Heterogeneous Multi-Core SOC (HMC-SOC) Architectures

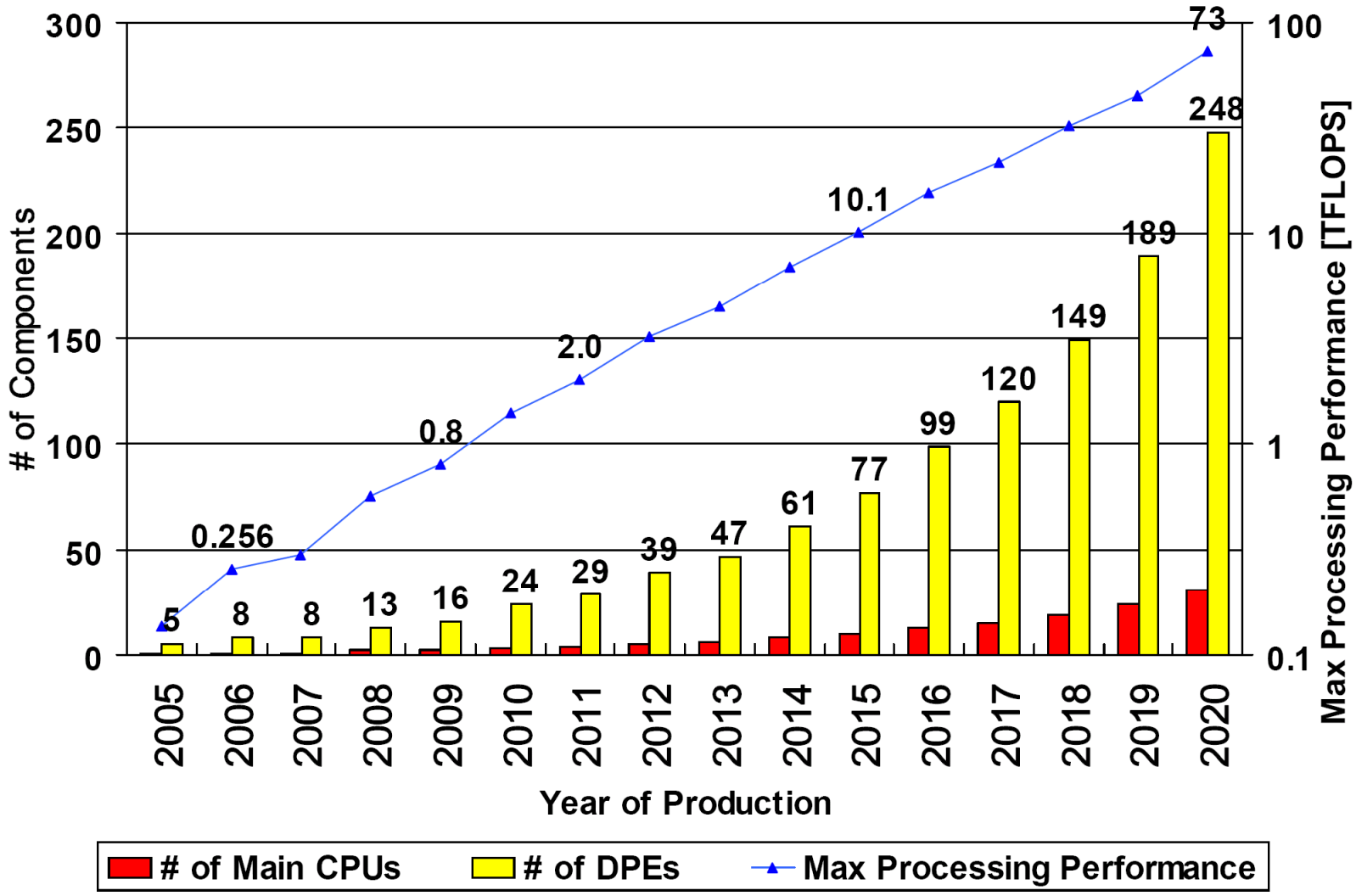
Mark McDermott

Fall 2009

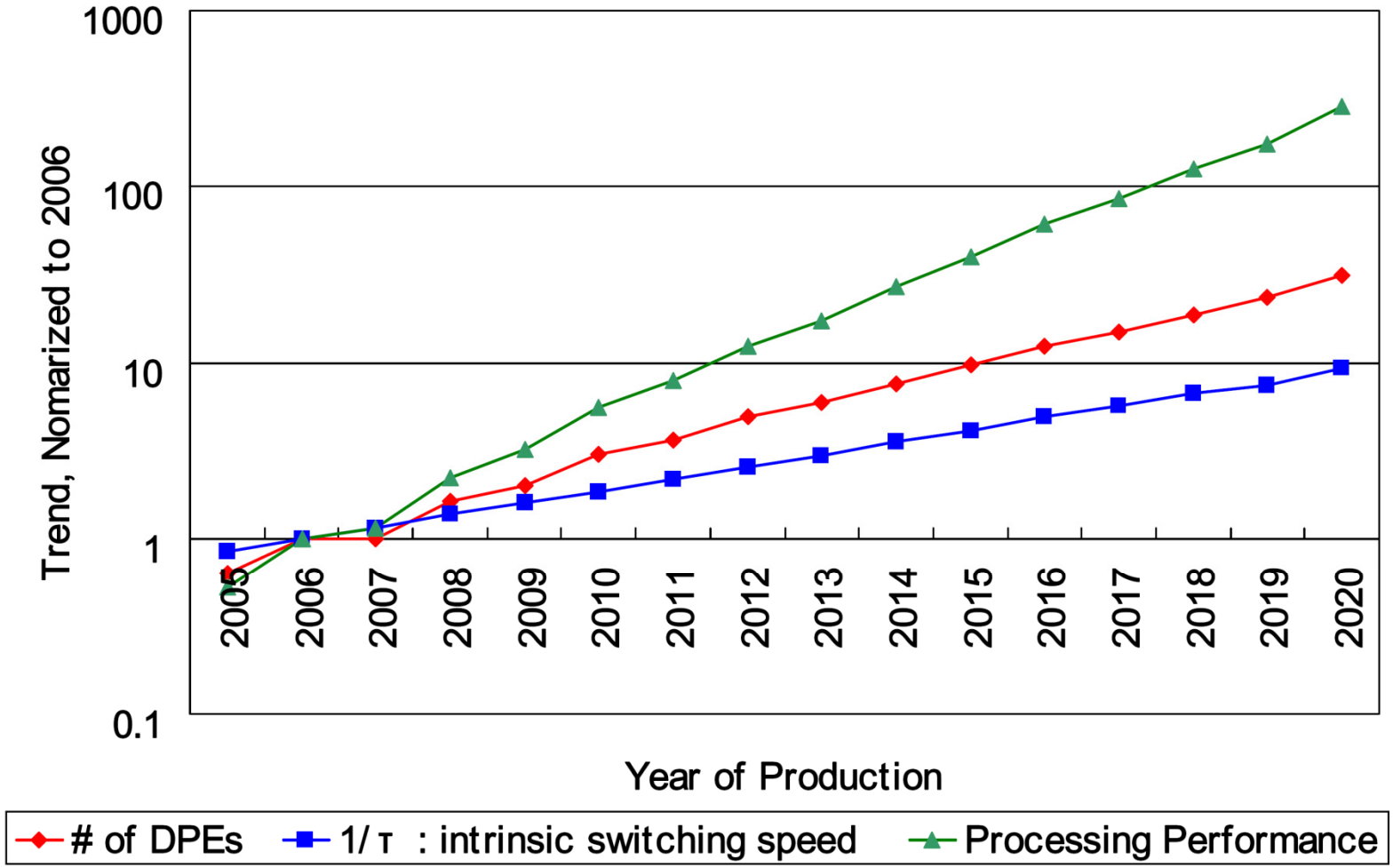
Agenda

- Market prediction
- Architectures of three HMC-SOC platforms:
  - Atmel DIOPSIS 940HF SOC
  - Texas Instruments OMAP
  - IBM Cell Broadband Processor
- HMC-SOC Bus Architectures
  - AMBA AXI
  - IBM Cell Element Interconnect Bus (EIB)
  - Network on Chip Architectures (NOC)

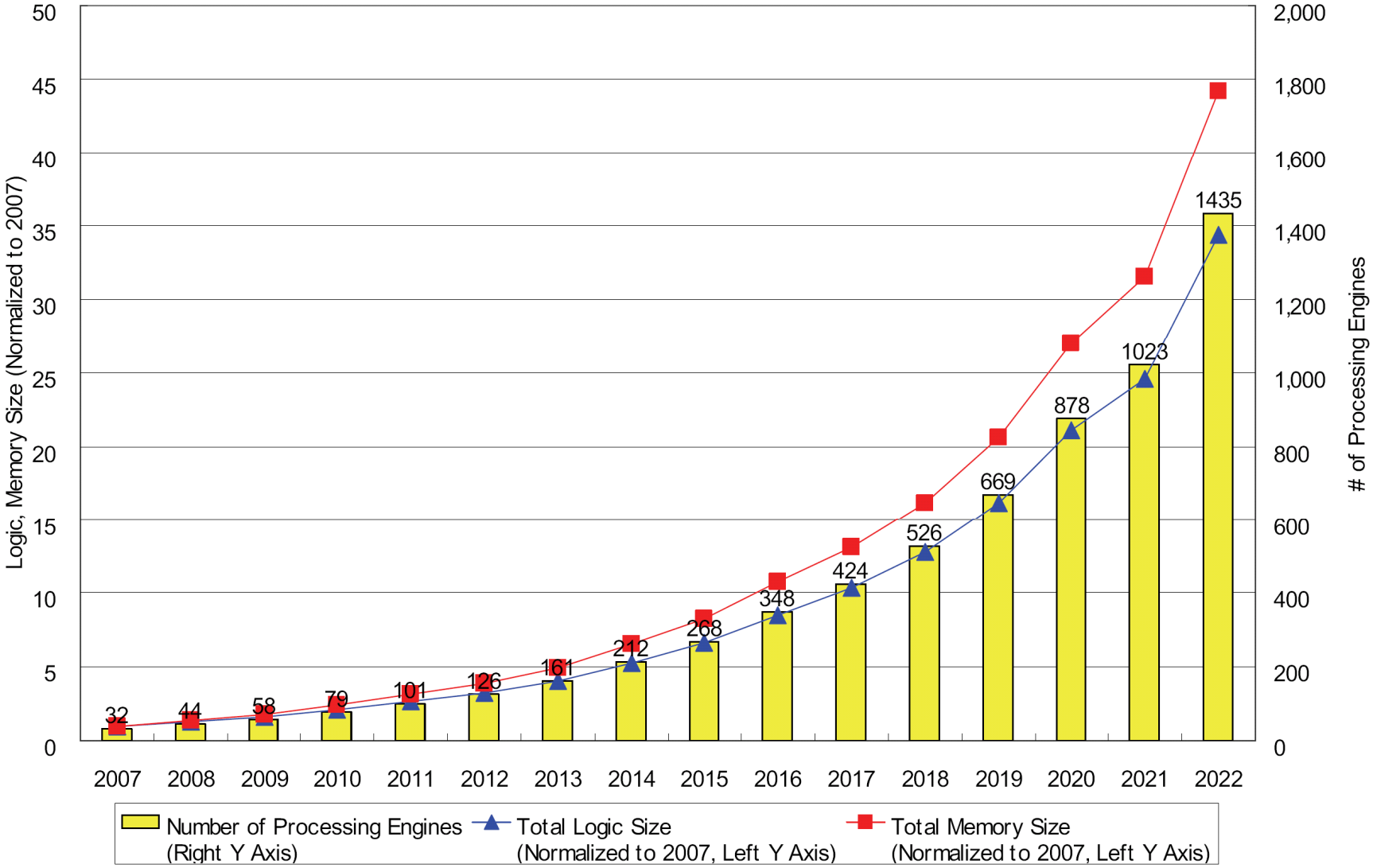
2006 ITRS Prediction: CPUs vs. Data Processing Engines



2006 ITRS Prediction: Processing Perf. vs. Transistor Perf.



2007 ITRS Prediction: Data Processing Engines

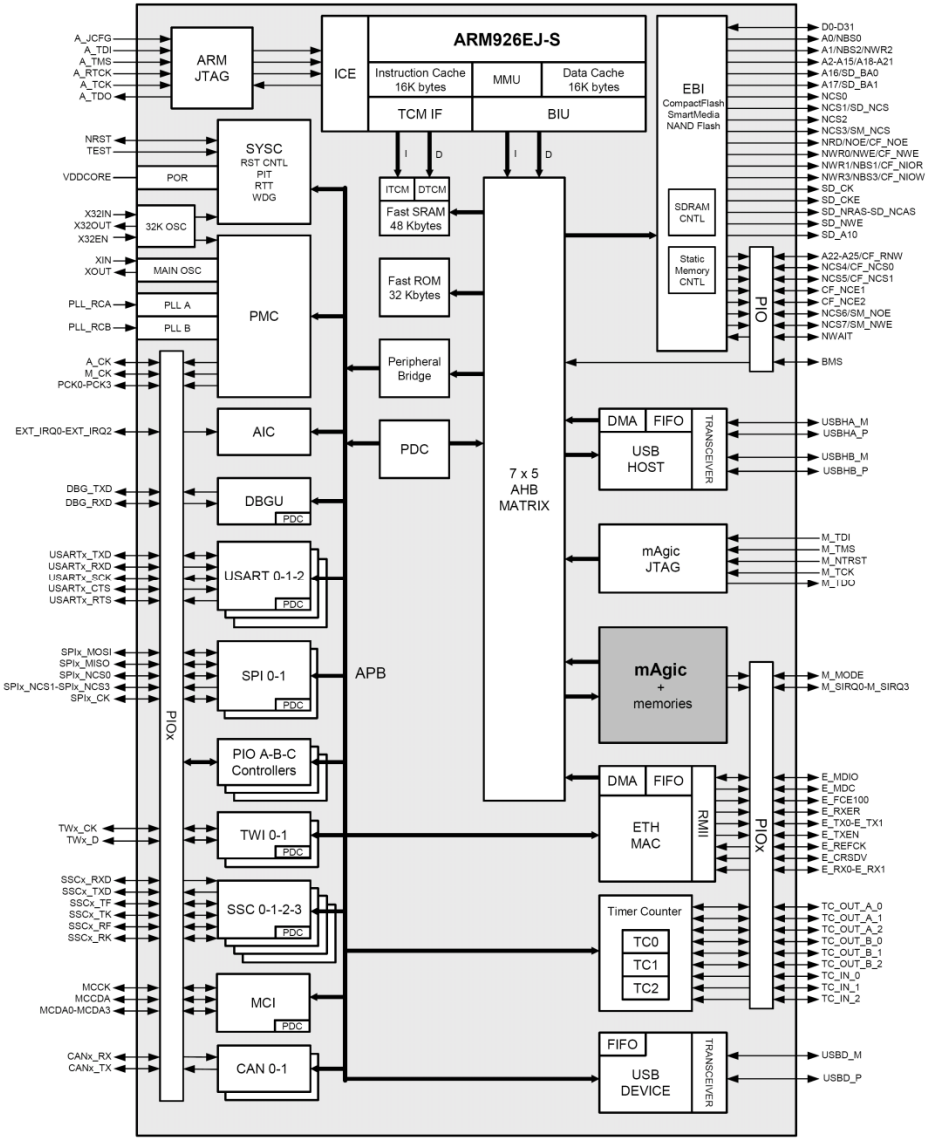


Agenda

- Market prediction
- **Architectures of three HMC-SOC platforms:**
  - **Atmel DIOPSIS 940HF SOC**
  - Texas Instruments OMAP
  - IBM Cell Broadband Processor
- HMC-SOC Bus Architectures
  - AMBA AXI
  - IBM Cell Element Interconnect Bus (EIB)
  - Network on Chip Architectures (NOC)

DIOPSIS 940 HF MPSOC

- Dual-core processing platform for audio, speech processing, automotive sound and robotics applications. The two cores:
  - ARM926EJ-S RISC microprocessor.
  - 40 bit Floating-point Magic DSP
    - Provides high dynamic range and maximum numerical precision.
- Synchronization between the two processors can be either based on interrupts, software polling or semaphores.



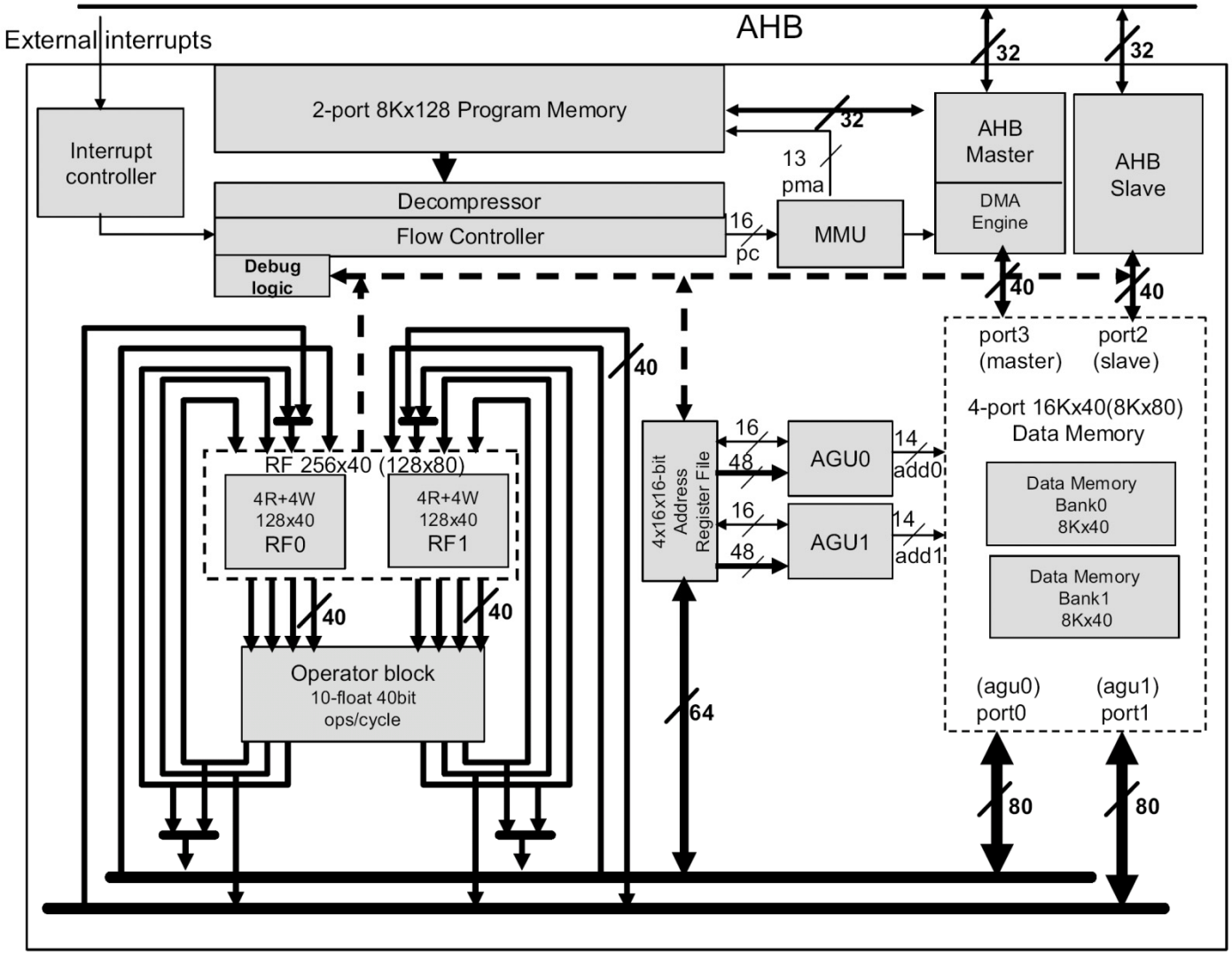
Courtesy Atmel, Inc.

