The IBM Power Edge of Network™ Processor:

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

Hubertus Franke
Research Staff Member, Mgr Scalable Systems Department
IBM T.J. Research Center
Yorktown Heights, NY 10598
frankeh@us.ibm.com

Co-authors: To many to mention (Research / STG)
Chief Architects: Charlie Johnson, Jeff Brown

Disruptive Innovations

Cloud Computing

Smarter Planet

SOA

Physics

Virtual Worlds

Game Players

Mashups

Mobile Devices

Crowdsourcing

Petaflop Supercomputers

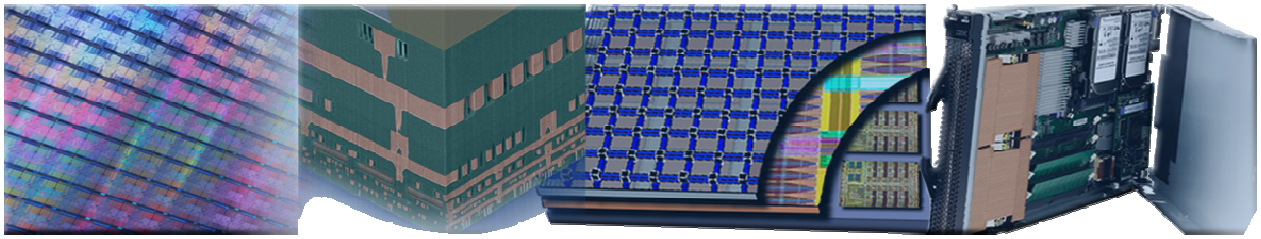
Software as a Service

Web 2.0

Globally Integrated Enterprise

Services Sciences

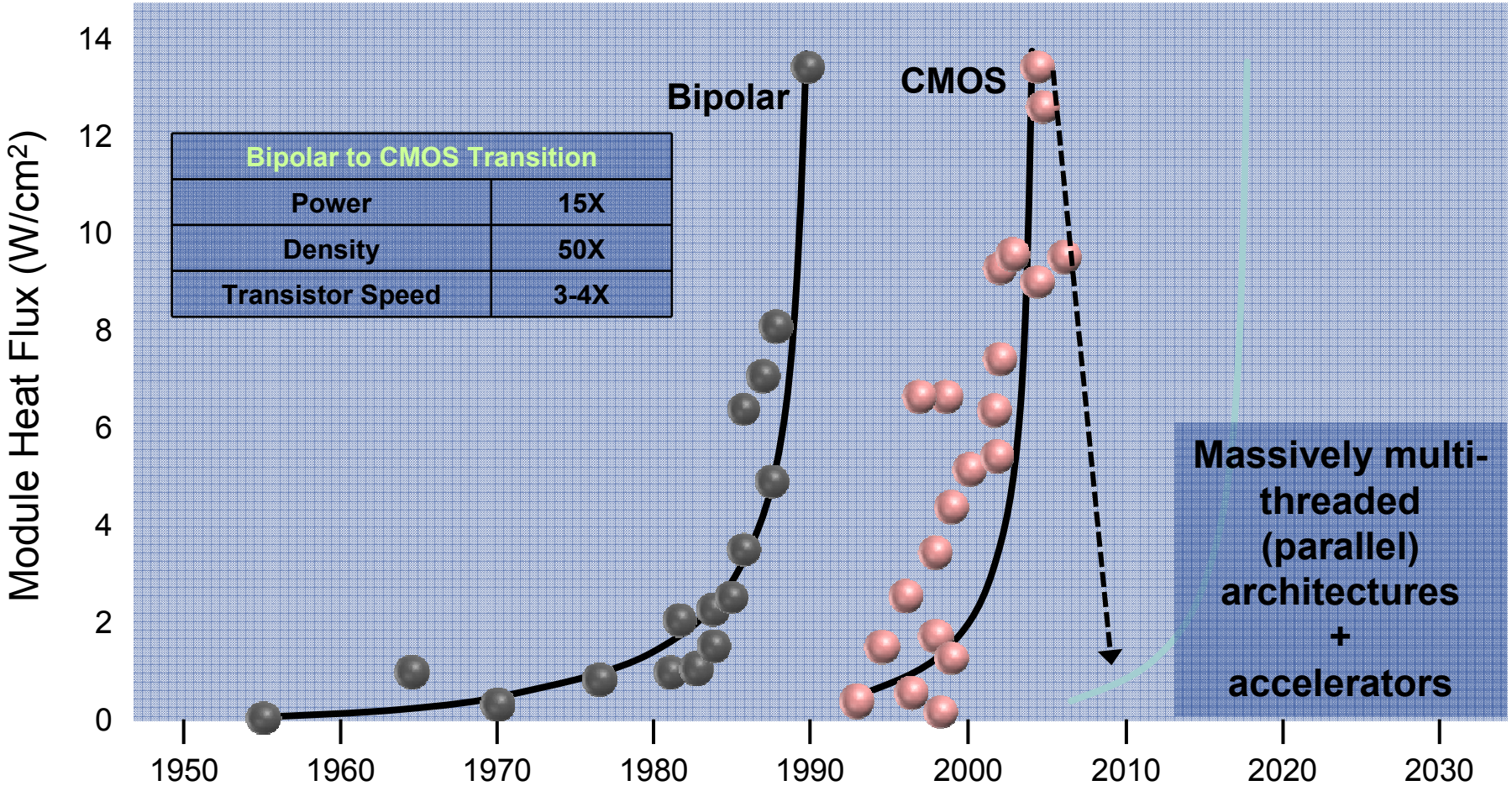
Reinventing How Systems are Built



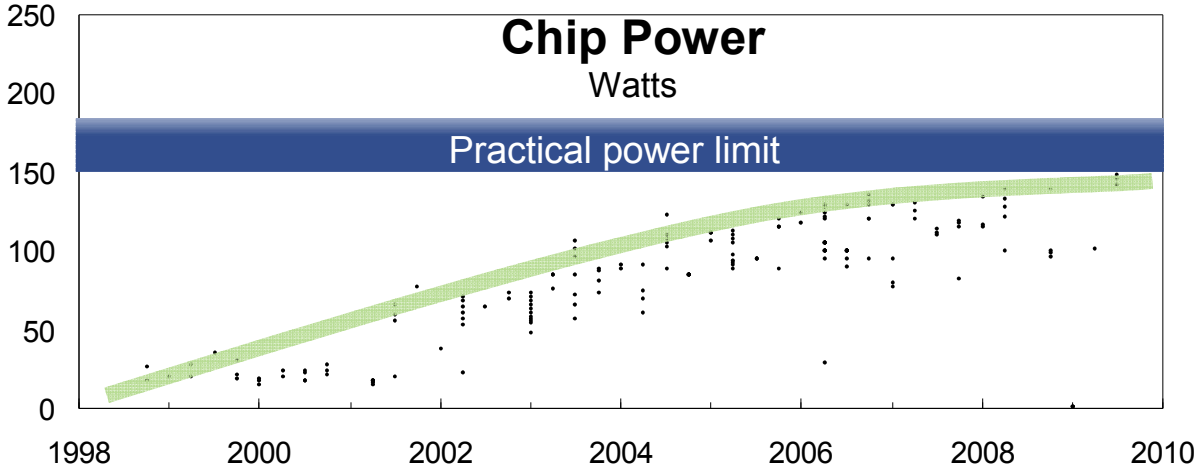
System Architecture

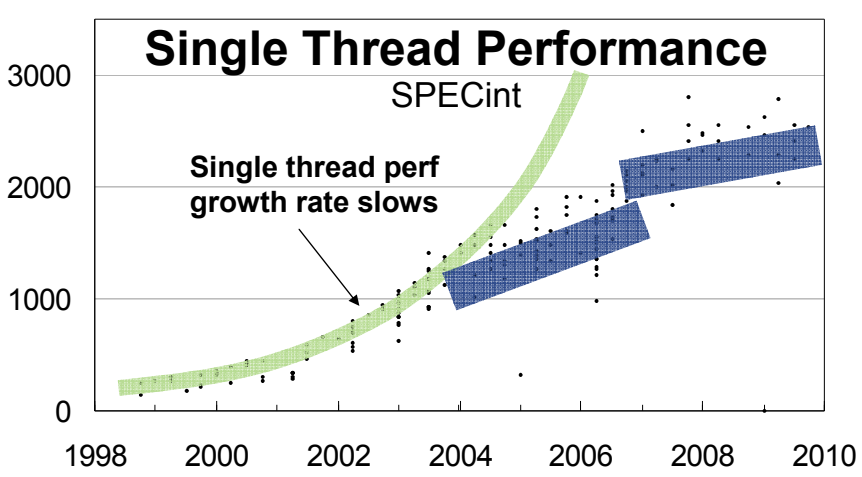
- **Technology discontinuities driven by *physics***
- **Business needs driving specialized hybrid systems**
- **Hybrid system future directions**

