- Shadowing
  - Extended Membership (XMBRS)
  - Other improvements

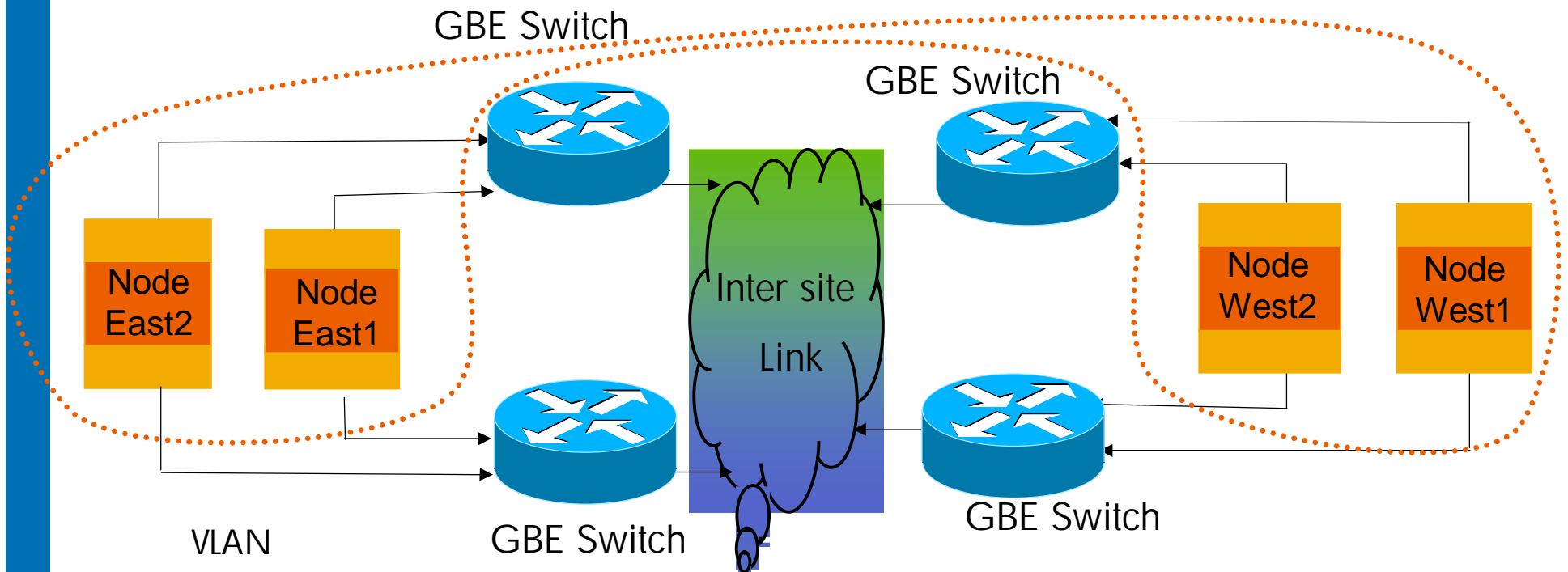
# IP Cluster Interconnect (OpenVMS cluster over IP)

- Introduction to OpenVMS Clustering Technology
- Cluster communication architecture.
- Need for Cluster over IP
- IPCI technical overview
- IPCI Demonstration.
- IPCI configuration details.
- Salient features of IPCI
- Customer advantage with IPCI

# Introduction to OpenVMS Clustering

- Reliability
  - HP OpenVMS is known as “gold standard” in disaster tolerance
- Scalability
  - Qualified for 96 nodes and also mixed architecture configuration (IPF-Alpha, Alpha-VAX)
  - OpenVMS supports clusters with up to 500 miles apart.
- Manageability
  - Shared-Everything model with Cluster wide file system
  - Single System Image, Cluster wide management facility

# Disaster Tolerant /Long Distance OpenVMS Clusters



LAN bridging/Extended LANs using switches

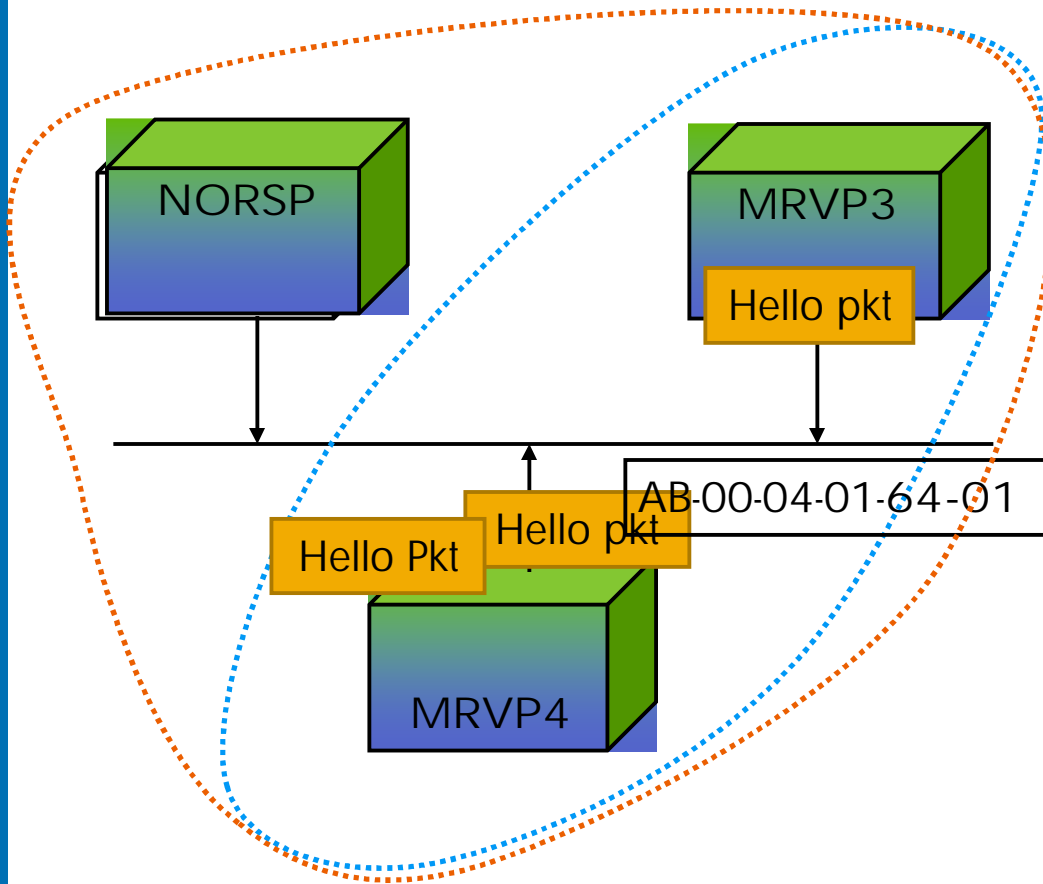
Nodes East1, East2, West1, West2 belong to same VLAN



