

PAPI for Blue Gene/Q: The 5 BGPM Components

Heike Jagode and Shirley Moore
Innovative Computing Laboratory
University of Tennessee-Knoxville

ESP Code for “Q” Workshop
Argonne National Laboratory
March 19-21, 2012

Overview

- Introduction
- Processor Unit (PUnit) Component
- L2 Unit Component
- I/O Unit Component
- Network Component
- Compute Node Kernel Unit (CNKUnit) Component

Introduction

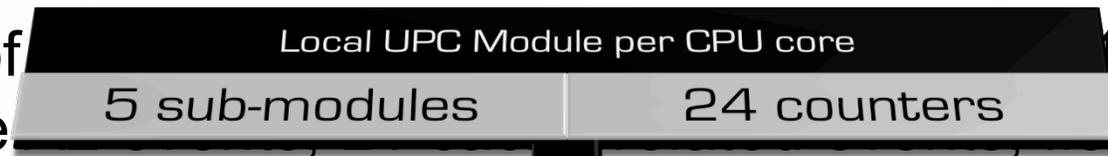
- Very little effort was put into hardware performance monitoring tools for the BG/Q predecessor BG/P.
- HPC community was left behind with rather poor and incomplete methods.
- To eliminate this limitation, for BG/Q we planned carefully and collaborate closely with IBM's Performance Group.

Result:

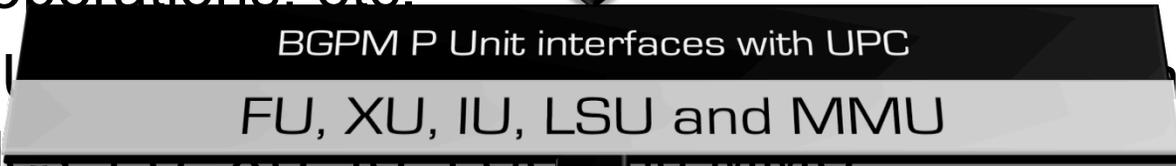
- Added 5 new components to PAPI to support hardware performance monitoring for the BG/Q network, the I/O system, and the Compute Node Kernel in addition to the processing cores

PUnit Component

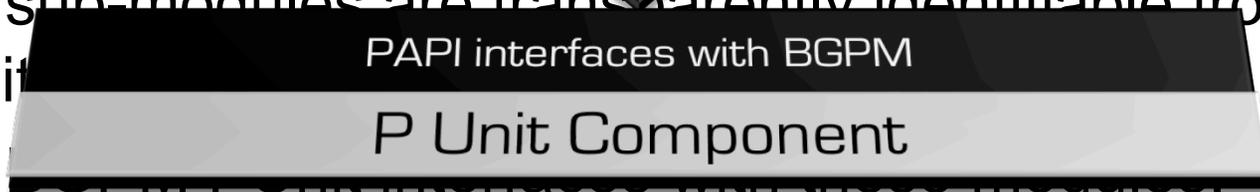
- Each of the 18 A2 CPU cores has a local UPC module.
- Each of the local UPC modules (14-bit) to sample point operations. etc.



- Local P Unit sub-modules: FU, XU, IU, LSU and MMU.



- The sub-modules are transparently identifiable from the PUnit.



- The PUnit interfaces with these modules.
- PAPI uses the BGPM interface.

PUnit Events (Native | Presets)

- Currently, there are 269 native PUnit events available:

