



Intel[®] QuickPath Interconnect Overview



presented by
Robert Safranek

Contributors:
Gurbir Singh,
Robert Maddox

Legal Disclaimer

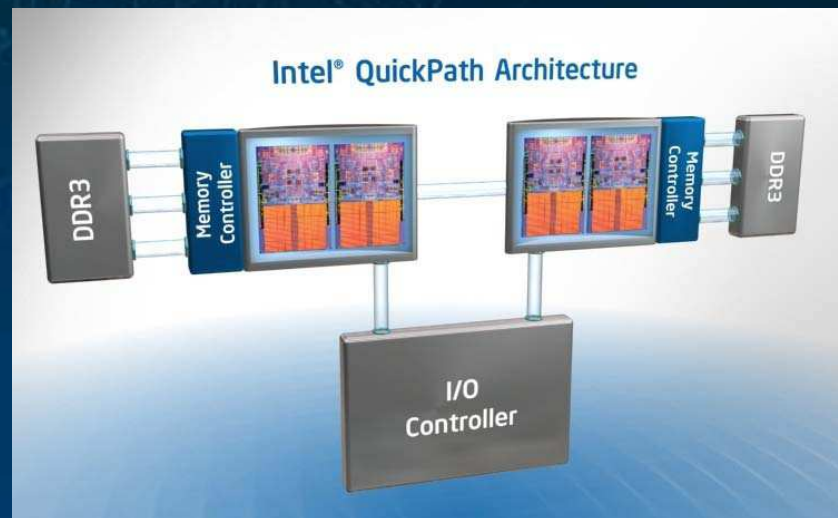
- INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL® PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT. INTEL PRODUCTS ARE NOT INTENDED FOR USE IN MEDICAL, LIFE SAVING, OR LIFE SUSTAINING APPLICATIONS.
- Intel may make changes to specifications and product descriptions at any time, without notice.
- All products, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.
- Intel, processors, chipsets, and desktop boards may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.
- Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.
- Intel, Intel Inside, and the Intel logo are trademarks of Intel Corporation in the United States and other countries.
- *Other names and brands may be claimed as the property of others.