# Current OpenVMS Cluster Interconnect Solution

- SCA (aka SCS) – System Communication architecture
  - Cluster communication protocol
- Cluster Interconnect
  - Alpha : LAN, Memory Channel, Shared Memory, CI
  - IPF (Integrity) :LAN
- LAN interconnect for long distance cluster communication
- Bridging and Extended LAN techniques for multi-site long distance clusters
- Nodes belong to same LAN/VLAN for cluster communications

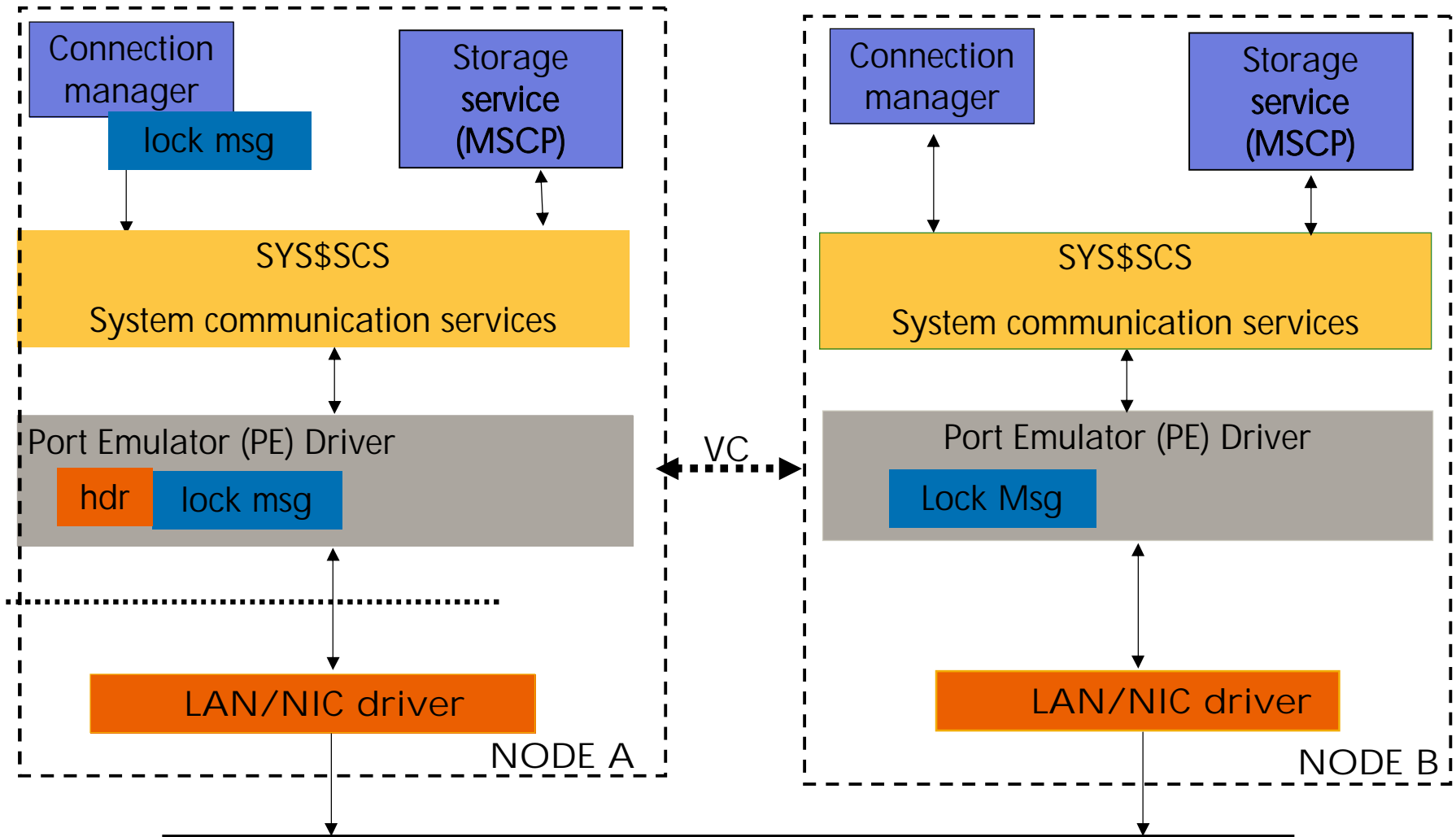
# OpenVMS Cluster Communication in LAN



MRVP3 and MRVP4 are in cluster  
Cluster group number 100  
The multicast address is  
AB-00-04-01-00-01 +  
the cluster group number  
AB-00-04-01-64-01

Configure NORSP as cluster and  
reboot NORSP

# OpenVMS Cluster Communication Architecture



- VC – Virtual Circuit consists of LAN Channels



# Port Emulator (PE) Driver

- Component that implements OpenVMS Cluster communications in LAN ( NISCA aka Network Interconnect System Communication Architecture)
- Transmits and receives datagram, sequenced messages and block transfer of data
- Multilayered architecture
- Consists of Transport layer, Channel Control Layer and Data Exchange layer



# Current Solution

- OpenVMS customers use LAN interconnect primarily for Cluster communications.
- NISCA protocol which is used for Cluster communication is LAN based.
- OpenVMS Nodes should be in same LAN or VLAN for cluster communications.
- To make OpenVMS cluster work between sites/Data center customers use bridging and Extended LAN techniques

# Technical Motivation for IPCI

- Network switches during higher loads can give priority to IP traffic than cluster (SCS) traffic
- Cluster instability during periods of heavy IP usage
- High router utilization for transporting cluster packets
- IP is de-facto industry standard
- Leveraging benefits of improvements in IP technology