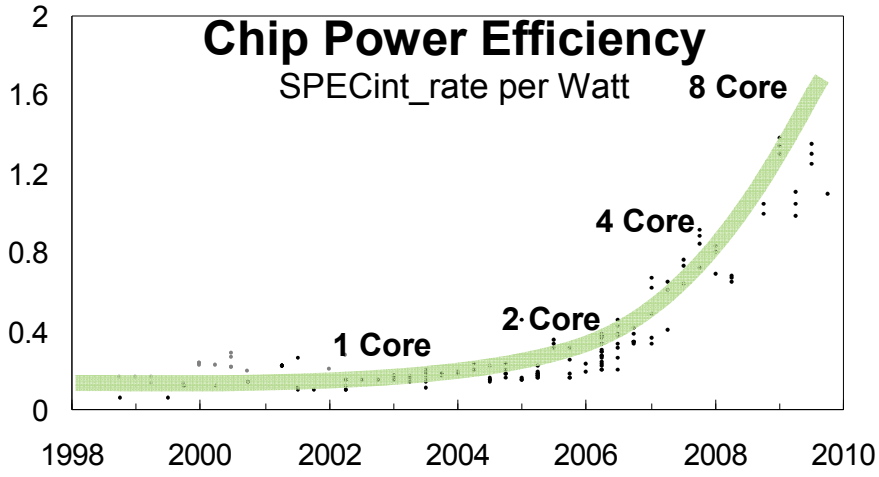
Key Technology Inflection Points



Microprocessor Technology Trends

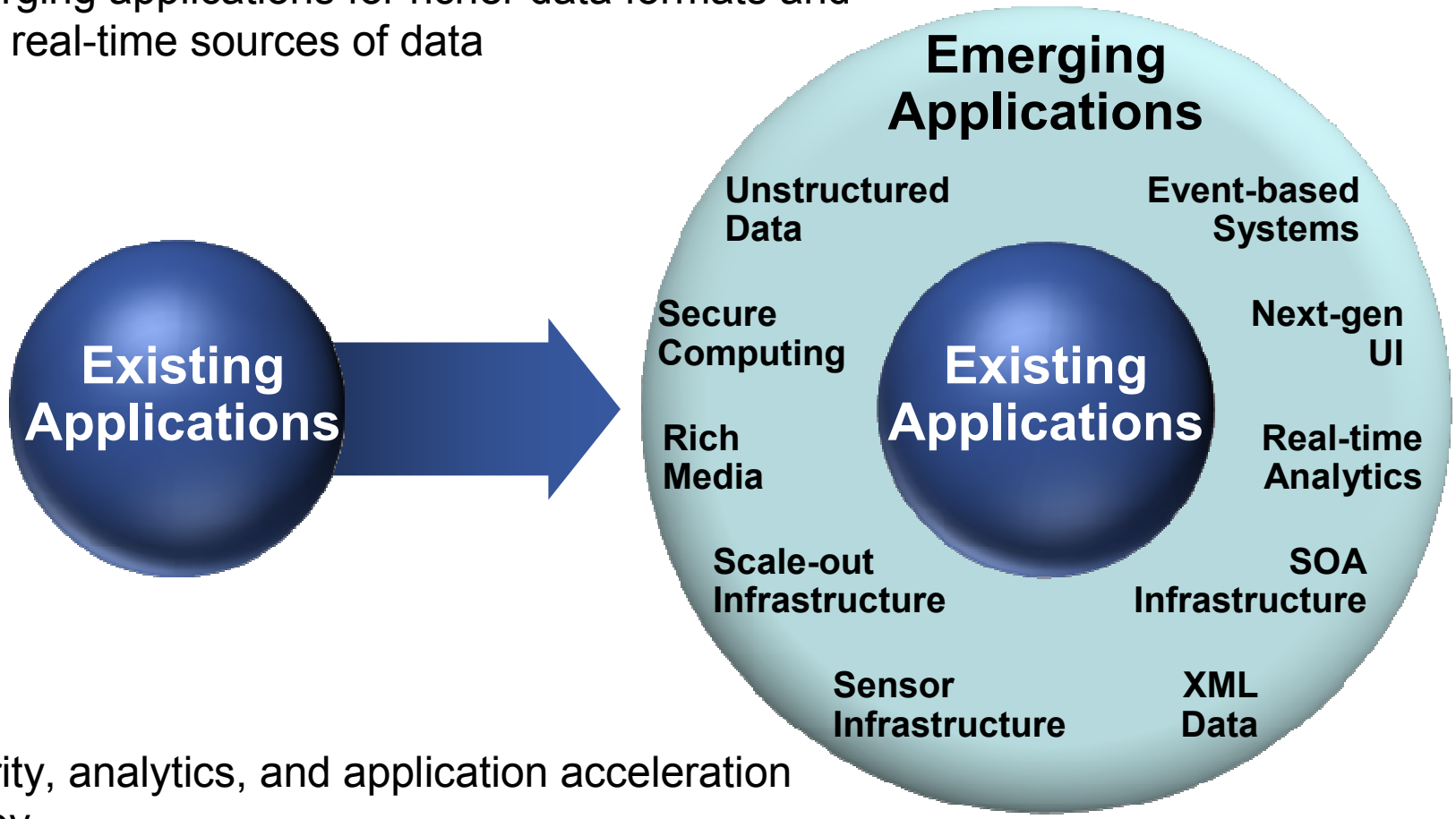






Future Systems and New Applications

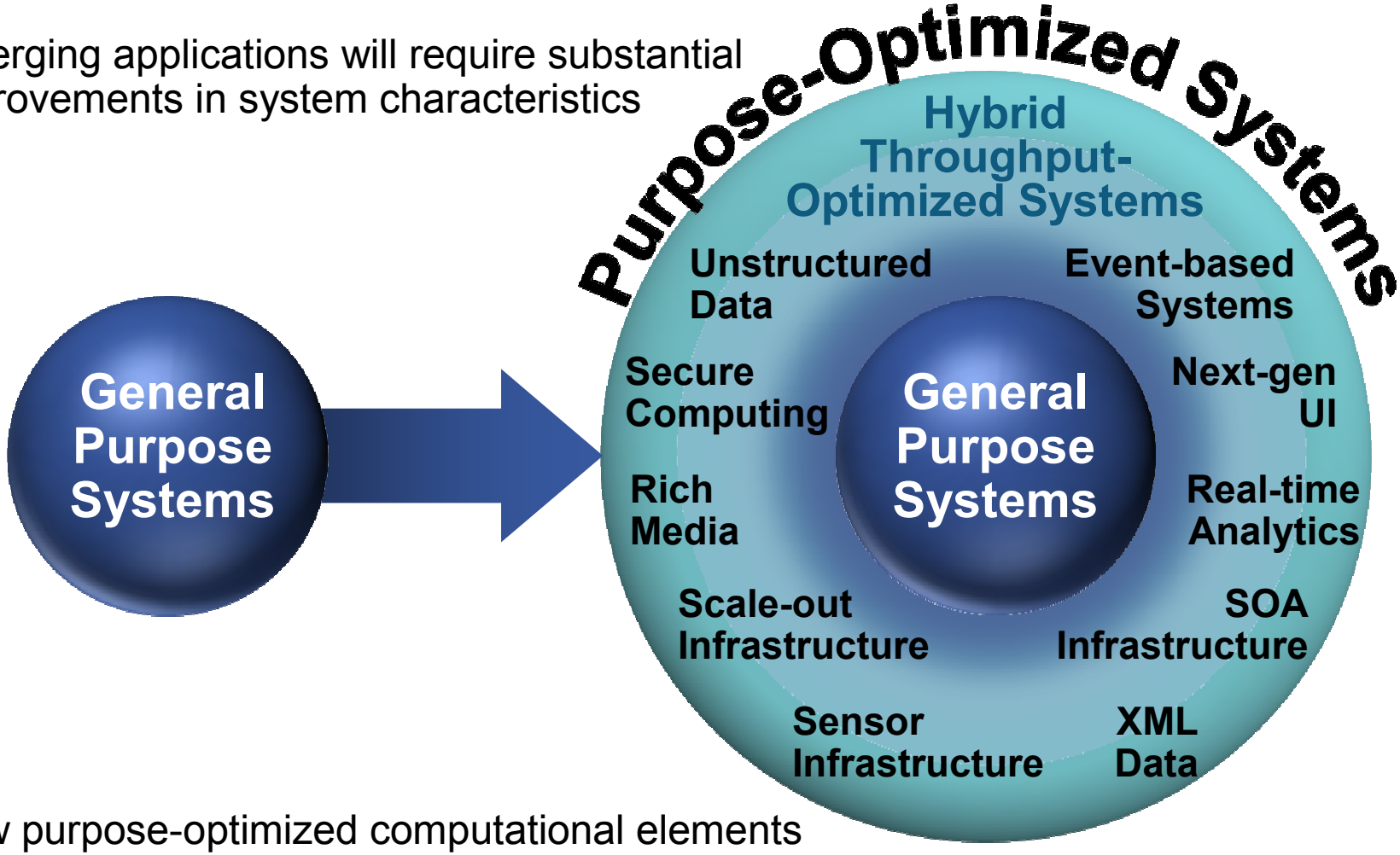
- Emerging applications for richer data formats and near real-time sources of data



- Security, analytics, and application acceleration are key

Future Systems and New Applications

- Emerging applications will require substantial improvements in system characteristics



- New purpose-optimized computational elements will fundamentally transform systems

Opportunity Exists for Optimization Across the Workload Space

System Workload Categorization

Transaction Processing and Database

- Data Warehousing
- OLTP
- Batch
- Business Processing DB

Analytics and High Performance Computing

- Data Mining Applications
- Numerical
- *Enterprise Search*
- *Gaming & Visualization*

Business Process Applications

- ERP
- SCM
- CRM

Web, Collaboration and Infrastructure

- Application Development
- Systems Management
- Infrastructure DB
- File & Print
- Web/Collaboration Content DB
- Web Serving/Hosting
- *Web/Networking Acceleration*
- Security
- Networking
- *IMS/VOIP Infrastructure*
- Proxy Caching
- Collaboration

Workload Attributes

Transaction Processing and Database

Scale
High Transaction Rates
High Quality of Service
Handle Peak Workloads
Resiliency and Security

Analytics and High Performance Computing

Compute intensive
High I/O Bandwidth
High Memory Bandwidth
Floating point
Scale-out Capable

Business Process Applications

Scale
High Quality of Service
Large Memory Footprint
Responsive Infrastructure

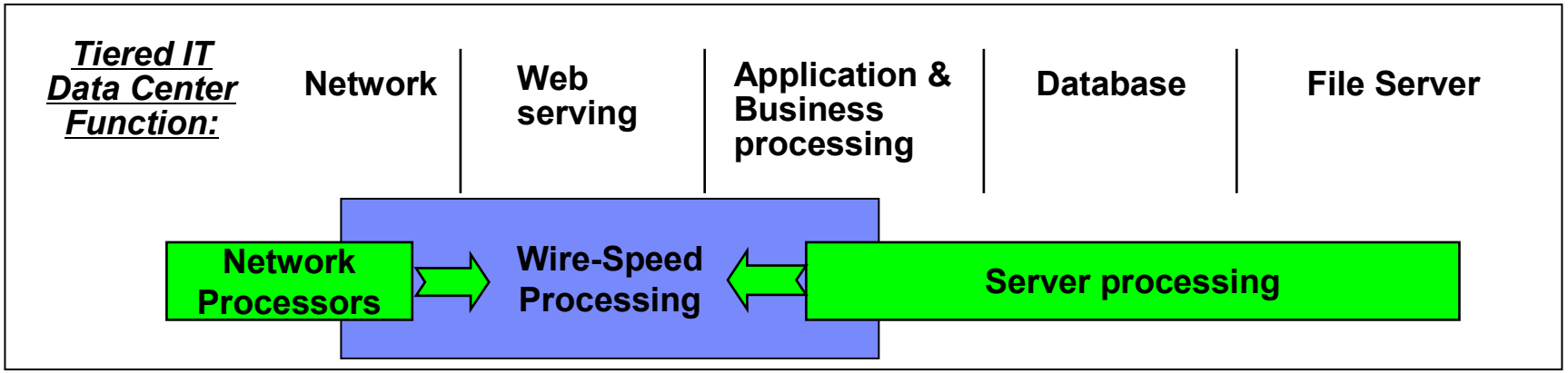
Web, Collaboration and Infrastructure

Highly Threaded
Throughput-oriented
Scale-out Capable

Wire-Speed Processing

A blurring of the Network and Server worlds

- Highly-multi-threaded low power cores with full PPC ISA
- Standard programming models with OS's & hypervisors
- Virtualization support for application consolidation
- Accelerators: for both Networking & Application tiers
- Integrated Network system & Memory I/O
- Server RAS & infrastructure
- Low total power solution based on throughput optimization

