

The IBM Power Edge of Network™ Processor:

A wire-speed System-on-a-Chip with 16 Power™ cores / 64 threads and optimized HW acceleration

Jeffrey D. Brown

Distinguished Engineer and IBM Academy of Technology Member
IBM Corporation

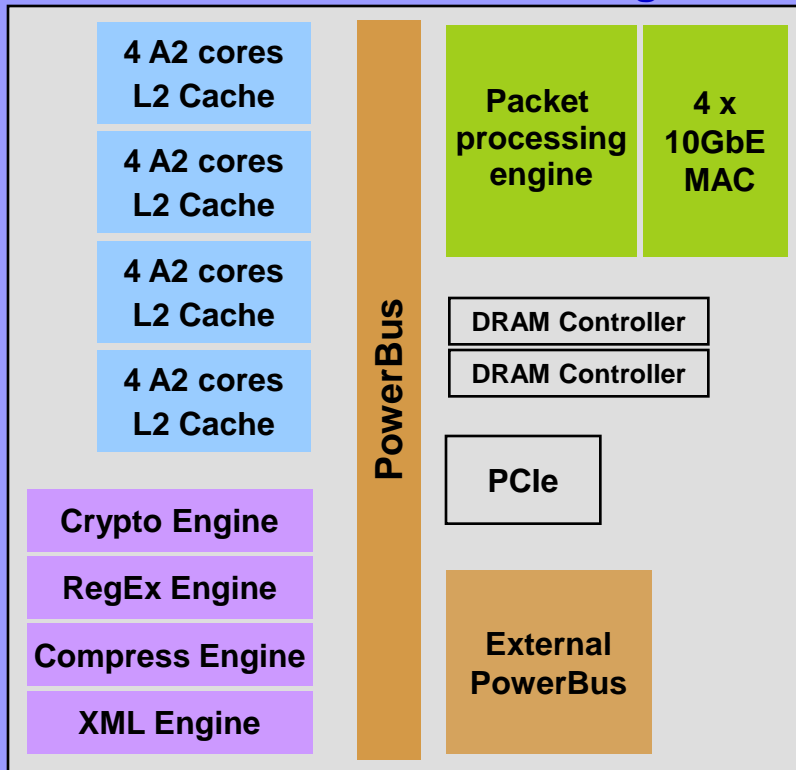
jeffdb@us.ibm.com

Co-authors: Sandra Woodward, Brian Bass, Charlie Johnson



Overview of IBM PowerEN™ Processor Chip

PowerEN™ SoC Block Diagram



- ✓ 64-bit PowerPC Architecture
- ✓ Virtualization Support
- ✓ Dual DDR3 DRAM Controllers
- ✓ Optimized Ethernet Offload Engine
- ✓ Integrated PCI-Express bus
- ✓ Cryptography Unit
- ✓ Regular Expression Unit
- ✓ XML Processing Unit
- ✓ Compression Unit
- ✓ Upward Scalability – 4 Chip
- ✓ Downward scalability - Subset

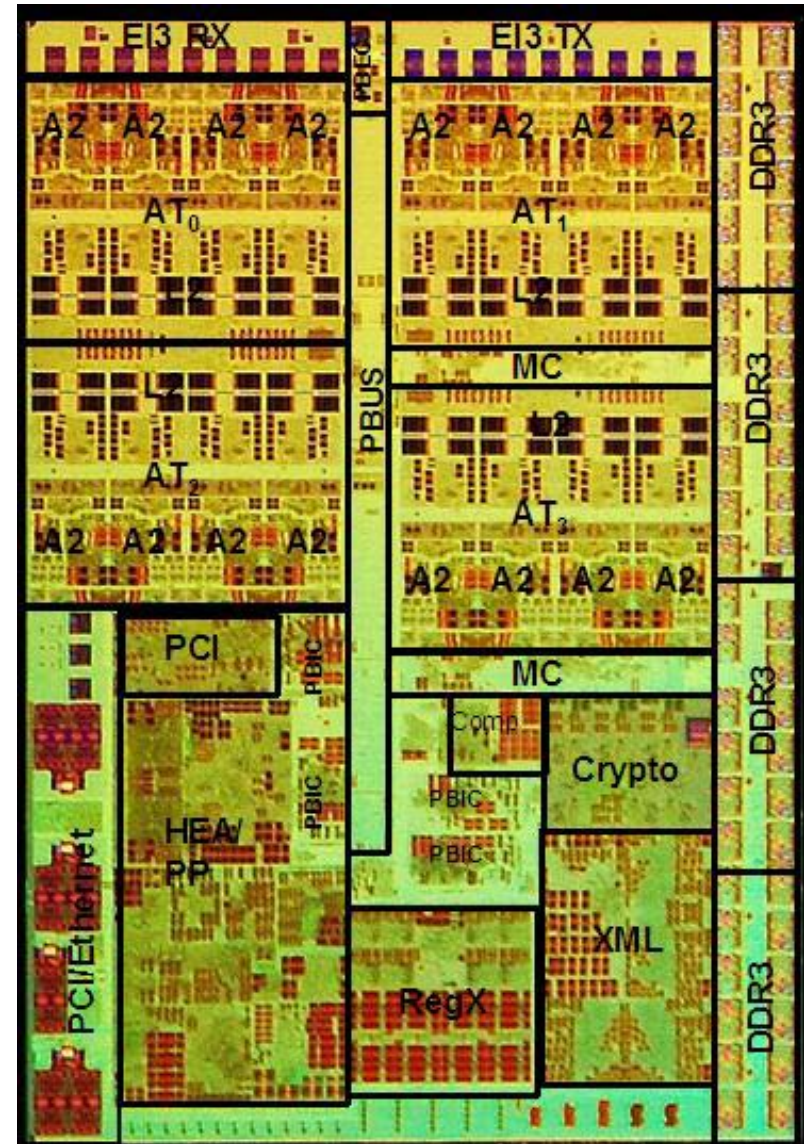
PowerEN™ has all basic IP needed for user/data plane and control plane traffic processing

IBM PowerEN™: Targeted at the Edge-of-Network

- Power efficient Throughput computing
 - Database Acceleration
 - Service Oriented Architecture Acceleration
 - Secure Multi-tenant Cloud Computing
- Enhanced processing of data payloads
 - Low latency message for Financial Information Exchange
- Deeper networking functions
 - Cyber Security
 - Network Intrusion Prevention
- Application targeted at "smarter planet" solutions
 - Compartmentalized streamed analytics
 - Data Reduction in storage subsystems