VLIW DSP

- The MagicV VLIW DSP is the numeric processor of D940HF. It operates on IEEE 754 40-bit extended precision floating-point and 32-bit integer numeric format.
- Main components of the DSP subsystem are:
  - Core processor
  - On-chip memories
  - DMA engine and its AHB master and slave interfaces.
  - Operators block
  - Register file
  - Multiple address generation unit
  - Program decoding and sequencing unit
- ARM processor boots the DSP from its Flash memory.

Courtesy Atmel, Inc.

VLIW DSP Block Diagram



Courtesy Atmel, Inc.

VLIW DSP (cont)

- The DSP is a Very Long Instruction Word engine, that works like a RISC machine by implementing triadic computing operations on data coming from the register file and data move operations between the local memories and the register file
  - Memory System contains 16K*40-bit on-chip memory locations supporting up to 6 accesses/cycle. 4-accesses/cycle are reserved for the activities driven by Multiple Address Generation unit
- Operators are pipelined for maximum performance. The pipeline depth depends on the operator used
  - Works on 32-bit signed integers and IEEE 754 extended precision 40-bit floating-point data
- Scheduling and parallelism operations are automatically defined and managed at compile time by the assembler-optimizer, allowing efficient code execution
- Architecture is designed for efficient C-language support

Courtesy Atmel, Inc.

ARM – DSP Inter-processor Communications

- DSP is connected to ARM processor through a master AHB IF and a slave AHB IF
- ARM Processor and DSP also exchange a set of discrete lines for the cores interconnection:
  - Three external interrupt input lines that go from the external pin (through PIO) to the AIC also go to the IRQ lines in the DSP
  - Four internal interrupt lines that go from the SSC to the AIC also go to IRQ lines in the DSP
  - NINT line goes from the AIC to PMC and is the AND of the fast interrupt NFIQ and NIRQ in the ARM

Courtesy Atmel, Inc.

Programming Environment for DIOPSIS 940 HF

- Uses the standard ARM programming environment for the 926EJS
- Very little commercial support for programming the DSP exists.
  - Typical of a good number of these kinds of platforms
  - Tools were generated for the SHAPES* effort by Target Compiler Technologies.
- DON'T FALL in LOVE with a HW ARCHITECTURE

* Software Hardware Architecture Platform for Embedded Systems

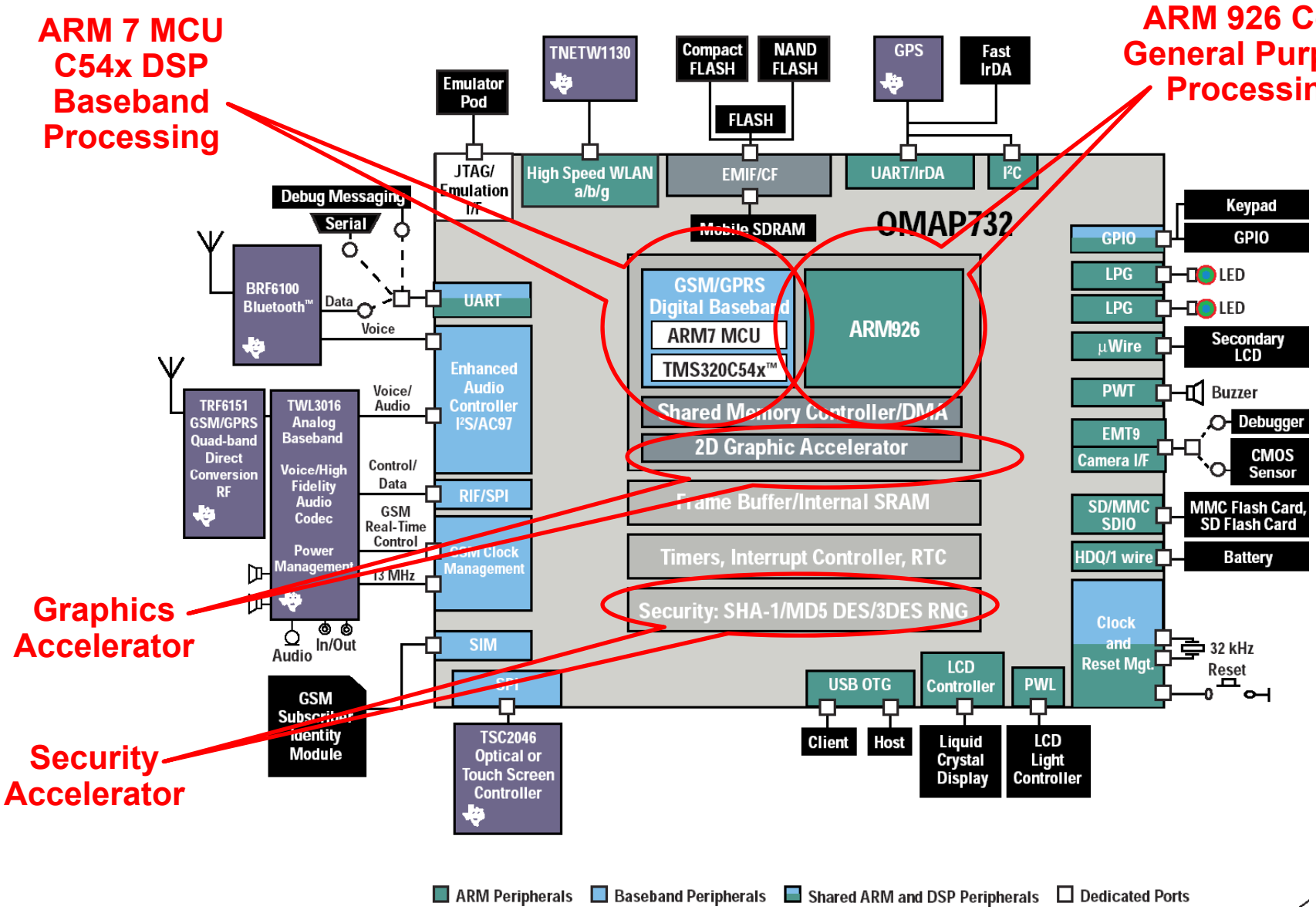
Agenda

- Market Prediction
- Architectures of three HMC-SOC platforms:
  - Atmel DIOPSIS 940HF SOC
  - **Texas Instruments OMAP**
  - IBM Cell Broadband Processor
- HMC-SOC Bus Architectures
  - AMBA AXI
  - IBM Cell Element Interconnect Bus (EIB)
  - Network on Chip Architectures (NOC)

TI OMAP (Open Multimedia Applications Platform)

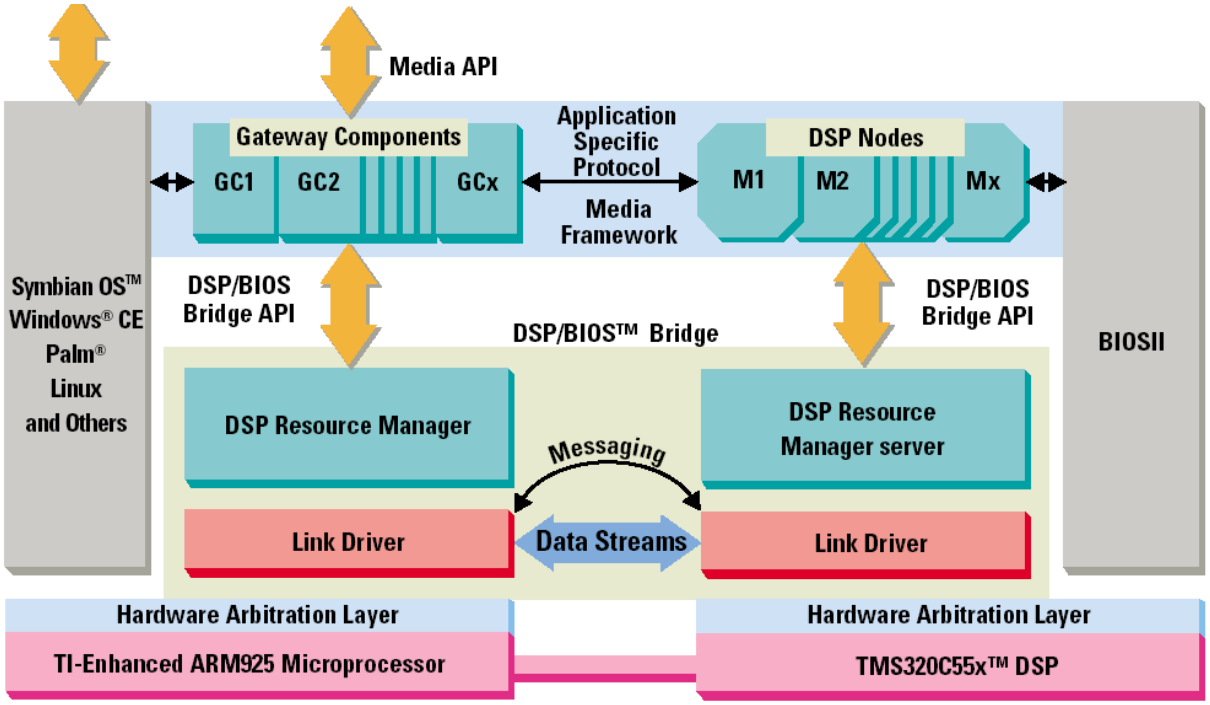
- 'Platform' = Processor + Software + Support
- Uses commercially available GPP and DSP components which are scalable to future generations
- Software from application to system software
  - DSP libraries; J2ME, Linux, MS WinCE, Palm, Symbian

TI OMAP732 Hardware Platform

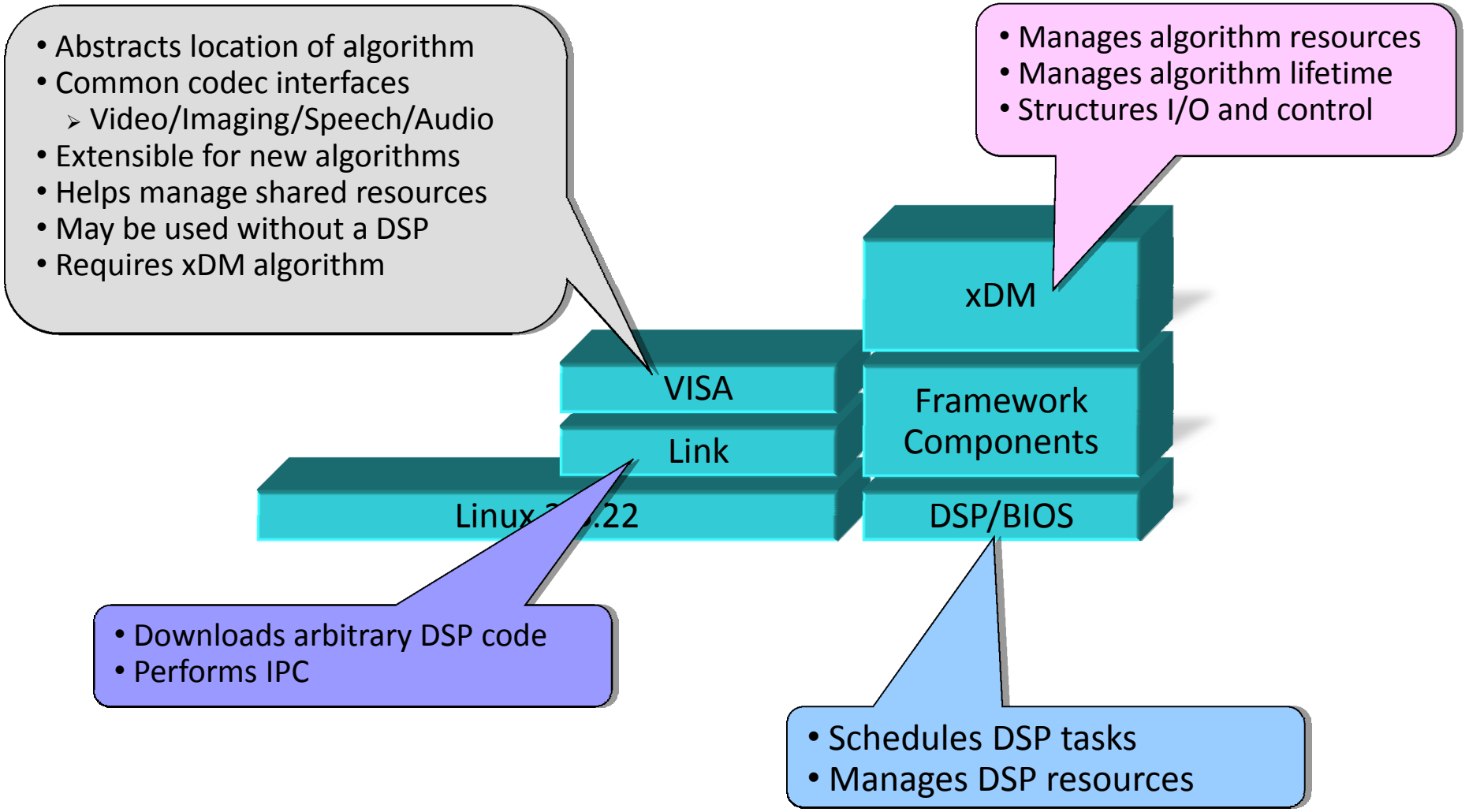


OMAP Programming Environment

- DSP/BIOS bridge to divide tasks between CPU, DSP: 'reroute' some tasks to DSP and run 'asynchronously'
- Allow CPU programmer to access and control DSP runtime environment (through API)
- Code developer sees as if only a single RISC processor is doing all the functions
  - Not two different programming environments.



Overview of SW stacks



OMAP Software

Drivers/OS/Apps

DSP Codecs

- MPEG4 SP Encode/Decode(D1)[3]
- MPEG2 MP Decode(D1/)[3]
- H.264 MP decode / BP encode (D1) [3]
- WMV/VC1 Decode (D1)[3]
- JPEG Encode/Decode
- AAC LC Decode
- WMA9 Decode
- MP3 Decode
- AAC HE Decode

Development Tools

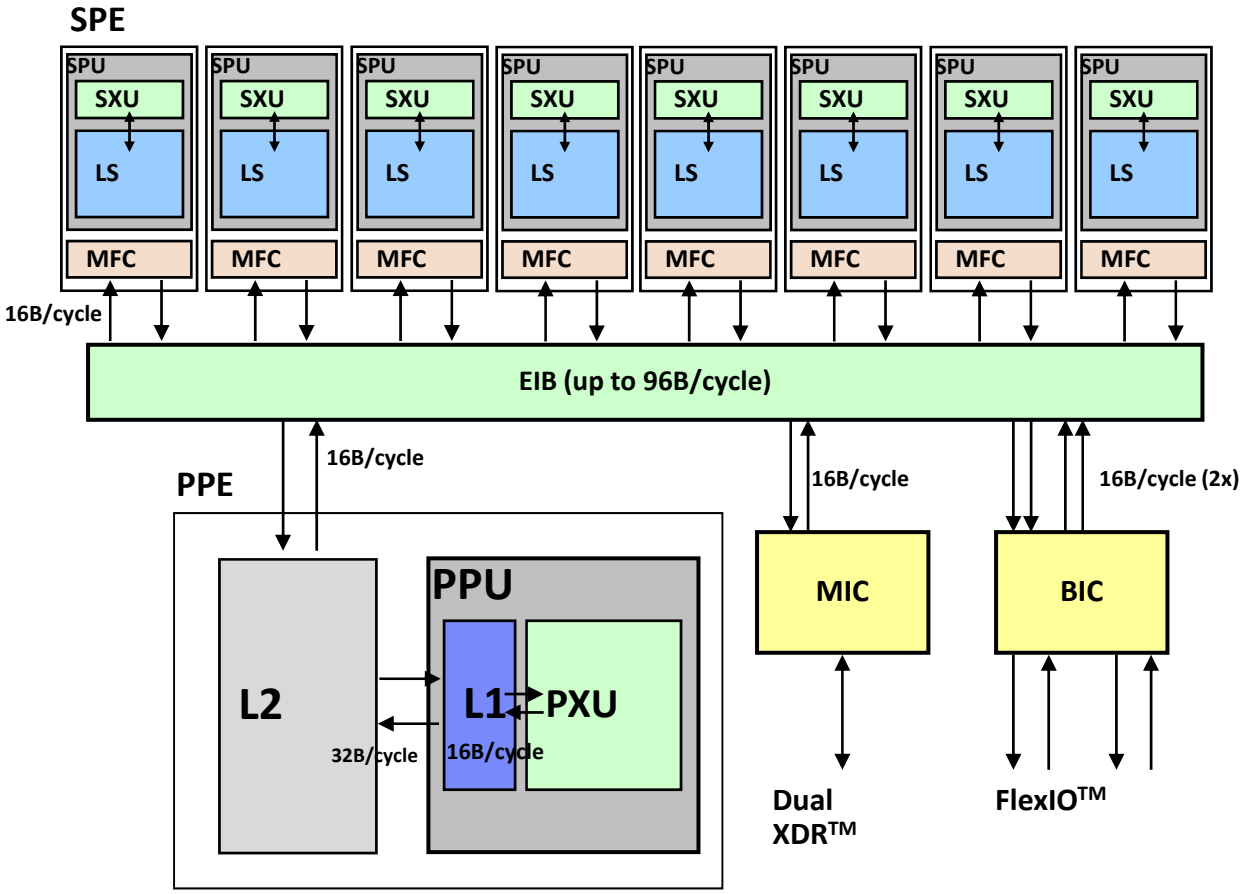
- Code Sourcery GNU gcc 4.2.1
- glibc
- Build-root “busybox” filesystem
- U-boot 1.1.4
- Platform Builder (WinCE only)

Agenda

- Market Prediction
- Architectures of three HMC-SOC platforms:
  - Atmel DIOPSIS 940HF SOC
  - Texas Instruments OMAP
  - **IBM Cell Broadband Processor**
- HMC-SOC Bus Architectures
  - AMBA AXI
  - IBM Cell Element Interconnect Bus (EIB)
  - Network on Chip Architectures (NOC)