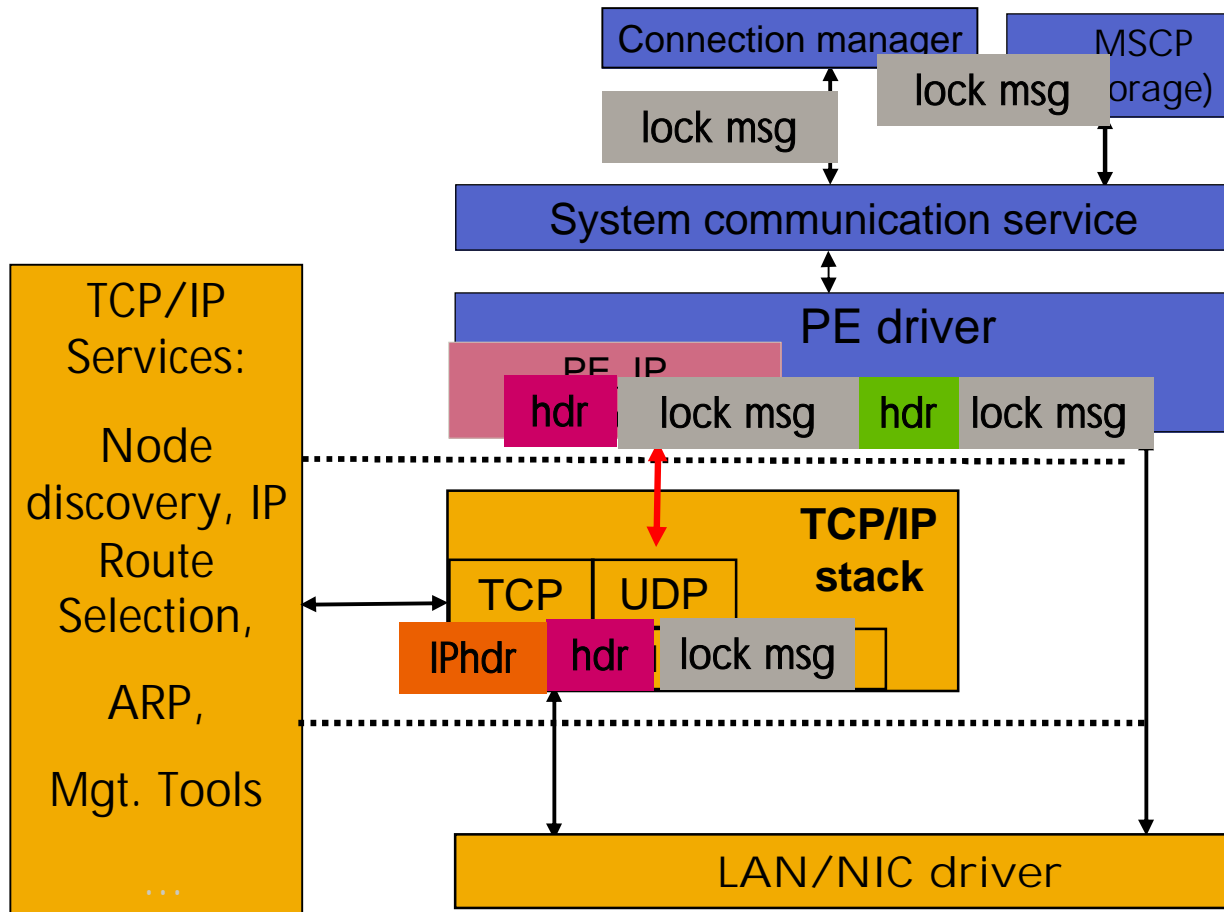
# Business/Market motivation

- Major Switch vendor dropping support for bridging
- Extra Cost/license for LAN bridging /layer 2 service.
- Corporate policies restricting scope of non-ip protocols
- Mandate of IP only network
- Specialized hardware/human resource costs for setting up multi site DT cluster with LAN bridging

# Solution: IP Cluster Interconnect (IPCI) aka OpenVMS Cluster over IP

- IPCI is the ability to make use of IP for OpenVMS clusters communications
- IPCI involves making OpenVMS cluster communication module (PE driver) to use IP services
- IPCI will coexist with LAN interconnect for Cluster communication
- IP unicast and optionally IP multicast for node discovery
- File based mechanism for unicast node discovery

# IPCI solution – PE driver over UDP



Major Parts:

PEdriver UDP Support:

TCP/IP Services boot time loading & initialization

 Existing Cluster Component

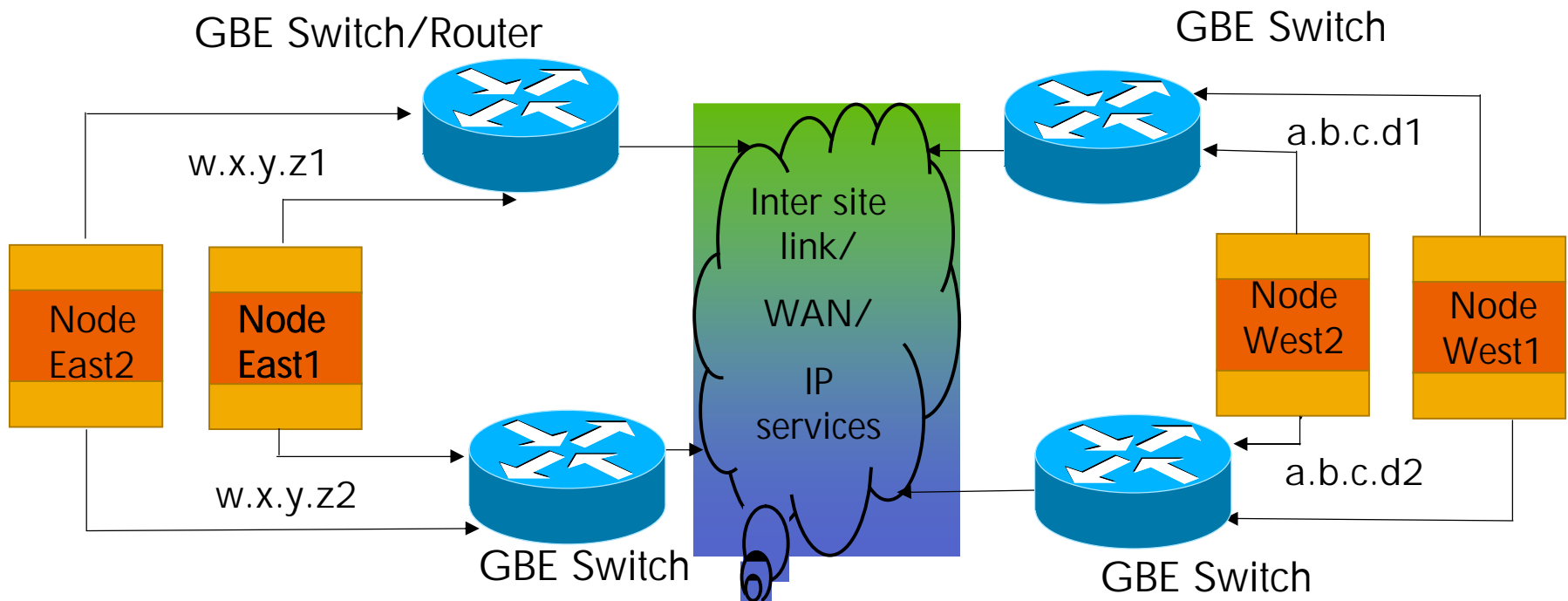
 New Component-Component interaction

 NEW PEdriver component

 Existing VMS/TCP/IP component



# Cluster using IPCI



- Node East1, East2, West1, West2 can be part of the same or different LAN for cluster communications using IPCI.
- East1 and West 2 has a Virtual Circuit (VC) VC consists of IP channels for SCS traffic

