**Performance Counter Monitoring for the Blue Gene/Q Architecture**

PAPI for BG/Q: 5 New Components

**PUnit Component**
- Each of the 18 A2 CPU cores has a local UPU module
- Each of these modules provides 24 counters (14-bit) to sample 2D events, 1L cache-related events, floating point operations, etc.
- Currently, there are 269 native PUnit events available
- Out of 107 possible predefined events, there are currently 43 events available of which 13 are derived events.

**Network Unit Component**
- The 3D-Torus network provides a local UPC network module with 65 counters - each of the 11 link has six 1M4 counter
- Currently, all network links are attached and this is hard-coded in the PAPI NWUnit component
- Currently, there are 31 Network Unit events available

**I/O Unit Component**
- The Message, PCIe, and DevBus module – which are collectively referred to as I/O module – provide together 43 counters (node-wide)
- Currently, there are 44 I/O Unit events available

**CNK Unit Component**
- CNK is the lightweight Compute Node Kernel that runs on all the 16 compute cores
- BGPM offers a "virtual" CNK Unit that has software counters collected by the kernel (kernel counter value are read via a system call)
- Currently, there are 28 CNK Unit events available

**Example: 3D-FFT on BG/Q**

- Why multi-dimensional FFTs?
  - The FFTs of multi-dimensional data are of particular importance in a number of different scientific applications but they are often one of the most computationally expensive components.
  - Parallel multidimensional FFTs are communication intensive, that’s why they often prevent the application from scaling to a very large number of processors.
  - A fundamental challenge of such numerical algorithms is the design and implementation to utilize efficiently thousands of nodes
  - Computation performed in three single stages:

<table>
<thead>
<tr>
<th>3D-FFT Decomposition</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D-Decomposition</td>
<td>MPI tasks organized in 2D virtual processor grid using MPI Cartesian grid topology construct</td>
</tr>
</tbody>
</table>

**Multiplexing**
- PAPI supports multiplexing for the BG/Q platform
- The BGPM PUnit does not directly implement multiplexing of event sets; but, it allows the simultaneous monitoring of multiple event sets by selecting the right counter slice in the in situ implementation.

<table>
<thead>
<tr>
<th>Multiplexing Events</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAPI_L2UNIT.MU_FIFO_PKT_RCV</td>
<td>A new FIFO packet has been received (The packet has been stored to L2. There is no pending switch request)</td>
</tr>
<tr>
<td>PAPI_L2UNIT.MU_MSG_INJ</td>
<td>A new message has been injected (All packets of the message have been stored to the ND FIFO)</td>
</tr>
<tr>
<td>PAPI_L2UNIT.MU_FIFO_PKT_Transmit</td>
<td>A new packet has been injected (Packet has been stored to ND FIFO)</td>
</tr>
</tbody>
</table>

**Conclusion**
- Performance analysis tools for parallel applications running on large scale systems rely on hardware-performance counters to gather performance relevant data from the system.
- PAPI’s five new components for BG/Q
  - Enable hardware performance counter monitoring for: Processing unit and L2 unit
  - 1D Teras
  - I/O system
  - Compute Node Kernel

**Example**
- As an example, we instrumented a parallel 3D-FFT kernel with PAPI for communication evaluation. With the default mapping of MPI tasks onto the torus network, the network counters detected a large amount of redundant inter-node communication. We were able to eliminate the unnecessary communication and achieved an overall 18% performance improvement of the 3D-FFT kernel by employing a custom mapping.

**Contact**
- Heike McCraw
- mccraw@utk.edu
- UT-Innovative Computing Laboratory (ICL)

**Acknowledgments**
- This material is based upon work supported by the U.S. Department of Energy Office of Science under Award DE-SC0008263.
- Access to the early access BG/Q system at Argonne National Laboratory was provided through the ASC Early Science Program.
- We would like to thank the IBM Performance Team for collaborating with us on the implementation of PAPI for BG/Q.

**BG/Q network overview**
- Compute nodes organized as a 3D-torus: every node is connected to ten neighbour nodes through bidirectional links.

**L2 Unit Component**
- Shared L2 cache is split into 16 separate slices; each of the 16 L2 memory slices has a L2 UPU module that provides 6 counters (node-wide)
- Currently, there are 32 L2 Unit events available:

**Overview**
- Only the local UPC module, L2 and I/O hardware support performance monitoring interrupts when a programmable counter overflow is detected.
- For that reason, only the PUnit, L2Unit, and I/OUnit provide overflow support in BGPM and PAPI