IBM Cell Features

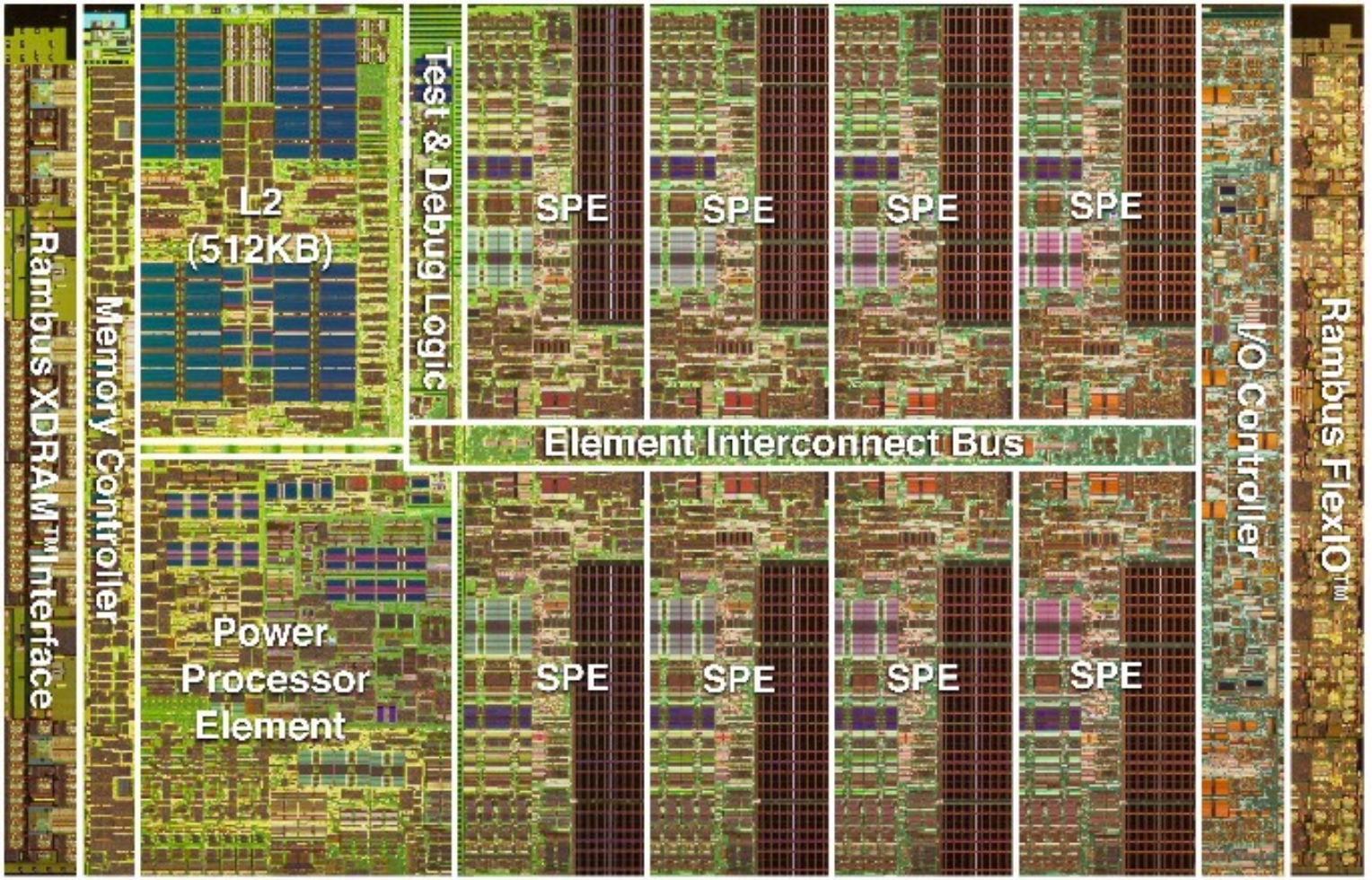
- Heterogeneous multi-core system architecture
  - Power Processor Element for control tasks
  - Synergistic Processor Elements for data-intensive processing
- Synergistic Processor Element (SPE) consists of
  - Synergistic Processor Unit (SPU)
  - Synergistic Memory Flow Control (MFC)
    - Data movement and synchronization
    - Interface to high-performance Element Interconnect Bus



64-bit Power Architecture with VMX

Courtesy IBM, Inc.

Cell Broadband Engine – 235mm^2

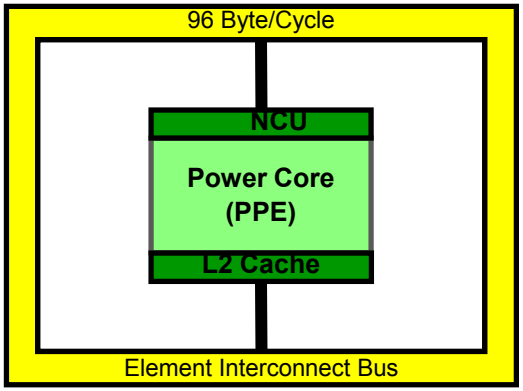


Courtesy IBM, Inc.

Cell Broadband Engine Components

Power Processor Element (PPE):

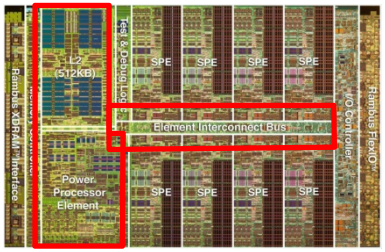
- General purpose, 64-bit RISC processor (PowerPC AS 2.0.2)
- 2-Way hardware multithreaded
- L1 : 32KB I ; 32KB D
- L2 : 512KB
- Coherent load / store
- VMX-32
- Realtime Controls
  - Locking L2 Cache & TLB
  - Software / hardware managed TLB
  - Bandwidth / Resource Reservation
  - Mediated Interrupts



Custom Designed
– for high frequency, space, and power efficiency

Element Interconnect Bus (EIB):

- Four 16 byte data rings supporting multiple simultaneous transfers per ring
- 96Bytes/cycle peak bandwidth
- Over 100 outstanding requests

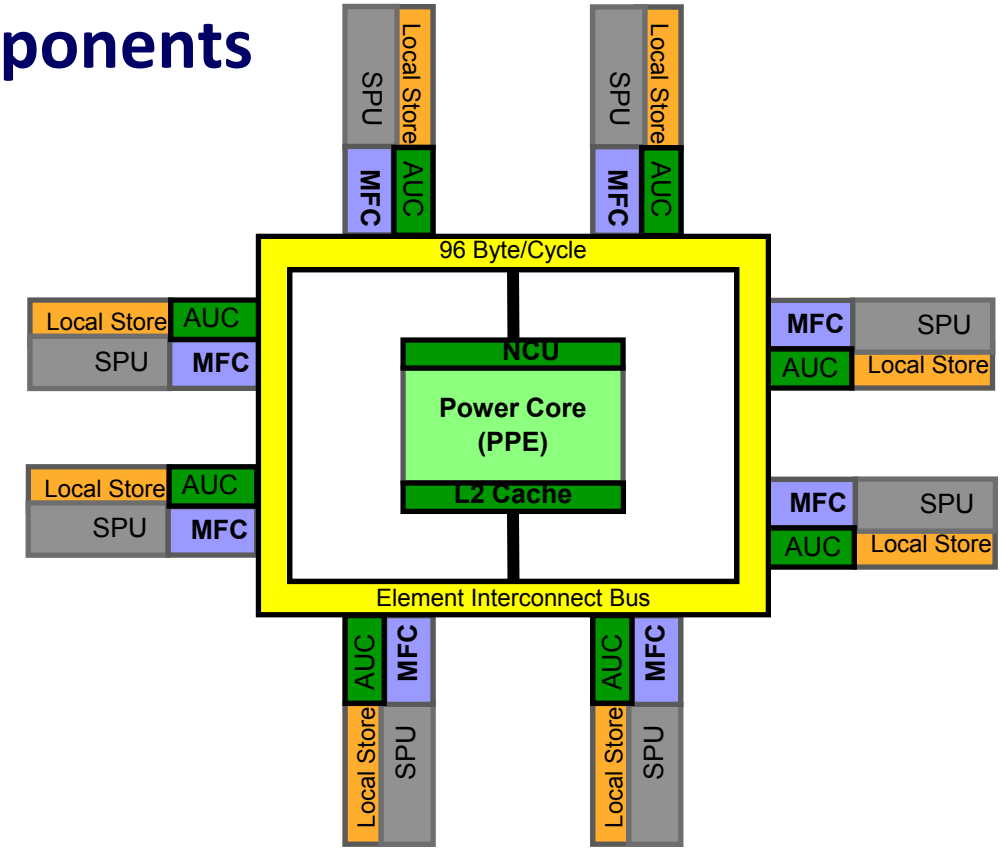


Courtesy IBM, Inc.

Cell Broadband Engine Components

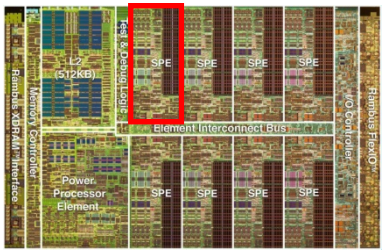
Synergistic Processor Element (SPE):

- Provides the computational performance
- Simple RISC User Mode Architecture
  - Dual issue VMX-like
  - Graphics SP-Float
  - IEEE DP-Float
- Dedicated resources: unified 128x128-bit RF, 256KB Local Store
- Dedicated DMA engine: Up to 16 outstanding requests



Memory Management & Mapping

- SPE Local Store aliased into PPE system memory
- MFC/MMU controls / protects SPE DMA accesses
  - Compatible with PowerPC Virtual Memory Architecture
  - SW controllable using PPE MMIO
- DMA 1,2,4,8,16,128 -> 16Kbyte transfers for I/O access
- Two queues for



Courtesy IBM, Inc.

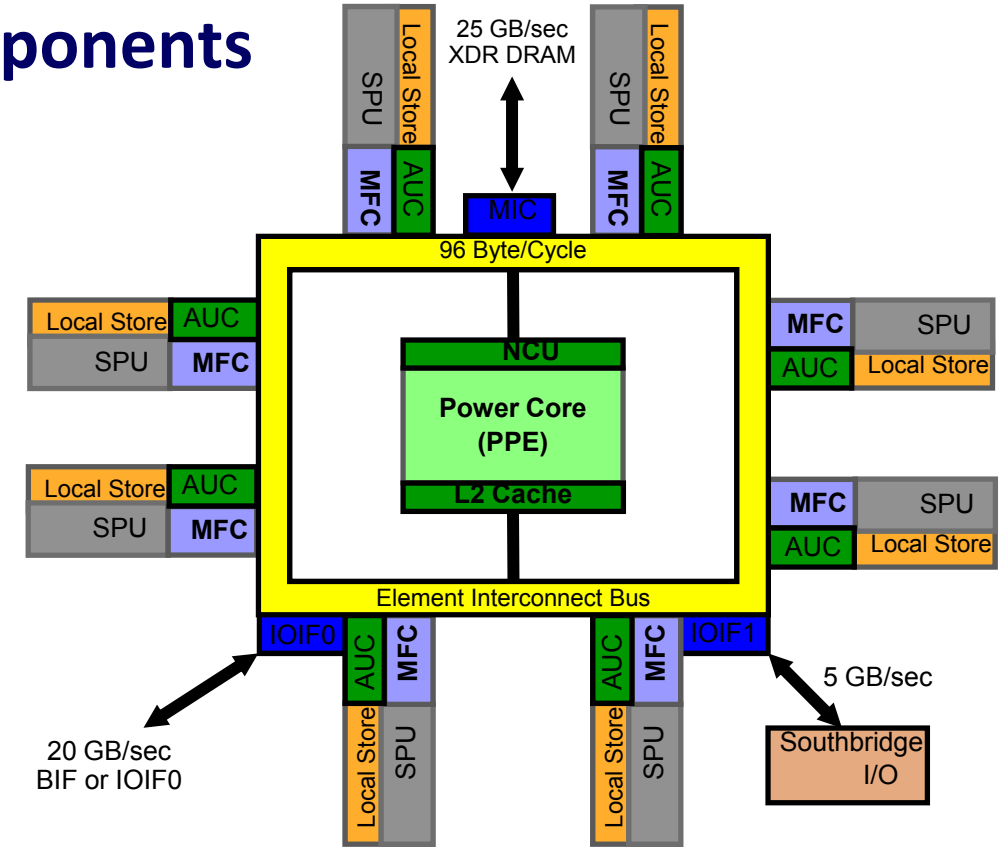
Cell Broadband Engine Components

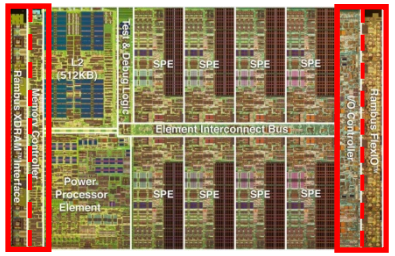
Broadband Interface Controller (BIC):

- Provides a wide connection to external devices
- Two configurable interfaces (60GB/s @ 5Gbps)
  - Configurable number of bytes
  - Coherent (BIF) and / or I/O (IOIFx) protocols
- Supports two virtual channels per interface
- Supports multiple

Memory Interface Controller (MIC):

- Dual XDR™ controller (25.6GB/s @ 3.2Gbps)
- ECC support
- Suspend to DRAM support





Courtesy IBM, Inc.

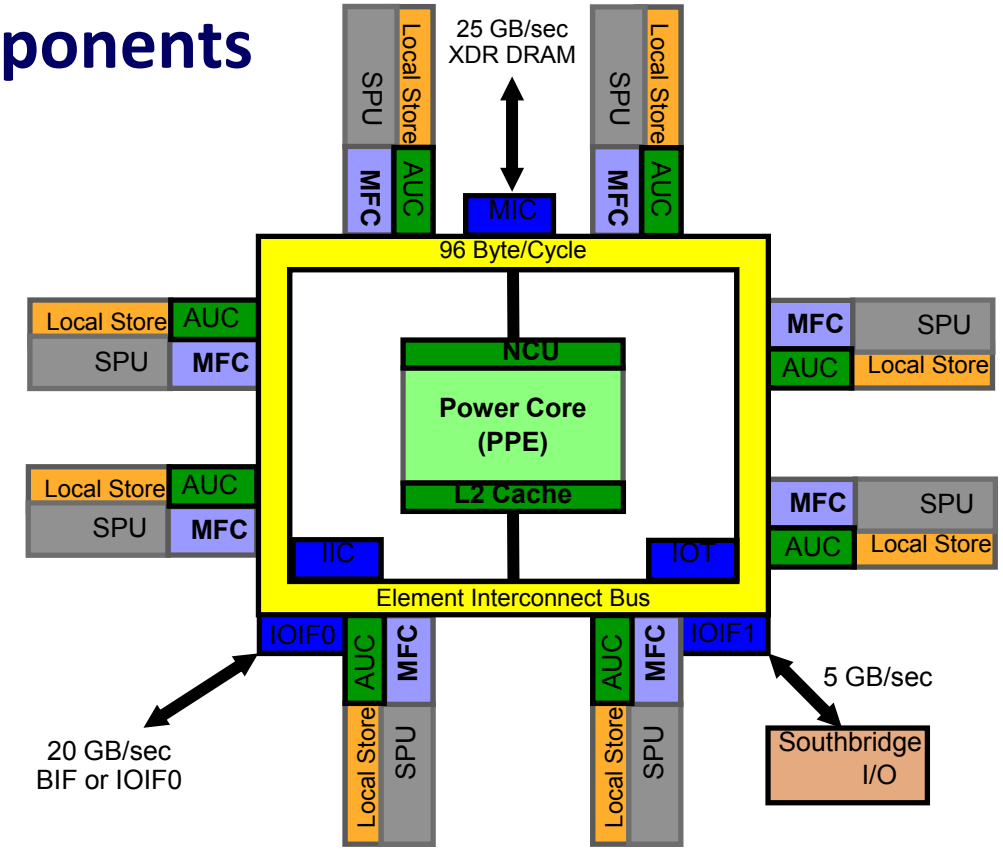
Cell Broadband Engine Components

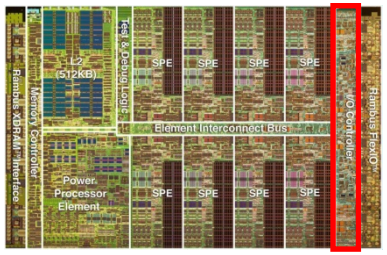
Internal Interrupt Controller (IIC)

- Handles SPE Interrupts
- Handles External Interrupts
  - From Coherent Interconnect
  - From IOIF0 or IOIF1
- Interrupt Priority Level Control
- Interrupt Generation Ports for IPI
- Duplicated for each PPE hardware thread

I/O Bus Master Translation (IOT)

- Translates Bus Addresses to System Real Addresses
- Two Level Translation
  - I/O Segments (256 MB)
  - I/O Pages (4K, 64K, 1M, 16M byte)
- I/O Device Identifier per page for LPAR
- IOST and IOPT Cache – hardware / software managed



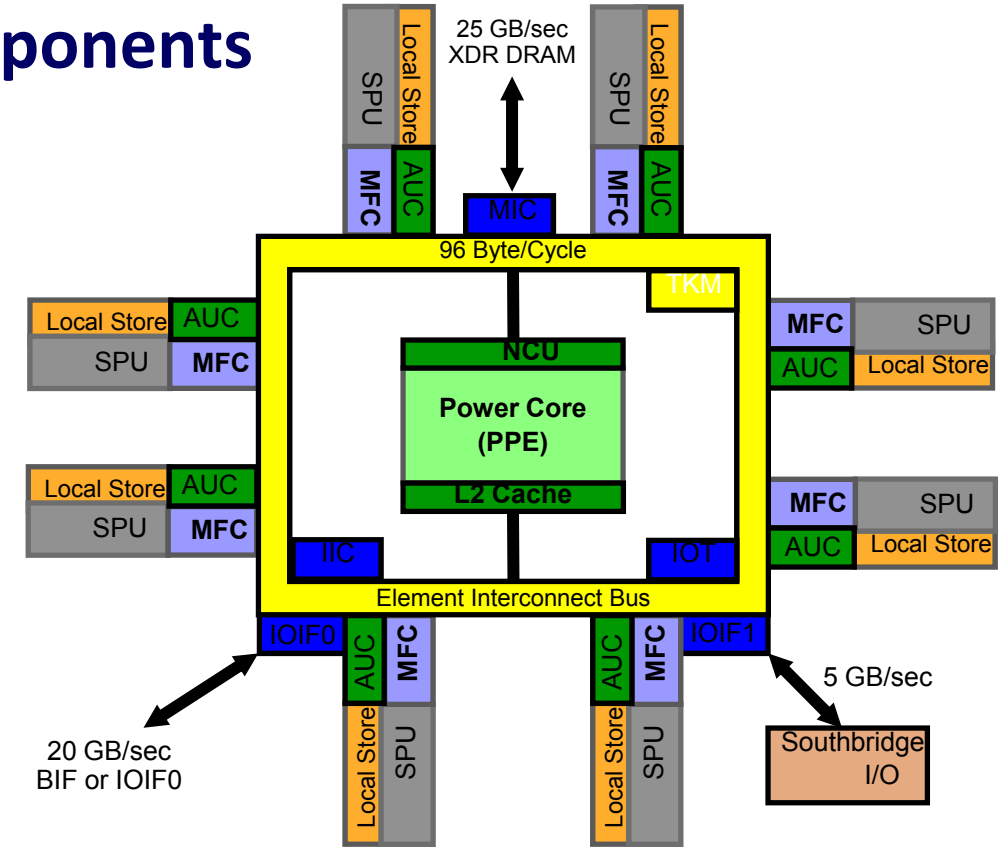


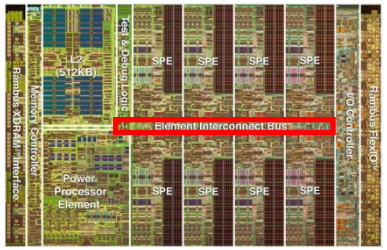
Courtesy IBM, Inc.