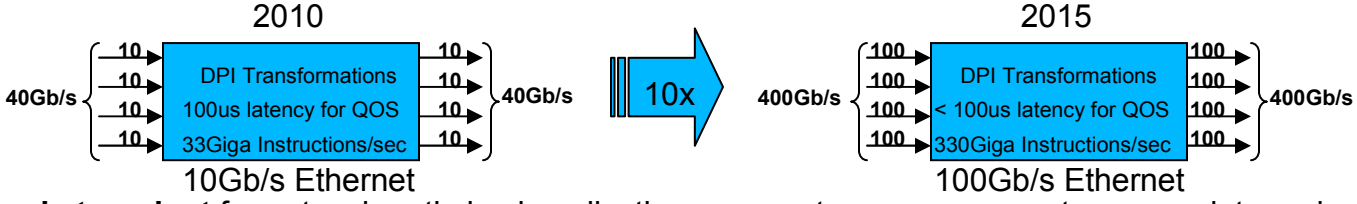


Wire-speed Differentiated Solutions

Emerging class of throughput and latency sensitive applications **whose performance and scaling require the optimization of both** networking and computing on specialized wire-speed processors

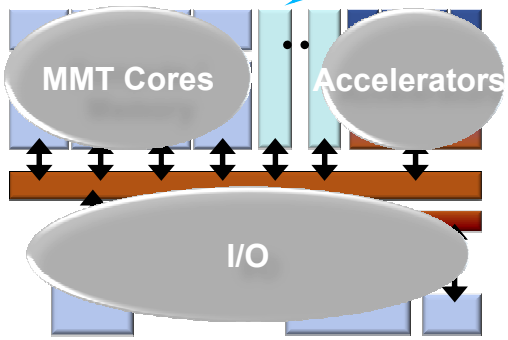
Applications require: **near real-time latency, high throughput, and ability to scale 1-1 with networking roadmaps**

Examples: **Deep-Packet Inspection** (DPI) as prototypical transformation function for security, monitoring, filtering, pattern matching, and data conversions

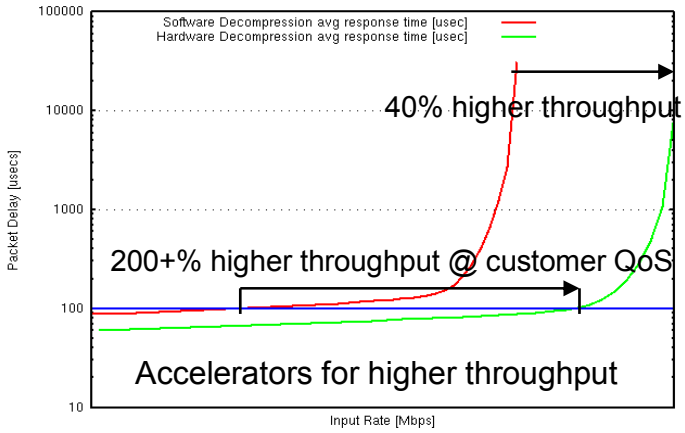


As **data flow is transient** for network-optimized applications, compute resources must process data and scale with networking rates with near real-time latency

Systems require: a **scalable wire-speed compute node** as basis engine which cannot be accomplished with commodity components due to compute density, throughput, latency, and power requirements



- Massively multi-threaded cores for concurrency, latency and power
- Accelerators for throughput, latency and power
- Tightly coupled integrated I/O for performance and latency

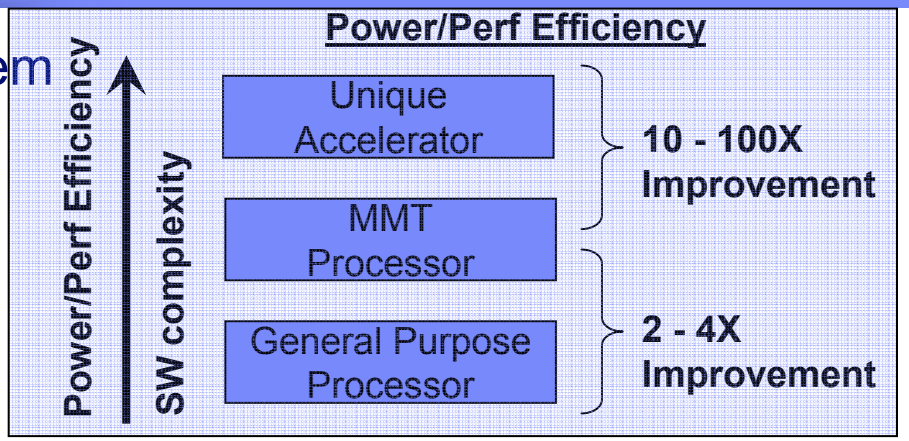


Pervasive wire-speed compute node forms the basis for a class of highly scalable and distributed computing and stream-based systems

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

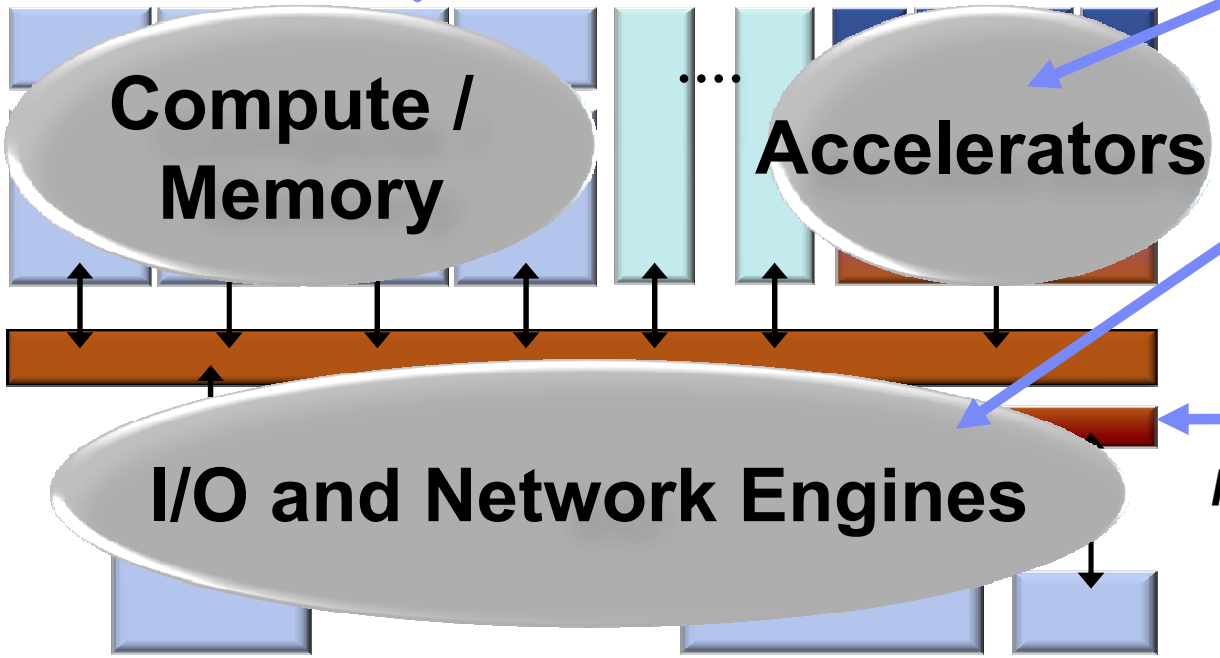
- Power efficient Throughput computing
  - Database Acceleration
  - Service Oriented Architecture Acceleration
  - Secure Multi-tenant Cloud Computing
- Enhanced processing of data payloads
  - Low latency message for Financial Information Exchange
- Deeper networking functions
  - Cyber Security
 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

Power EN Wire-speed processing system



Massive Multi-Threading

- High-efficiency design: Performance per watt & performance per cu. foot
- Simplicity of SMP programming model reduces S/W development costs



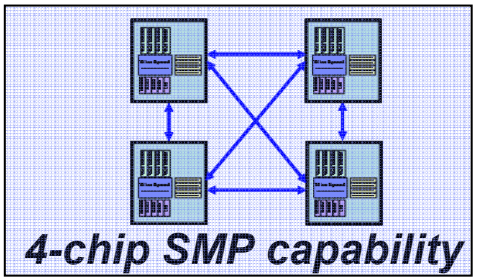
Accelerators

- Crypto for pervasive security
- Compression for BW opt.
- XML for SOA support
- Pattern matching for filtering

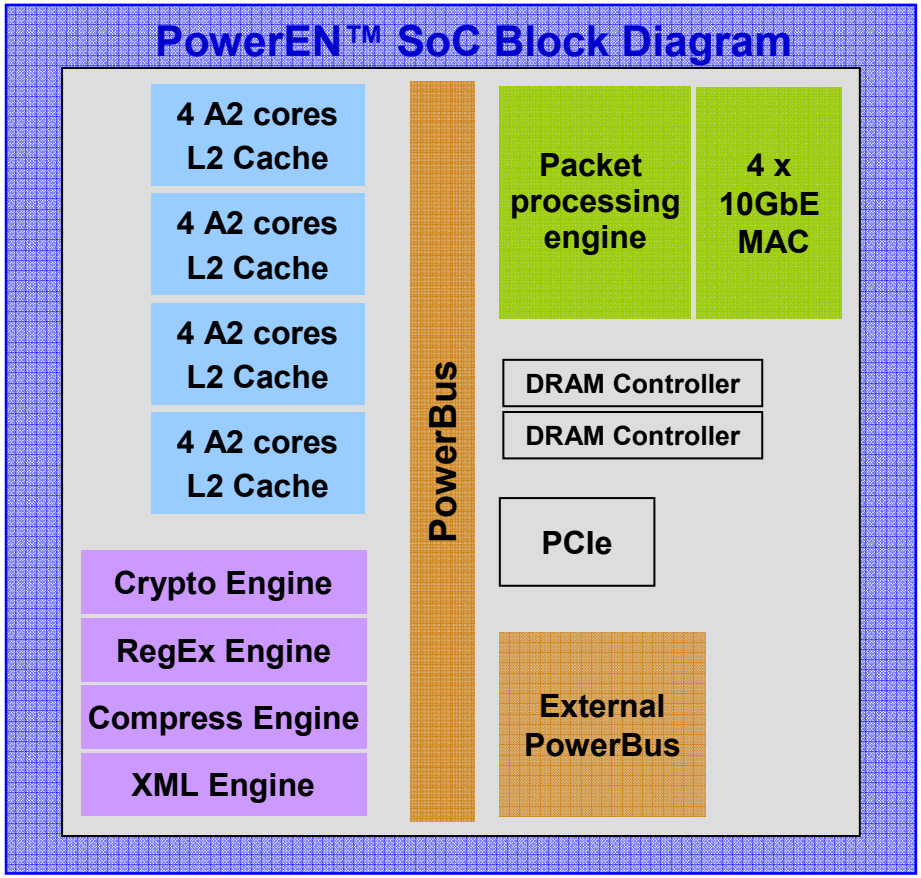
Integrated I/O

Efficient system-level I/O processing driven by emerging application requirements