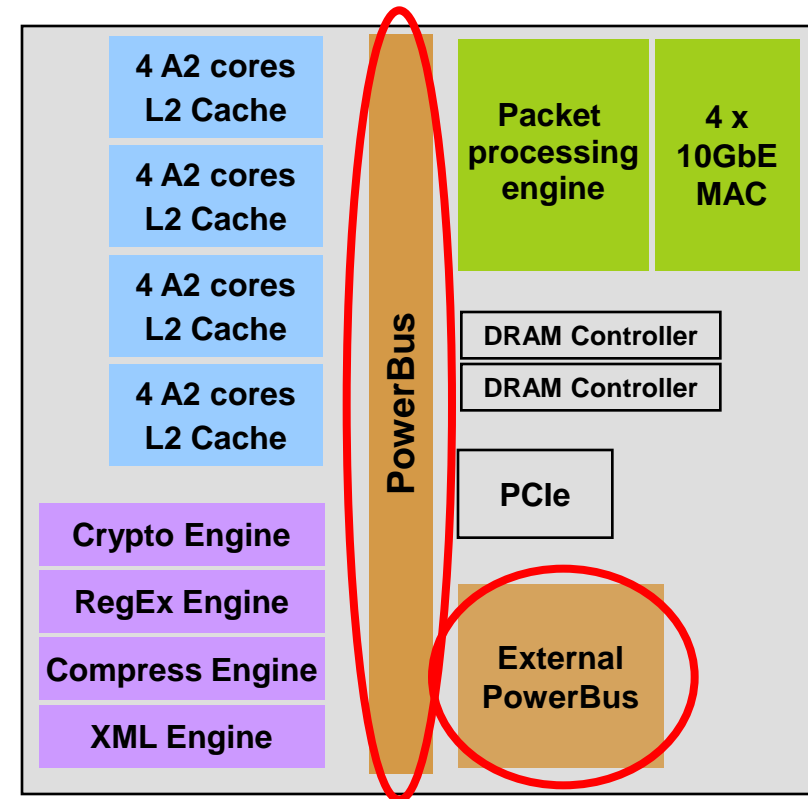
IBM PowerEN™ Chip

- Chip statistics:
 - 45nm - ultra-low k BEOL
 - Area: 410 mm²
 - 100 array types, ~300 instances
 - 124 RLMs, 1044 instances
 - 1.43B Transistors
 - 3.2 million latches
 - 5 PLLs, 10 concurrent frequencies
- Package Technology
 - 50mm FCPBGA (6) build-up layers
 - C4 Pitch: 148 x 185 mm pitch
- Power Delivery
 - Deep Trench Chip Capacitors
 - Adaptive Power Supply Strategy
 - 8 voltages domains



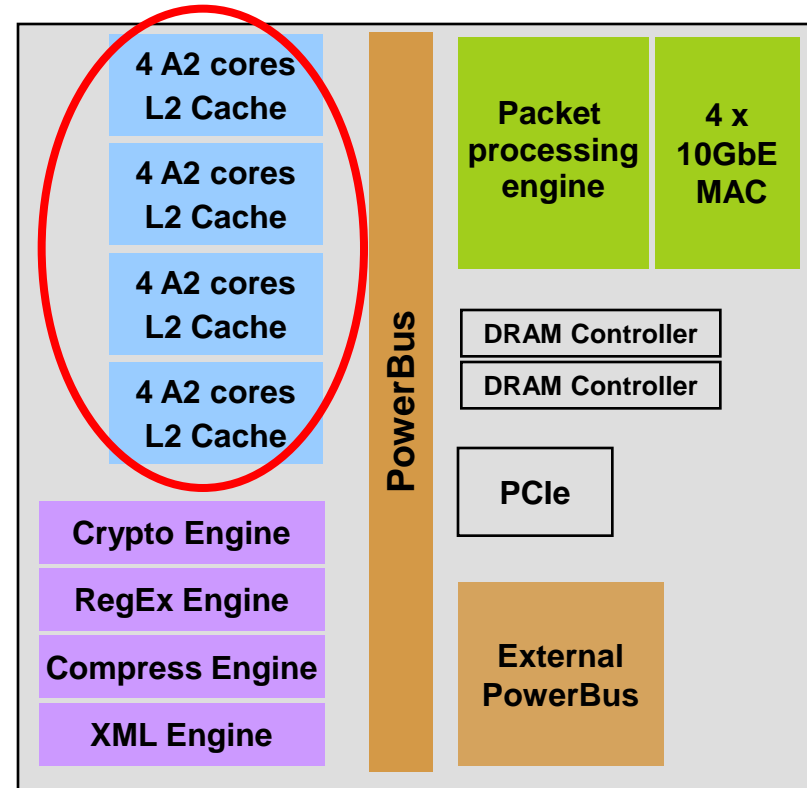
Interconnect Architecture

- “All Peers” architecture
 - Accel. and I/O are *first class citizens*
- Proven Power-Bus architecture
 - Independent CMD Network (one/cycle)
 - Two north, two south 16B data busses
 - ECC protected data paths
- 64 Byte Cache Line
- Cache Injection
 - Packets flow to / from Caches
 - New PBus commands
- 1.75 GHz operation
 - Asynchronous connection to AT Nodes and accelerators via PBICs
 - Synchronous connection to DRAM controllers
 - Three 4B 2.5 GHz EI3 external links (1,2, or 4 chip systems)



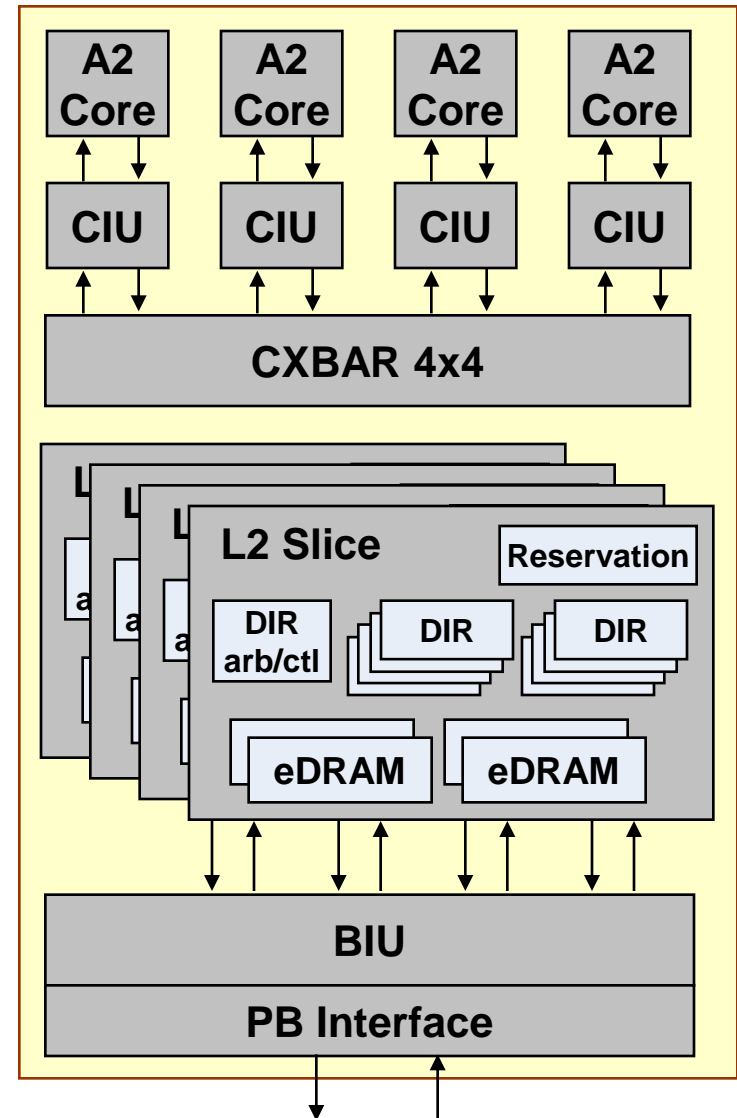
PowerPC™ Processing Element Architecture