Cell Broadband Engine Components

Token Manager (TKM):

- Bandwidth / Resource Reservation for shared resources
- Optionally enabled for RT tasks or LPAR
- Multiple Resource Allocation Groups (RAGs)
- Generates access tokens at configurable rate for each allocation group
  - 1 per each memory bank (16 total)
  - 2 for each IOIF (4 total)
- Requestors assigned RAG ID by OS / hypervisor
  - Each SPE
  - PPE L2 / NCU
  - IOIF 0 Bus Master
  - IOIF 1 Bus Master
- Priority order for using another RAGs unused tokens
- Resource over committed warning interrupt

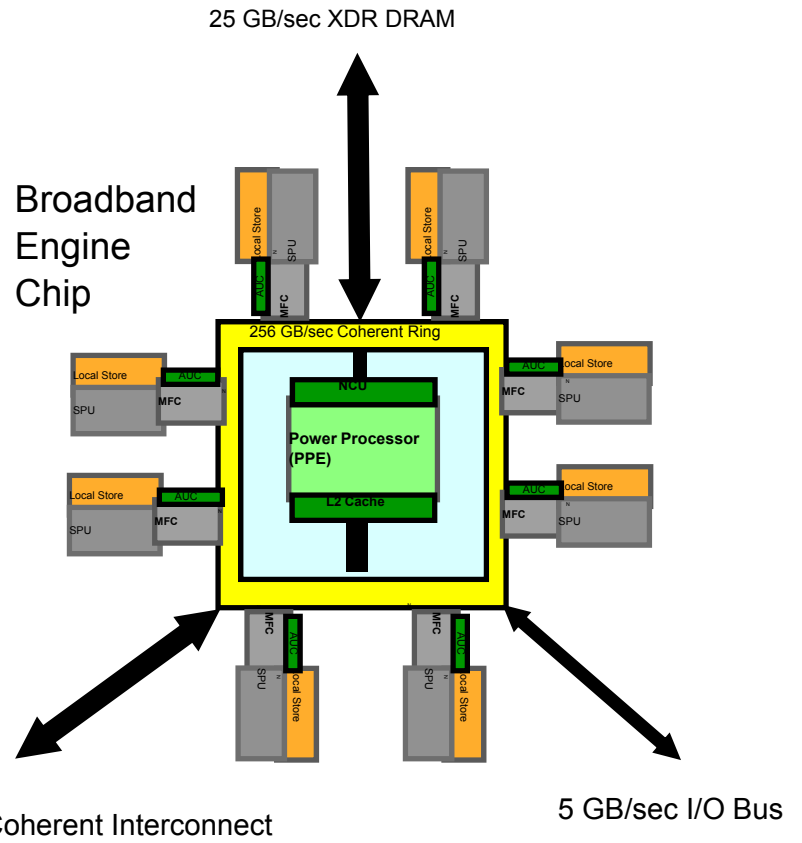




Courtesy IBM, Inc.

Cell Broadband Engine Software Overview

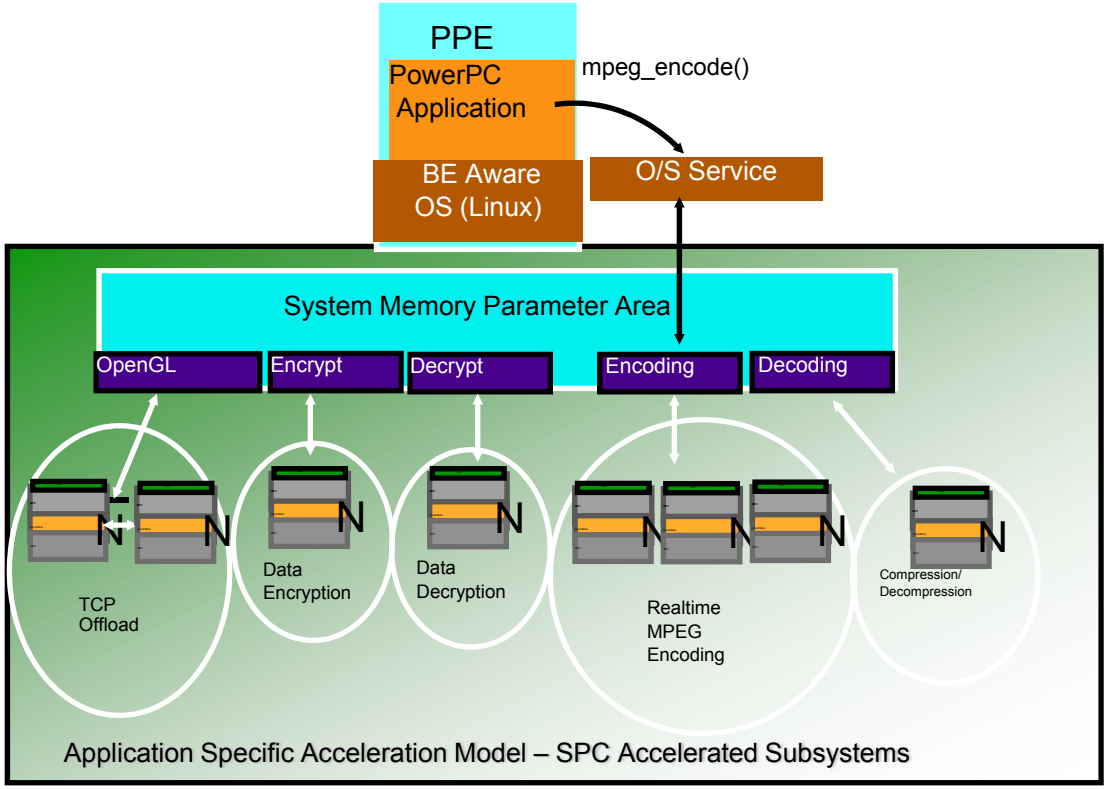
- **Flexible Program Models**
  - Application Accelerator Model
  - Function Offload Model
  - Computation Acceleration
  - Heterogeneous Multi-Threading Model



Courtesy IBM, Inc.

Programming Models

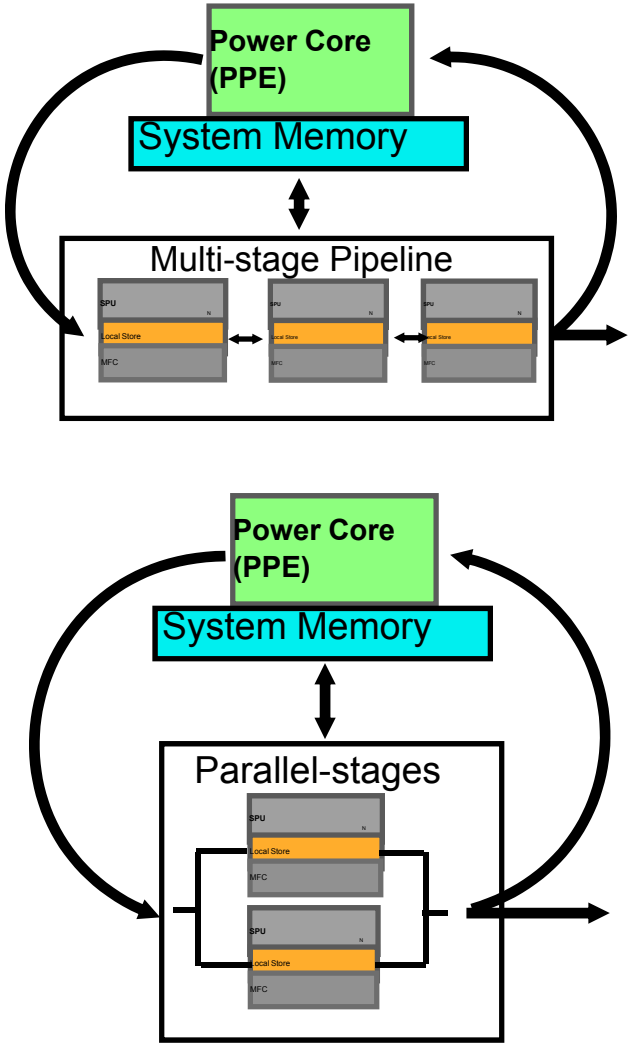
- **Application Specific Accelerators**
  - Acceleration provided by O/S services
  - Application independent of accelerators platform fixed



Courtesy IBM, Inc.

Subsystem Programming Model

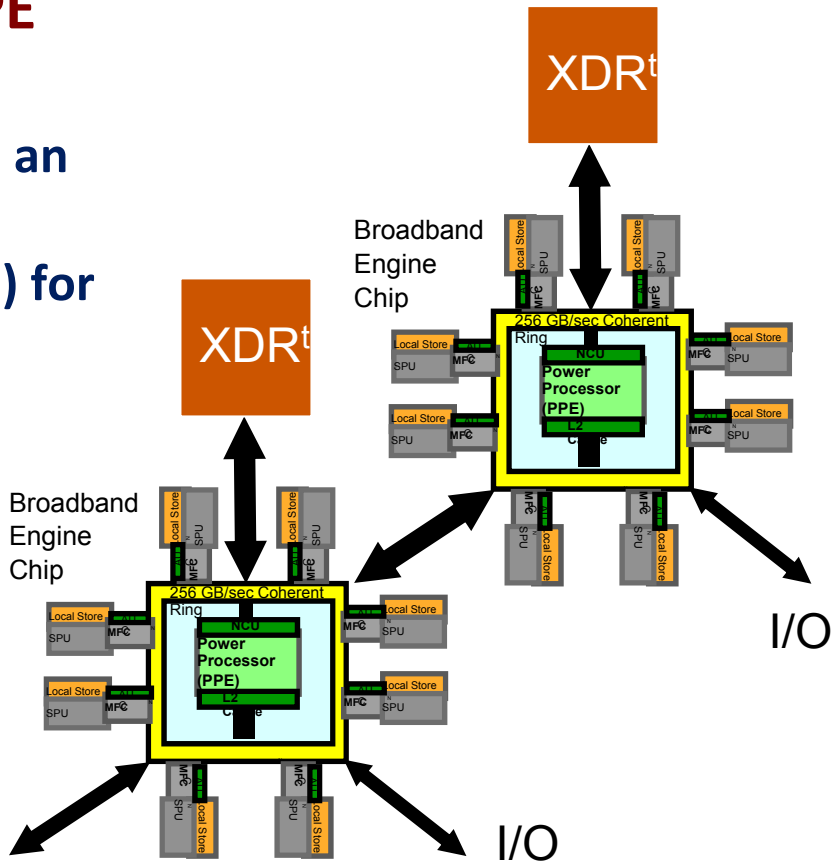
- **Function Offload**
  - Dedicated Function (problem/privileged subsystem)
    - Programmer writes/uses SPU "libraries"
      - Graphics Pipeline
      - Audio Processing
      - MPEG Encoding/Decoding
      - Encryption / Decryption
    - Main Application in PPE, invokes SPU bound services
      - RPC Like Function Call
      - I/O Device Like Interface (FIFO/ Command Queue)
    - 1 or more SPUs cooperating in subsystem
      - Problem State (Application Allocated)
      - Privileged State (OS Allocated)
    - Code-to-data or data-to-code pipelining possible
    - Very efficient in real-time data streaming applications



Courtesy IBM, Inc.

Parallel Computational Acceleration

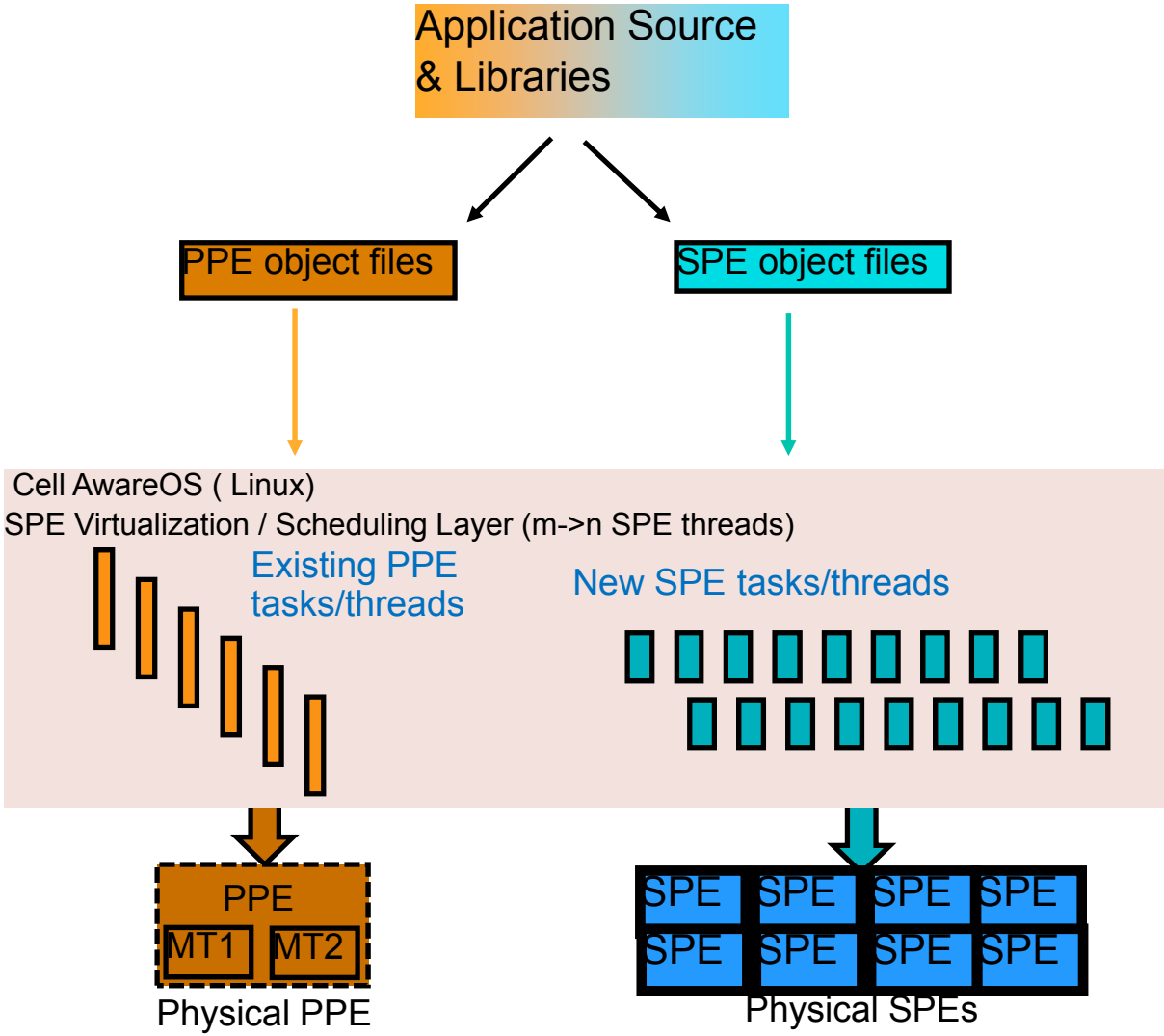
- **Single Source Compiler (PPE and SPE targets)**
  - Auto parallelization (treat target Cell as an Shared Memory MP)
  - Auto SIMD-ization (SIMD-vectorization) for PPE VMX and SPE
  - Compiler management of Local Store as Software managed cache (I&D)
- **Optimization Options**
  - OpenMP-like pragmas
  - MPI based Microtasking
  - Streaming languages
  - Vector.org SIMD intrinsics
  - Data/Code partitioning
  - Streaming / pre-specifying code/data use
    - Compiler or Programmer scheduling of DMAs
    - Compiler use of Local store as soft-cache



Courtesy IBM, Inc.

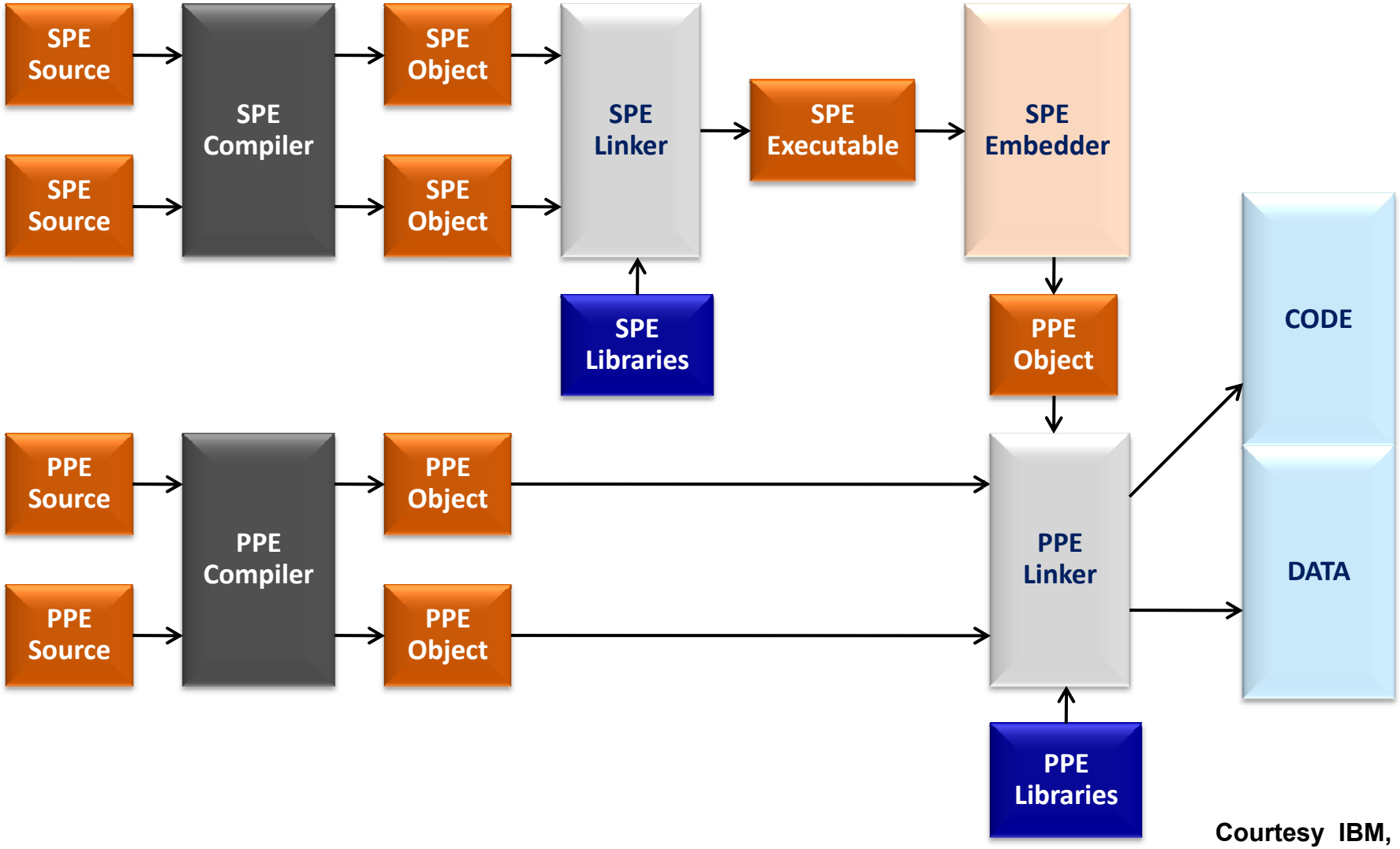
Heterogeneous Multi-Threading Model

- PPE Threads, SPE Threads
- SPE DMA EA = PPE Process EA Space
  - Or SPE Private EA space
- OS supports Create/Destroy SPE tasks
- Atomic Update Primitives used for Mutex
- SPE Context Fully Managed
  - Context Save/Restore for Debug
  - Virtualization Mode (indirect access)
  - Direct Access Mode (realtime)
- OS assignment of SPE threads to SPEs
  - Programmer directed using affinity mask
- SPE Compilers use OS runtime services



Courtesy IBM, Inc.