Network packet processing engine



Overview of IBM PowerEN™ Processor Chip

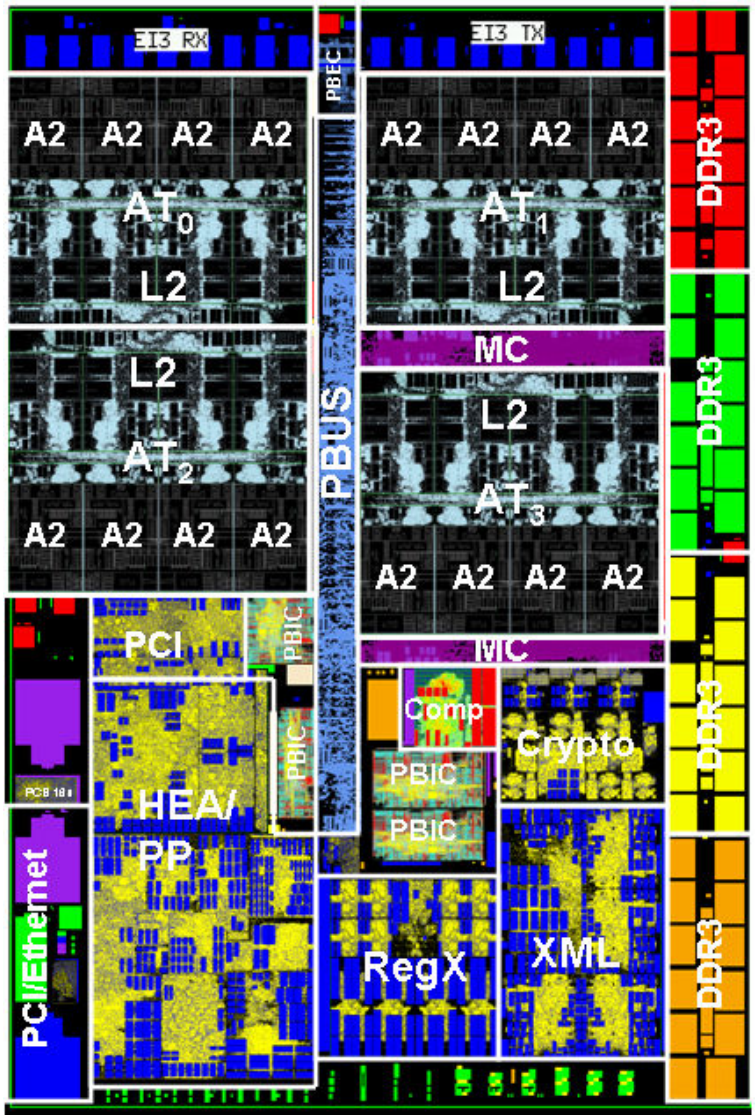


- ☑ 64-bit PowerPC Architecture
- ☑ Virtualization Support
- ☑ Dual DDR3 DRAM Controllers
- ☑ Optimized Ethernet Offload Engine
- ☑ Integrated PCI-Express bus
- ☑ Cryptography Unit
- ☑ Regular Expression Unit
- ☑ XML Processing Unit
- ☑ Compression Unit
- ☑ Upward Scalability – 4 Chip
- ☑ Downward scalability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

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

PowerEN: Wire-Speed Processor

Technology	IBM 45nm SOI
Core Frequency	2.3GHz @ 0.97V (Worst Case Process)
Chip size	428 mm2 (including kerf)
Chip Power (4-AT node) Chip Power (1-AT node)	65W @ 2.0GHz, 0.85V Max Single Chip 20W @ 1.4GHz, 0.77V Min Single Chip
Main Voltage (VDD)	0.7V to 1.1V
Metal Layers	11 Cu (3-1x, 2-1.3x, 3-2x, 1-4x, 2-10x)
Latch Count	3.2M
Transistor Count	1.43B
A2 Cores / Threads	16 / 64
L1 I & D Cache	16 x (16KB + 16KB) SRAM
L2 Cache	4 x 2MB eDRAM
Hardware Accelerators	Crypto, Compression, RegX, XML
Intelligent Network Interfaces	Host Ethernet Adapter/Packet Processor 2 Modes: Endpoint & Network
Memory Bandwidth	2x DDR3 controllers 4 Channels @ 800-1600MHz
System I/O Bandwidth	4x 10G Ethernet, 2x PCI Gen2
Chip-to-Chip Bandwidth	3 Links, 20GB/s per link
Chip Scaling	4 Chip SMP
Package	50mm FCPBGA (4 or 6 layers)



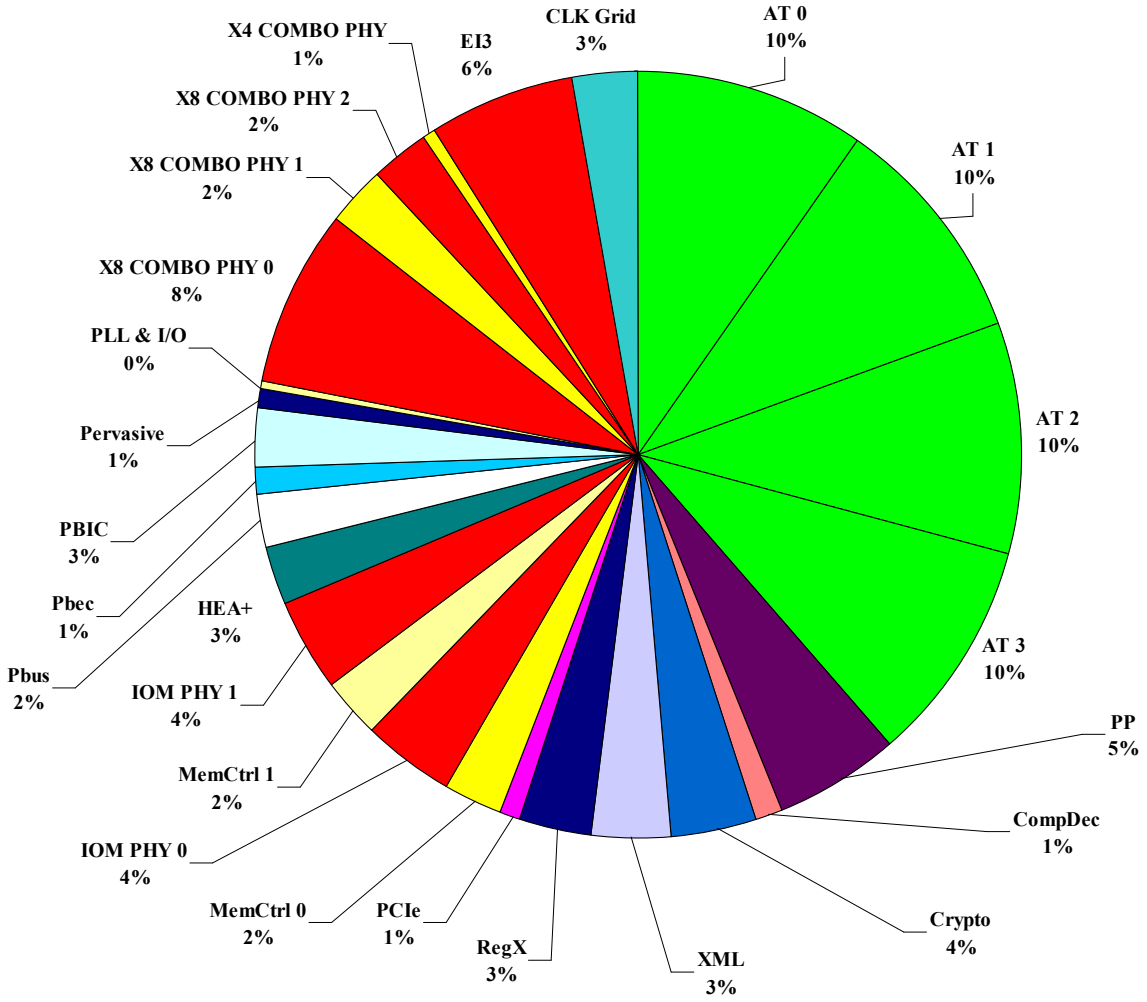
Special Purpose Hardware - PowerEN Accelerator Bandwidths

Accelerator Unit	# of Engines	Algorithm		Typical Bandwidth		Peak Bandwidth	
				single stream	total	single stream	total
HEA	4	Network node mode (Gbps)	256B packet	10	40	10	40
	4	Endpoint mode (Gbps)		10	40	10	40
De/Compression	1	Compression (Gbps input)		8	8	16	16
	1	Decompression (Gbps output)		8	8	16	16
Security Unit*	8	DES (Gbps)	256B payloads	4	32	6	48
	5	AES (Gbps)		5	25	8	40
	8	TDES (Gbps)		2	16	2.5	20
	2	ARC4 (Gbps)		1	2	2.5	5
	2	Kasumi (Gbps)		2.5	5	3	6
	6	SHA-1,256,512 (Gbps)		3.5 - 4	21 - 24	4 - 6	24 – 36
	6	MD5 (Gbps)		4	24	5	30
	5	AES/SHA-1,256,512 (Gbps)		2 - 3	10 - 15	4 - 6	20 - 30
	6	TDES/SHA-1,256,512 (Gbps)		1.5 - 2	9 - 12	2	12
	3	RSA/ECC exp. w/ 1024/2048 bits / key (ops / second)		15000 / 2000	45000 / 6000	1666 / 2420	50000 / 7260
RegX	8	(Gbps input)		2.5 - 5	20-40	9	72
XML	4	(Gbps input)		1 - 3	4 - 12	5	20

*Multiple algorithms can be processed in parallel but aggregate bandwidth may be effected

PowerEN:
Projected power breakdown by function @ 2.0GHz

- Cores/Caches: 40%
- Accelerators: 11%
- I/O & Bus Logic: 19%
- PHY: 27%
- Clock Grid: 3%



