Hewlett Packard
Enterprise

Assessing Your OpenVMS Cluster Interconnect

Part 1: Health

Keith Parris
Engineer, Hewlett Packard Enterprise

Factors in Cluster Interconnect Health

– LANCP network counters and device startup logs

– SCACP cluster counters and metrics

– High Availability configuration (redundancy)

– Monitoring of redundant link status

LANCP Counters and Logs

LANCP> SHOW DEVICE /COUNTERS

```
LANCP> SHOW DEVICE Exc /COUNTERS
NODE1 Device Counters EIA0 (25-SEP-2016 21:51:16.08):
                  Value  Counter
                  -----  -------
...
                      0 Unavailable station buffers
                      0 Unavailable user buffers
                      0 Alignment errors
                      0 Frame check errors
                      0 Frame size errors
                      0 Frame status errors
                      0 Frame length errors
                      0 Frame too long errors
                      0 Data overruns
                      0 Send data length errors
                      0 Receive data length errors
                      0 Transmit underrun errors
                      0 Transmit failures
                      0 Carrier check failures
                      0 Station failures
                      0 Initially deferred packets sent
                      0 Single collision packets sent
                      0 Multiple collision packets sent
                      0 Excessive collisions
                      0 Late collisions
                      0 Collision detect check failures
                      1 Link up transitions (24-AUG-2016 15:28:02.09)
                      0 Link down transitions
                   None Time of last generic transmit error
                   None Time of last generic receive error
```

LANCP Counters and Logs

LANCP> SHOW DEVICE /INTERNAL_COUNTERS

```
LANCP> SHOW DEVICE Exc /INTERNAL_COUNTERS
NODE1 Device Internal Counters EIA0 (25-SEP-2016 21:57:22.03):
                  Value  Counter
                  -----  -------
...
               000002A4  Requested link state <FlowControl, Fdx, 1000 mb,
                         Auto-negotiation>
               000002A5  Current link state <FlowControl, Fdx, 1000 mb,
                         Auto-negotiation, Link up>
...
             0:00:18.66  Total link downtime
             0:00:02.00  Shortest link uptime period acceptable
...
                         --- Driver Messages ---
24-AUG-2016 15:28:00.04  Link up: 1000 mbit, full duplex, flow control (receive
only)
```

Watch for possible duplex mismatch warnings.

SCACP Counters and Metrics

Transmit / Timeout Ratio

- For each Virtual Circuit, PEDRIVER tracks:
  - The average time it takes to acknowledge a sequenced packet
    - If a packet hasn't been acknowledged in twice the average time, PEDRIVER will assume it has been lost, and proactively retransmit it, to minimize the potential performance impact
  - The number of sequenced packets transmitted
  - The number of sequenced packets which time out
  - The ratio of sequenced packets transmitted divided by the number of such timeouts
- The OpenVMS Cluster Software Product Description says:
  - "The average packet-retransmit timeout ratio for OpenVMS Cluster traffic on the LAN from any system to another must be less than 1 timeout in 1000 transmissions."