- PowerPC 64 architecture – Embedded
- Enhanced for Co-processor/I/O interface
- 4 way Fine Grained SMT
- In Order Dispatch and Execution
- 2 way concurrent issue.
 - 1 Integer + 1 FPU instruction per cycle
 - Different threads
- Unified fixed point, ld/st, and branch unit
- 16KB L1 Data Cache – 4 Way Assoc.
- 16KB L1 Instr Cache – 8 Way Assoc.
- 12 Stage Pipe 27 FO4 design (7 XU, 5 IU, 6 FU)
- Fully associative I and D – ERAT
- MMU: 512 Entry TLB w/ Hardware Table Walk
- Hypervisor / Virtualization - One logical partition per core



PowerPC™ Processing Element Architecture

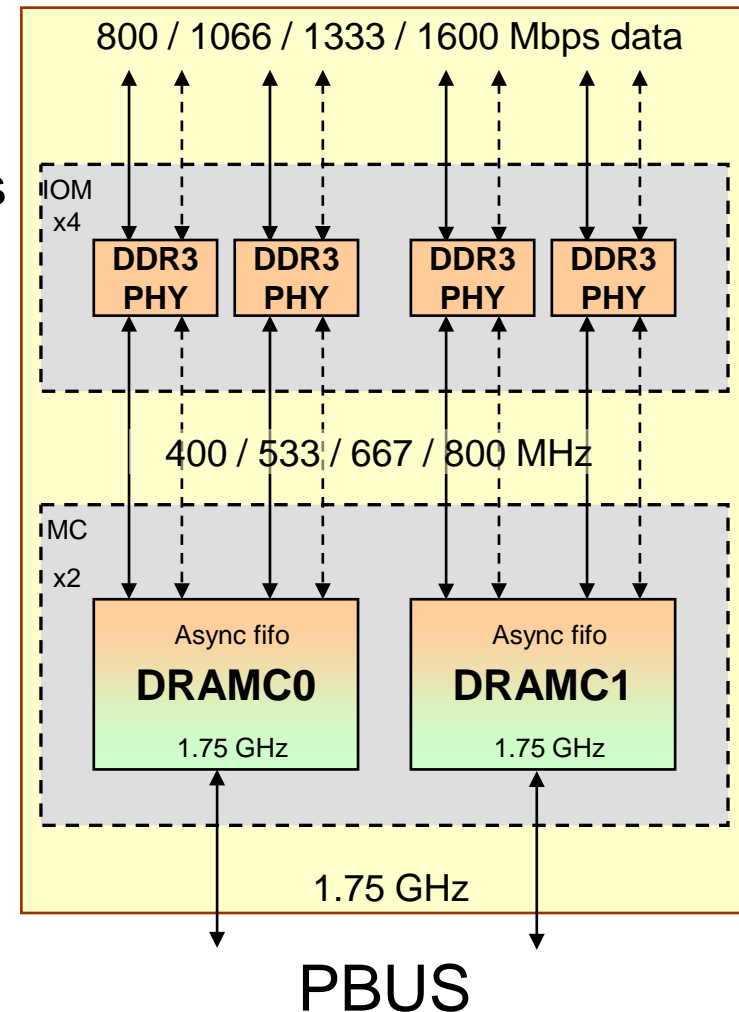
- 4 cores x 4 slices of shared L2
 - 512KB per slice
 - Concurrent reload data to all 4 cores
- 1:1 with processor cycle time
- 2MB eDRAM (total)
 - 64B cache lines
 - Inclusive of L1 I & D caches
 - 8 way set associative
- Fast core wake up on reservation loss
- ECC (SEC/DED) Data and on Directory
- Line locking & Way locking
- Slave memory region (non-coherent)
- Cache injection (full & partial line)
- Power Saving Mode: Rip Van Winkle



DDR3 DRAM Controller / PHY

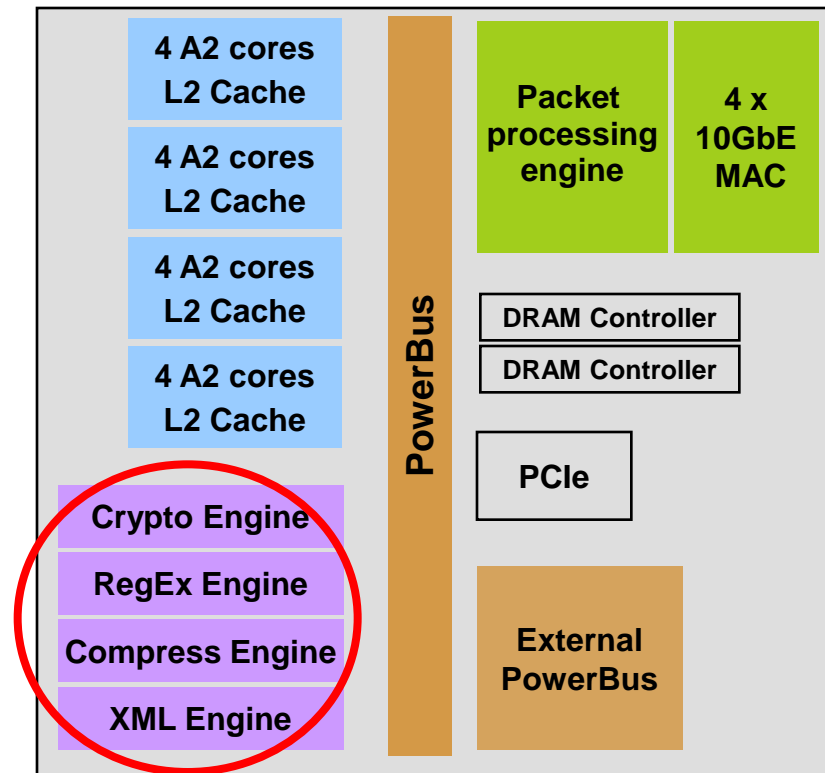
- Two independent DDR3 DRAM controllers
 - 2 independent channels / controller
 - Registered RDIMMs or unbuffered UDIMMs
 - Up to two DIMMs per channel and 1, 2 or 4 ranks on DIMM
 - Attach up to 32 GB per DRAM controller
 - 64 byte block ECC matches cacheline size
 - 800 MHz, 1.066 GHz, 1.33 GHz, 1.6GHz DRAMS
- PBus Interface
 - One PBus interface (C/D) per controller
 - Capable of 16 Bytes outbound per PBus cycle and 16 Bytes inbound per PBus cycle