Evolution of a Wire-Speed Processor

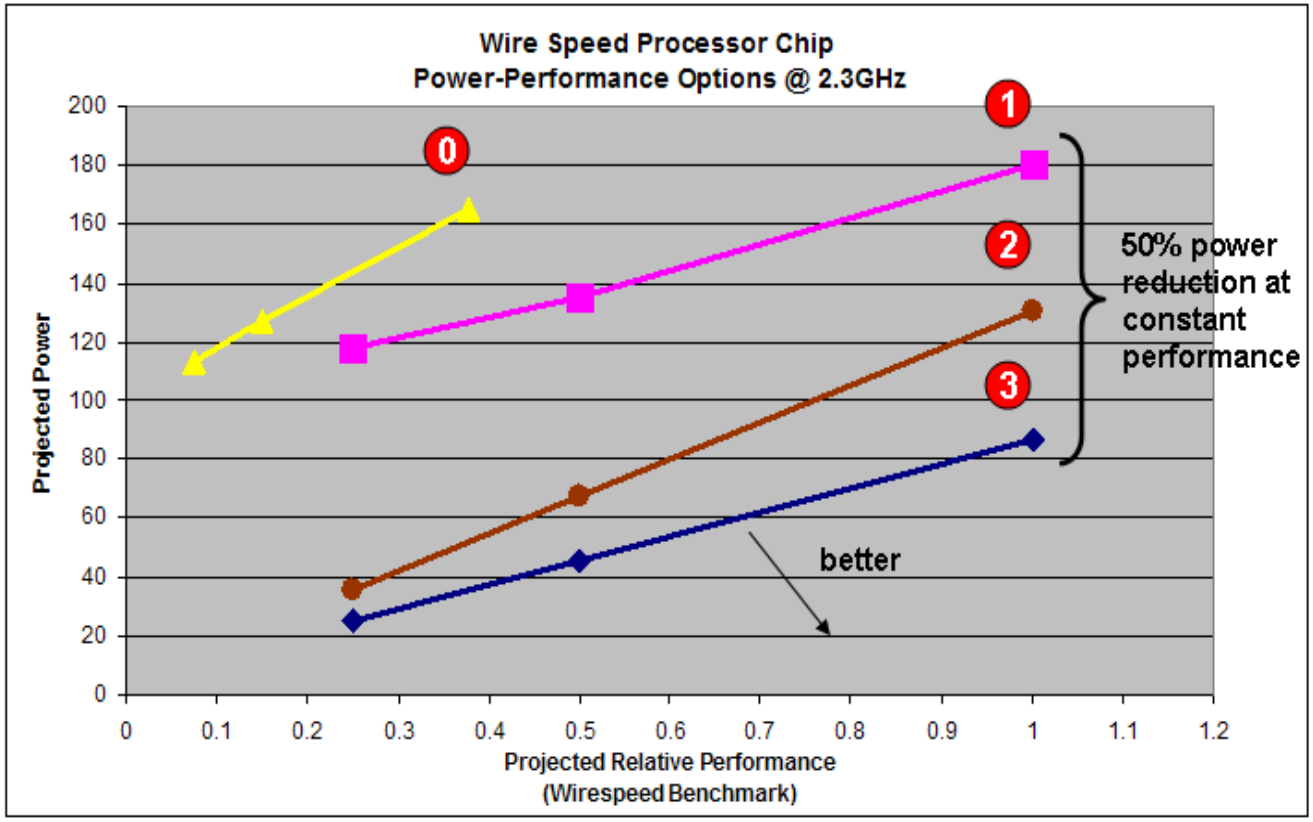
Power Savings from Architecture / Integration

- Low Power Integrated Design
  - -- Accelerators
  - -- I/O (Ethernet/HEA)
  - -- Cores/Threads
  - -- In 25-75 W Sockets

1 - Eliminates I/O Bottlenecks

2 - Improves Latency & State Replication

3 - Scalable Solution
  - -- 10 => 40 => 100 GbE



[General Purpose Processor chip]

⇩

[WireSpeed]

0 : 8- GP cores

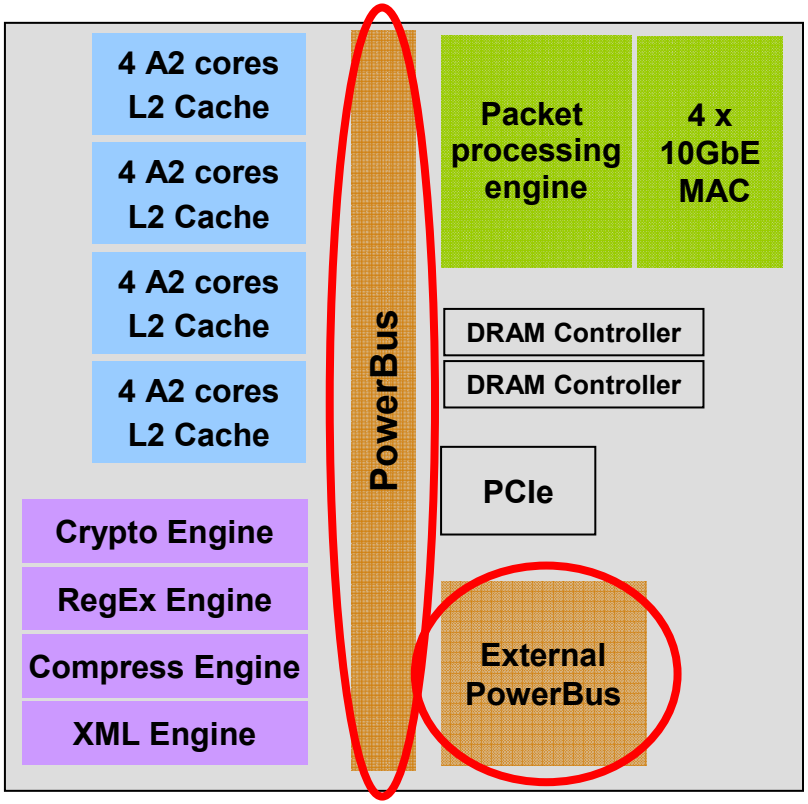
1 : 8- GP cores + 4 Accelerators

2 : 8- GP cores + 4 Accelerators + Integrated I/O (DDR3, 10Ge, HEA)

3 : 16- A2 cores + 4 Accelerators + Integrated I/O (DDR3, 10Ge, HEA)

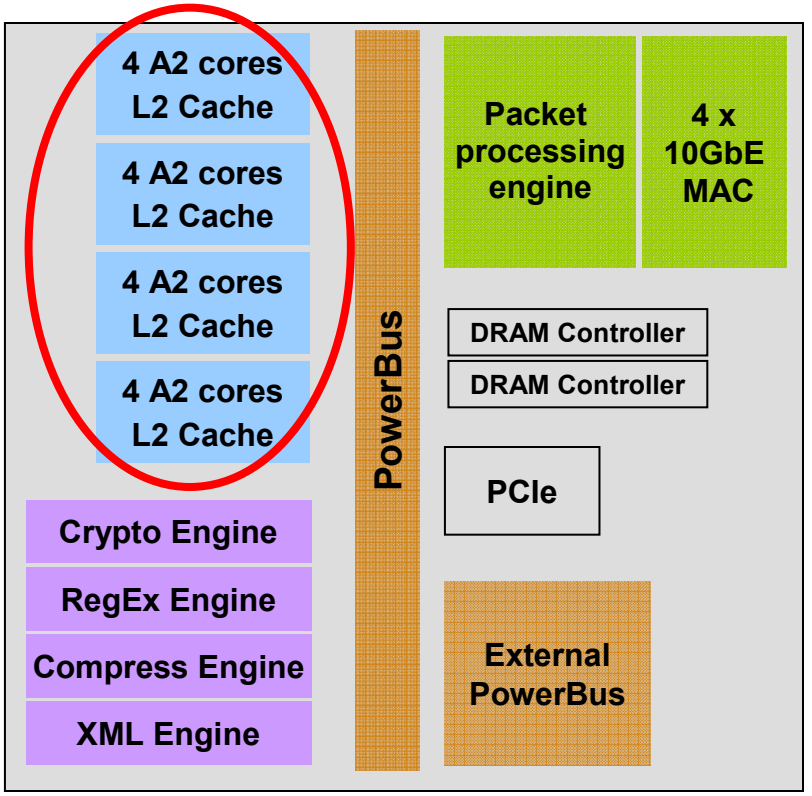
Interconnect Architecture

- *"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

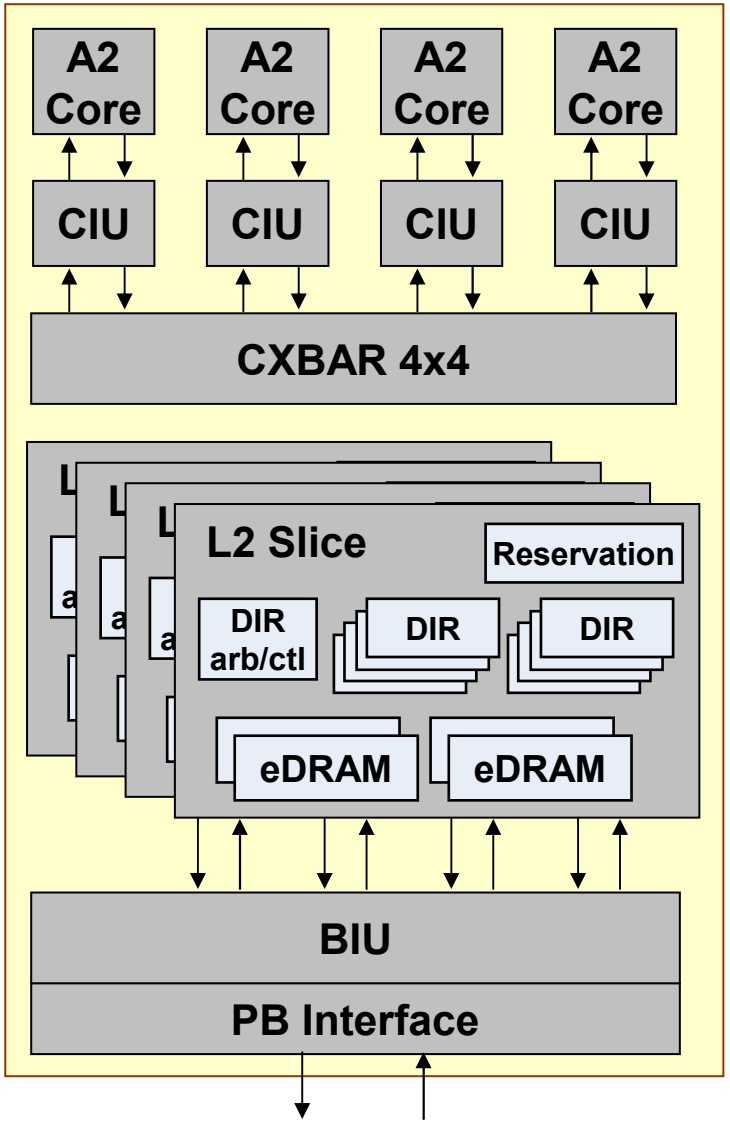
- PowerPC 64 architecture – Embedded
- Enhanced for Co-processor/I/O interface
- 4 way Fine Grained SMT
- In Order Dispatch and Execution
- 2 way concurrent issue.
  - 1 Integer + 1 FPU instruction per cycle
  - Different threads
- Unified fixed point, ld/st, and branch unit
- 16KB L1 Data Cache – 4 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