Intel® QuickPath Interconnect

High Performance / Low pin count system interface

New System Choices

- Efficient scaling
- Number of processors
- Platform Topologies

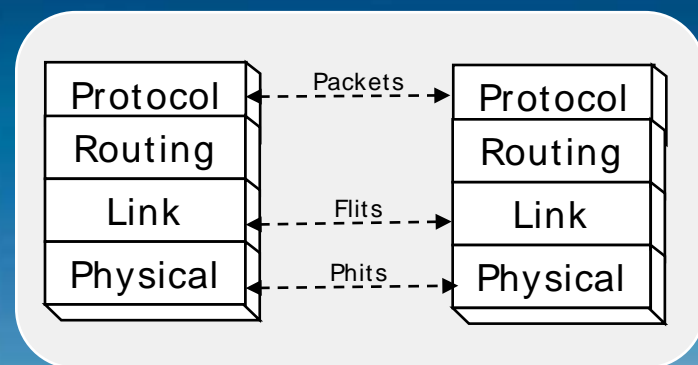


Improved System Robustness

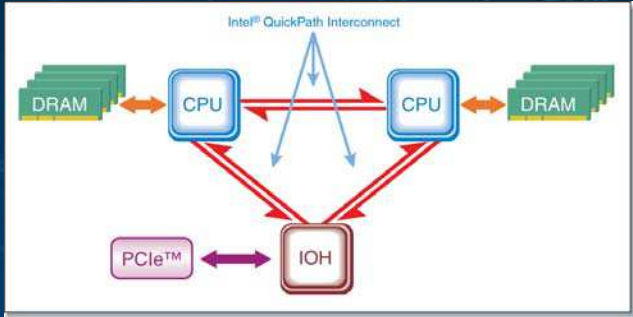
- Error Detection & Recovery

Layered Architecture

- Modular, Flexible
- Applied to Xeon® and Itanium® processor-based systems