DDR3 Memory



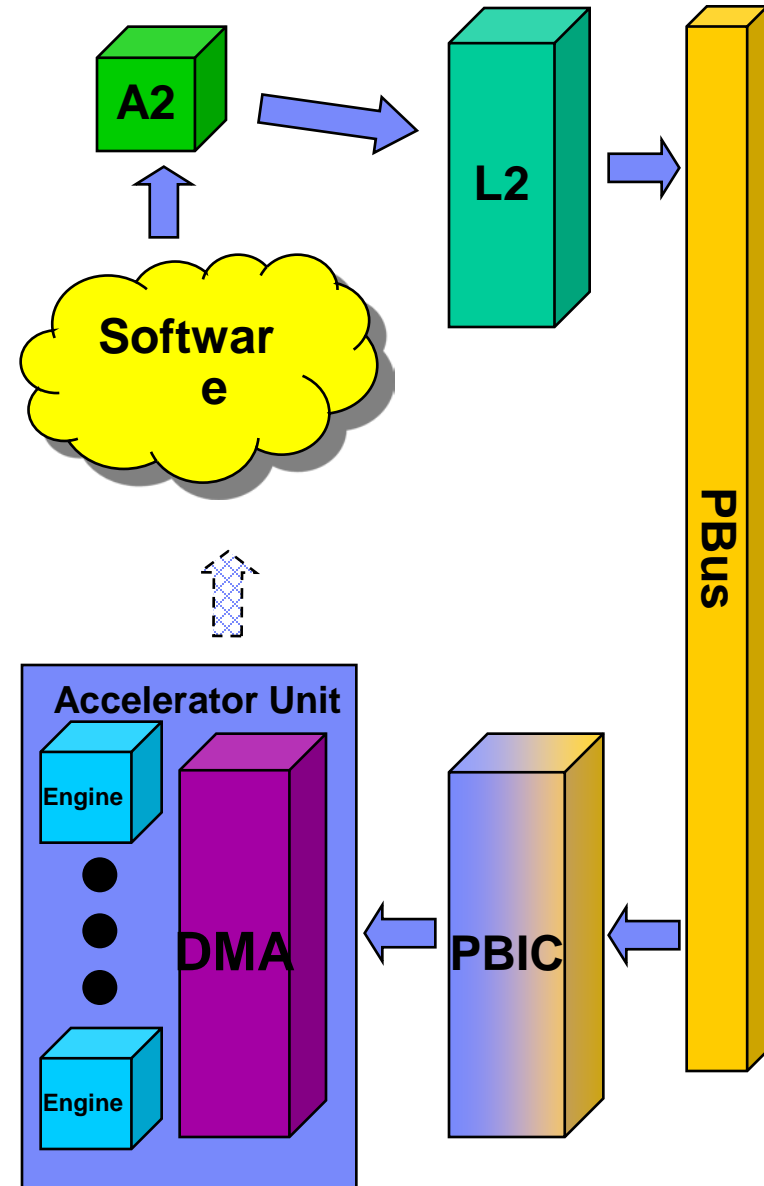
Accelerator (Co-Processor) Architecture

- Objectives
 - Performance
 - Common API (Ease of Use)
 - QOS
 - Virtualization – Protection
- Common Architecture for all Accelerators
- Integrated in Power Architecture
 - New Initiate Co-Processor Instruction
 - New Wait on Loss of Reservation
 - (Thread wake up)
 - Application Context communicated to Co-Processor
 - Access Control
- L2 Cache Intervention / Injection
- Accelerator MMU derived from Processor MMU
 - Accelerators operate in Application Address Space