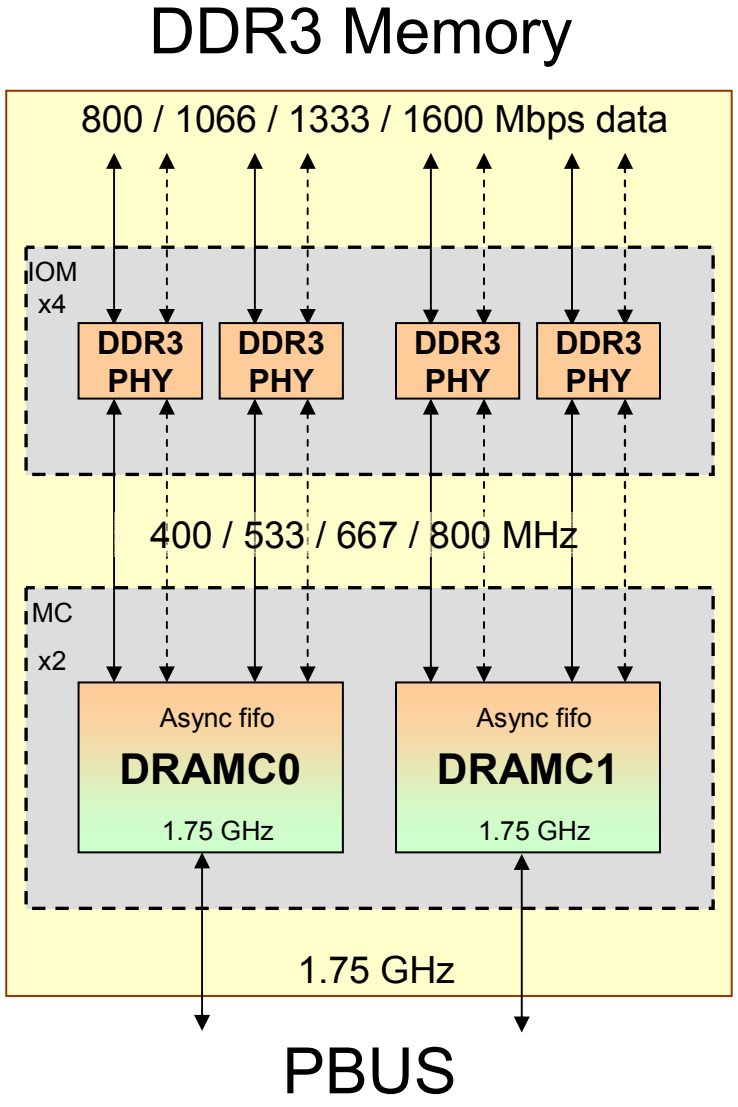
PowerPC™ Processing Element Architecture