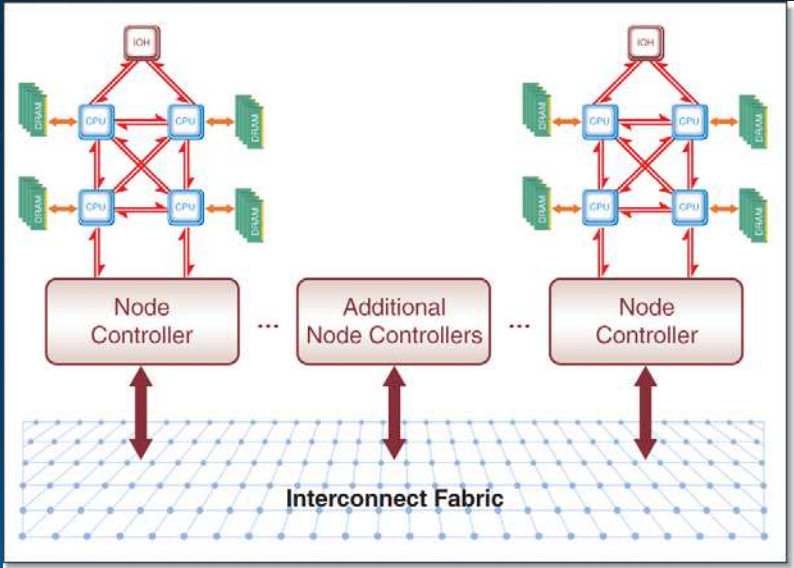


Common Platform Configurations

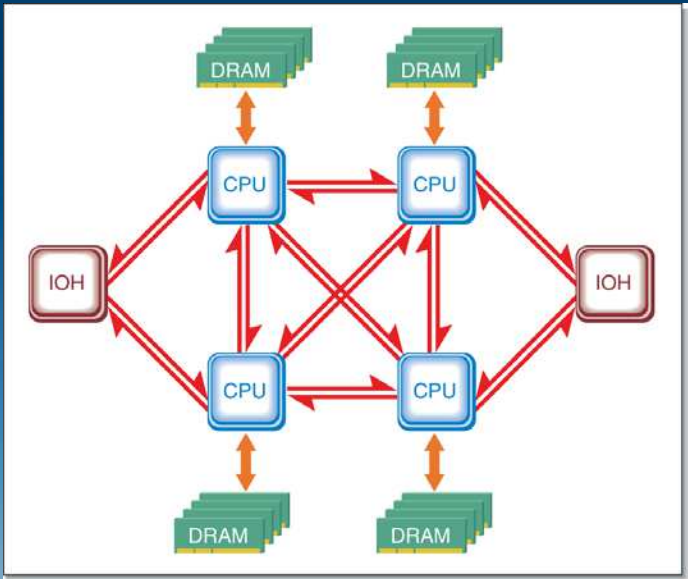


1 and 2 Socket Server Topologies

Node Controller based Topologies

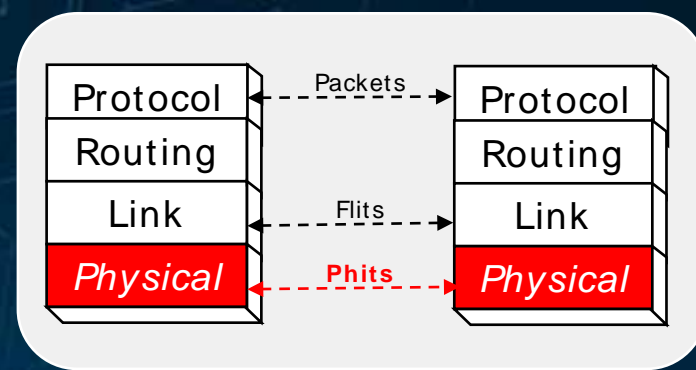


4 and 8 Socket Server Fully (and Partially) Connected Topologies

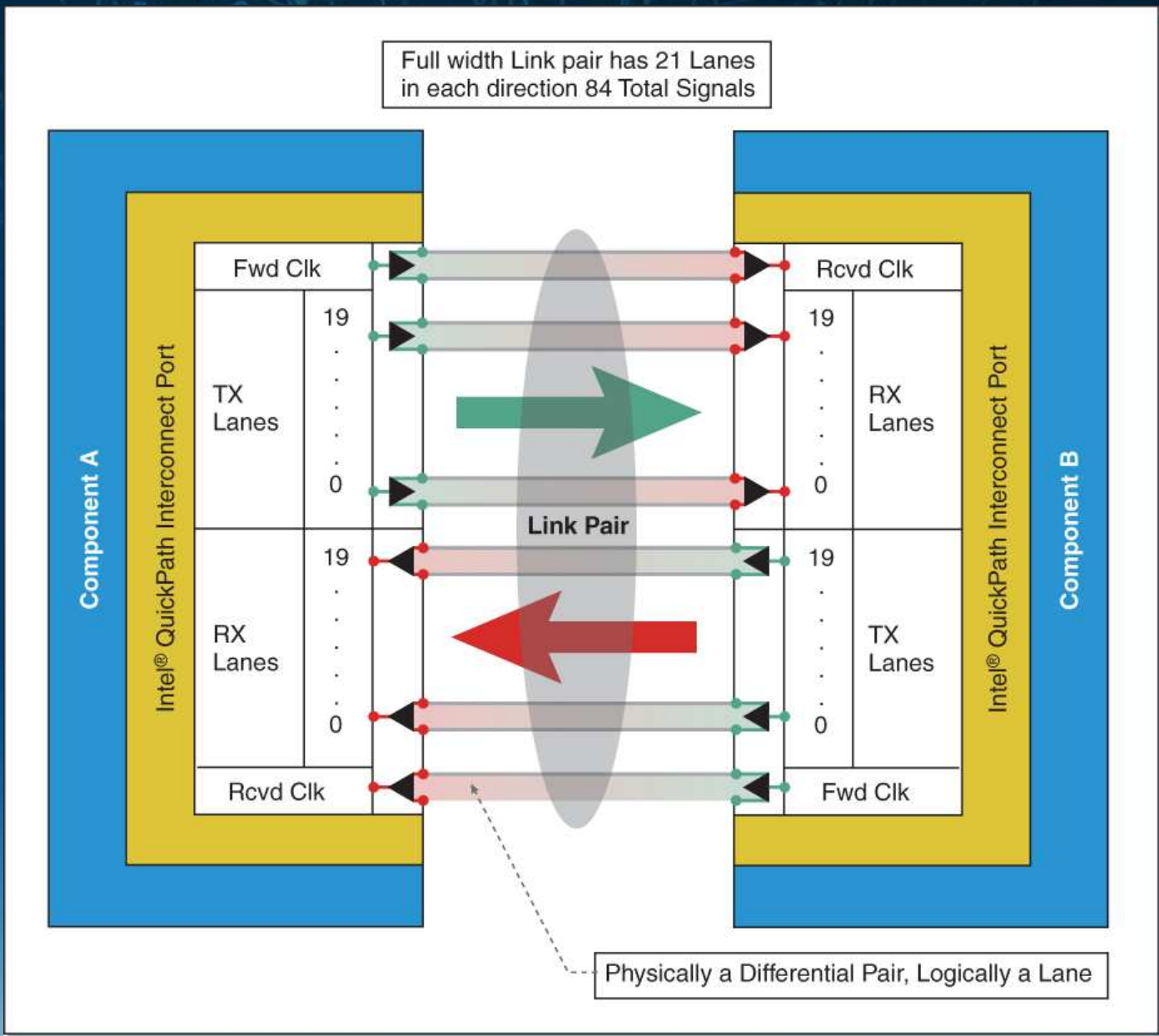


Physical Layer

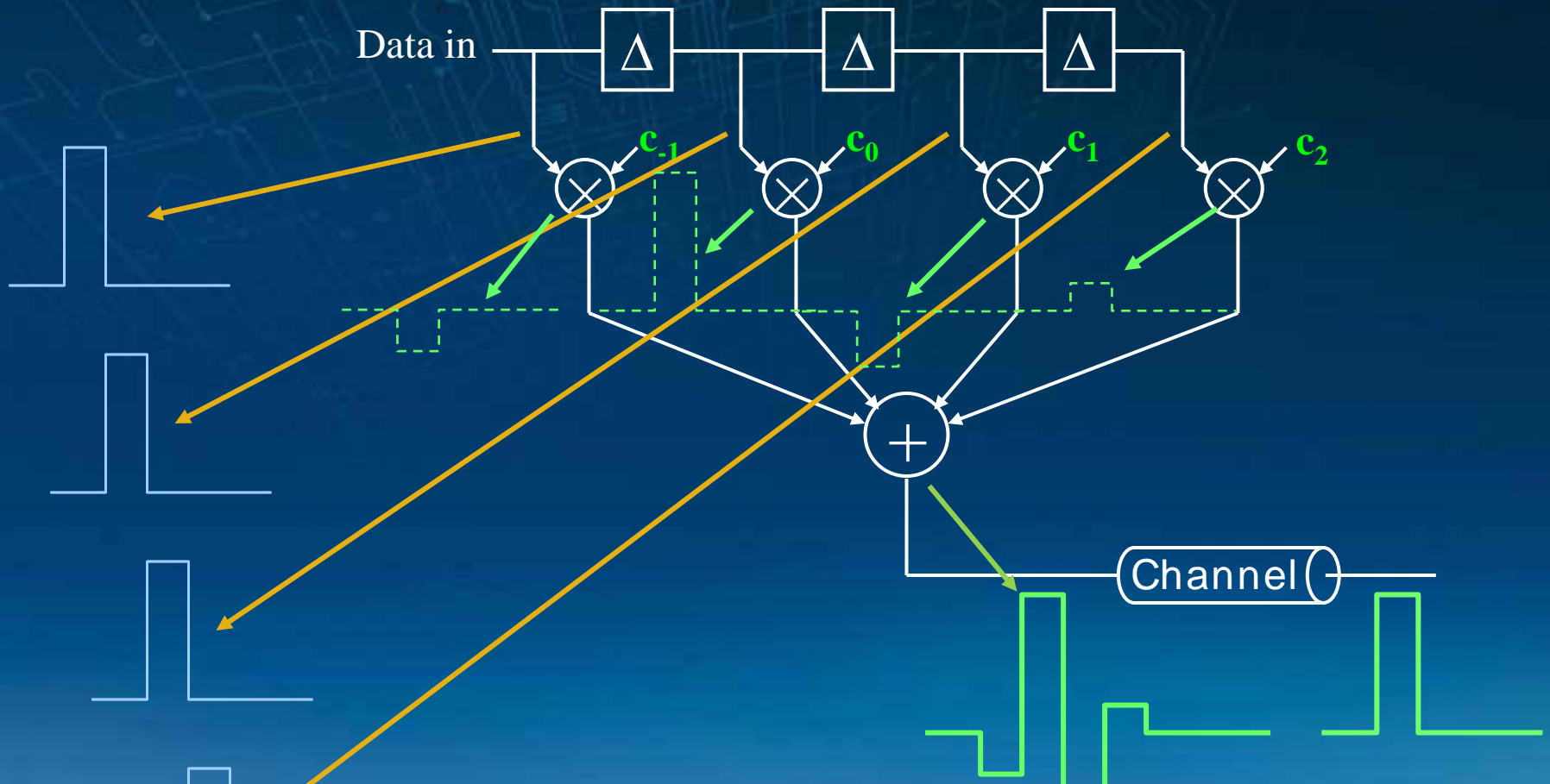
- Intel® QPI Link-Pair
 - Two sets of Unidirectional links
 - Transmitter provides a Forwarded Clock
 - Full width Link-Pair is 84 signals
- Link Widths - 20 or 10 or 5 lanes
- Data Rate: up to 6.4 GT/s
 - Full Width Link is 12.8GB/s (or 25.6GB/s for a Link-Pair)
 - Link BW Calculation is based on data payload only
- Other Physical Layer features
 - Polarity Inversion and Lane Reversal
 - Transmitter Equalization
 - Probe-less Testing