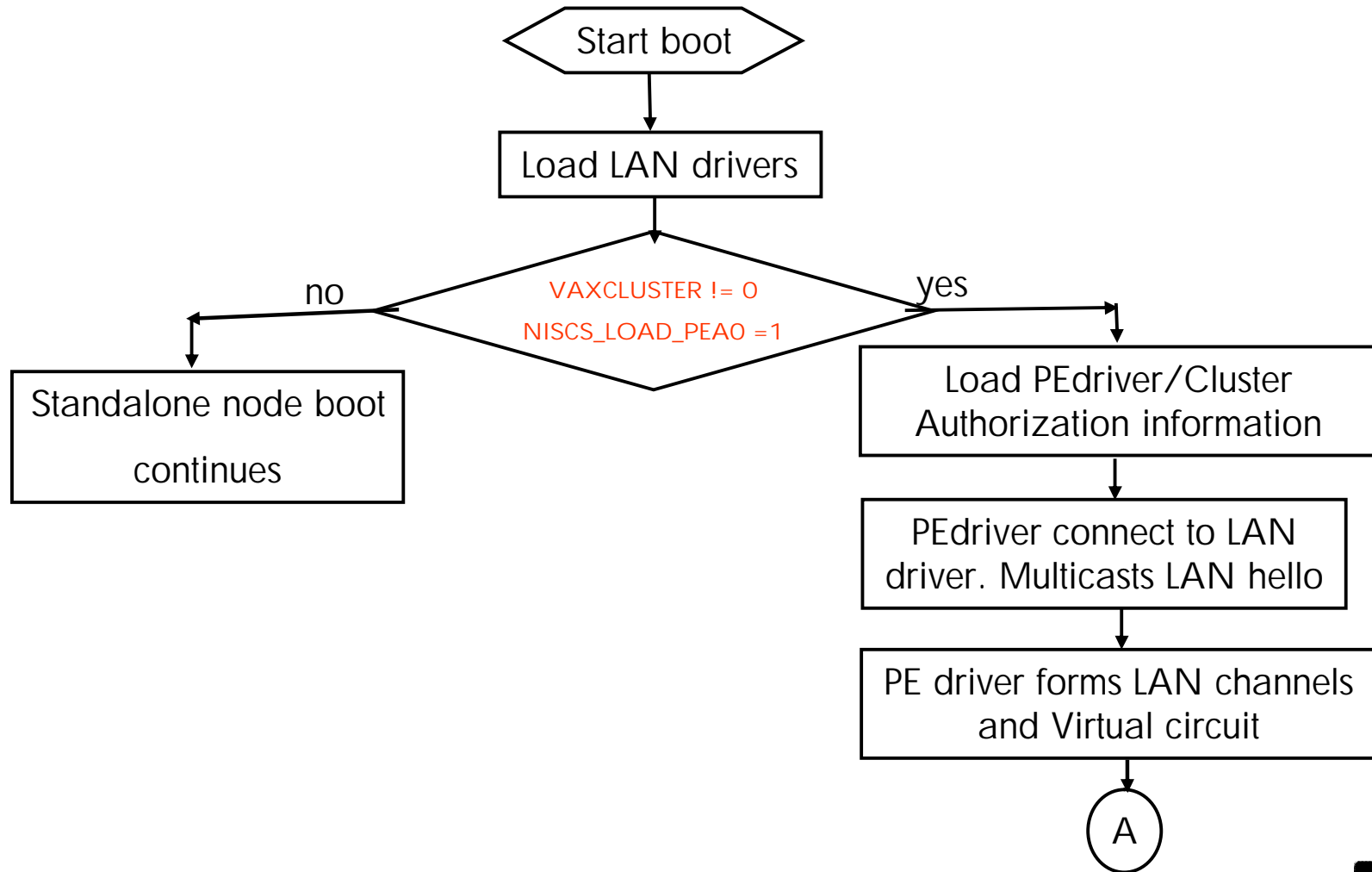
# PE driver over UDP

- The IP UDP service has the same packet delivery characteristics as 802 LANs
- PEdriver uses the IP based UDP datagram service as another LAN device
- Only the lowest layer of PEdriver has extension to locate and connect to the TCP/IP stack
- Uses Kernel mode Interface to talk to IP stack.
- Node discovery using IP Unicast through configuration file
- Alternate and Additional Mechanism: IP Multicast

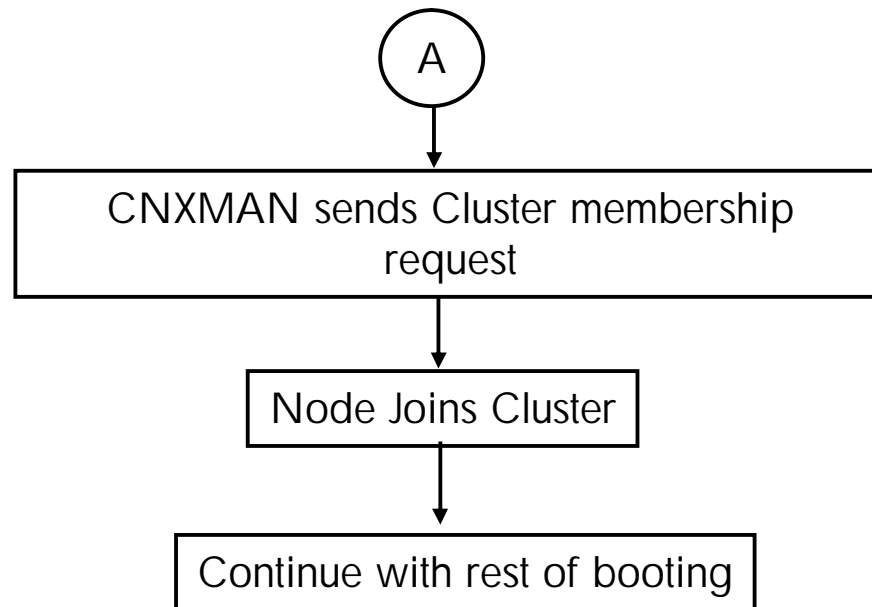
# TCP/IP Services boot time loading and Initialization

- Cluster communications are available in an IP only network environment
- Existing boot sequence – LAN, PE driver, TCP/IP
- Boot Sequence with IPCI – LAN, TCP/IP, PE driver
- Ability to make use of boot time configuration information to initialize TCP/IP services