SCACP Counters and Metrics

Transmit / Timeout Ratio

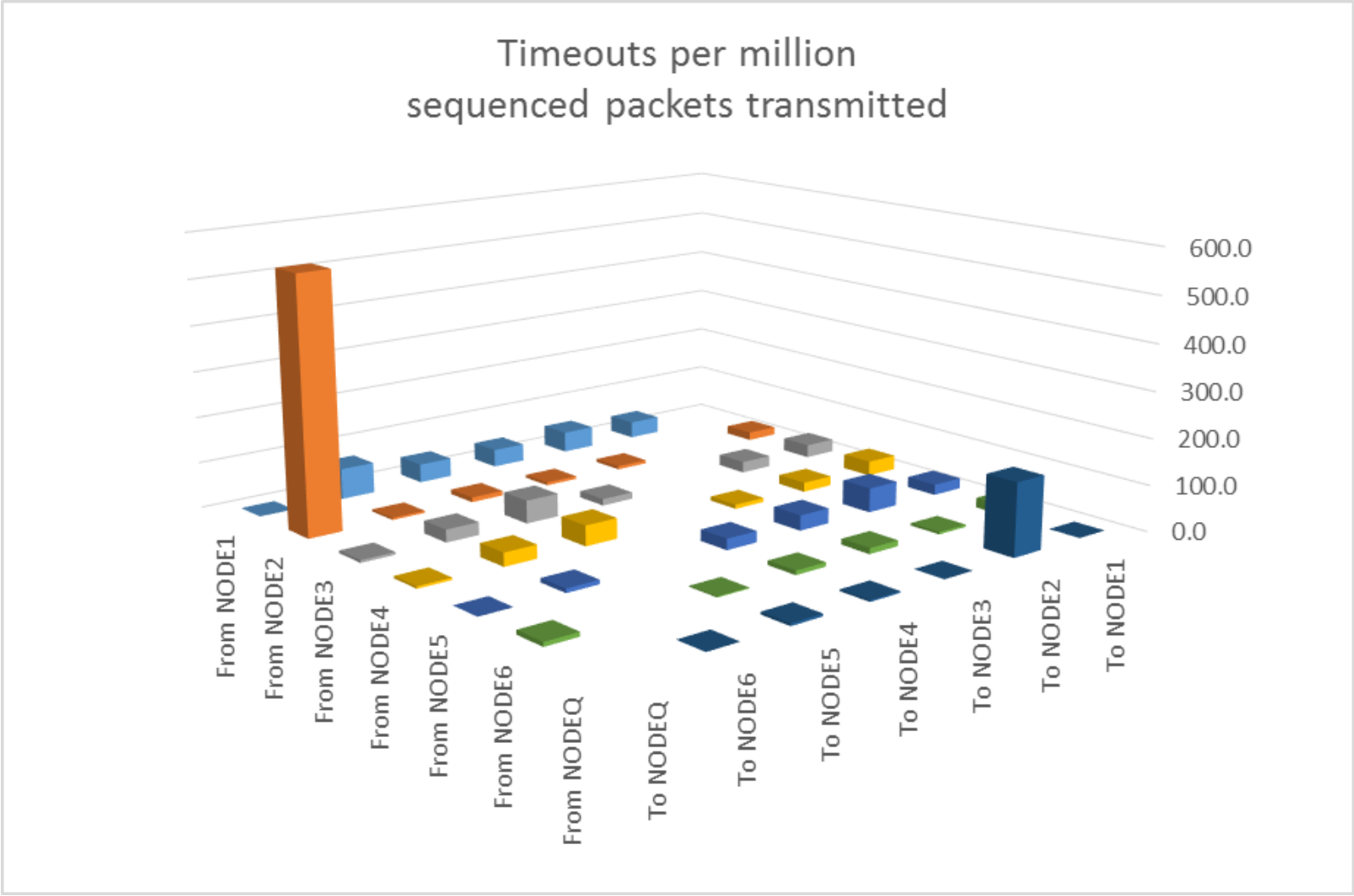
```
SCACP> SHOW VC

NODE1 PEA0 VC Summary  5-SEP-2016 16:44:08.94:
```

Remote Node	VC State	Total Errors	Xmt:TMO	Channels Open	Channels ECS	ECS Pri	MaxPkt Size	ReXmt TMO(uSec)	--XmtWindow-- Cur	Max	Mgt	Xmt Options	Total Pkts(S+R)	Most Recent VC Opened Time	Most Recent VC Closed Time
NODEQ	Open	0	Infinite	2	2	0	1426	756321.4	73	128	0		409949	03-SEP 09:22:12.96	(No time)
NODE6	Open	3028	14847	2	2	0	1426	17580.5	128	128	0		124743326	03-SEP 09:17:43.67	(No time)
NODE5	Open	2396	24306	2	2	0	1426	54867.4	128	128	0		153002334	03-SEP 09:17:42.66	(No time)
NODE4	Open	4247	25567	2	2	0	1426	2891.5	128	128	0		231996292	03-SEP 09:17:40.49	(No time)
NODE3	Open	2586	21708	2	2	0	1426	5455.2	128	128	0		144285982	03-SEP 09:17:32.83	(No time)
NODE2	Open	2514	26897	2	2	0	1426	17370.2	47	128	0		169211170	03-SEP 09:17:32.83	(No time)
NODE1	Open	0	Infinite	1	1	0	1426	3000000.0	1	8	0		5	03-SEP 09:17:28.65	(No time)

SCACP Counters and Metrics

Transmit / Timeout Ratio



SCACP Counters and Metrics

Transmit / Retransmit Ratio

- For each Channel (Path), PEDRIVER tracks:
  - The number of sequenced packets transmitted
  - The number of sequenced packets which must be retransmitted
  - The ratio of sequenced packets transmitted divided by the number of retransmits
- The OpenVMS Cluster Software Product Description says:
  - “The average packet-retransmit timeout ratio for OpenVMS Cluster traffic on the LAN from any system to another must be less than 1 timeout in 1000 transmissions.”