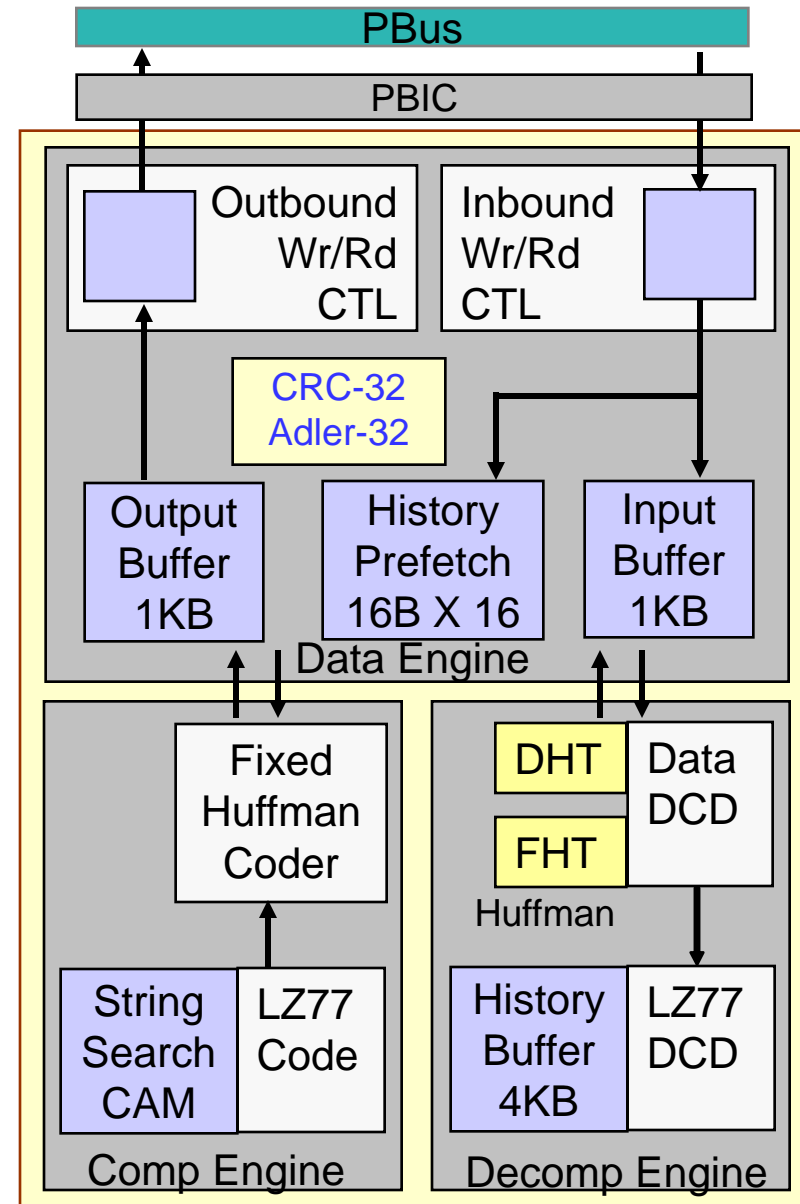
Accelerator Interface

1. Software receives an input packet
2. Software builds the CPB and CRB in cache
3. Software issues the ICSWX instruction
4. L2: ICSWX => Cop_Req PBus command
5. The PBus transports the Cop_Req
6. The PBIC passes the CRB and Cop_Req to Accelerator
7. The DMA logic assigns the Algorithm Engine and fetches data and parameters
8. The DMA logic and the Algorithm Engine work together to process the data and generate the output data
9. The DMA logic stores the output data
10. The DMA logic stores the status into the CSB
11. The DMA logic performs the final handshake
12. The SW retrieves the output status and data



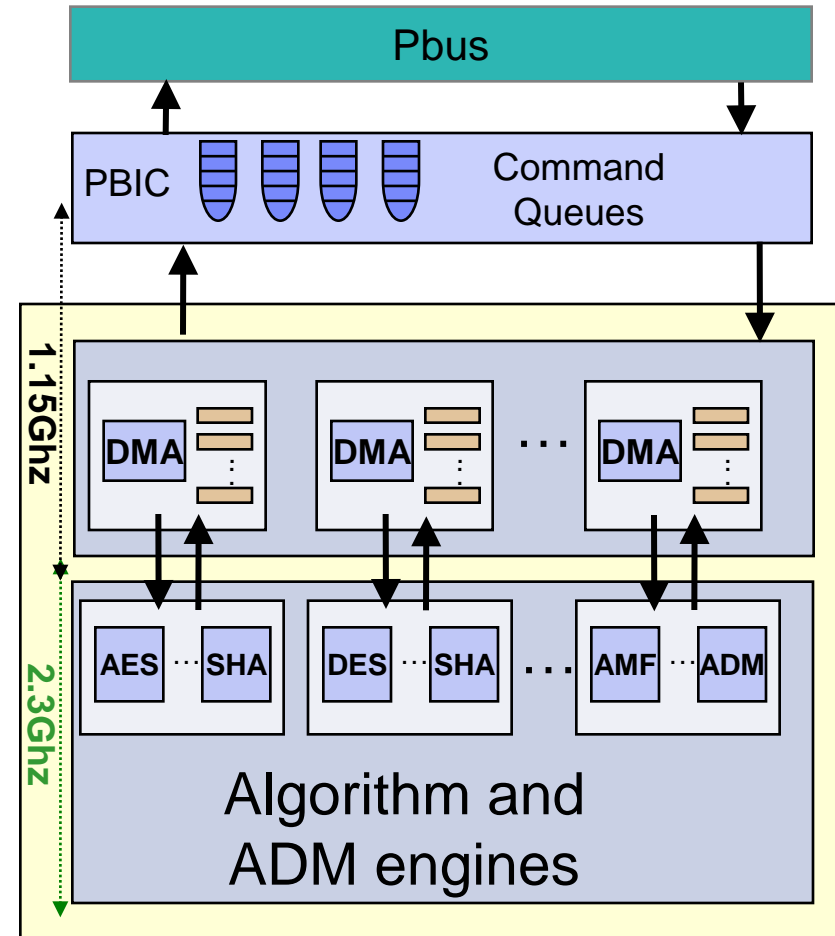
Compression/Decompression

- Standards
 - Supports file formats defined by RFC1950 (ZLIB) and RFC1952 (GZIP)
 - Compliant to RFC1951 and DEFLAT
- Pipelined data engine
 - Deep pipelines to minimize latency and increase bandwidth
 - 8 Gbps output from engine (Decomp)
 - 8 Gbps input to engine (Comp)
- Decompression
 - Supports interleaved messages, packet by packet decompression
 - Static & Dynamic Huffman decoding supported
- Compression
 - Supports single messages to be compressed
 - Static Huffman coding support