Compiling and binding a Cell BE program



Courtesy IBM, Inc.

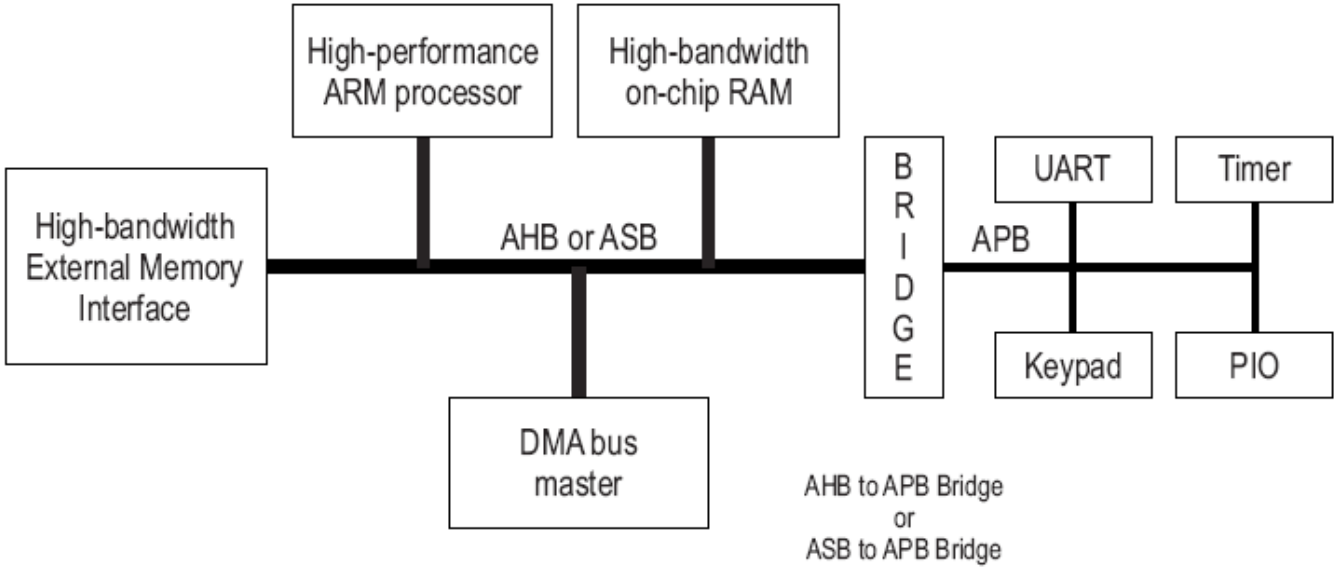
Agenda

- Market Prediction
- Architectures of three HMC-SOC platforms:
  - Atmel DIOPSIS 940HF SOC
  - Texas Instruments OMAP
  - IBM Cell Broadband Processor
- **HMC-SOC Bus Architectures**
  - **AMBA AXI**
  - IBM Cell Element Interconnect Bus (EIB)
  - Network on Chip Architectures (NOC)

AMBA Introduction

- Advanced Microcontroller Bus Architecture (AMBA), created by ARM as an interface for their microprocessors.
- Easy to obtain documentation (free download) and can be used without royalties.
- Very common in commercial SoC's
  (e.g. Qualcomm Multimedia Cellphone SoC)
- AMBA 2.0 released in 1999, includes APB and AHB
- AMBA 3.0 released in 2003, includes AXI

AMBA 2.0 System-Level View



AMBA AHB

* High performance
* Pipelined operation
* Multiple bus masters
* Burst transfers
* Split transactions

AMBA ASB

* High performance
* Pipelined operation
* Multiple bus masters

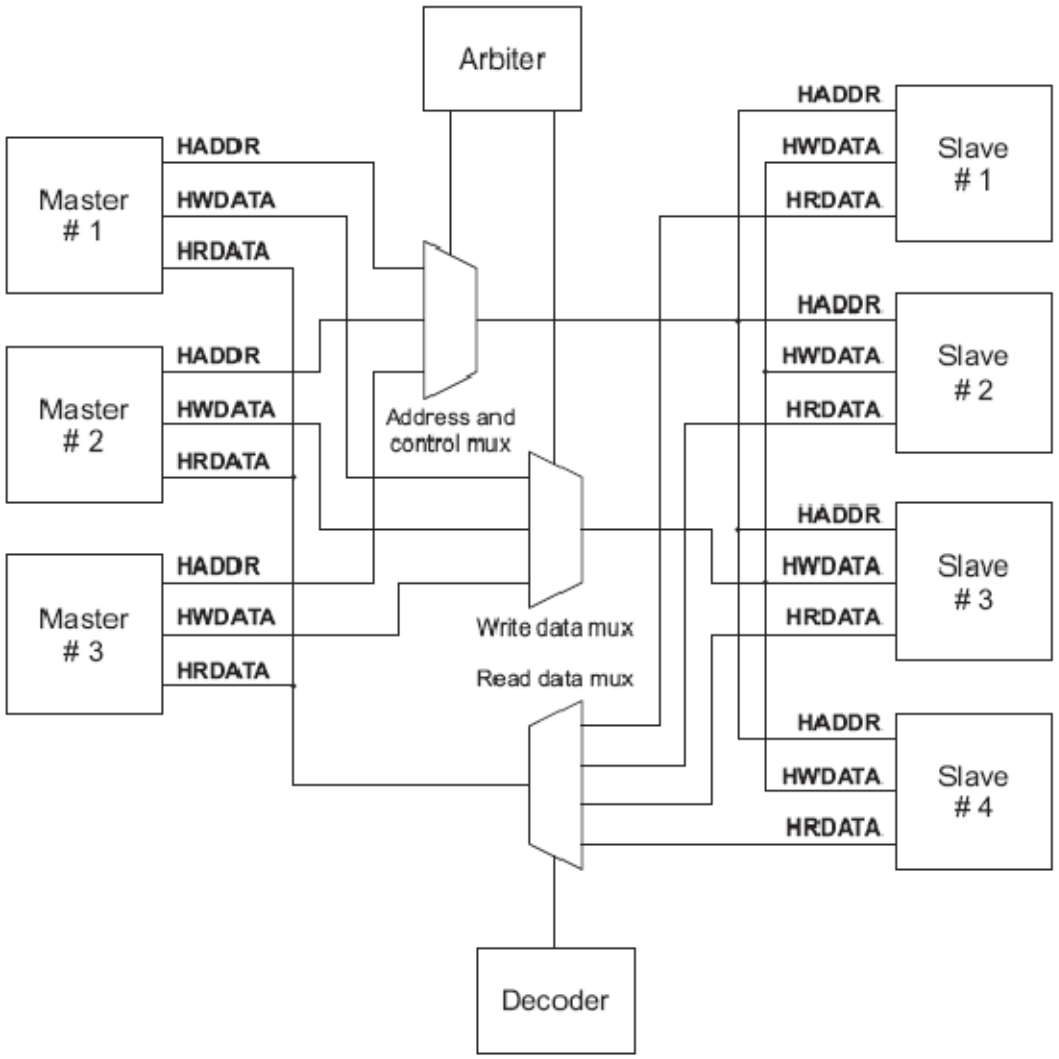
AMBA APB

* Low power
* Latched address and control
* Simple interface
* Suitable for many peripherals

Source: AMBA Specification, Rev. 2.0

AHB Architecture

- Central MUX is used, rather than a bus
- Achieves smaller delays than a single wire w/ tri-state buffers



Source: AMBA Specification, Rev. 2.0

AMBA 3.0

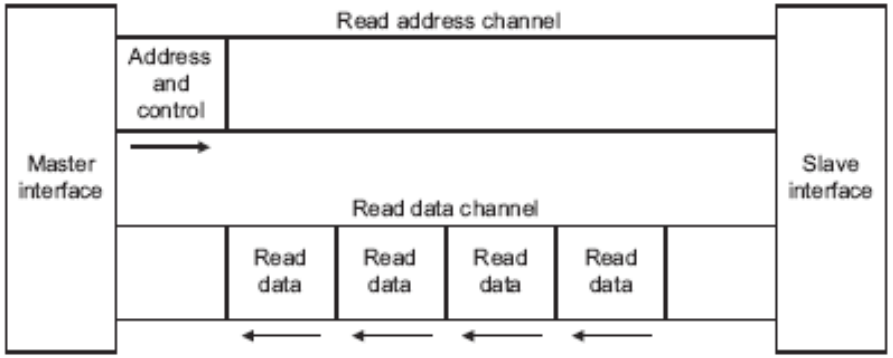
- **AXI high performance protocol**
  - Support for separate read address, write address, read data, write data, write response channels
  - Up to 16 masters allowed
  - Requires ~77 control signals
  - Out of order (OO) transaction completion
  - Fixed mode burst support
    - Useful for I/O peripherals
  - Advanced system cache support
    - Specify if transaction is cacheable/bufferable
    - Specify attributes such as write-back/write-through
  - Enhanced protection support
    - Secure/non-secure transaction specification
  - Exclusive access (for semaphore operations)
  - Register slice support for high frequency operation

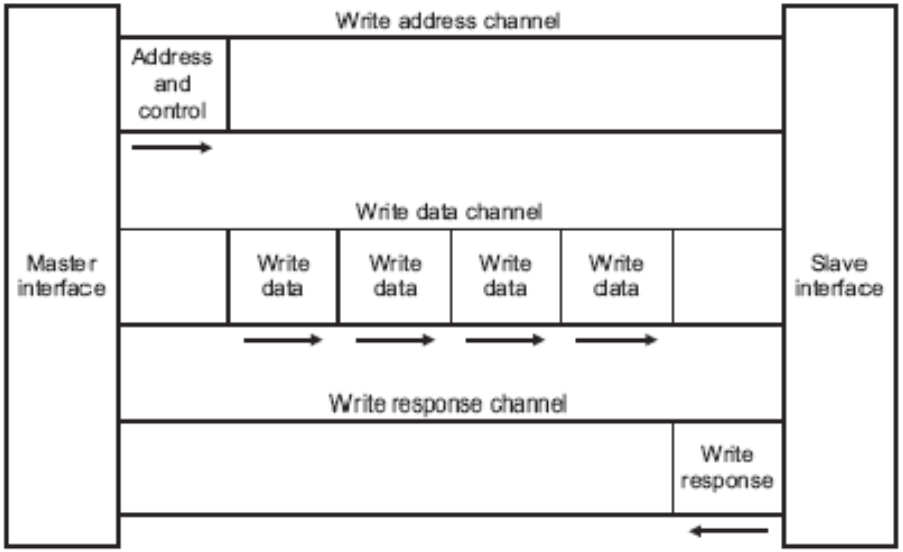
Source: AMBA AXI Protocol Specification

On-Chip Communication Architectures: Sudeep Pasricha & Nikil Dutt

Multi-Channel Support

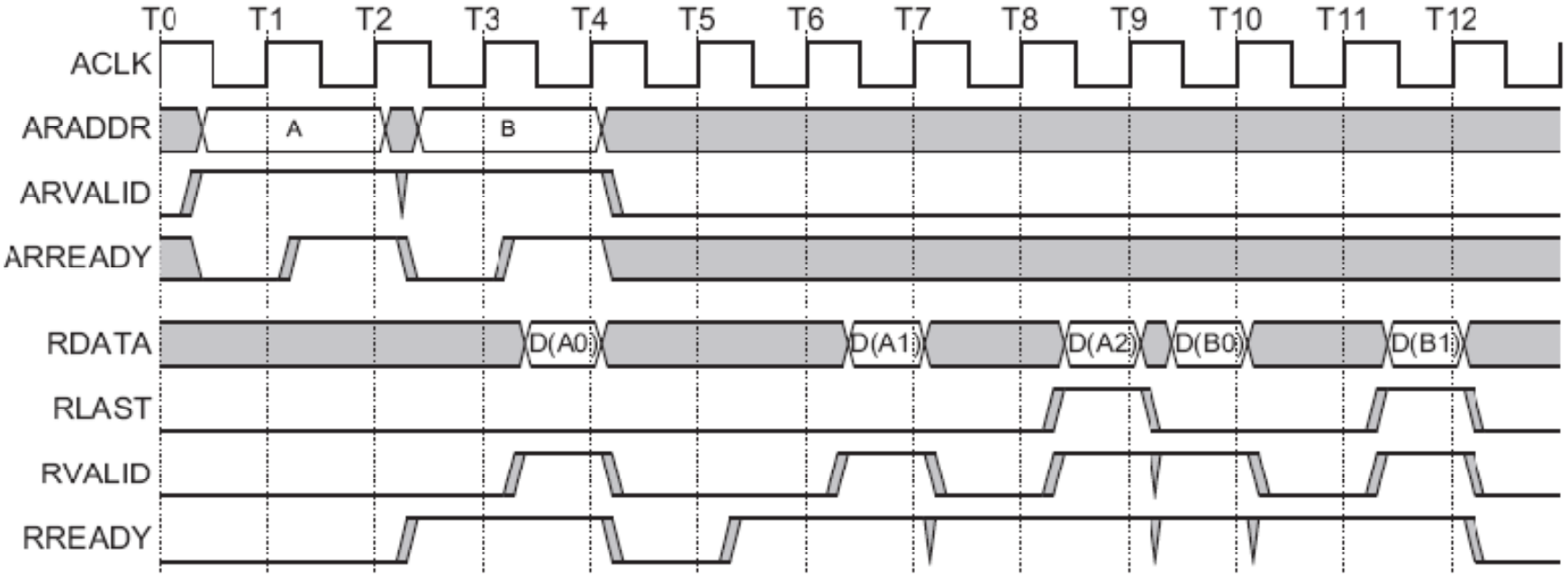
- Address, Data, and Response split between channels, rather than phases
- Allows simultaneous reads and writes





Source: AMBA AXI Protocol Specification

AXI Read Transactions

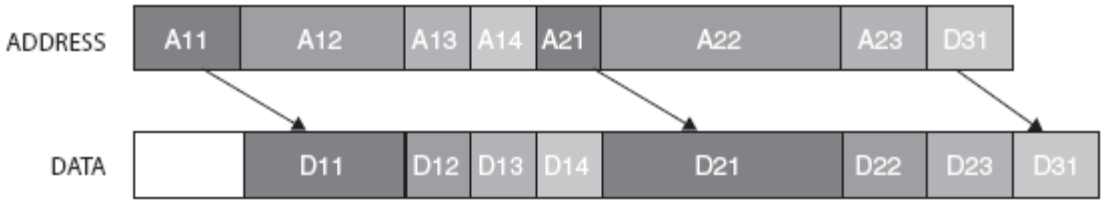


- Up to 16 transactions can be queued at once

Source: AMBA AXI Protocol Specification

AHB vs. AXI Burst

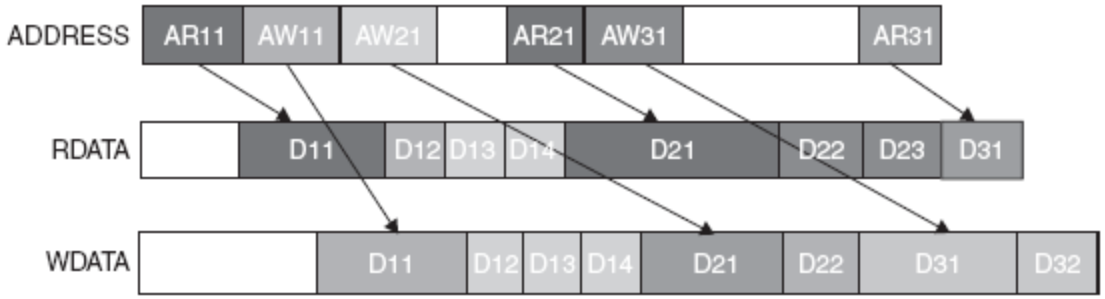
- **AHB Burst**
  - Address and Data are locked together (single pipeline stage)
  - HREADY controls intervals of address and data



- **AXI Burst**
  - One Address for entire burst

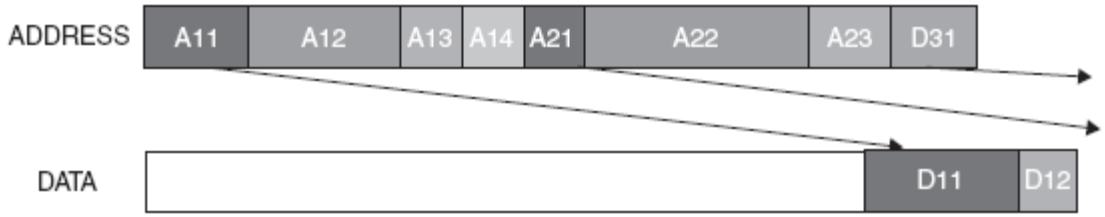
AHB vs. AXI Burst

- **AXI Burst**
  - Simultaneous read, write transactions
  - Better bus utilization

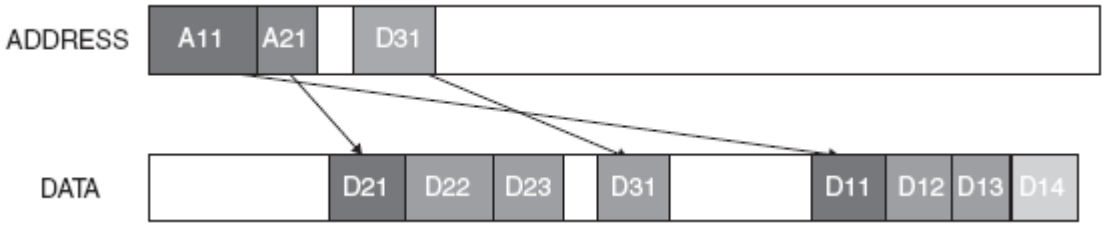


AXI Out of Order Completion

- **With AHB**
  - If one slave is very slow, all data is held up
  - SPLIT transactions provide very limited improvement



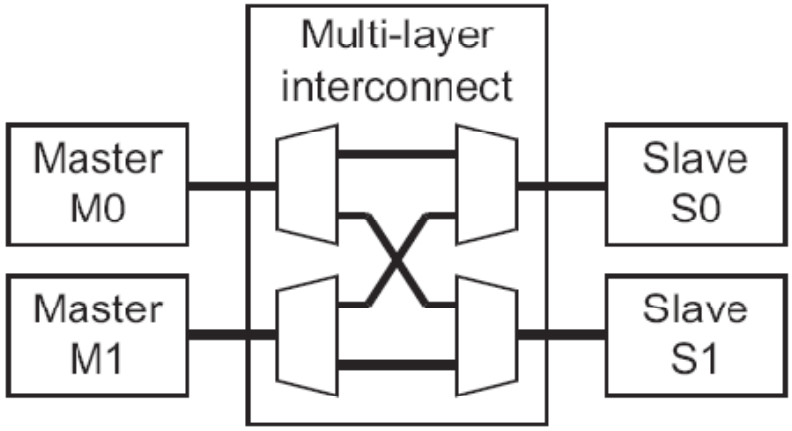
- **With AXI Burst**
  - Multiple outstanding addresses, out of order (OO) completion allowed
  - Fast slaves may return data ahead of slow slaves