SCACP Counters and Metrics

Transmit / Retransmit Ratio

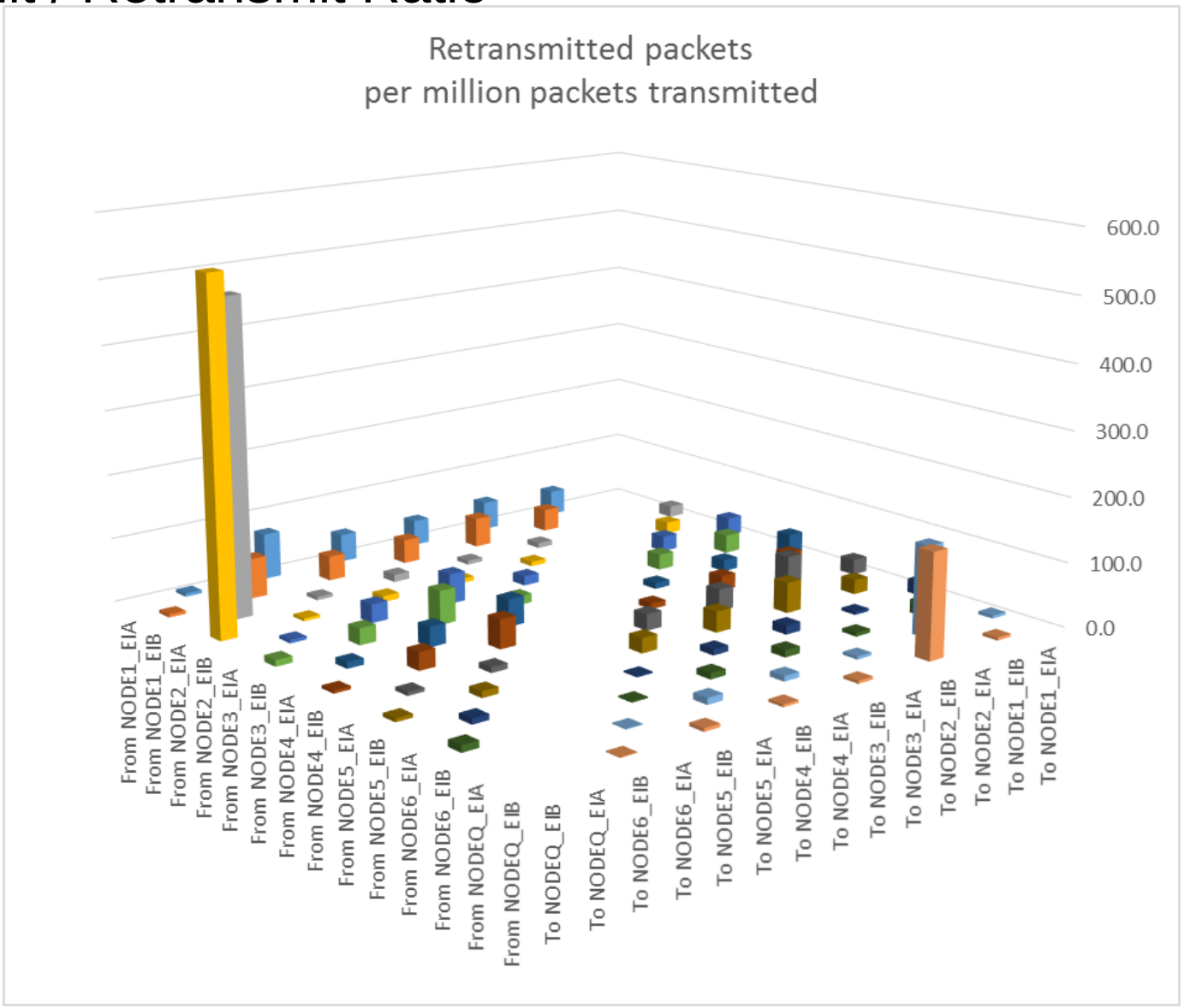
```
SCACP> SHOW CHANNEL /COUNTERS

NODE1 PEA0 Channel Counters and Errors  5-SEP-2016 16:44:08.95:
```