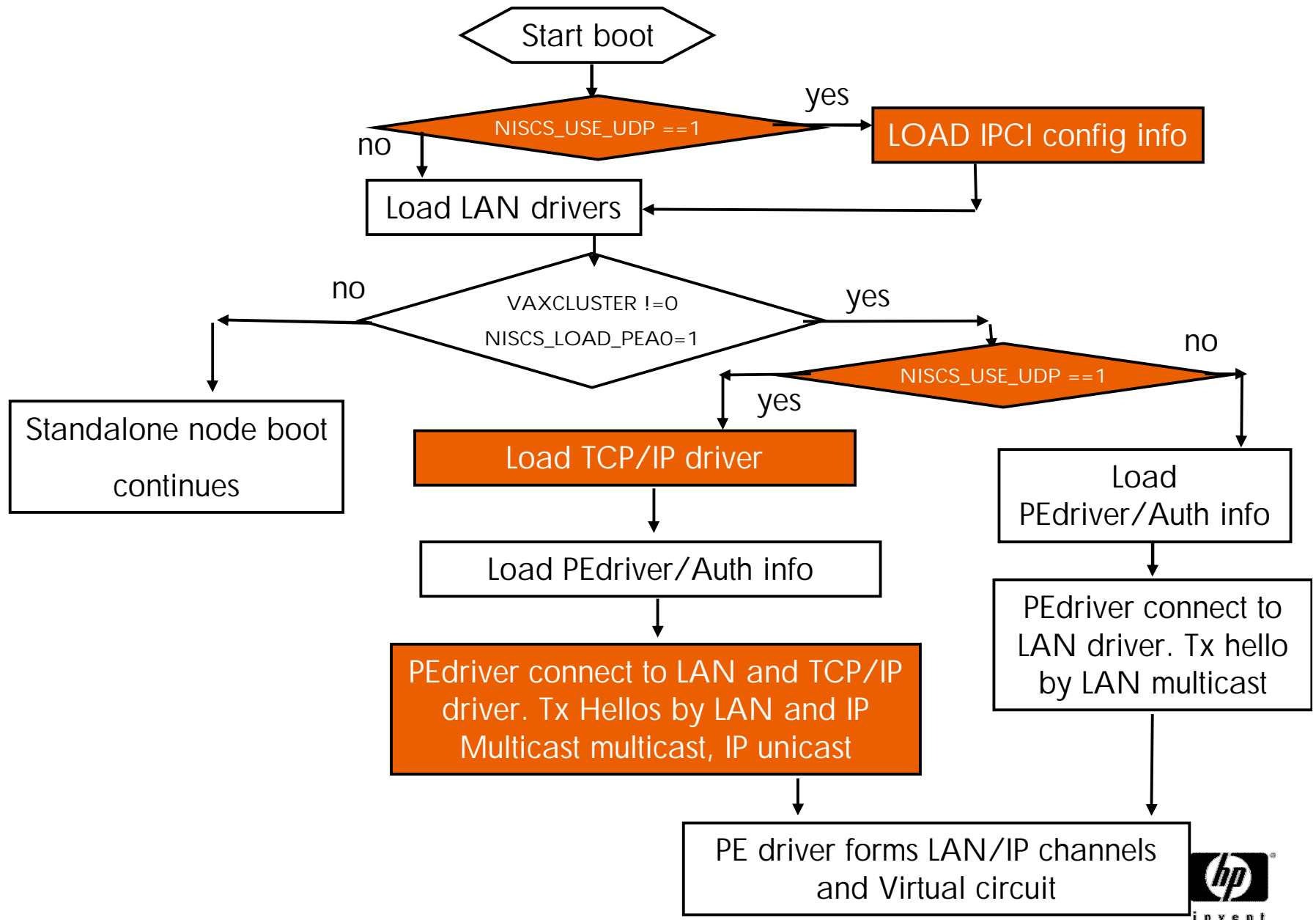
# OpenVMS Cluster formation (LAN)



# OpenVMS Cluster formation (LAN)



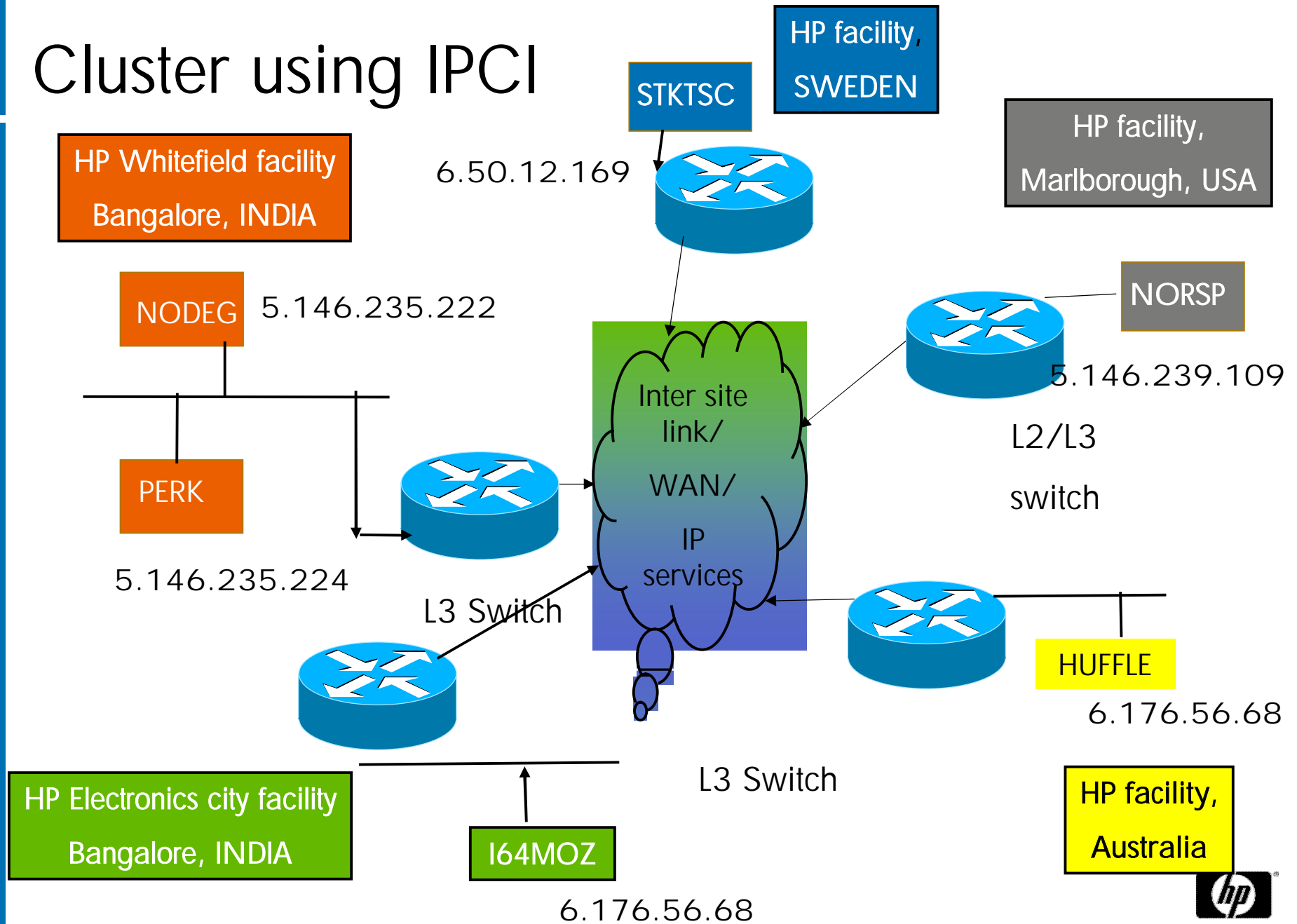
# OpenVMS Cluster formation (IPCI)



# Node Discovery – Unicast and Multicast

- IP unicast used for node discovery and hello packets.
- IP multicast can also be used (Administratively scoped IP multicast address)
- Remote nodes not in IP multicast domain use IP unicast technique to join Cluster and send hello packets.
- Basic Cluster principle: All nodes must see all.