On-Chip Communication Architectures: Sudeep Pasricha & Nikil Dutt

Multi-Layer Connectivity

- PL300 Interconnect is implemented as a crossbar:
- Multiple masters can talk to multiple slaves simultaneously



Source: PL300 Technical Reference Manual

Comparison of AMBA Bus Types

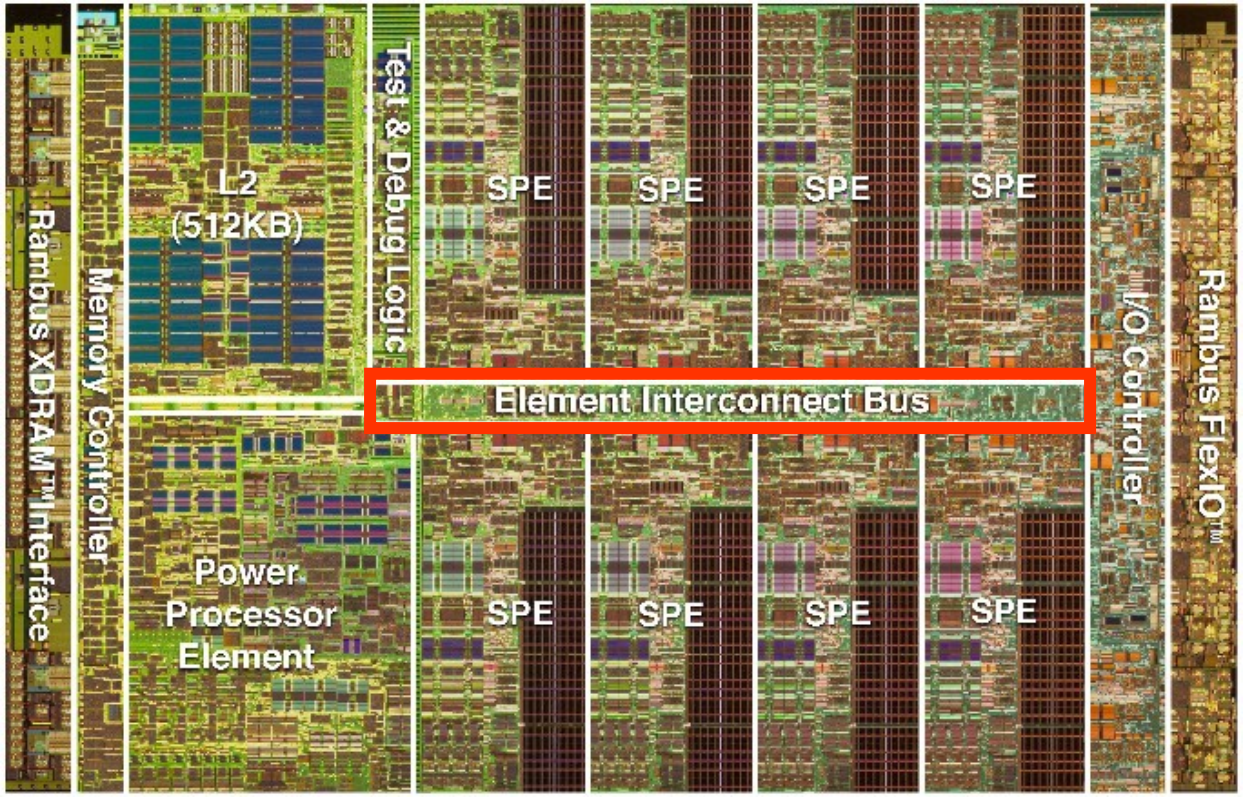
	APB	AHB	AXI / PL300
Processors	all	ARM7,9,10	ARM11/Cortex
Control Signals	4	27	77
No. of Masters	1	1-15	1-16
No. of Slaves	1-15	1-15	1-16
Interconnect Type	Central MUX?	Central MUX	Crossbar w/ 5 channels
Phases	Setup, Enable	Bus request, Address, Data	Address, Data, Response
Xact. Depth	1	2	16
Burst Lengths	1	1-32	1-16
Simultaneous Read & Write	no	no	yes

Agenda

- Market Prediction
- Architectures of three HMC-SOC platforms:
  - Atmel DIOPSIS 940HF SOC
  - Texas Instruments OMAP
  - IBM Cell Broadband Processor
- HMC-SOC Bus Architectures
  - AMBA AXI
  - **IBM Cell Element Interconnect Bus (EIB)**
  - Network on Chip Architectures (NOC)

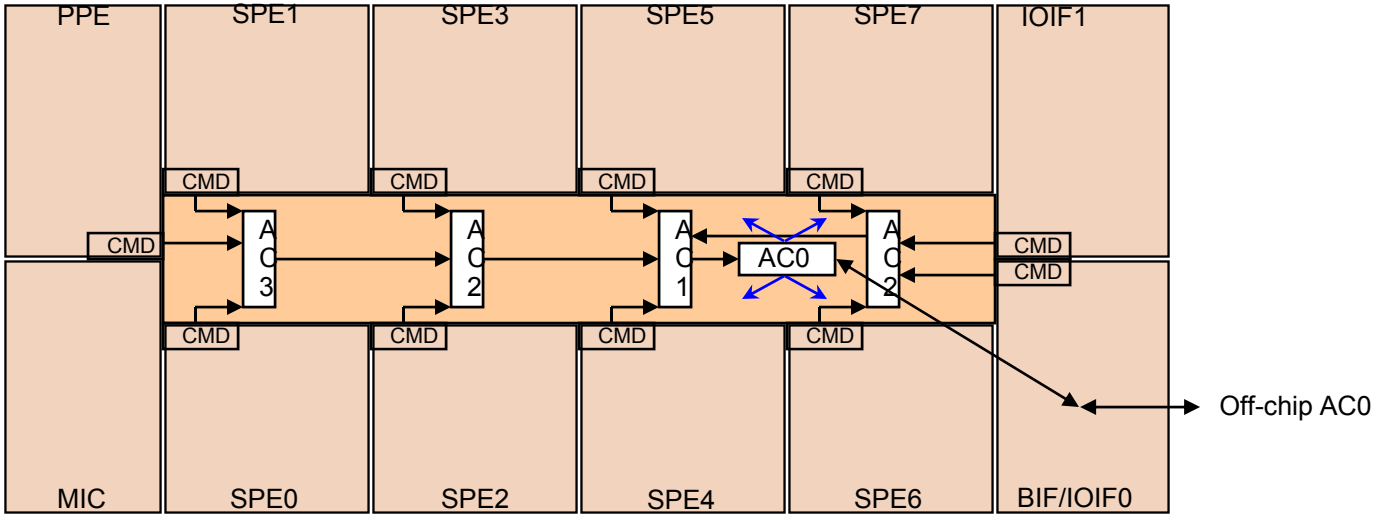
Cell Broadband Processor Element Interconnect Bus (EIB)

- **EIB data ring for internal communication**
  - Four 16 byte data rings, supporting multiple transfers
  - 96B/cycle peak bandwidth
  - Over 100 outstanding requests



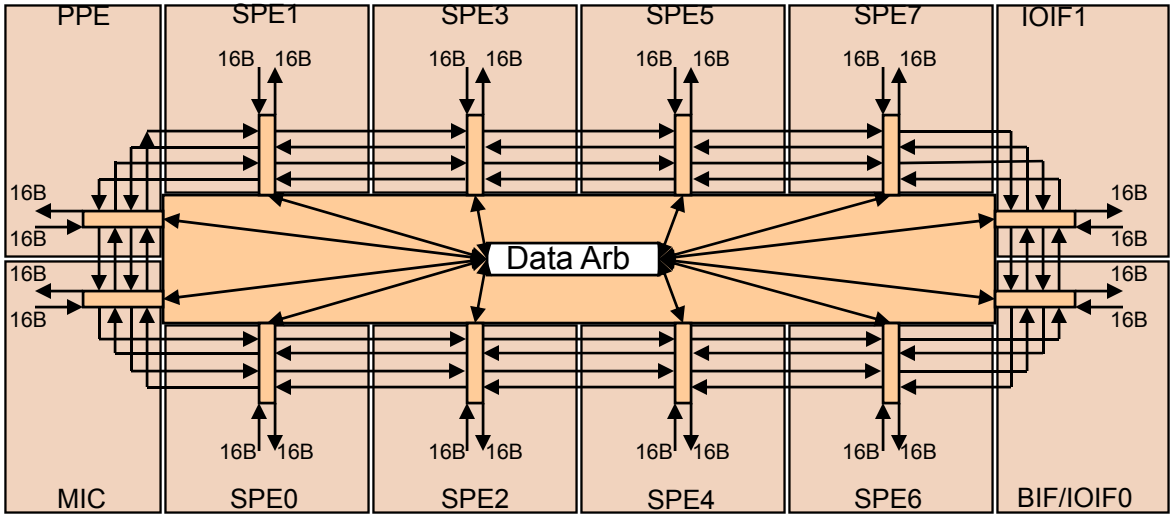
Element Interconnect Bus – Command Topology

- "Address Concentrator" tree structure minimizes wiring resources
- Single serial command reflection point (AC0)
- Address collision detection and prevention
- Fully pipelined
- Content –aware round robin arbitration
- Credit-based flow control



Element Interconnect Bus - Data Topology

- Four 16B data rings connecting 12 bus elements
  - Two clockwise / Two counter-clockwise
- Physically overlaps all processor elements
- Central arbiter supports up to three concurrent transfers per data ring
  - Two stage, dual round robin arbiter
- Each element port simultaneously supports 16B in and 16B out data path
  - Ring topology is transparent to element data interface

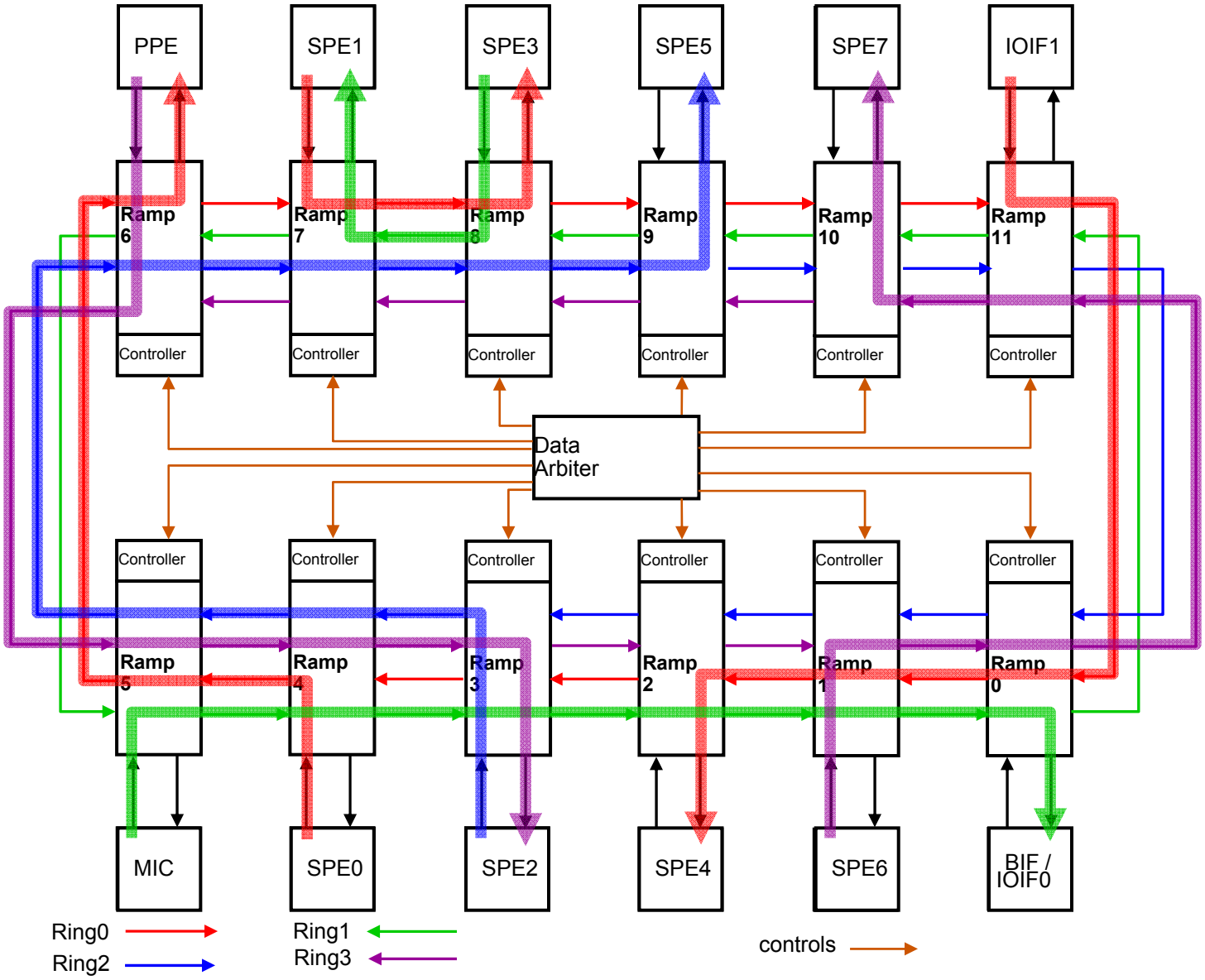


Internal Bandwidth Capability

- Each EIB Bus data port supports 25.6GBytes/sec* in each direction
- The EIB Command Bus streams commands fast enough to support 102.4 GB/sec for coherent commands, and 204.8 GB/sec for non-coherent commands.
- The EIB data rings can sustain 204.8GB/sec for certain workloads, with transient rates as high as 307.2GB/sec between bus units

* The above numbers assume a 3.2GHz core frequency – internal bandwidth scales with core frequency

Example of eight concurrent transactions

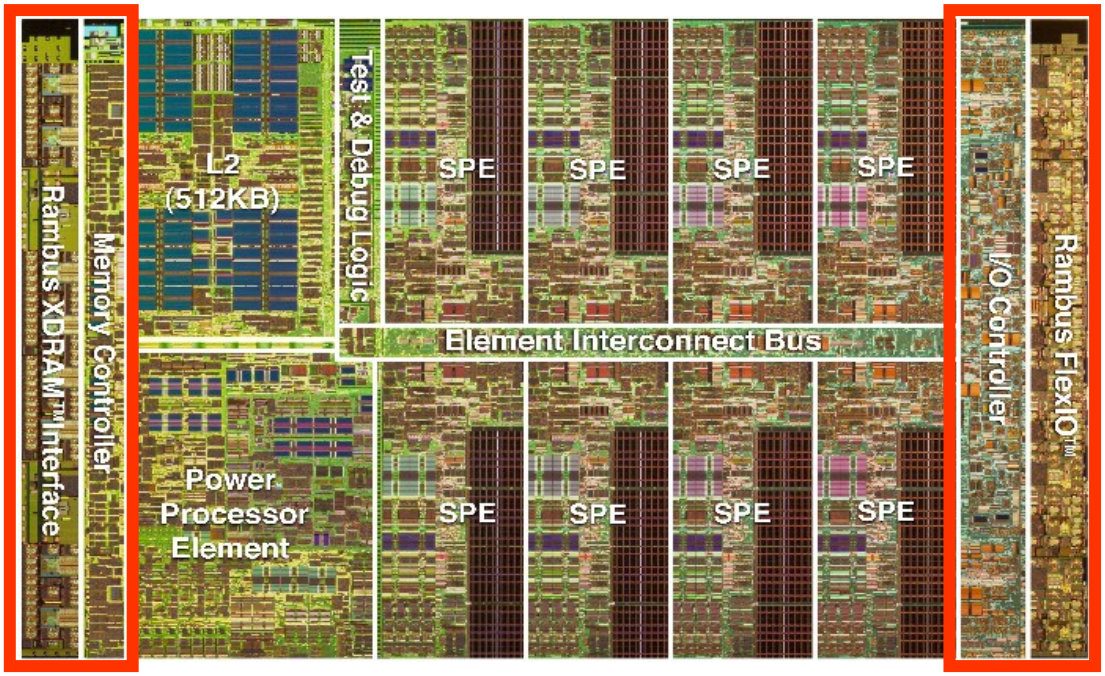


Resource Allocation Management

- **Optional facility used to minimize over-allocation effects of critical resources**
  - Independent but complementary function to the EIB
  - Critical (managed) resource's time is distributed among groups of requestors
- **Managed resources include:**
  - Rambus XDR™ DRAM memory banks (0 to 15)
  - BIF/IOIF0 Inbound and BIF/IOIF0 Outbound
  - IOIF1 Inbound and IOIF1 Outbound
- **Requestors Allocated to Four Resource Allocation Groups (RAG)**
  - 17 requestors – PPE, SPEs, I/O Inbound (4 VCs), I/O Outbound (4 VCs)
- **Central Token Manager controller**
  - Requestors ask permission to issue EIB commands to managed resources
  - Tokens granted across RAGs allow requestor access to issue command to the EIB
  - Round robin allocation within RAG
  - Dynamic software configuration of the Token Manager to adjust token allocation rates for varying workloads
  - Multi-level hardware feedback from managed resource congestion to throttle token allocation

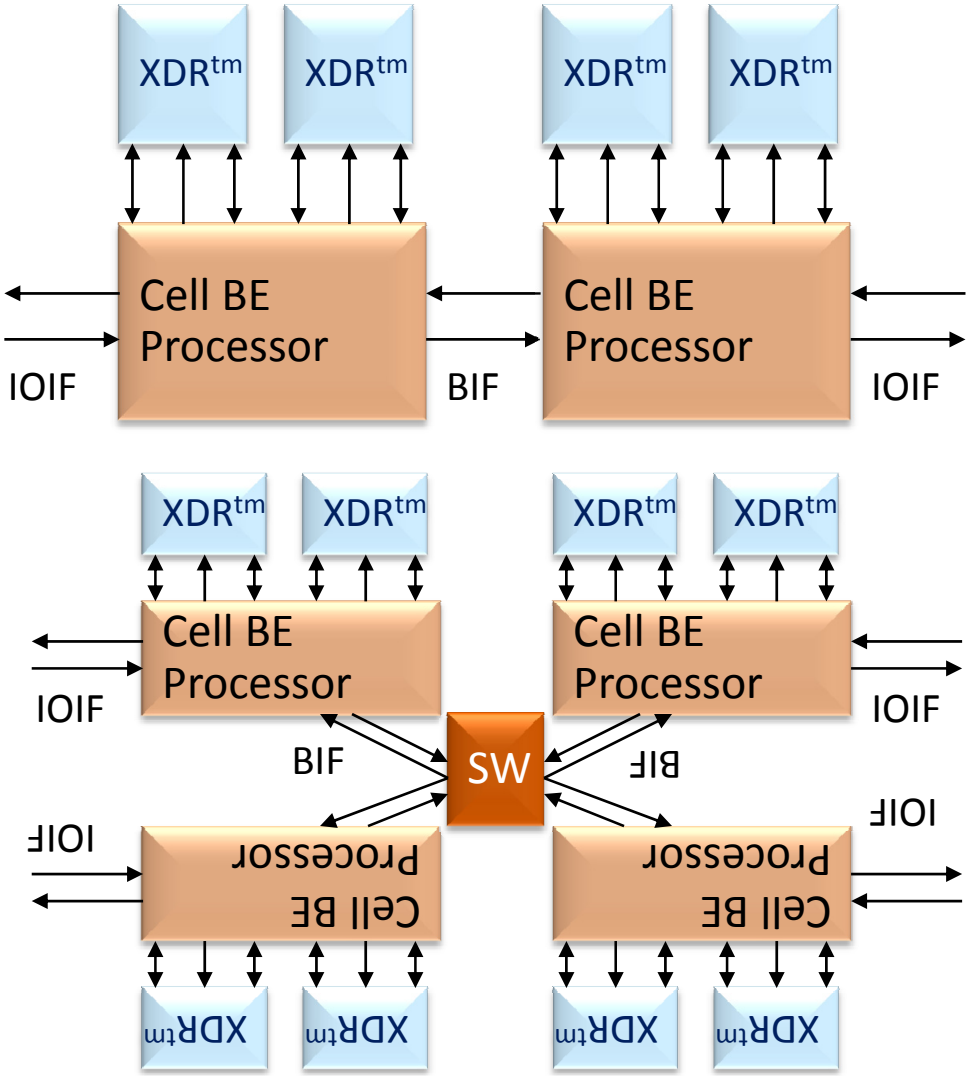
I/O and Memory Interfaces

- **I/O Provides wide bandwidth**
  - Dual XDRTM controller (25.6GB/s @ 3.2Gbps)
  - Two configurable interfaces (76.8GB/s @6.4Gbps)
    - Configurable number of Bytes
    - Coherent or I/O Protection
  - Allows for multiple system configurations



Cell BE Processor Can Support Many Systems

- Game console systems
- Blades
- HDTV
- Home media servers
- Supercomputers



Agenda

- Market Prediction
- Architectures of three HMC-SOC platforms:
  - Atmel DIOPSIS 940HF SOC
  - Texas Instruments OMAP
  - IBM Cell Broadband Processor
- HMC-SOC Bus Architectures
  - AMBA AXI
  - IBM Cell Element Interconnect Bus (EIB)
  - **Network on Chip Architectures (NOC)**

NOC Architectures

- **Paradigm shift in Multi-core SOC design**

- **Communication challenges**
  - Synchronous communication infeasible
  - Errors due to integrity issues
    - RLC effects
    - Cross-coupling effects
  - Delay Insensitive may be a better approach.

