13th ANNUAL WORKSHOP 2017

FABRIC PERFORMANCE MANAGEMENT AND MONITORING

Todd Rimmer, Omni-Path Lead Software Architect

Intel Corporation

March, 2017
LEGAL DISCLAIMERS

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

A "Mission Critical Application" is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL'S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS' FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENCE IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The cost reduction scenarios described in this document are intended to enable you to get a better understanding of how the purchase of a given Intel product, combined with a number of situation-specific variables, might affect your future cost and savings. Circumstances will vary and there may be unaccounted-for costs related to the use and deployment of a given product. Nothing in this document should be interpreted as either a promise of or contract for a given level of costs.

Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families: Go to: Learn About Intel® Processor Numbers.

All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or go to: http://www.intel.com/design/literature.htm

The High-Performance Linpack (HPL) benchmark is used in the Intel® FastFabrics toolset included in the Intel® Fabric Suite. The HPL product includes software developed at the University of Tennessee, Knoxville, Innovative Computing Libraries.

Intel, Intel Xeon, Intel Xeon Phi™ are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States or other countries.
MANAGEMENT OF AN OMNI-PATH CLUSTER

• Intel® OPA leverages existing stacks for each type of management
• Assorted 3rd party unified management consoles
• Intel® OPA provides a scalable centralized fabric management stack
Fabric utilization and performance monitoring is critical to fabric operations.

**Intel® OPA Fabric Statistics**
  - Over 130 performance counters per port
  - Including utilization, packet rate and congestion per VL
  - 64-bit counters (many decades to rollover)
## OMNI-PATH PORT COUNTERS

### Performance: Transmit
- Xmit Data
- Xmit Pkts
- MC Xmt Pkts

### Performance: Receive
- Rcv Data
- Rcv Pkts
- MC Rcv Pkts

### Performance: Congestion
- Congestion Discards
- Rcv FECN
- Mark FECN
- Rcv BECN
- Xmit Time Congestion
- Xmit Wait

### Performance: Bubbles
- Rcv Bubble
- Xmit Wasted BW
- Xmit Wait Data

### Link Quality Indicator
- LinkWidthDowngrade

### Errors: Signal Integrity
- Uncorrectable Errors
- Link Downed
- Rcv Errors
- Exc. Buffer Overrun
- FM Config Errors
- Link Error Recovery (retrain)
- Local Link Integ Err (replay)

### Errors: Security
- Xmit Constraint
- Rcv Constraint

### Errors: Other
- Rcv Sw Relay Err
- Xmit Discards
- Rcv Rmt Phys Err

---

**Utilization**

**Congestion**

**Errors**

**Statistics**
PM DATA GATHERING

- PM gathers data at a configurable fixed interval
  - Counters gathered from all ports
  - Can exclude HFI ports to avoid compute jitter
  - Can exclude per VL counters
  - Parallelized across devices and within devices
  - Progressive back-off algorithm for retries

```
<SweepInterval>10</SweepInterval> <!-- time between sweeps in seconds -->
<ProcessHFICounters>1</ProcessHFICounters> <!-- process HFI Counters -->
<ProcessVLCounters>1</ProcessVLCounters> <!-- process Per-VL Counters -->
<PmaBatchSize>2</PmaBatchSize> <!-- max parallel requests to a given PMA -->
<MaxParallelNodes>10</MaxParallelNodes> <!-- max devices in parallel -->
<MaxAttempts>3</MaxAttempts>
<RespTimeout>250</RespTimeout>
<MinRespTimeout>35</MinRespTimeout> <!-- in milliseconds -->
```
- Each PM Sweep creates a “PA Image”
- PA Image contains
  - Timestamp
  - Topology graph at time of image
    - Node names, LIDs, GUIDs, link speeds
    - FMs
  - vFabrics at time of image
    - Name, VLs used, membership
  - Counter values
  - Results of data analysis computations
- Counters on each link are consolidated for each direction
- Counters in each direction are consolidated into categories
  - Utilization
  - Signal Integrity
  - Congestion
  - Bubble
  - Routing
  - Security
  - SMA Congestion
- Utilization based percentages computed for some counters
  - Such as packet and time based congestion counters
- Configurable weighted sum applied to generate a summary value for some categories
- Summary value compared to configurable threshold
  - Histogram bucketing of number of ports near or beyond threshold
  - PM logging when exceed threshold
PM DEVICE GROUPS