# Cluster using IPCI



(E) TELNET (16.138.182.13) - PowerTerm 525

File Edit Terminal Communication Options Script Help

STOP [Icons]

SYSTEMS		MEMBERS
NODE	SOFTWARE	STATUS
I64MOZ	VMS XBXH-J2I	MEMBER
NODEG	VMS XBXH-J2I	MEMBER
HUFFLE	VMS XBYE-BL1	MEMBER
STKTSC	VMS XBYE-BL1	MEMBER
NORSP	VMS XBYD-J2I	MEMBER
PERK	VMS XBXI-BL1	MEMBER

\$  
\$ █

F1 F2 F3 F4 F5 F6 F7 F8 F9 F10 F11 F12

VT420-7 24:3 Caps Hold On Line

start [Taskbar icons]



# Cluster using IPCI

- 2 nodes in Bangalore Whitefield facility
- 1 node in Bangalore (India) Electronics city facility
- 1 node in Marlborough, USA (for demonstration)
- 1 node in HP Australia, HP SWEDEN (for demo)
- Distance between Bangalore facilities <50 miles. PING latency 4 ms. Cluster Latency approx 4ms.
- Distance approx 8000 miles. Ping Latency between Bangalore and Marlborough 350ms. Cluster Latency approx 350ms.

# Add a node into a IPCI cluster

\$ @SYS\$MANAGER:CLUSTER\_CONFIG\_LAN

Cluster Configuration Procedure  
CLUSTER\_CONFIG\_LAN Version V2.80  
Executing on an IA64 System

DECnet Phase IV is installed on this node.

IA64 satellites will use TCP/IP BOOTP and TFTP services for downline loading

TCP/IP is installed and running on this node.

Enter a "?" for help at any prompt. If you are familiar with the execution of this procedure, you may want to mute extra notes and explanations by invoking it with "@CLUSTER\_CONFIG\_LAN BRIEF".

This IA64 node is not currently a cluster member.

MAIN Menu

1. ADD I64MOZ to existing cluster, or form a new cluster.
2. MAKE a directory structure for a new root on a system disk.
3. DELETE a root from a system disk.
4. EXIT from this procedure.

1. ADD I64MOZ to existing cluster, or form a new cluster.
2. MAKE a directory structure for a new root on a system disk.
3. DELETE a root from a system disk.
4. EXIT from this procedure.

Enter choice [4]: 1

Is the node to be a clustered node with a shared SCSI/FIBRE-CHANNEL bus (Y/N)? n

IA64 node, using LAN for cluster communications. PEDRIVER will be loaded.

No other cluster interconnects are supported for IA64 nodes.

Enter this cluster's group number: 1985

Enter this cluster's password:

Re-enter this cluster's password for verification:

ENABLE IP for cluster communications (Y/N)? Y

UDP port number to be used for Cluster Communication over IP[49153]?

Enable IP multicast for cluster communication(Y/N)[Y]?

What is IP the multicast address[239.242.7.193]?

What is the TTL (time to live) value for IP multicast packets [32] ?

Do you want to enter unicast address(es)(Y/N)[Y]?

What is the unicast address[Press [RETURN] to end the list]? 6.118.162.109

What is the unicast address[Press [RETURN] to end the list]? 6.50.12.169

What is the unicast address[Press [RETURN] to end the list]? 6.176.56.68

What is the unicast address[Press [RETURN] to end the list]? 5.146.235.222

What is the unicast address[Press [RETURN] to end the list]? 6.138.182.6

What is the unicast address[Press [RETURN] to end the list]?

What is the unicast address[Press [RETURN] to end the list]?

\*\*\*\*\*

Cluster Communications over IP has been enabled. Now  
CLUSTER\_CONFIG\_LAN will run the SYS\$MANAGER:TCPIP\$CONFIG  
procedure. Please select the IP interfaces to be used for  
Cluster Communications over IP (IPCI). This can be done  
selecting "Core Environment" option from the main menu  
followed by the "Interfaces" option. You may also use  
this opportunity to configure other aspects.

\*\*\*\*\*

Press Return to continue ...

Checking TCP/IP Services for OpenVMS configuration database files.



## HP TCP/IP Services for OpenVMS Configuration Menu

Configuration options:

- 1 - Core environment
- 2 - Client components
- 3 - Server components
- 4 - Optional components
  
- 5 - Shutdown HP TCP/IP Services for OpenVMS
- 6 - Startup HP TCP/IP Services for OpenVMS
- 7 - Run tests
  
- A - Configure options 1 - 4
- [E] - Exit configuration procedure

Enter configuration option: 1

## HP TCP/IP Services for OpenVMS Core Environment Configuration Menu

Configuration options:

- 1 - Domain
- 2 - Interfaces
- 3 - Routing
- 4 - BIND Resolver
- 5 - Time Zone

A - Configure options 1 - 5

[E] - Exit menu

Enter configuration option: 2

## HP TCP/IP Services for OpenVMS Interface & Address Configuration Menu

Hostname Details: Configured=i64moz, Active=i64moz

Configuration options:

- 0 - Set The Target Node (Current Node: I64MOZ)
- 1 - IE0 Menu (EIA0: TwistedPair 100mbps)
- 2 - 6.138.182.6/24 i64moz Configured
- 3 - IE1 Menu (EIB0: TwistedPair 1000mbps)

[E] - Exit menu

Enter configuration option: 2

HP TCP/IP Services for OpenVMS Address Configuration Menu (Node: I64MOZ)

IEO 6.138.182.6/24 i64moz Configured IEO

Configuration options

- 1 - Change address
- 2 - Set "i64moz" as the default hostname
- 3 - Delete from configuration database
- 4 - Add to IPCI database
- 5 - Make active on live system
- 6 - Add standby aliases to configuration database (for failSAFE IP)