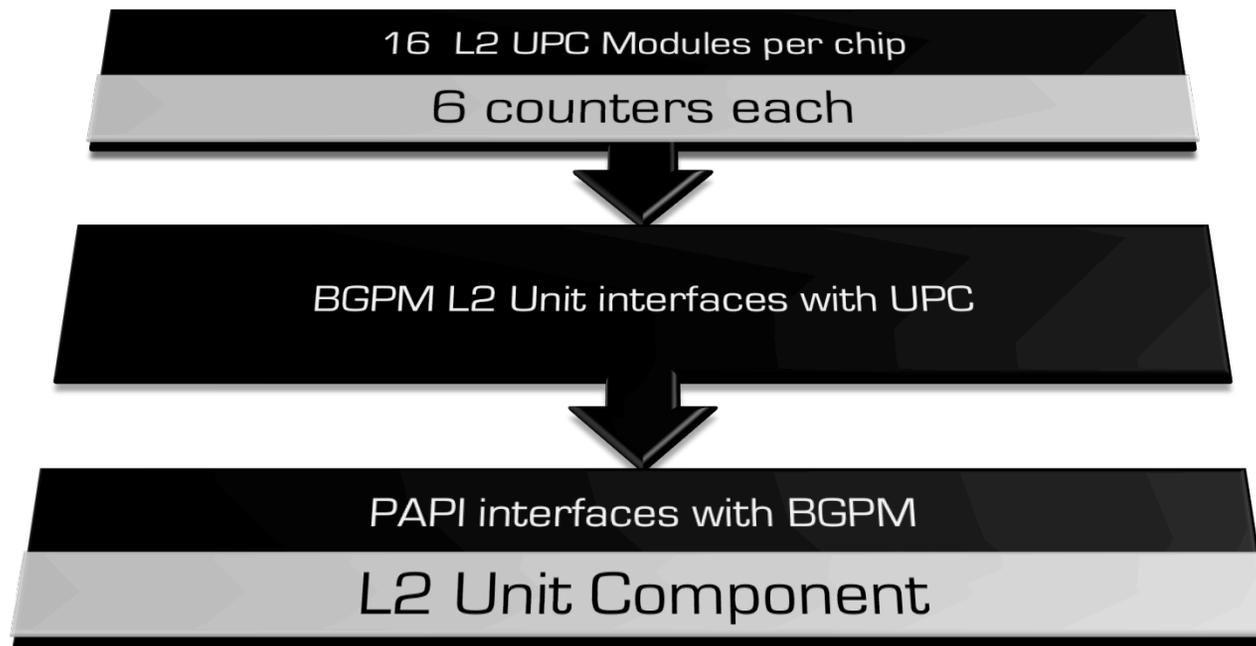
PUnit Event	Description
PEVT_AXU_INSTR_COMMIT	A valid AXU (non-load/store) instruction is in EX6, past the last flush point.
PEVT_AXU_CR_COMMIT	A valid AXU CR (update) instruction is in EX6, past the last flush point.
PEVT_AXU_IDLE	No valid AXU instruction is in the EX6 stage.

- Out of 107 possible PAPI predefined events, there are currently 41 events available of which 12 are derived events:

Name	Code	Avail	Deriv	Description (Note)
PAPI_L1_ICM	0x80000001	Yes	No	Level 1 instruction cache misses
PAPI_FXU_IDL	0x80000011	Yes	No	Cycles integer units are idle
PAPI_TLB_DM	0x80000014	Yes	Yes	Data translation lookaside buffer misses
PAPI_TLB_IM	0x80000015	Yes	No	Instruction translation lookaside buffer misses
PAPI_TLB_TL	0x80000016	Yes	Yes	Total translation lookaside buffer misses
PAPI_L1_LDM	0x80000017	Yes	No	Level 1 load misses
PAPI_L1_STM	0x80000018	Yes	No	Level 1 store misses
PAPI_BTAC_M	0x8000001b	Yes	No	Branch target address cache misses
PAPI_PRF_DM	0x8000001c	Yes	No	Data prefetch cache misses
PAPI_TLB_SD	0x8000001e	Yes	No	Translation lookaside buffer shutdowns
PAPI_CSR_FAL	0x8000001f	Yes	No	Failed store conditional instructions
PAPI_CSR_SU	0x80000020	Yes	Yes	Successful store conditional instructions
PAPI_CSR_TOT	0x80000021	Yes	No	Total store conditional instructions
PAPI_CSR_FAL	0x80000022	Yes	No	Failed store conditional instructions

L2 Unit Component

- Shared L2 cache is split into 16 separate slices
- Each of the 16 L2 memory slices (per chip) has a L2 UPC module that provides 6 counters (node-wide)



L2 Unit Native Events

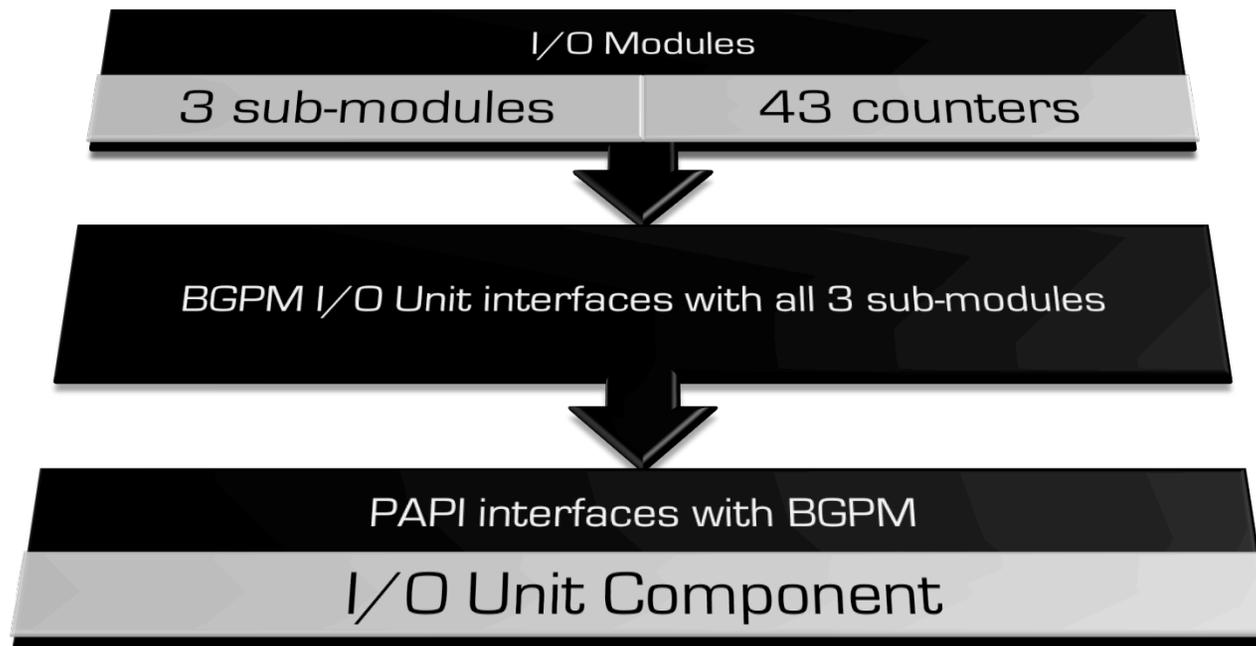
- Currently, there are 32 L2 Unit events available:

