connect

OpenVMS Boot Camp

> Bedford-Glen, MA

OpenVMS Cluster Case Studies

Keith Parris

Thank you to our Global, Platinum & Gold Sponsors!

Global:

Platinum:

NEMONIX™ Engineering, Inc.

STROMASYS

eCube Systems

Software INTERNATIONAL Concepts

Gold:

HP Renew

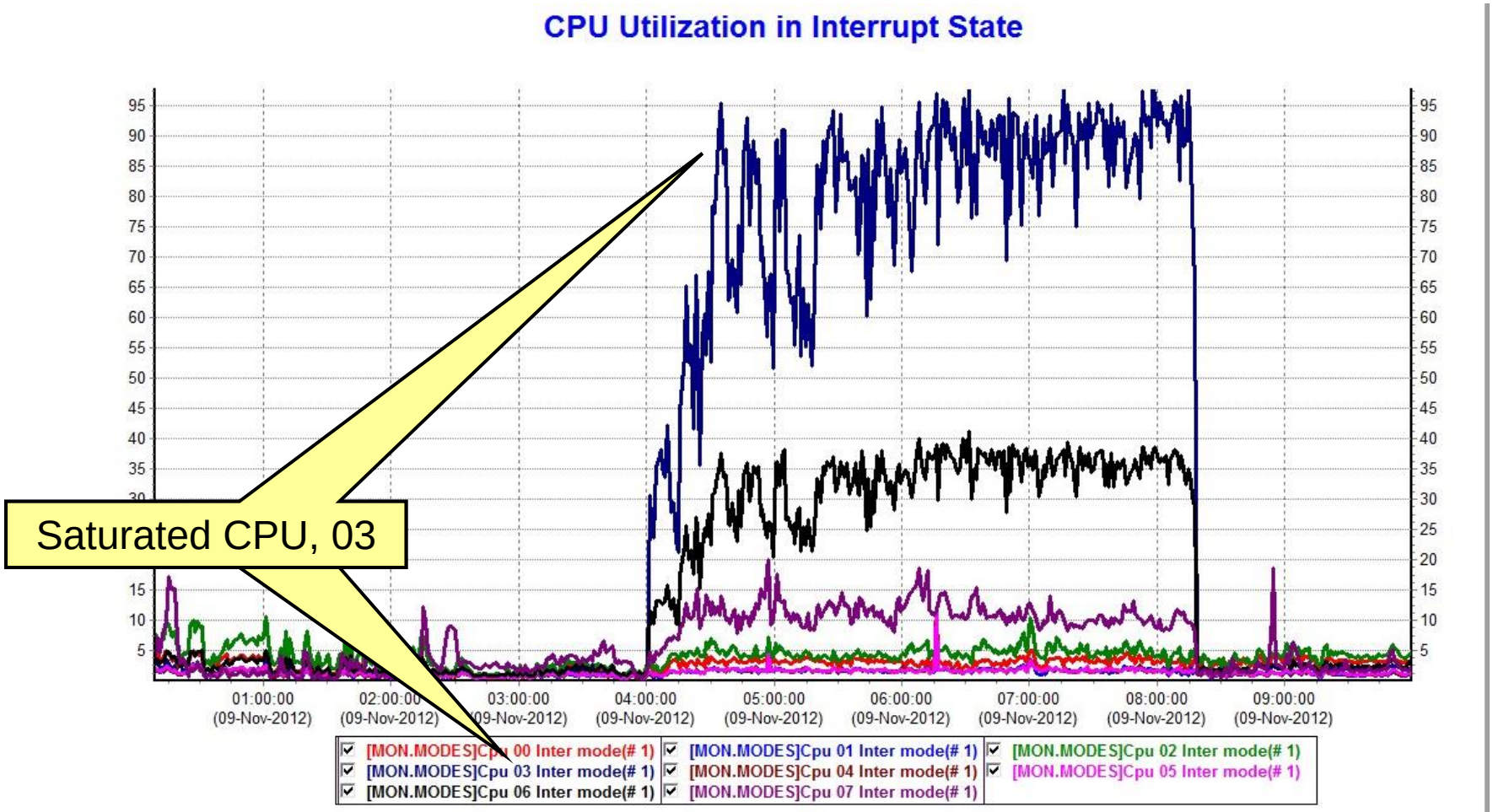
ATTUNITY

Cluster Case Study 1

- Performance anomaly a few months previously
- More-recent history of intermittent pauses, 6-7 seconds in duration
  - Some pauses, for example, doing $DIRECTORY command accessing a given disk
  - Some pauses in character echoing to terminal emulator connected via Telnet