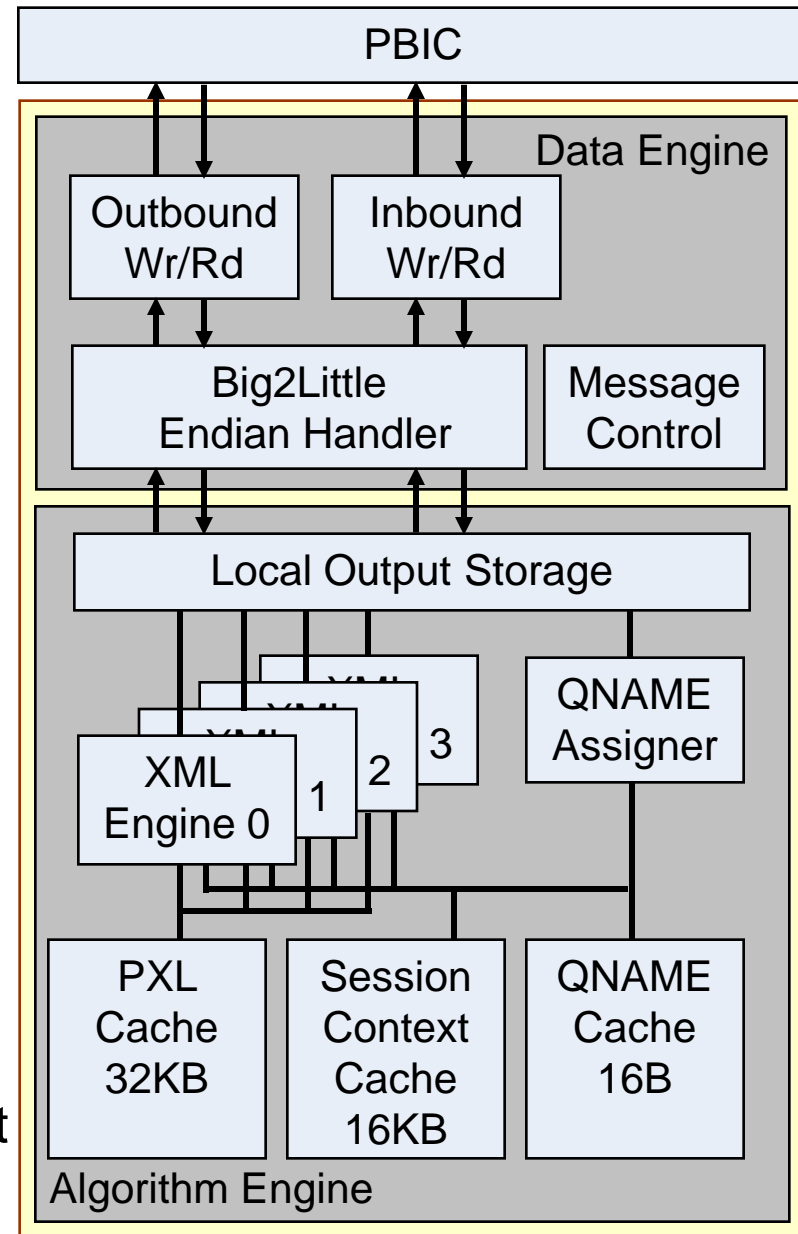
Crypto Data Mover

- Symmetric Algorithm Acceleration
 - AES Modes:
 - Key Lengths: 128b, 192b, 256b
 - DES & 3DES Modes
 - ARC4, Kasumi
 - HASH: SHA-1, SHA-256, SHA-512, MD5
 - HMAC supported for SHA
 - Combined 3DES/AES and SHA
- Asymmetric Algorithm Acceleration
 - Modular Math Functions for RSA/ECC
 - Point Functions for ECC
 - RSA lengths: 512b, 1024b, 2048b, 4096b
- Asynchronous Data Mover (ADM)
 - Any Source Byte offset, Any Dest Byte offset, Any length up to 16M bytes
- Random Number Generator (RNG)
 - Supplies a 64b random number, Supports FIPS 140 compliance



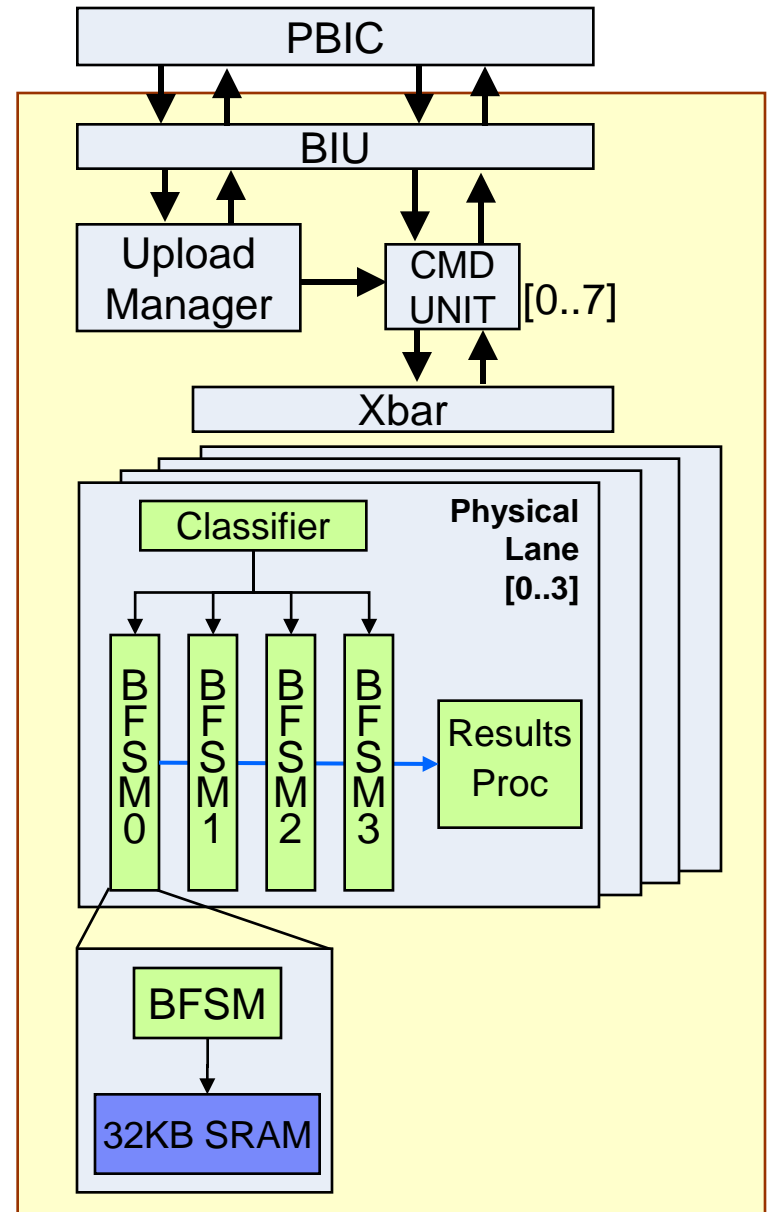
XML Unit / XML Engines

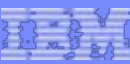
- Fully asynchronous offload operation
- Four Parsing Engines
 - Performs lexical analysis
 - Checks well-formedness
 - Normalizes whitespace
 - Seamlessly switches state to process interleaved document fragments
 - Processes multiple characters simultaneously
 - Supports multiple character encodings
- Four Post Processing Engines
 - XPATH evaluation
 - Schema validation
 - Filtering in hardware (reject)
 - Process a fragment of incoming document
 - XSLT processing, XML Routing