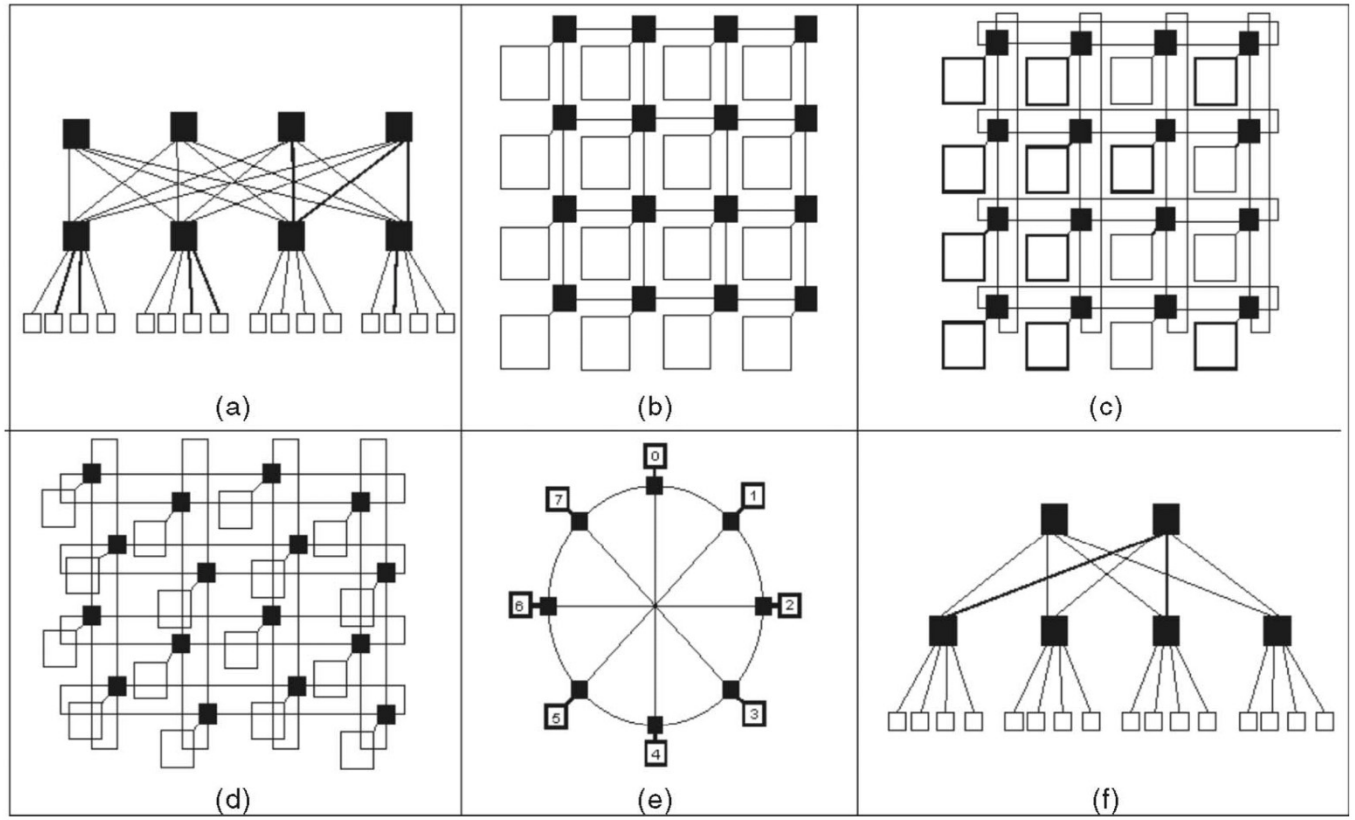
- **Network-on-Chip (NoC)**
  - Packet switching based communication.
  - Extremely high bandwidth by pipelined signal transmission.
  - Asynchronous (delay insensitive) communication between routers.
  - Support for error control schemes.

Chatha, ASU, 2005

Topologies

- **Heritage of networks with new constraints**
  - Need to accommodate interconnects in a 2D layout
  - Cannot route long wires (clock frequency bound)

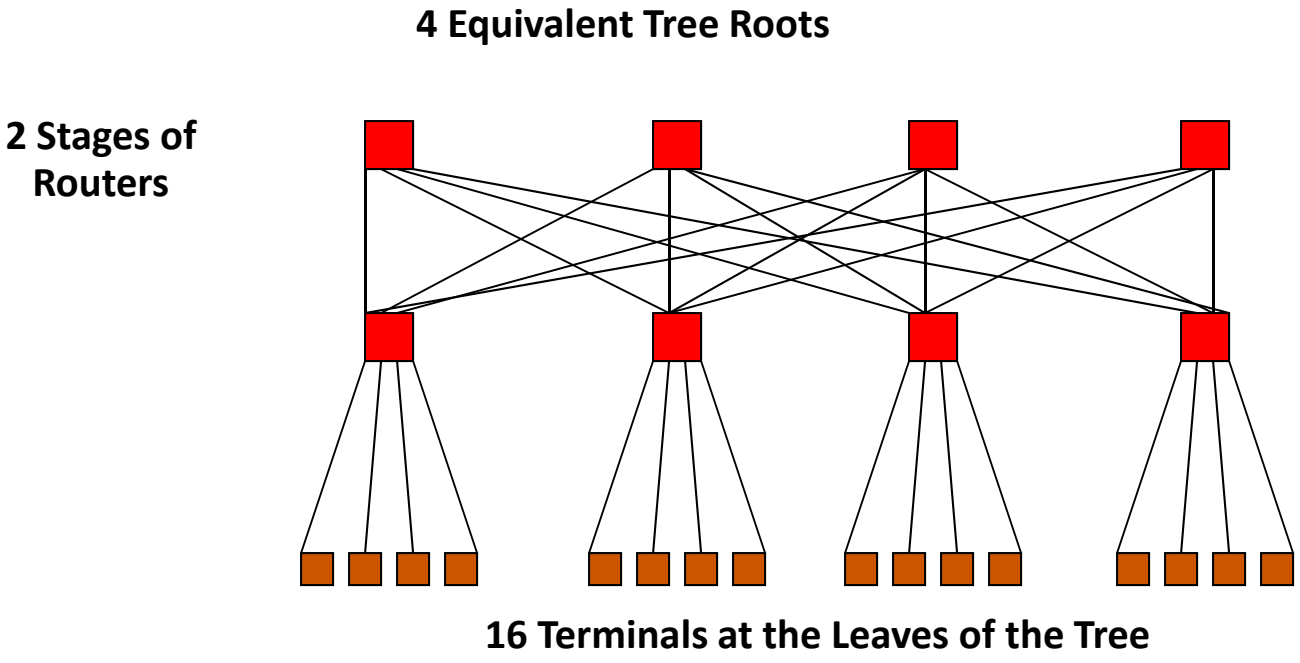
a) SPIN,
b) CLICHE' & Mesh
c) Torus
d) Folded torus
e) Octagon
f) BFT



Pande, et al 2005

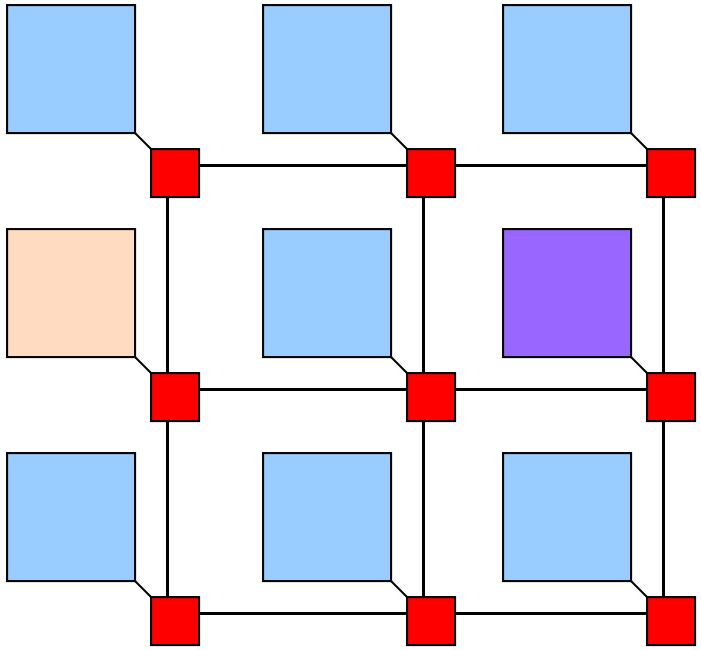
Architectures: SPIN

- **SPIN: Scalable, Programmable, Integrated Network**
  - Every level has same number switches
  - Network grows as (NlogN)/8
  - Trades area overhead and decreased power efficiency for higher throughput
  - Illustrative of performance vs. power consumption



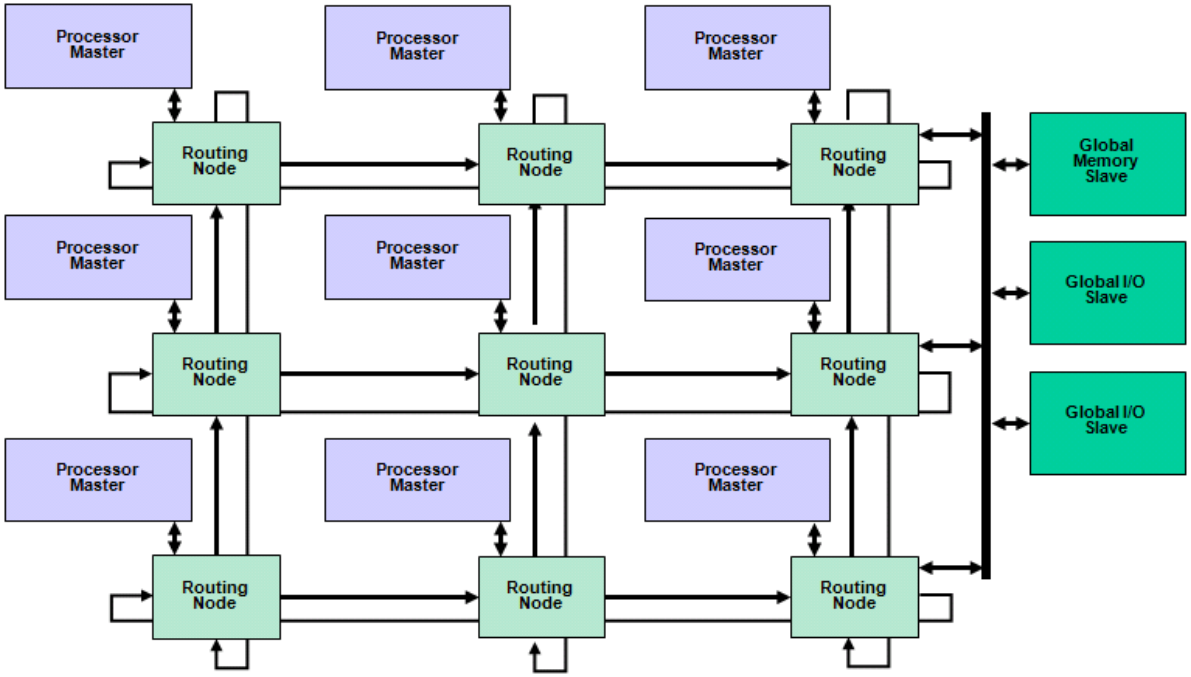
Architectures: CLICHE

- **CLICHÉ: Chip-Level Integration of Communicating Heterogeneous Elements**
  - Two-dimensional mesh network layout for NoC design
  - All switches are connected to the four closest other switches and target resource block, except those switches on the edge of the layout
  - Connections are two unidirectional links



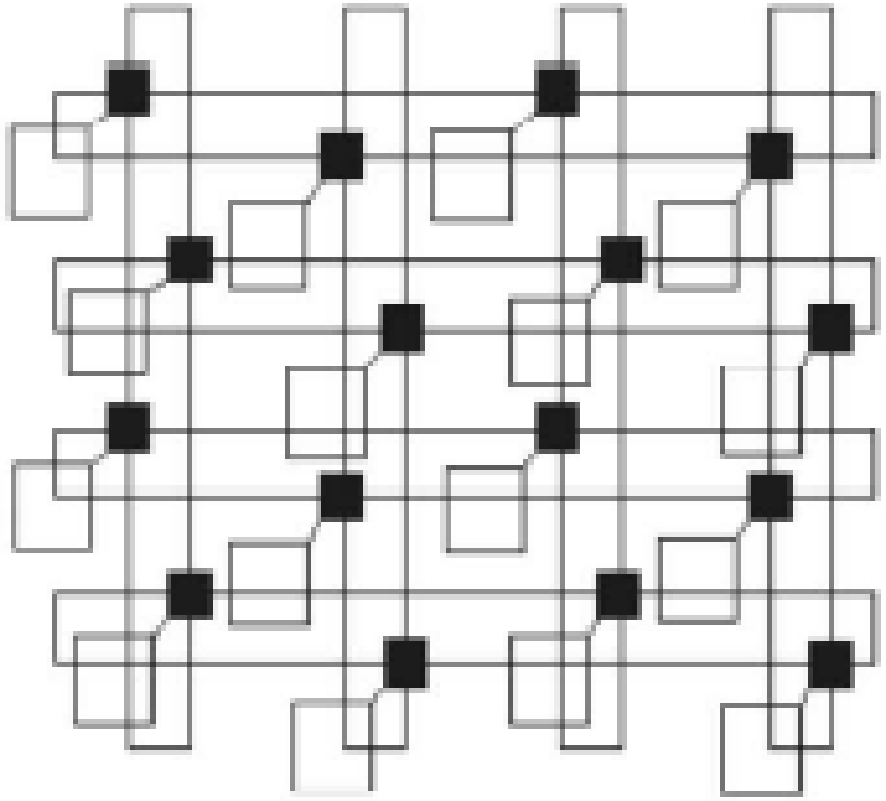
Architectures: Torus

- Similar to mesh based architectures
  - Wires are wrapped around from the top component to the bottom and rightmost to leftmost
  - Smaller hop count
  - Higher bandwidth
  - Decreased Contention
  - Increased chip space usage



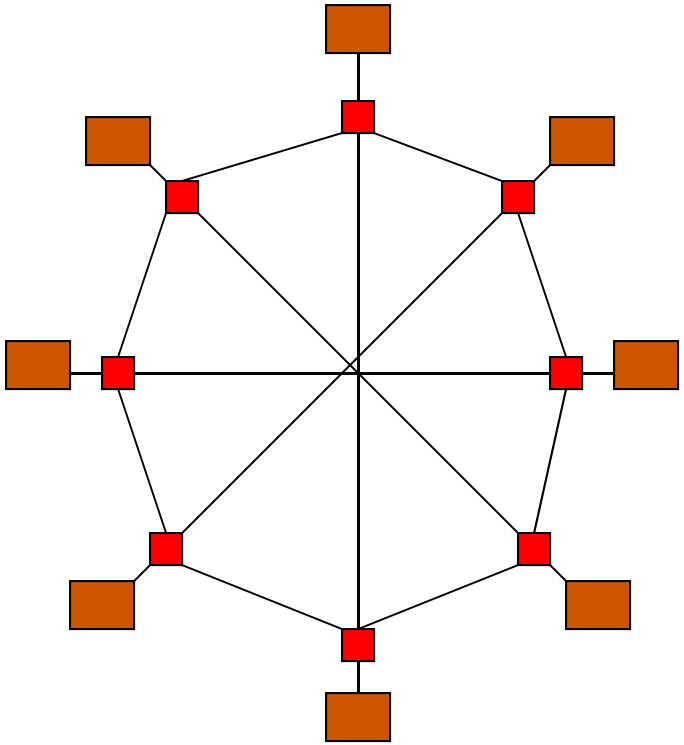
Architectures: Folded Torus

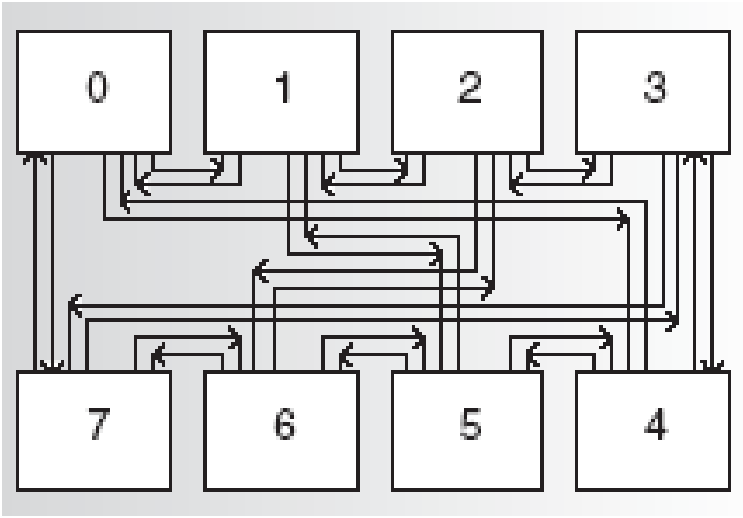
- Similar to Torus
  - Torus, the long end-around connections can yield excessive delays
  - Avoided by folding the torus