[E] - Exit menu

Enter configuration option: 4

Updated Interface in IPCI configuration file: SYS\$SYSROOT:[SYSEXE]TCPIP\$CLUSTER.DAT;

Added address IEO:6.138.182.6 to IPCI database

## HP TCP/IP Services for OpenVMS Interface & Address Configuration Menu

Hostname Details: Configured=i64moz, Active=i64moz

Configuration options:

- 0 - Set The Target Node (Current Node: I64MOZ)
- 1 - IEO Menu (EIAO: TwistedPair 100mbps)
- 2 - 6.138.182.6/24 i64moz Configured,IPCI
- 3 - IE1 Menu (EIBO: TwistedPair 1000mbps)

[E] - Exit menu

Enter configuration option: E

## HP TCP/IP Services for OpenVMS Configuration Menu

Configuration options:

- 1 - Core environment
- 2 - Client components
- 3 - Server components
- 4 - Optional components
  
- 5 - Shutdown HP TCP/IP Services for OpenVMS
- 6 - Startup HP TCP/IP Services for OpenVMS
- 7 - Run tests
  
- A - Configure options 1 - 4
- [E] - Exit configuration procedure

Enter configuration option: E

Will I64MOZ be a boot server [Y]?N

Enter a value for I64MOZ's ALLOCLASS parameter [17]:

Does this cluster contain a quorum disk [N]?

# SYSSYSTEM:PE\$IP\_CONFIG.DAT

## Configuration File for IPCI.

! CLUSTER\_CONFIG\_LAN creating for CHANGE operation on 8-NOV-2008 10:46:19.26

multicast\_address=239.242.7.193

ttl=32

udp\_port=49153

unicast=6.118.162.109

unicast=6.50.12.169

unicast=6.176.56.68

unicast=5.146.235.222

unicast=6.138.182.6

# SYSSYSTEM:PE\$IP\_CONFIG.DAT

- Generated by CLUSTER\_CONFIG\_LAN.COM
- Read early in the boot sequence.
- Provides information to PEdriver.
- Can be common through out cluster.
- Remote node IP address should be present in local node PE\$IP\_CONFIG.DAT in order to allow remote node join the cluster.
- Best practice for IP unicast: Include all IP address and have one copy throughout the cluster.
- "\$MC SCACP reload" to be used to refresh IP unicast list on a live system.

# SYSSYSTEM:TCPIP\$CLUSTER.DAT

Configuration File for IPCI.

- default\_route=6.138.182.1
- interface=IE0,EIA0,6.138.182.6,255.255.255.0

# SYS\$SYSTEM:TCPIP\$CLUSTER.DAT

- Generated by TCPIP\$CONFIG which is invoked by CLUSTER\_CONFIG\_LAN.COM
- Read early in the boot sequence.
- Provides information to PEdriver to use the correct TCP/IP interface (WE0 OR WE1) for Cluster traffic.
- Provides information to TCP/IP stack to initialize the interface with IP address and default route.

# Console Messages

HP OpenVMS Industry Standard 64 Operating System, Version XBXH-J2I

© Copyright 1976-2008 Hewlett-Packard Development Company, L.P.

%DECnet-I-LOADED, network base image loaded, version = 05.16.00

%VMScIuster-I-LOADIPCICFG, loading the IP cluster configuration files

%VMScIuster-S-LOADEDIPCICFG, Successfully loaded IP cluster configuration files

%SMP-I-CPUTRN, CPU #1 has joined the active set.

%SYSINIT-I- waiting to form or join an OpenVMS Cluster

%VMScIuster-I-LOADSECDB, loading the cluster security database

%EIA0, Auto-negotiation mode assumed set by console

# Console Messages

```
%EWEO, Link up: 1000 mbit, full duplex, flow control (txrx)
%EWDO, Link up: 1000 mbit, full duplex, flow control (txrx)
%PEAO, Configuration data for IP clusters found
%PEAO, IP Multicast enabled for cluster communication, Multicast address, 239.242.7.193
%PEAO, Cluster communication enabled on IP interface, WEO
%PEAO, Successfully initialized with TCP/IP services
%PEAO, Remote node Address, 6.138.185.6,!INDIA added to unicast list of IP bus, IEO
%PEAO, Remote node Address, 5.146.235.222, !INDIA added to unicast list of IP bus, IEO
%PEAO, Remote node Address, 5.146.239.109,!USA added to unicast list of IP bus, IEO
%PEAO, Remote node Address, 5.146.235.224,INDIA added to unicast list of IP bus, IEO
%PEAO, Remote node Address, 6.176.56.68,AUSTRALIA added to unicast list of IP bus, IEO
%PEAO, Remote node Address, 6.50.12.169 ! SWEDEN, added to unicast list of IP bus, IEO
%PEAO, Hello sent on IP bus IEO
%PEAO, Cluster communication successfully initialized on IP interface , IEO
%CNXMAN, Sending VMSccluster membership request to system NORSP
%CNXMAN, Now a VMSccluster member -- system I64MOZ
%STDRV-I-STARTUP, OpenVMS startup begun at 29-OCT-2008 15:20:41.10
```

# SCACP commands

```
$ MC SCACP SHOW CHANNEL PERK
```

```
NODEG PEA0 Channel Summary 10-MAY-2008 05:09:51.38:
```

Remote Node	LAN/Dev	IP/Dev	Channel State	ECS state	Buffer Size	Delay uSec	Packets (S+R)
PERK	WE0	IE0	Open	N(T,I,F)	1394	708.9	30410576
PERK	EWA	EIA	Open	Y(T,P,F)	1426	551.9	26586732
PERK	WE0	IE1	Open	N(T,I,F)	1394	784.2	38475399
PERK	EWA	EIB	Open	Y(T,P,F)	1426	572.4	23780101
PERK	EIA	EIB	Open	Y(T,P,F)	1426	694.1	15288091



IP channels between nodes



LAN channels between nodes



# NEW SCACP commands

\$ MC SCACP SHOW CHANNEL <nodename>/IP and /LAN

\$ MC SCACP SHOW IP\_INTERFACE <ip\_interface>

Example. MC SCACP SHOW IP\_INTERFACE we0

\$ MC SCACP START IP\_INTERFACE <ip\_interface>

Example. MC SCACP START IP WE0

\$MC SCACP START IP\_INTERFACE <ip\_interface>

\$ MC SCACP SET IP\_INTERFACE <ip\_interface>

Example MC SCACP SET IP\_INTERFACE WE0/PRIORITY=4

\$MC SCACP STOP IP\_INTERFACE <ip\_interface>

Example MC SCACP STOP IP\_INTERFACE WE0

\$ MC SCACP RELOAD

# Virtual Circuit (VC)

- VC is logical connection between two nodes
- VC consists of Channels (LAN or IP)
- VC check summing and compression applicable to Cluster over IP also.