- 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



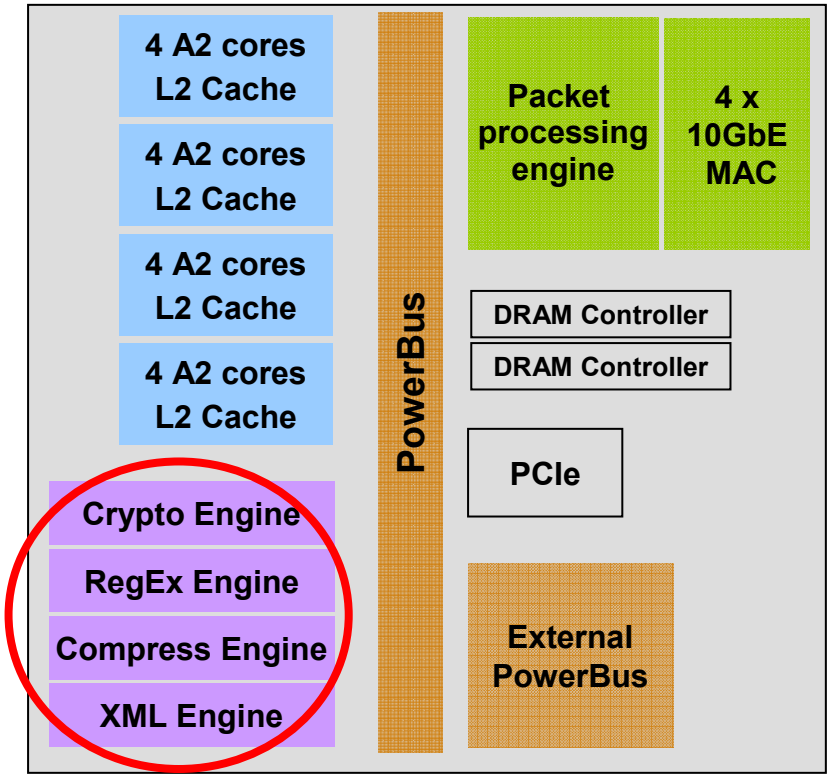
DDR3 DRAM Controller / PHY

- 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



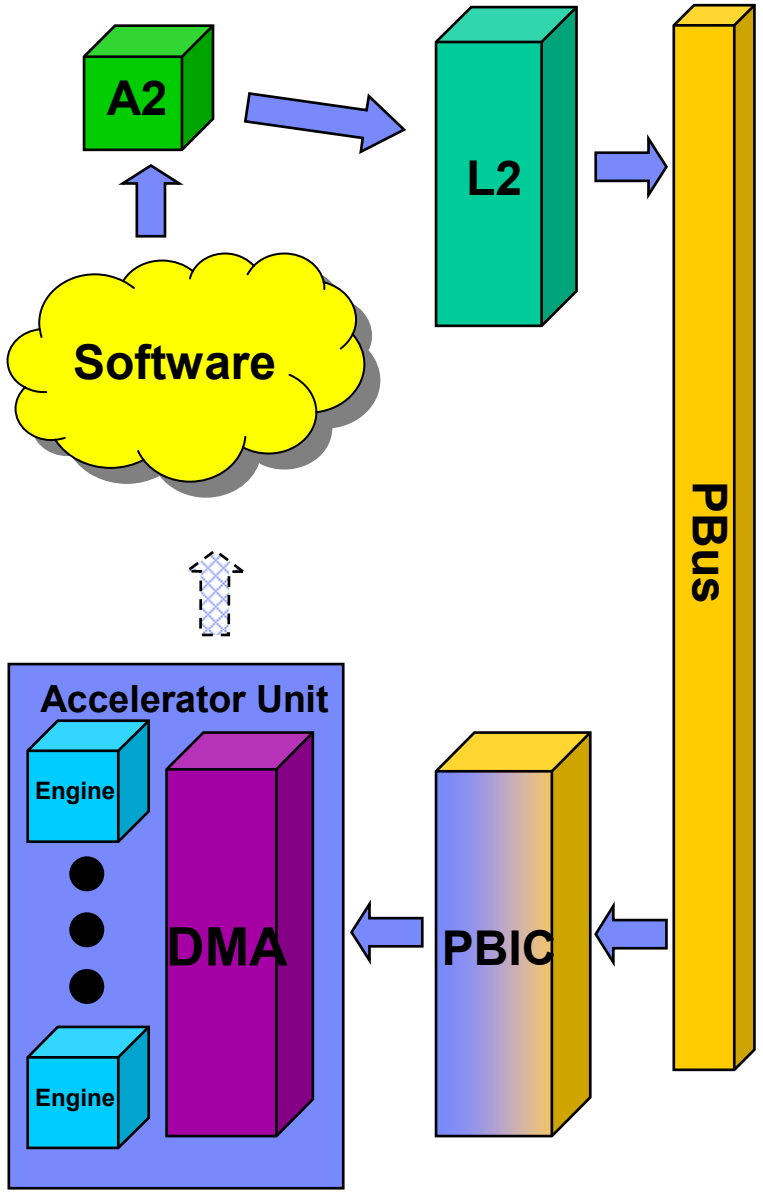
Accelerator (Co-Processor) Architecture

- 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



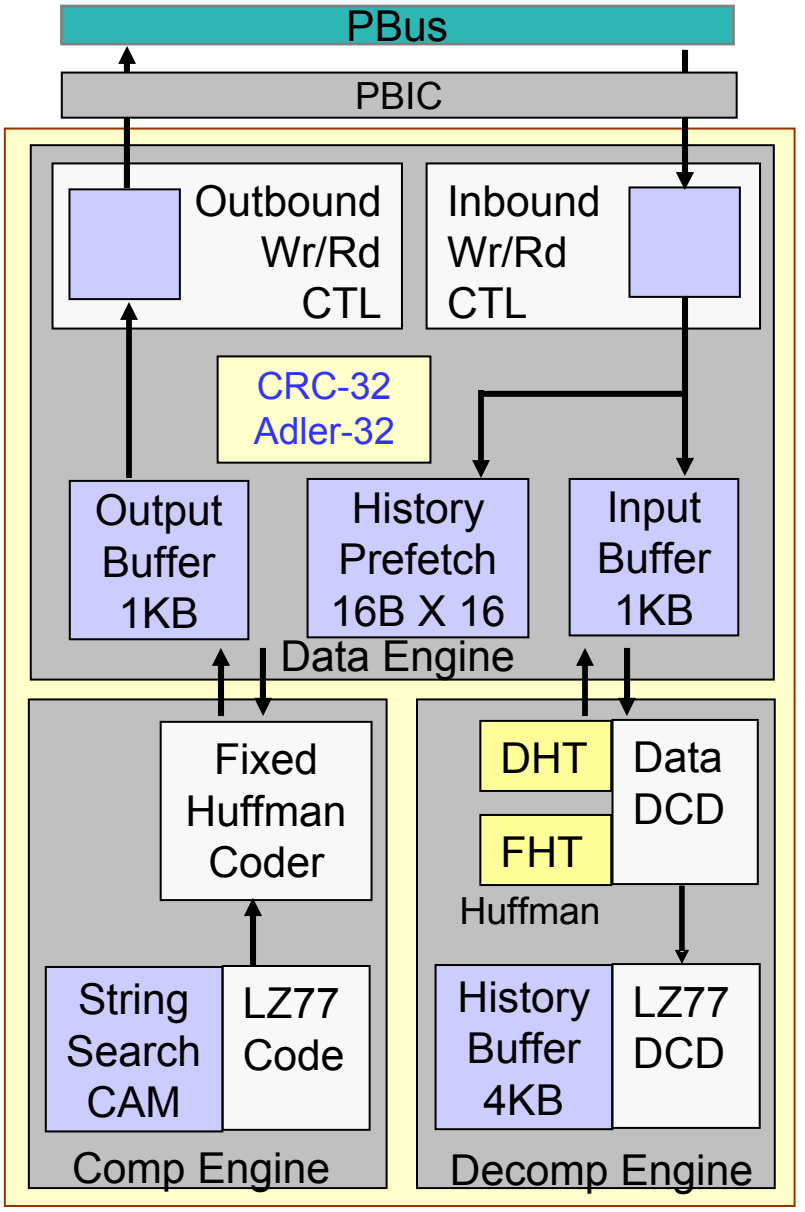
Accelerator Interface

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



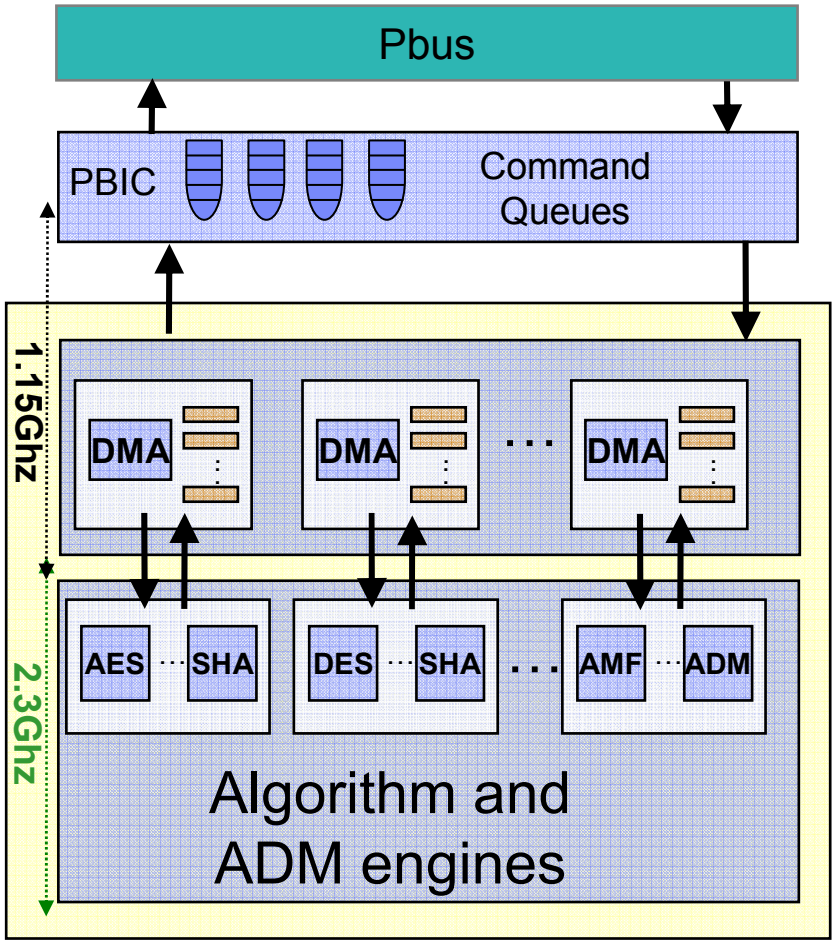
Compression/Decompression

- 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

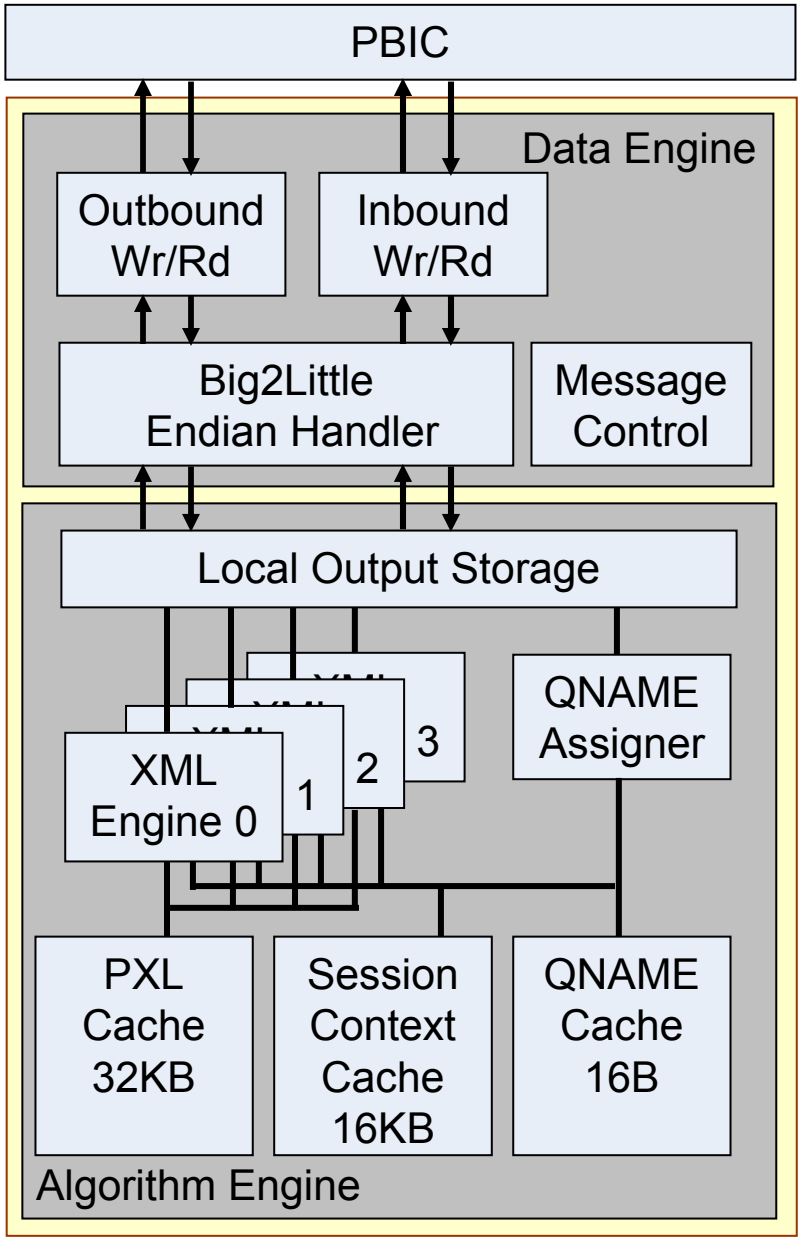
- Symmetric Algorithm Acceleration
  - AESModes:
    - Key Lengths: 128b,192b,256b
  - DES & 3DES Modes:
  - ARC4, Kasumi
  - HASH: SHA-1,SHA-256,SHA-512, MD5
    - HMAC supported for SHA
    - Combined 3DES/AES and SHA
- Asymmetric Algorithm Acceleration
  - Modular Math Functions for RSA/ECC
  - Point 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

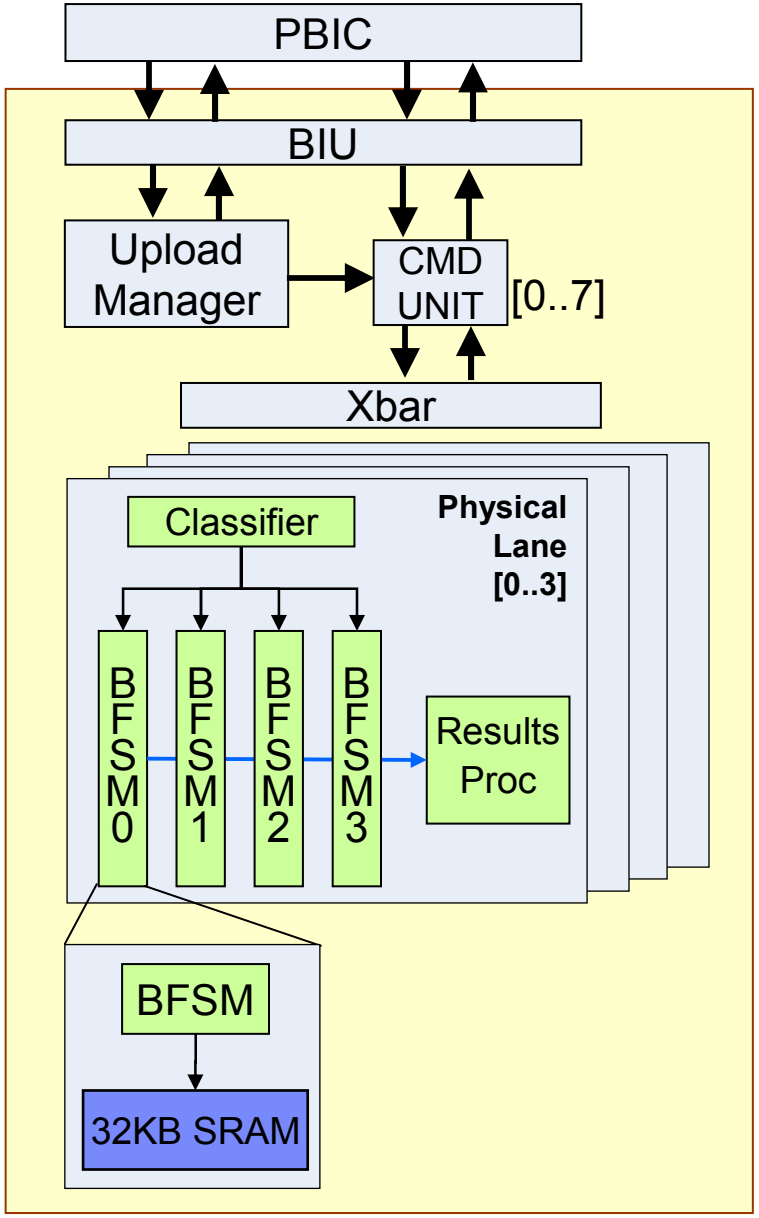
XML Unit / XML Engines

- Fully asynchronous offload operation
- Four Parsing Engines
  - Performs lexical analysis
  - Checks well-formness
  - 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