Analyzing the past performance anomaly

- Event lasted from about 1:30 in the morning until about 11:00 AM
- Symptoms reported included batch jobs being slow, connects via SSH failing due to timeouts, and inputs like the username or password prompt or any command entered taking 3-4 minutes to respond
- CPUs and disks did not seem particularly busy under MONITOR
- Some improvement was noted for about 20 minutes after adding RMS Global Buffers to some files, but that improvement was only temporary
- At 9:30-10:30 a PURGE command was issued on a log file directory which contained 120,000 files, some quite fragmented
- Available data:
  - T4 performance data
  - Existing systems to re-examine
- Unavailable data:
  - Availability Manager data (Event and Lock Contention log files)
  - HP PerfDat data
  - Locking activity, or lock queue data


CPU Utilization in Interrupt State
Saturated CPU, 03
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
01:00:00 (09-Nov-2012)
02:00:00 (09-Nov-2012)
03:00:00 (09-Nov-2012)
04:00:00 (09-Nov-2012)
05:00:00 (09-Nov-2012)
06:00:00 (09-Nov-2012)
07:00:00 (09-Nov-2012)
08:00:00 (09-Nov-2012)
09:00:00 (09-Nov-2012)
[MON.MODES]Cpu 00 Inter mode(# 1)
[MON.MODES]Cpu 01 Inter mode(# 1)
[MON.MODES]Cpu 02 Inter mode(# 1)
[MON.MODES]Cpu 03 Inter mode(# 1)
[MON.MODES]Cpu 04 Inter mode(# 1)
[MON.MODES]Cpu 05 Inter mode(# 1)
[MON.MODES]Cpu 06 Inter mode(# 1)
[MON.MODES]Cpu 07 Inter mode(# 1)

Locating source of high interrupt-state time on CPU 03

- From TLviz, File → Properties

```
File # 1 :
      Name : T:\R2\PROJECTS\T4\09NOV\T4_NODE8_09NOV2012_0001_2359_COMP.CSV
      Node : NODE8
...
      Monitor Comment: HP BL870c i2  (1.60GHz/5.0MB) (32764Mb with 8 cpu(s))
      MpCpus: 8
...
      LckMgr_Mode: 0
      LckMgr_CpuId: 0

      Pref CPU for _FGB0: 7
      Pref CPU for _FGA0: 7
      Pref CPU for _PEA0: 2
      Pref CPU for _PKA0: 5
      Pref CPU for _PKB0: 4
      Pref CPU for _EWA0: 6
      Pref CPU for _EWB0: 5
      Pref CPU for _EWC0: 4
      Pref CPU for _EWD0: 3
      Pref CPU for _EWE0: 2
      Pref CPU for _EWF0: 1
      Pref CPU for _EWG0: 0
      Pref CPU for _EWH0: 6
```

Saturated CPU, 03

Locating source of high interrupt-state time on CPU 03

- From SDA> SHOW LAN

```
-- EWD Unit Summary 26-NOV-2012 13:02:46 --

UCB      UCB Addr  Fmt   Value           Client     State
---      --------  ---   -----           ------  -----------
EWD0     90635780
EWD2     908D5200  Eth   60-07           SCA     0017 STRTN,LEN,UNIQ,STRTD
EWD4     914F8540  802E  08-00-2B-80-3C  DNAME   0017 STRTN,LEN,UNIQ,STRTD
EWD5     914FA700  Eth   80-3C           DNAME   0017 STRTN,LEN,UNIQ,STRTD
EWD8     91706AC0  Eth   08-00           IP      0015 STRTN,UNIQ,STRTD
EWD9     91707700  Eth   08-06           ARP     0015 STRTN,UNIQ,STRTD
EWD10    91708340  Eth   86-DD           IPV6    0015 STRTN,UNIQ,STRTD
EWD11    9174B280  Eth   60-00                   0004 UNIQ
```

- Could be OpenVMS Cluster (SCA) or IP traffic – don't know which
- Could examine counters in the SCACP and TCPIP utilities, or SDA; or
- Shift one type of traffic off EWD0 and re-examine T4 data afterward

Locating source of high interrupt-state time on CPU 03

• From SDA> SHOW LAN /FULL

```
LAN Data Structures
-------------------
         -- EWD2 60-07 (SCA) Counters Information 26-NOV-2012 13:02:46 --

Octets received          204635757213    Octets sent              266652850752
PDUs received              1021083224    PDUs sent                  1117053926
Mcast octets received     21503002412    Mcast octets sent           424973496
Mcast PDUs received         182228834    Mcast PDUs sent               3171444
Unavail user buffer                0    Multicast not enabled               0
Last UUB time                   None    User buffer too small               0
...
LAN Data Structures
-------------------
          -- EWD8 08-00 (IP) Counters Information 26-NOV-2012 13:02:46 --

Octets received            2331218644    Octets sent                      270
PDUs received                23772348    PDUs sent                          5
Mcast octets received      2331218398    Mcast octets sent                  0
Mcast PDUs received          23772345    Mcast PDUs sent                    0
Unavail user buffer                0    Multicast not enabled              0
Last UUB time                   None    User buffer too small              0
```