Remote Node	Device Loc	Device Rmt	-- Transmit -- Messages	-- Transmit -- Bytes	-- Receive -- Messages	-- Receive -- Bytes	Xmt:Rexmit	Rexmit Errors	TransmitFail Penalties	Receive Errors	Other Errors
NODEQ	EIA	EIA	194045	22138929	305392	37587038	194045	1	0	0	0
NODEQ	EIB	EIB	200057	23082418	303599	37383942	200057	1	0	0	0
NODE6	EIA	EIA	22513988	532278900	40013589	355380837	13898	1620	0	0	0
NODE6	EIB	EIB	22513013	531182295	40013687	355114230	15967	1410	0	0	0
NODE5	EIA	EIA	29146855	2013508884	47496370	2049511410	22860	1275	0	0	0
NODE5	EIB	EIB	29146943	2017471213	47497592	2050665097	25955	1123	0	0	0
NODE4	EIA	EIA	54316651	3282297220	61818239	804781227	24746	2195	0	0	0
NODE4	EIB	EIB	54316237	3282388311	61818901	804438553	26444	2054	0	0	0
NODE3	EIA	EIA	28092240	1770826373	44182982	1287060229	21999	1277	0	0	0
NODE3	EIB	EIB	28092358	1775735873	44184219	1286546028	21428	1311	0	0	0
NODE2	EIA	EIA	33843756	3026485654	50915674	2760293609	25542	1325	0	0	0
NODE2	EIB	EIB	33842879	3025006978	50915955	2760101485	28416	1191	0	0	0
NODE1	LCL	LCL	132118	13741550	217085	22578118	Infinite	0	0	0	0

SCACP Counters and Metrics

Transmit / Retransmit Ratio



Hewlett Packard Enterprise

SCACP Counters and Metrics

PEDRIVER Transmit Window

- For each Virtual Circuit, PEDRIVER keeps a Current and Maximum Transmit Window Size
  - Current Transmit Window Size can grow and shrink over time:
    - Starts out at 1
    - Grows with successful acknowledgements of sequenced packets
      - Tends to grow fairly slowly over time
      - The more the traffic, the faster it can grow
    - Cut back if packets need to be retransmitted, either because
      - Sequenced packet itself is lost, or
      - Acknowledgement for the sequenced packet is lost on the way back
    - If 1 retransmission occurs, Current transmit window is cut back to ½ of Maximum
    - If multiple retransmissions occur in a short timeframe, Current transmit window is cut back down to 1
- Ideally, we'd like to see the Current Transmit Window Size grow to and stay at the Maximum
- Current Transmit Window sizes less than the Maximum represent throttled-back performance and negative performance impact

SCACP Counters and Metrics

PEDRIVER Transmit Window

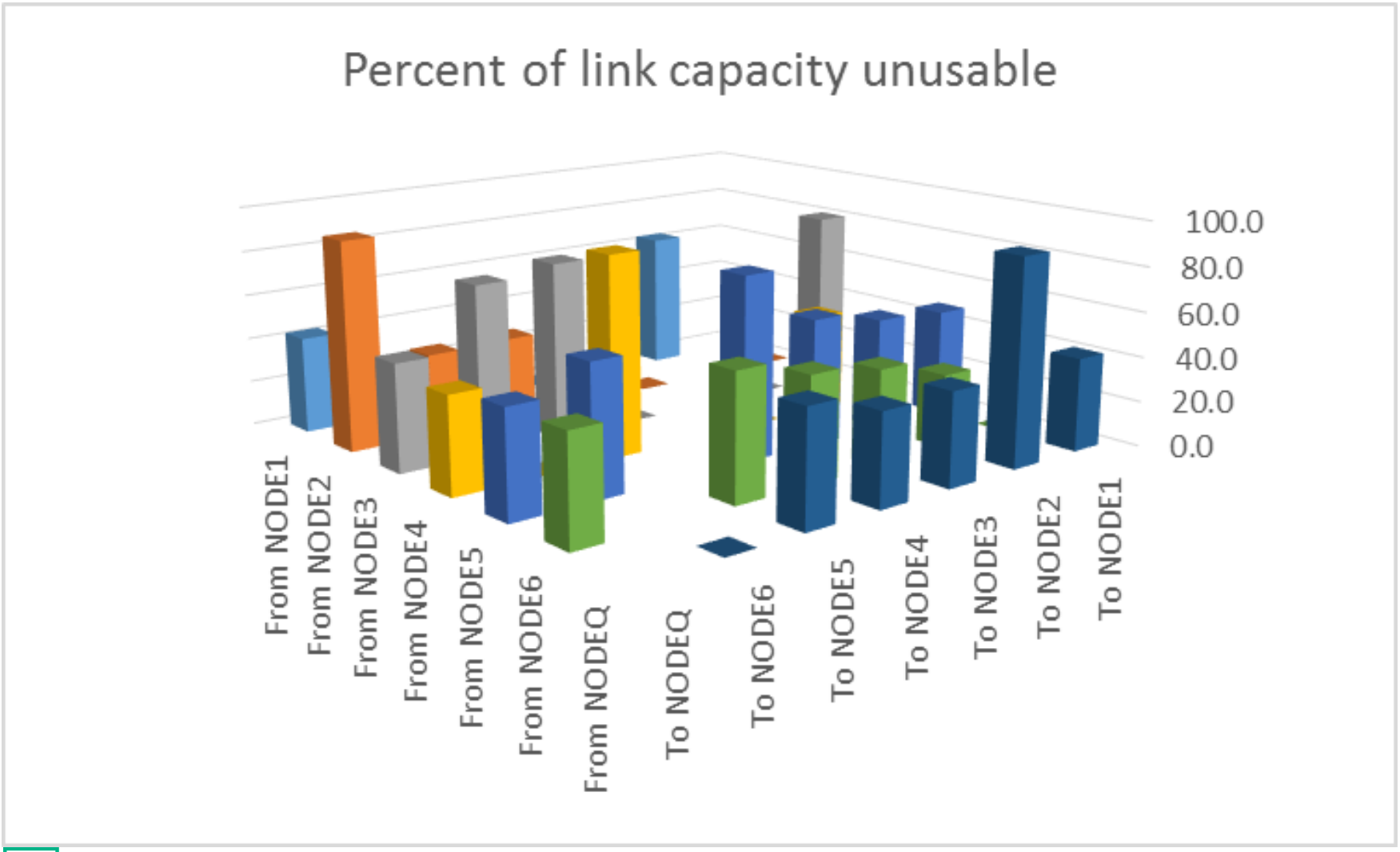
```
SCACP> SHOW VC

NODE1 PEA0 VC Summary  5-SEP-2016 16:44:08.94:
```