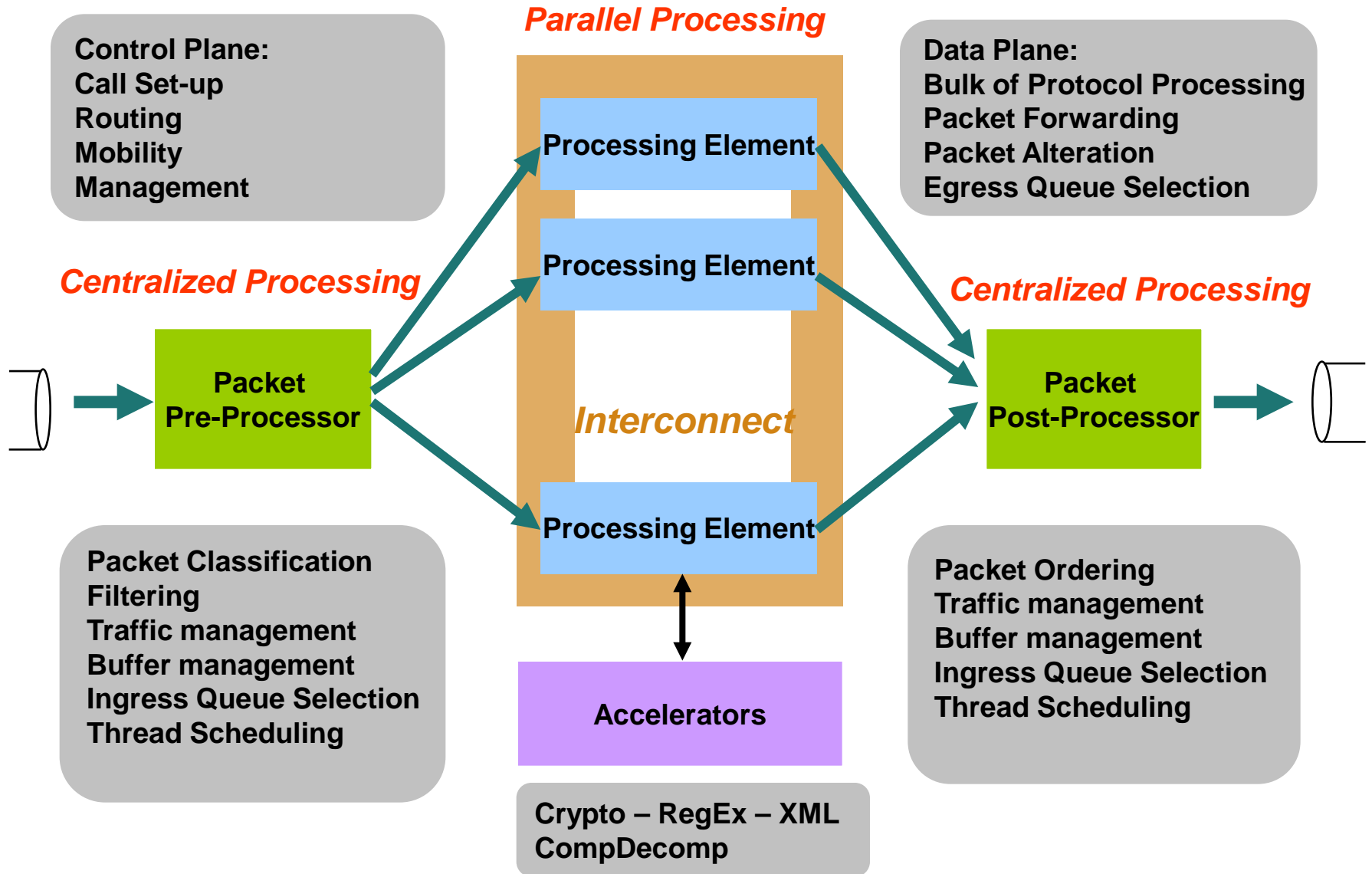
RegX (Pattern Matching Engine)

- Processes 8 CRB's in parallel
- Four independent Physical Lanes; each composed of 4 programming state machines (BFSM)
 - Each lane is time multiplex by two logical lanes
 - Each BFSM is connected to 32KB of SRAM (total of 512KB)
- SRAM holds resident rules (SW managed cache) and temporary rules (HW managed cache)
- Strong dependency between hardware, compiler, patterns and workload