• Cumulative totals say 1,021 million packets received were SCA and only 23 million were IP: so overall EWD0 workload is more SCA traffic, but this may be misleading because of averaging over time

- Determining source of EWD0 interrupt-state workload during specific period

Locating source of high interrupt-state time on CPU 03

- From TLviz Correlate function on T4 data:

```
Correlations against [MON.MODES]Cpu 03 Inter mode for the Data Collection on
node NODE5 for the period between
11/9/2012 12:08:10 AM and 11/9/2012 9:59:36 AM
are as follows :
0.998  [MON.MODES]Cpu 06 Inter mode
0.992  [MON.MODE]Interrupt State
0.977  [MON.MODES]Cpu 03 Busy
0.937  [MON.DISK]QLen
0.927  [MON.DISK]RespT
0.922  [MON.MODES]Cpu 06 Busy
0.915  [MON.SYST]Cpu Busy
0.901  [MON.MODES]Cpu 01 Busy
0.900  [MON.MODES]Cpu 01 Kernel mode
0.899  [NET.TCP]RxPkt
0.899  [NET.TCP]RxPk/s
0.897  [NET.TCP]TxPkt
0.897  [NET.TCP]TxPk/s
0.897  [NET.TCP]TPkSz
0.895  [MON.MODES]Cpu 07 Inter mode
0.895  [MON.MODES]Cpu 00 Kernel mode
0.890  [NET.TCP]TxMb/s
0.890  [NET.TCP]TxByte
…
```

Best guess: Disk backup operation with data sent out over TCP/IP

Other findings

- Evidence of CPU saturation in interrupt state at other times
- Many large directory files
  - many files with identical long prefixes in file names
- Evidence of lock tree remastering, which can cause pauses
  - Recommended enabling Jumbo Packets and using 10 GbE instead of 1 GbE uplinks from blade chassis to make remastering operations quicker
- High lock rates to RMS indexed files
  - Recommended adding RMS Global Buffers to reduce lock rates

Recommendations we made for performance anomaly

- Replace default Fast Path settings with optimal settings to reduce interrupt state and MP_Synch time
- Add monitoring tools: Console Management, Availability Manager, and tools to gather lock activity and lock queue data
- Reduce lock rates by increased use of RMS Global Buffers on hot files to reduce chances of saturation of the CPU handling cluster traffic or the Lock Manager Dedicated CPU

Recommendations we made for performance anomaly

- Hyperthreads make CPU interrupt-state saturation analysis harder. While troubleshooting, consider disabling hyperthreads (and if capacity is of concern, perhaps replace the extra co-thread “cores” with real physical processors plugged into the empty sockets), or else avoid using (i.e do a $STOP/CPU command on) the co-thread CPUs for these critical functions:
  - The Primary CPU, which handles such critical functions as timekeeping and Timer Queue Entries (TQEs)
  - Any CPUs handling device interrupts for busy devices
  - The CPU handling PEA0 (PEDRIVER)
  - The CPU handling TCP/IP (BG0) or the TCP/IP Packet Processing Engine (PPE)
  - The Lock Manager Dedicated CPU

Blade Configuration

- BL870c i2 and BL890c i2 Blades each had one populated processor socket and one empty socket
  - Result: MaxNUMA configuration
    - All Socket Local Memory (SLM)
    - No InterLeaved Memory (ILM)
    - OpenVMS data cannot go into ILM; goes into RAD 0, overflows if needed into RAD 1, etc.
- Evidence of some CPU saturation in interrupt state
- Dedicated-CPU Lock Manager enabled on BL890c i2
- Hyperthreads enabled on BL870c i2 and BL890c i2
- How to optimize Fast_Path assignments?

Cluster Case Study 2

- Disaster-tolerant OpenVMS Cluster
- IPCI as cluster interconnect
- Question: How to best configure IPCI for redundancy?

Path Failover Considerations with IPCI

- failSAFE IP does not work with IPCI
  - OpenVMS 8.4 Release Notes say “the IP address and interface used for cluster communication must not be used for Failsafe configuration”
  - Failover is controlled by a process, and during a quorum loss no process can be scheduled to run
- failSAFE IP can use only one NIC at a time

Path Failover Considerations with IPCI

- LAN Failover works fine with Clusters using LAN or IPCI, but…
- LAN Failover is triggered only by loss of Link signal
  - Loss of connectivity which doesn't cause a loss of Link signal will not be detected (i.e. router to which link is connected loses its uplink to the outside world)
- LAN Failover likewise uses only one NIC at a time

Path Failover Considerations with IPCI

- PEDRIVER can track individual paths between any pair of NICs
- PEDRIVER can use multiple paths in parallel at once
- Conclusion:
  - Give each NIC its own unique IP address and let PEDRIVER track connectivity and do optimal path selection

Questions?

Speaker contact info:

- Keith Parris
- E-mail: keith.parris@hp.com or keithparris@yahoo.com
- Website: http://www2.openvms.org/kparris/