- Sysadmin may define device groups
  - All, SWs and HFIs groups available by default
- Category data and histograms organized per group

```
<DeviceGroup>
  <Name>storage</Name>
  <NodeDesc>oss*</NodeDesc>
  <NodeDesc>mds*</NodeDesc>
</DeviceGroup>

<DeviceGroup>
  <Name>compute</Name>
  <NodeDesc>mycomp*</NodeDesc>
</DeviceGroup>

<DeviceGroup>
  <Name>xenon_phi</Name>
  <NodeDesc>mycomp[50-99]</NodeDesc>
</DeviceGroup>
```

SWs Group
VFABRIC DATA

- Category information per vFabric
- Histograms and summary data per vFabric
- Cross references vFabric to specific ports, VL(s) and per VL counters

Advanced QoS Virtual Fabrics Configuration

<table>
<thead>
<tr>
<th>Compute A</th>
<th>Compute B</th>
<th>Storage</th>
<th>Switch PO</th>
<th>FM</th>
<th>Admin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>Full</td>
<td>Full</td>
<td>Full</td>
<td>Full</td>
<td>Full</td>
</tr>
<tr>
<td>Limited</td>
<td>Limited</td>
<td>Limited</td>
<td>Full</td>
<td>Full</td>
<td>Limited</td>
</tr>
<tr>
<td>Full</td>
<td>Full</td>
<td>Full</td>
<td>N/A</td>
<td>Full</td>
<td>Full</td>
</tr>
<tr>
<td>Full</td>
<td>Full</td>
<td>Full</td>
<td>N/A</td>
<td>Full</td>
<td>Full</td>
</tr>
<tr>
<td>Full</td>
<td>Full</td>
<td>Full</td>
<td>N/A</td>
<td>Full</td>
<td>Full</td>
</tr>
<tr>
<td>Full</td>
<td>Full</td>
<td>Full</td>
<td>N/A</td>
<td>Full</td>
<td>Full</td>
</tr>
</tbody>
</table>

Applications

- Networking
- SA (SM implicit)
- Compute
- Storage
- Checkpoint
- AllOthers
- PA, PM
PM IMAGE RETENTION

- Recent PM Images cached in RAM
- Short term history storage to disk
- Images age out to keep recent history
- Images on disk compressed
- May composite images on disk
  - Trade off interval vs storage needs vs duration of history

```xml
<TotalImages>10</TotalImages> <!-- total in RAM images for history and freeze -->
<ShortTermHistory>
  <Enable>1</Enable>
  <StorageLocation>/var/lib/opafm/pahistory</StorageLocation>
  <TotalHistory>24</TotalHistory> <!-- in hours -->
  <ImagesPerComposite>3</ImagesPerComposite>
  <MaxDiskSpace>10240</MaxDiskSpace> <!-- in MiB -->
  <CompressionDivisions>8</CompressionDivisions>
</ShortTermHistory>
```
PM DATA ACCESS

• PA defines an in-band protocol to query the PM
• CLI tools
  • opareport – text or XML output
  • opaextract* - CSV/spreadsheet output
• TUI tools
  • opatop – interactive TUI
• FM GUI
  • Out of band access via FE
OPATOP TUI

Top view shows fabric and per group summaries

Multiple Levels of Drill Down
• Study areas of interested, drill down to the port

Full access to PM on-line history
• Review performance hours or days ago

Can freeze/bookmark an image and study it as long as desired
FM GUI
• PM Failover
  • Active passive redundancy
  • PM integrated with the SM, follows SM failover rules
• PM Data Synchronization
  • PM in RAM and on disk images synchronized across redundant PMs
  • Synchronization rate can be throttled
SUMMARY

- Omni-Path FM includes a powerful performance monitoring and analysis subsystem
  - Monitors >130 hardware counters
  - Consolidates data in categories
  - Cross references data against device groups and virtual fabrics
  - Retains short term history
  - Monitoring and data retention redundancy
- PA protocol allows access to PM data
- Assorted CLI and GUI tools to display data

https://github.com/01org/opa-fm
https://github.com/01org/opa-fmgui
https://github.com/01org/opa-ff
THANK YOU

Todd Rimmer, Omni-Path Lead Software Architect

Intel Corporation