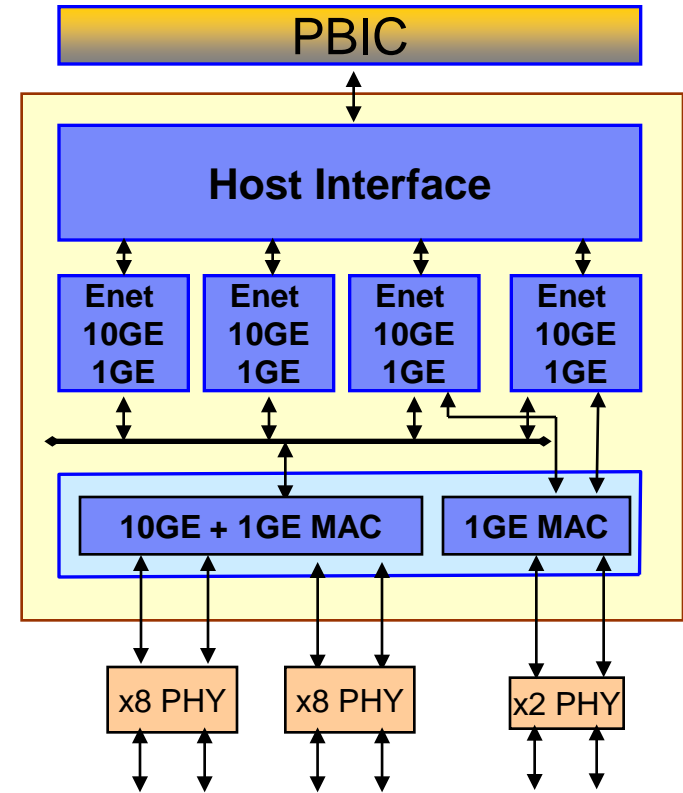


IBM PowerEN™ Packet Processing Framework



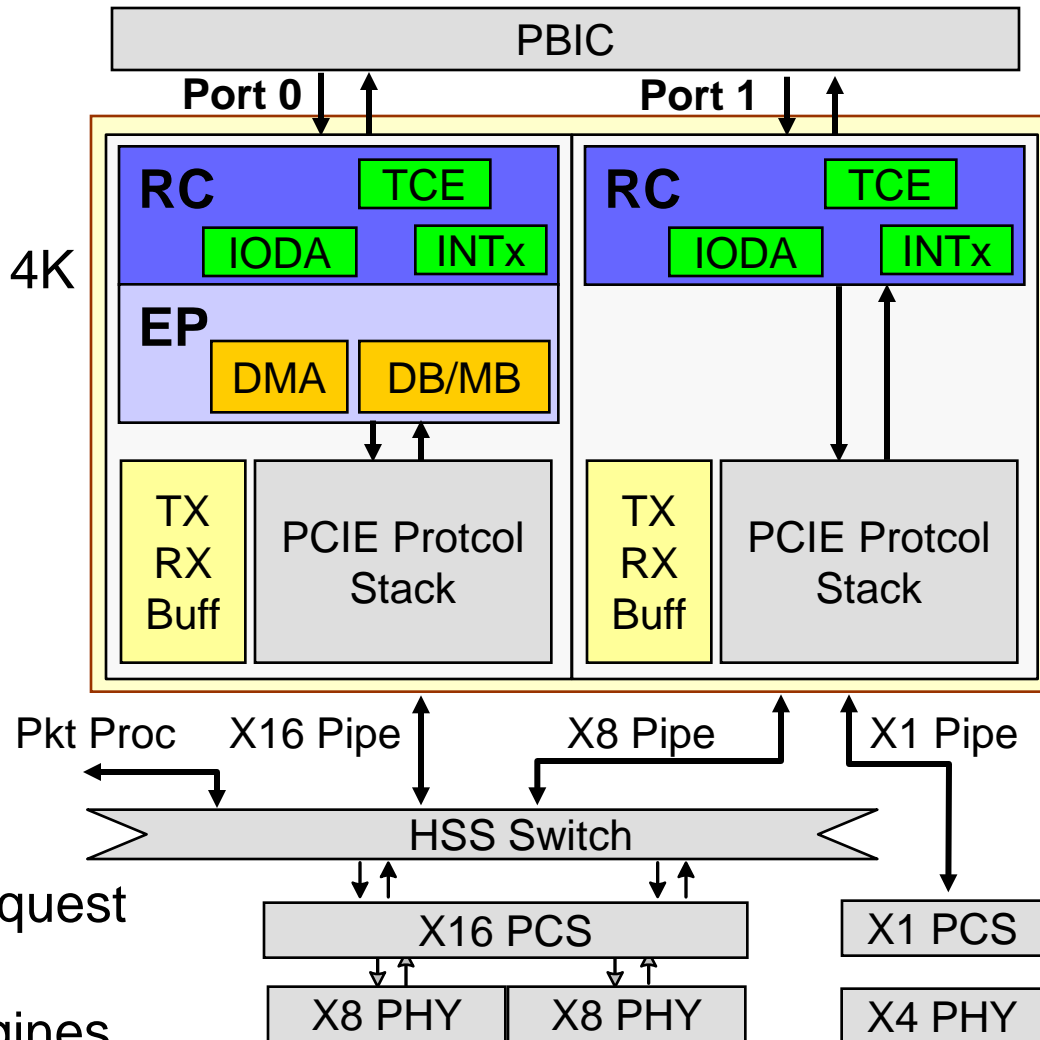
Packet Processor Architecture

- Offload Centralized Media Speed Functions
 - Packet Classification / Distribution / Ordering
 - BFSM based Parser
 - Virtualization - Up to 16 LPARs, Integrated L2 virtual Switch, PVID
- END POINT MODE: L4+ termination
 - Pull model Software interface (128 Queues)
 - Scatter/gather descriptors
 - Low latency Queues, Header separation
 - TCP/UDP IPv4, v6 Checksum assist
 - QOS support (ingress Queue selection)
- NETWORK NODE MODE: Packet forwarding
 - Push model Software interface (64 Queues)
 - HW managed queues
 - Ingress – Egress Scheduler (Flexible ingress queue) selection
 - Completion Unit (Packet ordering – 16K packets)
 - Thread to Thread messaging with ordering

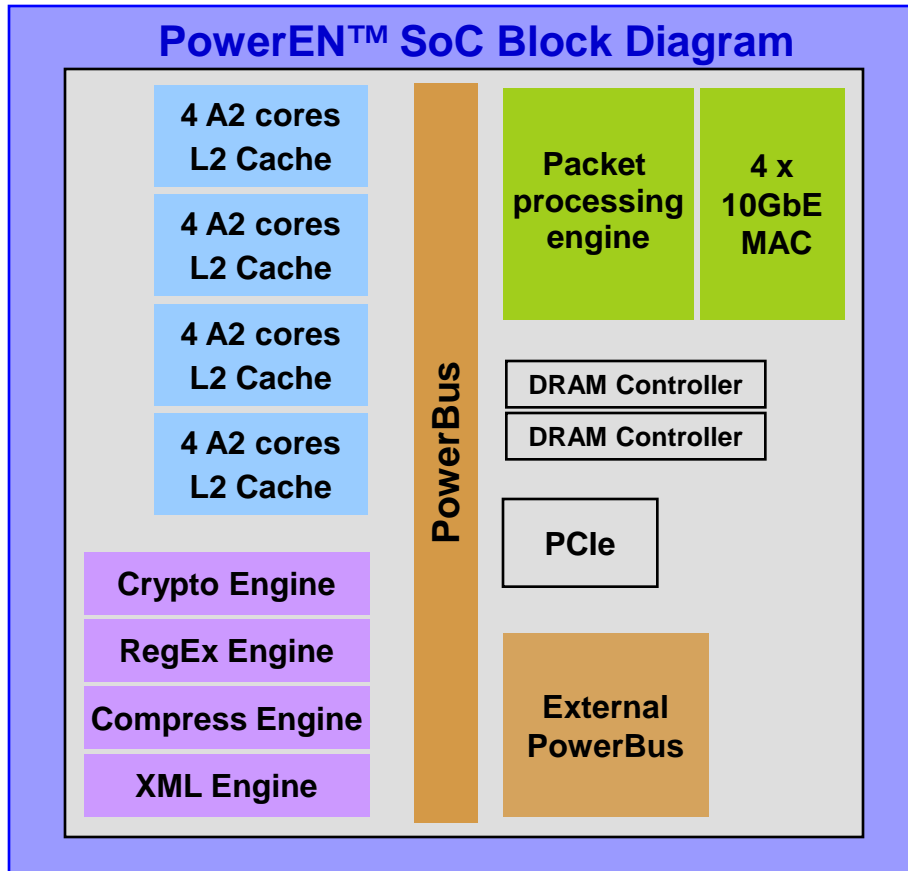


PCI-Express

- Two PCIe ports
- PCIe Gen 2 Features
 - 5Gb/sec per lane/direction
 - Max Payload: 512B, Max Read: 4K
- Root Port Mode - PHB
 - IODA-based definition
 - TCE based Address Translation
 - TCE cache: 64-entry 4-way
 - Inbound MSI Validation
- Endpoint Mode
 - PCIe SRIOV Virtualization
 - DMA Follows accel. SW model
 - Engaged via Coprocessor Request
 - Similar compl/status reporting
 - Tx and Rx data streaming engines
 - Separate Doorbell and Interrupts per PCIe PF/VF
 - Dedicated Mailbox space

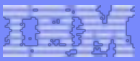


IBM PowerEN™ Processor Chip



- ✓ Targeted at the Edge-of-network
- ✓ Power efficient Throughput computing
- ✓ Enhanced processing of data payloads
- ✓ Deeper networking functions
- ✓ Application targeted at "smarter planet" solutions

PowerEN™ has all basic IP needed for user/data plane and control plane traffic processing at the Edge of Network



Thank You



Glossary

- **BFSM** – **B**inary **F**inite **S**tate **M**achine
- **BIU** – **B**us **I**nterface **U**nit
- **CCB** – **C**oprocessor **C**ompletion **B**lock
- **CIU** – **C**ore **I**nterface **U**nit
- **Cop_Req** – **C**oprocessor Request Power Bus Command
- **CPB** – **C**oprocessor **P**arameter **B**lock
- **CRB** – **C**oprocessor **R**equest **B**lock
- **CSB** – **C**oprocessor **S**tatus **B**lock
- **DHT** – **D**ynamic **H**uffman **T**able
- **FHT** – **F**ixed **H**uffman **T**able
- **HSS** – **H**igh **S**peed **S**erial
- **ICSWX** – **I**nitiate **C**oprocessor **S**tore **W**ord **I**nde**X**ed
- **LPAR** – **L**ogical **P**artition
- **PBIC** – **P**ower **B**us **I**nterface **C**ontroller
- **PBus** – **P**ower **B**us
- **PVID** – **P**artition **V**irtual **I**dentifier