L2Unit Event	Description
PEVT_L2_HITS	hits in L2, both load and store. Network Polling store operations from core 17 on BG/Q pollute in this count during normal use
PEVT_L2_MISSES	cacheline miss in L2 (both loads and stores
PEVT_L2_PREFETCH	fetching cacheline ahead of L1P prefetch
...	...

- BG/Q processor has two DDR3 memory controllers, each interfacing with eight slices of the L2 cache to handle their cache misses (one controller for each half of the 16 cores on the chip).
- The counting hardware can either keep the counts from each slice separate, or combine the counts from each slice into single values (which is the default).

I/O Unit Component

- The Message, PCIe, and DevBus module – which are collectively referred to as I/O modules – provide together 43 counters (node-wide)



I/O Unit Native Events

- Currently, there are 44 I/O Unit events available.
- The three I/O sub-modules are transparently identifiable from the I/O Unit event names.

IOUnit Event	Description
PEVT_MU_PKT_INJ	A new packet has been injected (Packet has been stored to ND FIFO)
PEVT_MU_MSG_INJ	A new message has been injected (All packets of the message have been stored to ND FIFO)
PEVT_MU_FIFO_PKT_RCV	A new FIFO packet has been received (The packet has been stored to L2. There is no pending switch request)
...	...
PEVT_PCIE_INB_RD_BYTES	Inbound Read Bytes Request
PEVT_PCIE_INB_RDS	Inbound Read Request
PEVT_PCIE_INB_RD_CMPLT	Inbound Read Completion
...	...
PEVT_DB_PCIE_INB_WRT_BYTES	PCIe inbound write bytes written
PEVT_DB_PCIE_OUTB_RD_BYTES	PCIe outbound read bytes requested
PEVT_DB_PCIE_OUTB_RDS	PCIe outbound read request
...	...

Network Unit Component

- The 5D-Torus network provides a local UPC network module with 66 counters - each of the 11 links has six 64-bit counters.
 - As of right now, a PAPI user cannot select which network link to which to attach.
 - Currently, all network links are attached and this is hard-coded in the PAPI NWUnit component.
- The BGPM NWUnit interfaces with the network modules.
- PAPI Network Unit Component interfaces with BGPM.

Network Unit Native Events

- Currently, there are 31 Network Unit events available

NWUnit Event	Description
PEVT_NW_USER_PP_SENT	Number of 32 byte user point to point packet chunks sent. Includes packets originating or passing through the current node
PEVT_NW_USER_DYN_PP_SENT	Number of 32 byte user dynamic point to point packet chunks sent. Includes packets originating or passing through the current node
PEVT_NW_USER_ESC_PP_SENT	Number of 32 byte user escape point to point packet chunks sent. Includes packets originating or passing through the current node
...	...

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 values are read via a system call).
- Currently, there are 29 CNK Unit events available.

CNKUnit Event	Description
PEVT_CNKHWT_SYSCALL	External Input Interrupt
PEVT_CNKHWT_CRITICAL	Critical Input Interrupt
PEVT_CNKHWT_FIT	Fixed Interval Timer Interrupt
...	...

Overflow and Multiplexing

Overflow:

- Only the local UPC module, L2 and I/O UPC hardware support performance monitor interrupts when a programmed counter overflows
- For that reason, only the PUnit, L2Unit, and I/OUnit provide overflow support in BGPM and PAPI.

Multiplexing:

- PAPI supports multiplexing for the BG/Q platform.
- The BGPM PUnit does not directly implement multiplexing of event sets, but it does indirectly support multiplexing by supporting a multiplexed event set type.