# Equivalent Channel Set (ECS)

- Set of Eligible Channels used by PEdriver for communication
- ECS technique applicable to IP channels also
- Tight (T), Peer (P), Fast(F) characteristics are required for a channel to be part of ECS
- TPF characteristics applicable to IP channels also.
- LAN channels preferred over IP channels

# Distance and Latency

- Current Distance limitation will still be applicable.
- Speed of Light causes approx 1 millisecond for 50 mile roundtrip.
- Distance more than 500 miles require site specific configuration and we suggest contact HP DTCS. (HP Disaster Tolerant Cluster services) or Product management.

# Security

- Normal intranet and Internet Security principles
- VPN (virtual Private Network)
- TTL (Time to live)
- Firewalls.

# Performance

- Engineering will conduct some performance test to recommend configurations for optimal performance.
- PING – PONG tests will be conducted and recommendations will be based on the results.
- Observations reveal tcp/ip ping latency close to latency reported by PEdriver

# Feature details

- Available with OpenVMS V8.4 only (Alpha and Integrity)
  - Will be available with V8.4 Field Test
  - No Prior version support
- Requires HP TCP/IP services for OpenVMS V5.7
  - Not available with other TCP/IP stacks at this time
  - Initial release supports IPv4 only; no IPv6
  - Requires static IP addresses and IP Unicast; optionally uses IP Multicast
- Coexists with LAN interconnect for Cluster communication
- Support for Satellite nodes included
- Existing intra-node distance/latency limitation (500 miles) applies

# Salient Features

- Ability to Discover nodes and form cluster in an IP only network
- Perform rolling upgrades to the new version without a cluster reboot
- Dynamically load balance using all the available healthy interconnect interfaces
- Interoperability with prior versions
- Delay measurement technique for discovering path with minimal latency

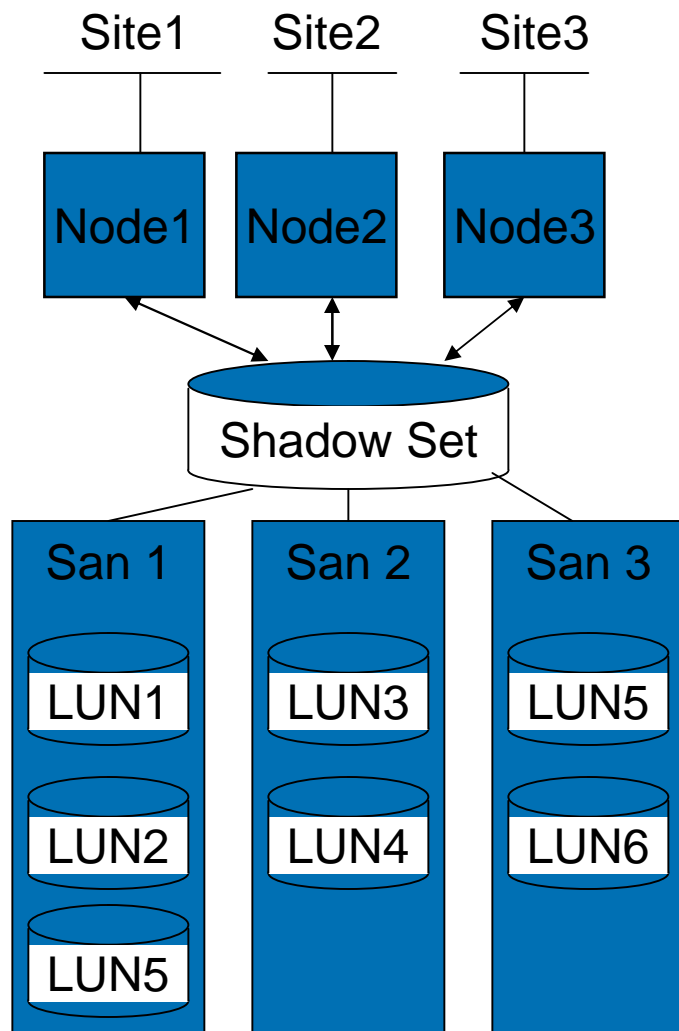
# Customer Advantage

- Only IP services (No LAN bridging) are provided by some Telco Vendors. Customers can now use OpenVMS clusters with IPCI
- Lower infrastructural and Operational costs
- No extra license/cost for LAN bridging (Layer 2 service)
- Leverage the benefits from the improvements in IP and LAN interconnect technology
- One network infrastructure in Data center for all purposes

# Shadowing

- Shadowing Extended Membership
  - Increases the number of member disks in a host based volume shadowing set from 3 to 6 disks

# OpenVMS Clusters



- 6 Member Shadow Sets
  - Use case 1: Backup without putting redundancy at risk
  - Use Case 2: 3 Site Cluster

# XMBRS

show dev DSA6

Device Name	Device Status	Error Count	Volume Label	Free Blocks	Trans Count	Mnt Cnt
DSA6:	Mounted	0	XMBRS	49954	1	1
\$1\$DGA1:	(CSGF2)	ShadowSetMember	0	(member of DSA6:)		
\$1\$DGA2:	(CSGF2)	ShadowSetMember	0	(member of DSA6:)		
\$1\$DGA3:	(CSGF2)	ShadowSetMember	0	(member of DSA6:)		
\$1\$DGA4:	(CSGF2)	ShadowSetMember	0	(member of DSA6:)		
\$1\$DGA5:	(CSGF2)	ShadowSetMember	0	(member of DSA6:)		
\$1\$DGA6:	(CSGF2)	ShadowSetMember	0	(member of DSA6:)		

# XMBRS

- To turn it on, just specify more than 3 members for the shadow set

```
$ MOUNT/SHADOW DSA16 -
```

```
_ $ /SHADOW=($1$dga1, $1$dga2, $1$dga3,  
$1$dga4) -
```

```
_ $ DSA6
```

- No new qualifiers
- No changes to DISMOUNT

# XMBRS Compatibility

- Mixed version, 3 member shadow sets will continue to work
- To use 4 or more members, all systems that have the VU MOUNTed, must have the new software
- MOUNT/INCLUDE of a once XMBR virtual unit on an older version may not find all of the members

# XMBRS Performance

- Performance testing in progress
- Reads could be faster
- Writes \*WILL\* be slower
  
- Important to use SET SHADOW commands
  - Noticeable difference with /COPY\_SOURCE

# Write Bitmap Enhancements

- Fix buffered mode transitions
- Multicast “set bit” messages
- Optimize sequential writes

# Field test Kit

- Field test with OpenVMS 8.4 field test
- Scheduled for Feb 2009
- Contact Details
  - Engineering Manager: Jim Lanciani  
([Jim.lanciani@hp.com](mailto:Jim.lanciani@hp.com))
  - Product Manager: Leo Demers  
([Leo.demers@hp.com](mailto:Leo.demers@hp.com))