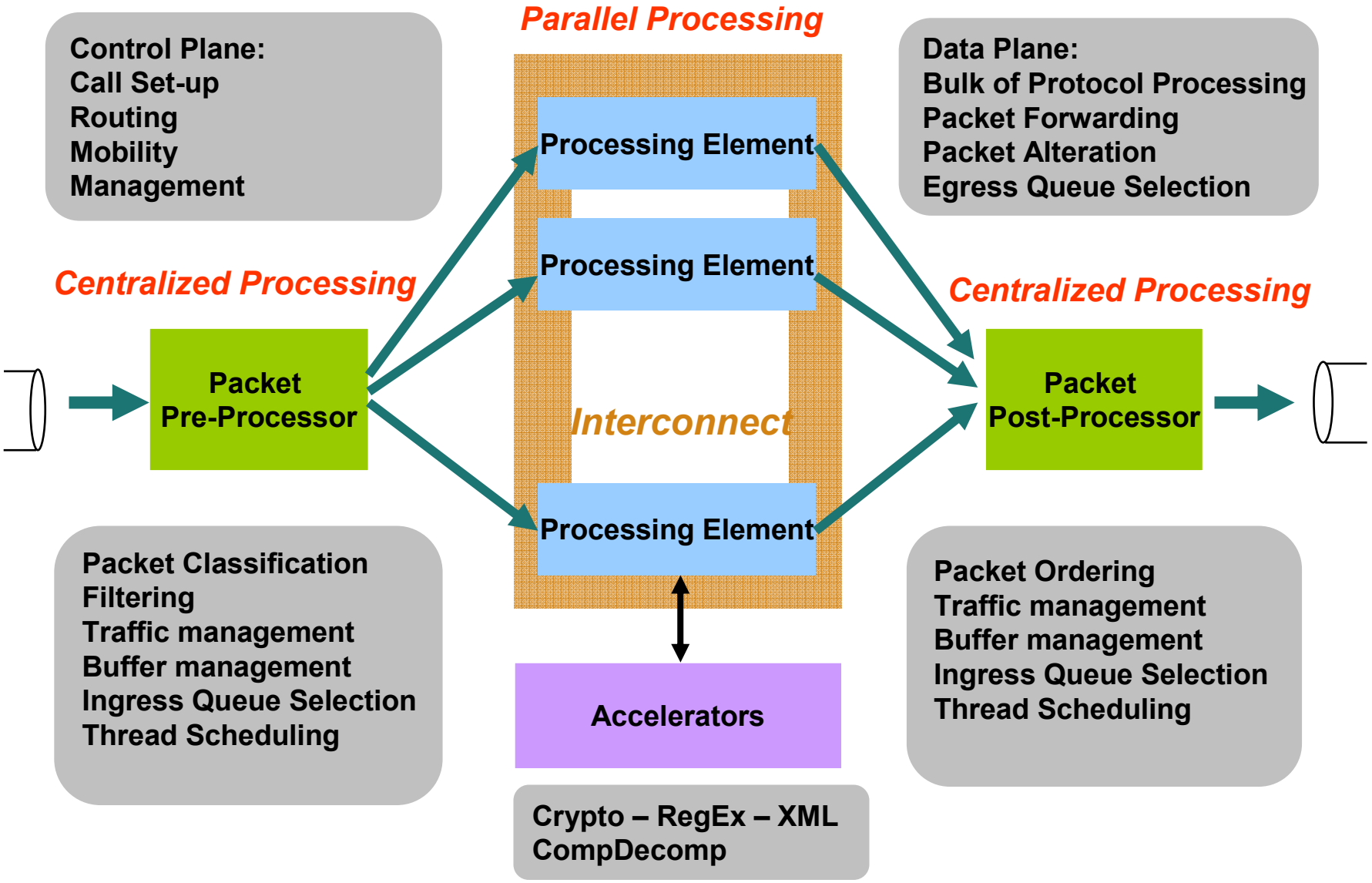


RegX (Pattern Matching Engine)

- 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, complier, patterns and workload

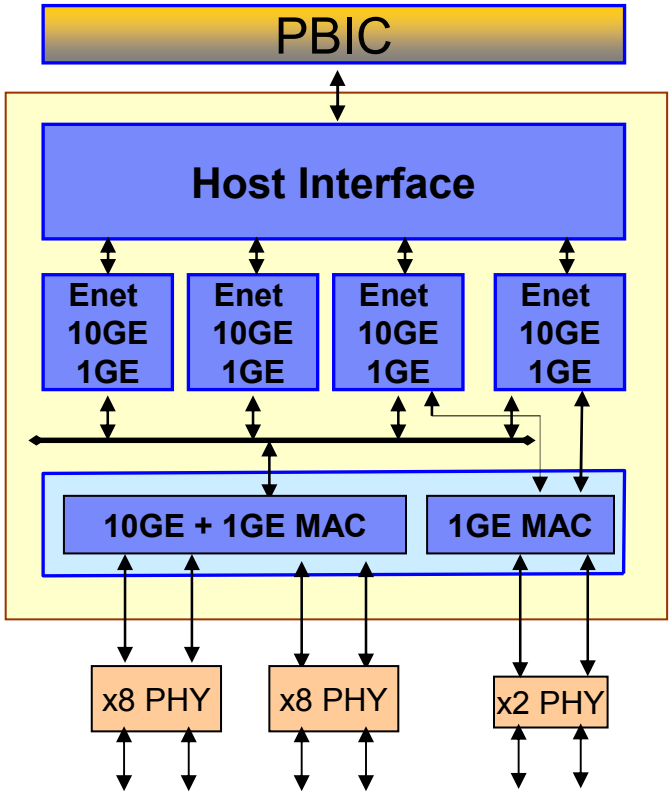


IBM PowerEN™ Packet Processing Framework



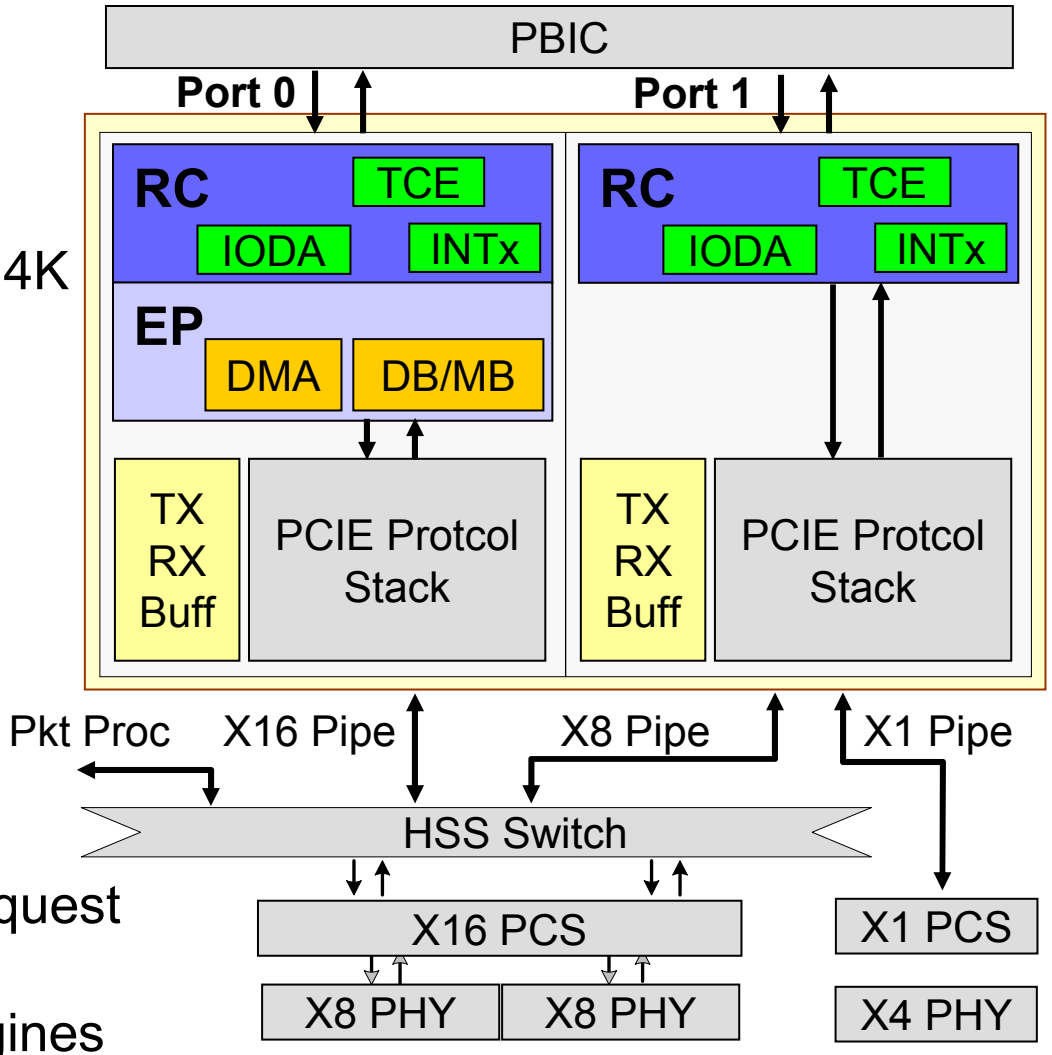
Packet Processor Architecture

- 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

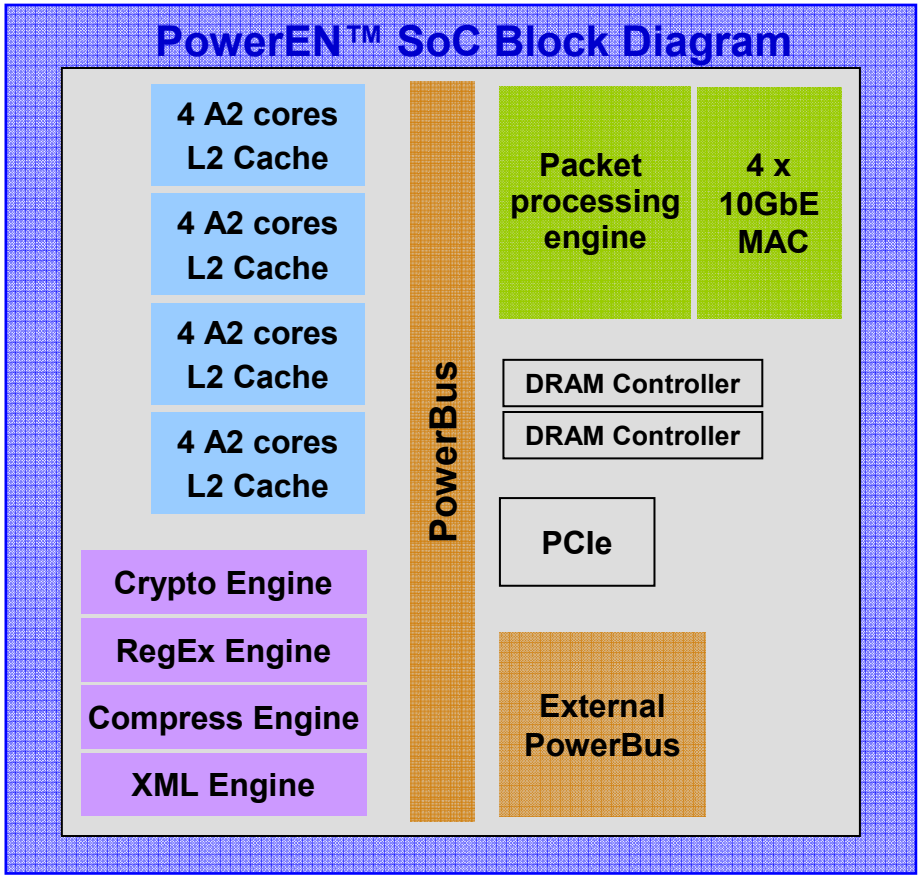


PCI-Express

- 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