Architectures: Octagon

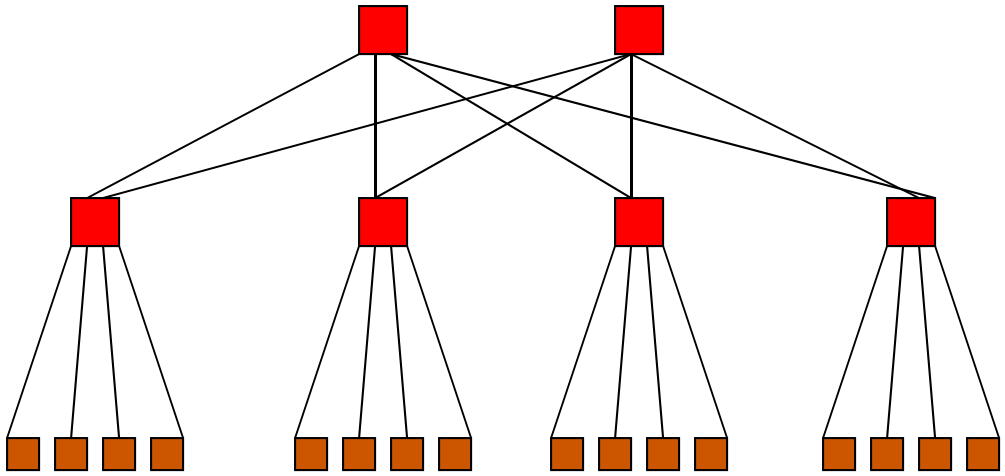
- **Standard model: 8 components, 12 interconnects**
  - Design complexity increases linearly with number of nodes
  - Largest packet travel distance is two hops
  - High throughput
  - Shortest path routing easy to implement



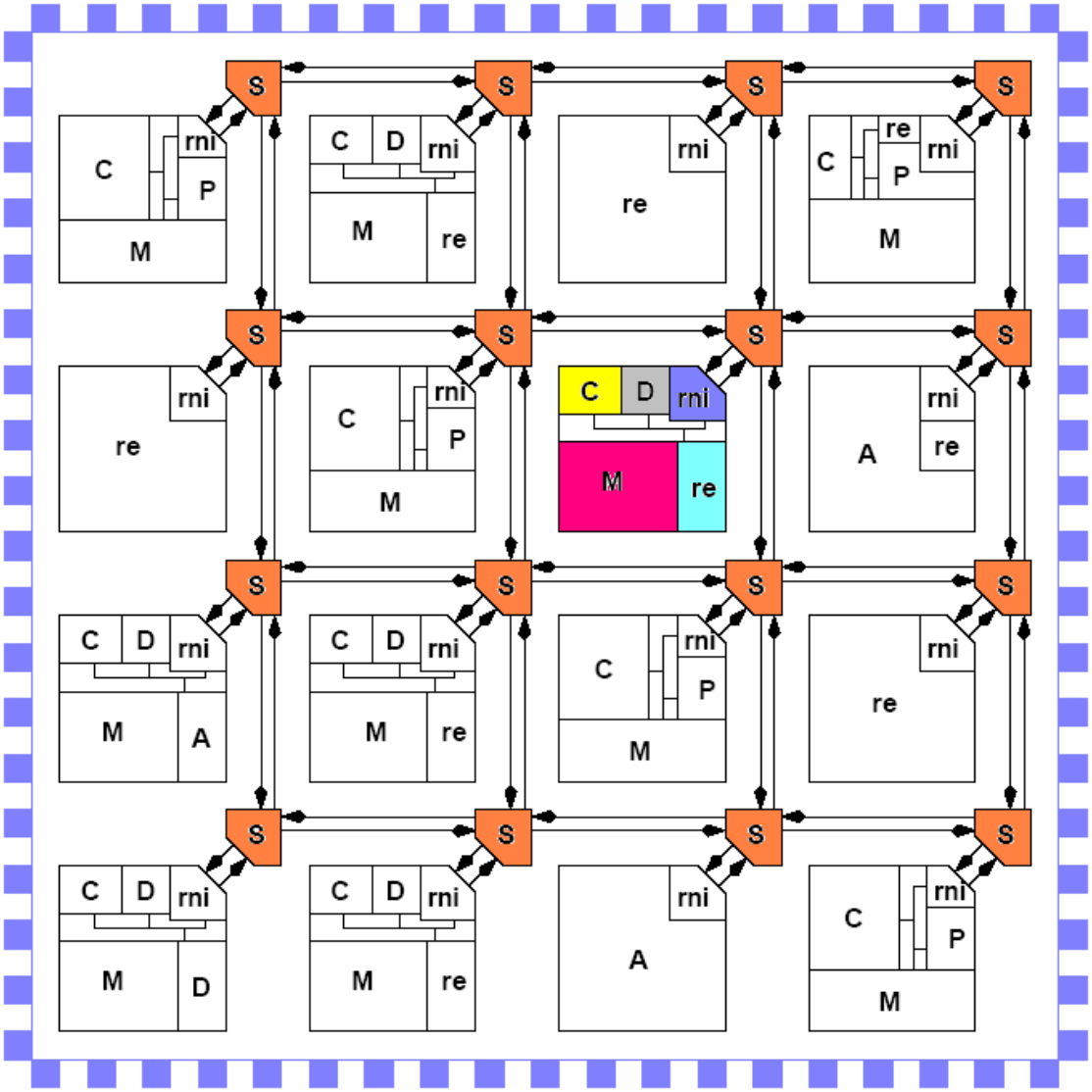


Architectures: BFT

- **BFT: Butterfly Fat Tree**
  - Each node in tree model has coordinates (level, position) where level is depth and position is from left to right
  - Leaves are component blocks
  - Interior nodes are switches
  - Four child ports per switch and two parent ports
  - LogN levels, ith level has n/(2^i+1) switches, n = leaves (blocks)
  - Use traffic aggregation to reduce congestion



Mesh based router architecture



S = switch/router

rni = resource network interface

C = cache

P = processor

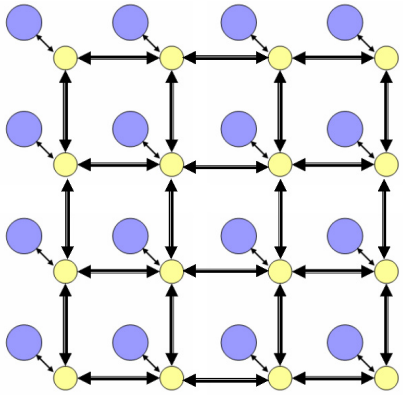
M = memory

D = DSP

re = reconfigurable logic

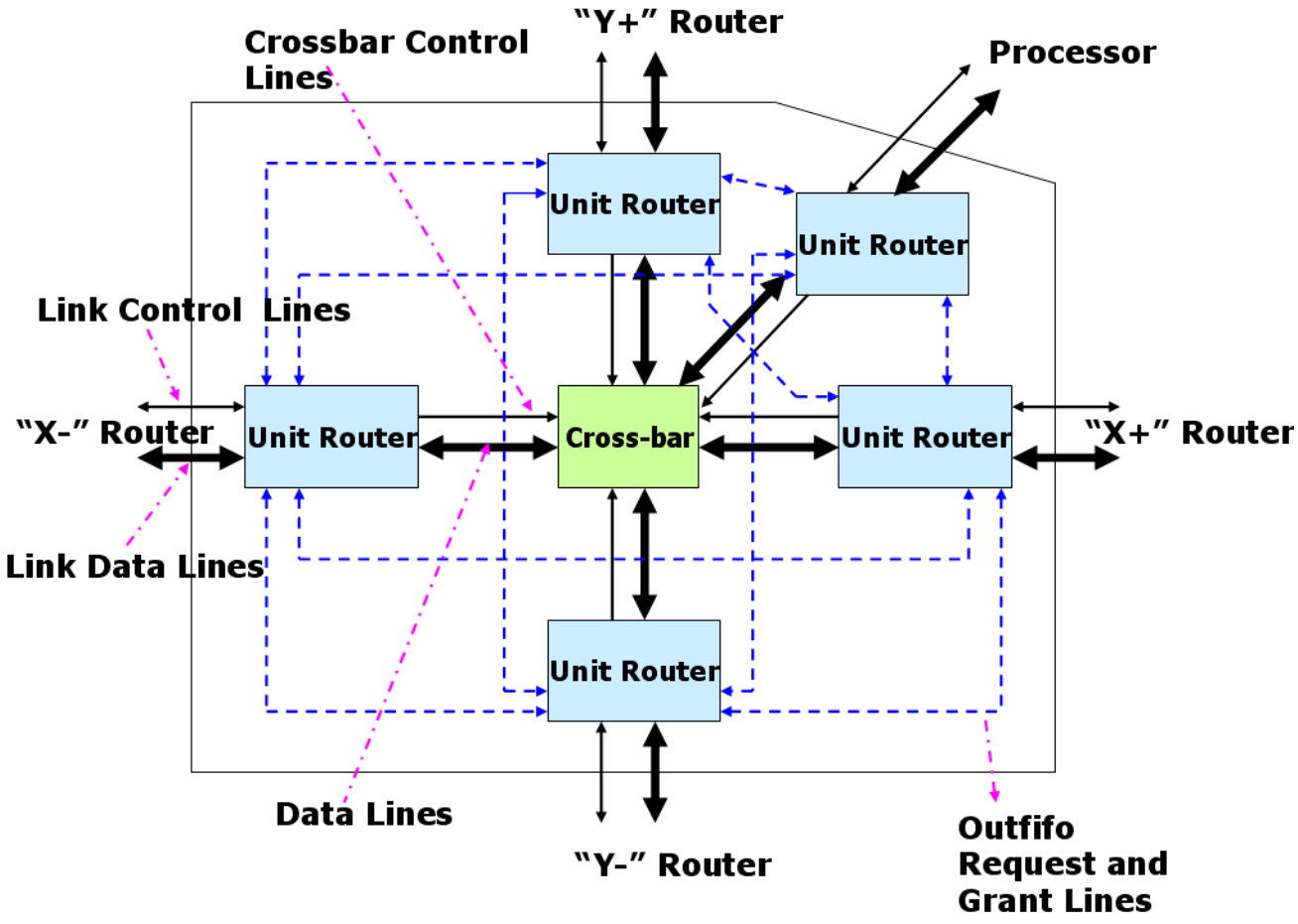
Chatha, ASU, 2005

Mesh based router architecture



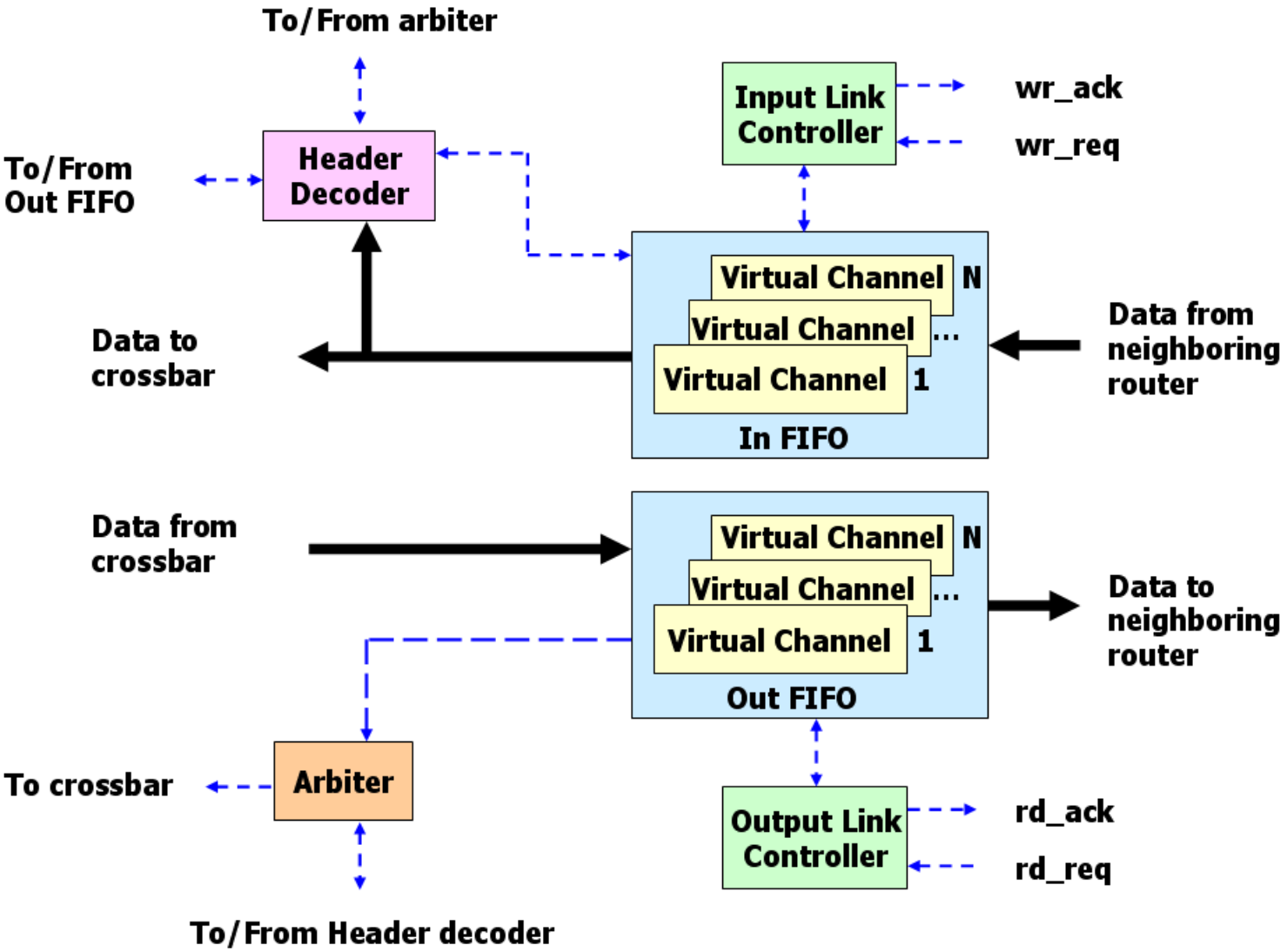
Router

Processor



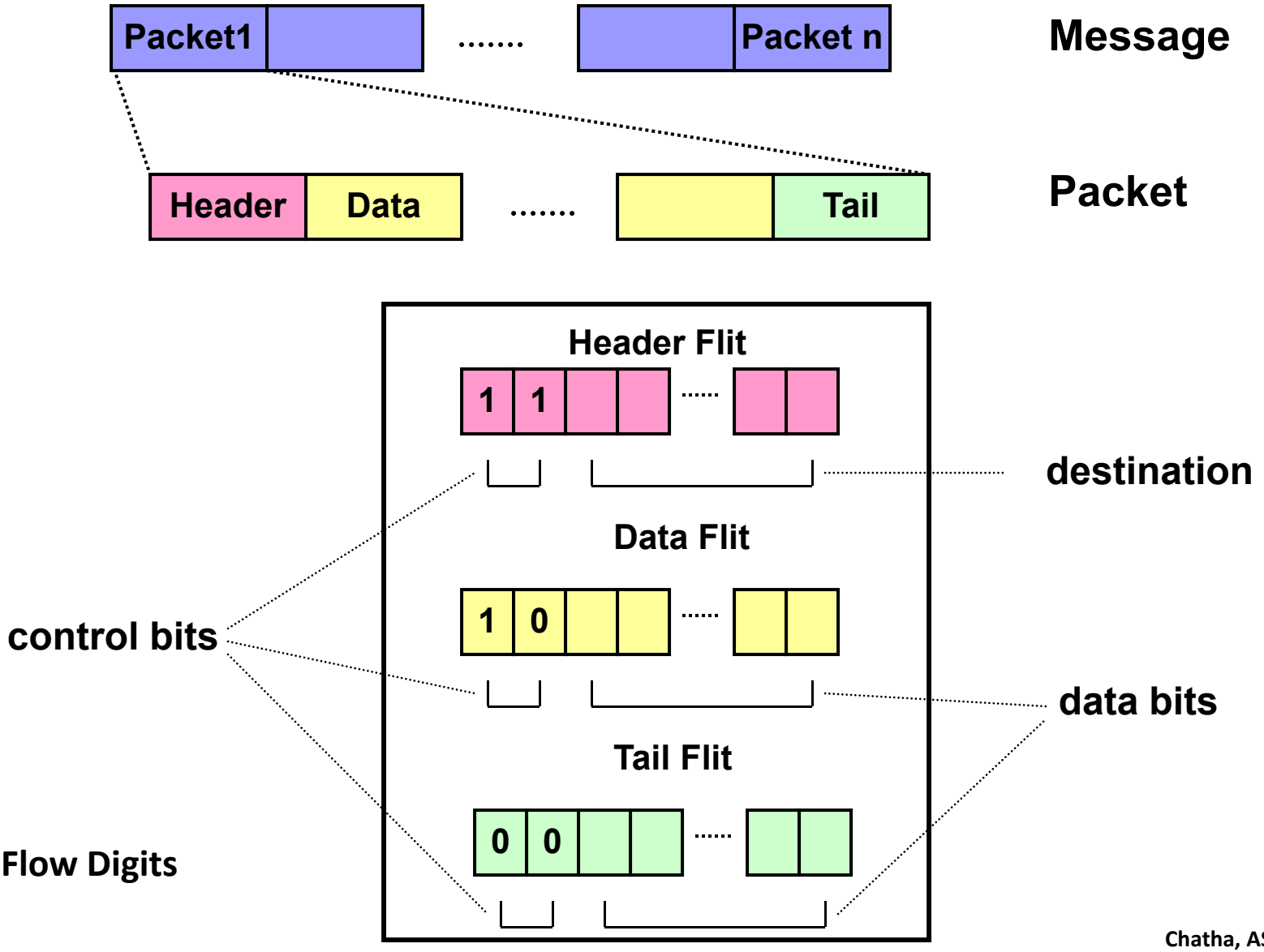
Chatha, ASU, 2005

Unit Router Architecture



Chatha, ASU, 2005

Packet Format



Chatha, ASU, 2005

Routing Performance metrics

- **Injection Rate**
  - Number of packets that are injected from the source into the network per unit time
- **Average Network Latency**
  - Average delay experienced by packets as they traverse from source to destination.
- **Average Setup Latency**
  - Average time required to reserve the virtual channels for the GT traffic from source to destination.
- **Acceptance rate**
  - Number of packets reaching each of the destination nodes per clock cycle at a particular injection rate.
- **Average power consumption**
  - Sum of average dynamic and leakage power consumed by the network per clock cycle.

Chatha, ASU, 2005

Quality-of-Service levels

- **Guaranteed Throughput (GT)**
  - Throughput and latency guarantees
  - Supports bursty traffic
  - 3 stages – setup, transmit, tear-down
    - setup – virtual channels reserved from source to destination
    - transmit – packets transferred with maximum throughput
    - tear-down – path is set free
- **Best Effort (BE)**
  - no time guarantees

Chatha, ASU, 2005

Quality-of-Service architecture

- Virtual channels divided into two sets
  - Guaranteed throughput and best effort
  - GT traffic can take over BE virtual channels
- Higher priority given to GT traffic
  - Header decoder
  - Arbiter
  - Output link controller
- Round-robin priority mechanism within each class

Chatha, ASU, 2005