Remote Node	VC State	Total Errors	Xmt:TMO	Channels Open	Channels ECS	ECS Pri	MaxPkt Size	ReXmt TMO(uSec)	--XmtWindow-- Cur	--XmtWindow-- Max	--XmtWindow-- Mgt	Xmt Options	Total Pkts(S+R)	Most Recent VC Opened Time	Most Recent VC Closed Time
NODEQ	Open	0	Infinite	2	2	0	1426	756321.4	73	128	0		409949	03-SEP 09:22:12.96	(No time)
NODE6	Open	3028	14847	2	2	0	1426	17580.5	128	128	0		124743326	03-SEP 09:17:43.67	(No time)
NODE5	Open	2396	24306	2	2	0	1426	54867.4	128	128	0		153002334	03-SEP 09:17:42.66	(No time)
NODE4	Open	4247	25567	2	2	0	1426	2891.5	128	128	0		231996292	03-SEP 09:17:40.49	(No time)
NODE3	Open	2586	21708	2	2	0	1426	5455.2	128	128	0		144285982	03-SEP 09:17:32.83	(No time)
NODE2	Open	2514	26897	2	2	0	1426	17370.2	47	128	0		169211170	03-SEP 09:17:32.83	(No time)
NODE1	Open	0	Infinite	1	1	0	1426	3000000.0	1	8	0		5	03-SEP 09:17:28.65	(No time)

SCACP Counters and Metrics

PEDRIVER Transmit Window



High Availability configuration (redundancy)

- Cluster Interconnect network should have complete redundancy
  - No single point of failure
- This means:
  - At least 2 NICs per node
  - At least 2 independent networks
  - And more is better, up to a point:
    - PEDRIVER has to track all paths and may have difficulty as number of NICs per node and number of nodes per cluster grows; 4 to 8 NICs per server is probably plenty

Monitoring of redundant link status

- Redundant links must be monitored, or failures will be missed and failure of last path will cause an outage
- LAVC$FAILURE_ANALYSIS is recommended:
  - Generates OPCOM console (and OPERATOR.LOG) messages when NICs, nodes, and paths fail (and are repaired)
  - EDIT_LAVC.COM tool from OpenVMS Freeware CD Volume 6 directory [KP_CLUSTERTOOLS] automates setup of LAVC$FAILURE_ANALYSIS from SYS$EXAMPLES:LAVC$FAILURE_ANALYSIS..MAR template file
  - See article in OpenVMS Technical Journal Volume 2, “Local Area Network Cluster Interconnect Monitoring”
  - Also requires console management system or equivalent to scan for %LAVC error messages, preferably in real time

Case Study

– Silent packet corruption in the network

– 2-site OpenVMS Cluster

– Bugchecks due to silent packet corruption in the inter-site network

– Enabled Virtual Circuit checksumming:

  – Set NISCS_PORT_SERV = 3 in SYSGEN parameters, or

  – SCACP> SET VC ***nodename*** /CHECKSUMMING

– NISCS Checksum is only 16 bits

  – 1 chance of 65,536 that checksum will randomly match

– Still crashes

– Solution: Customer changed to different network provider