Logical View of a Link Pair



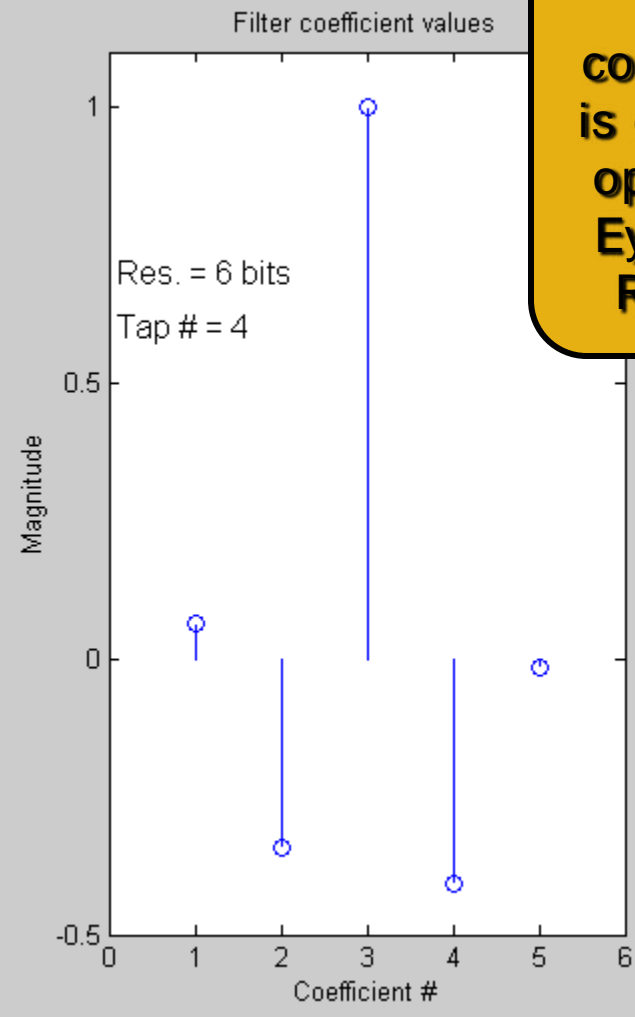
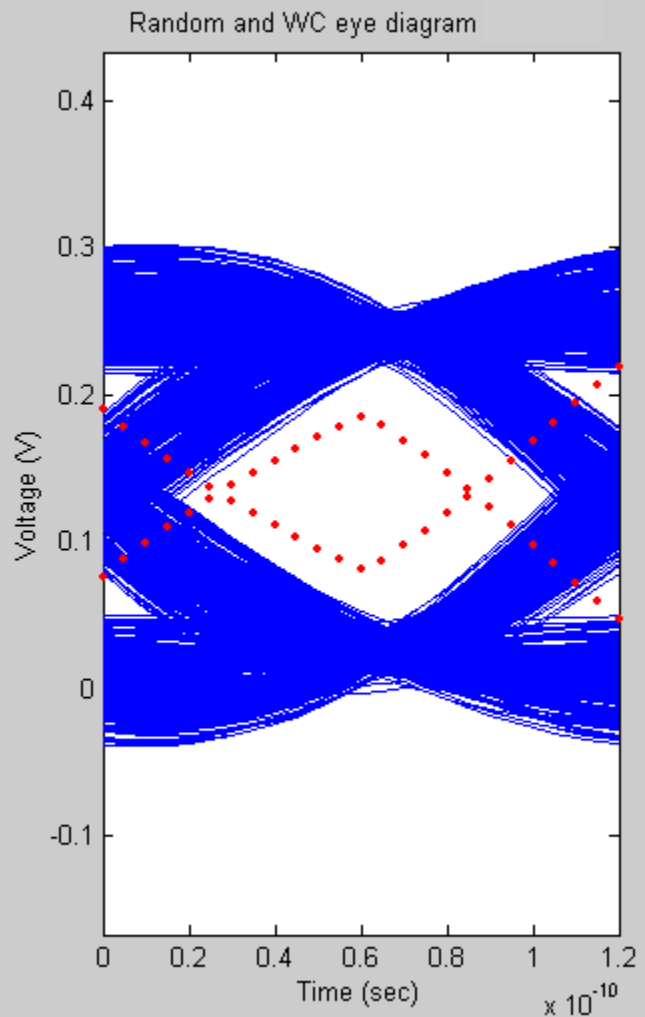
Transmitter Discrete-time Linear Equalizer



QPI Tx circuit reshapes waveform to match channel characteristics

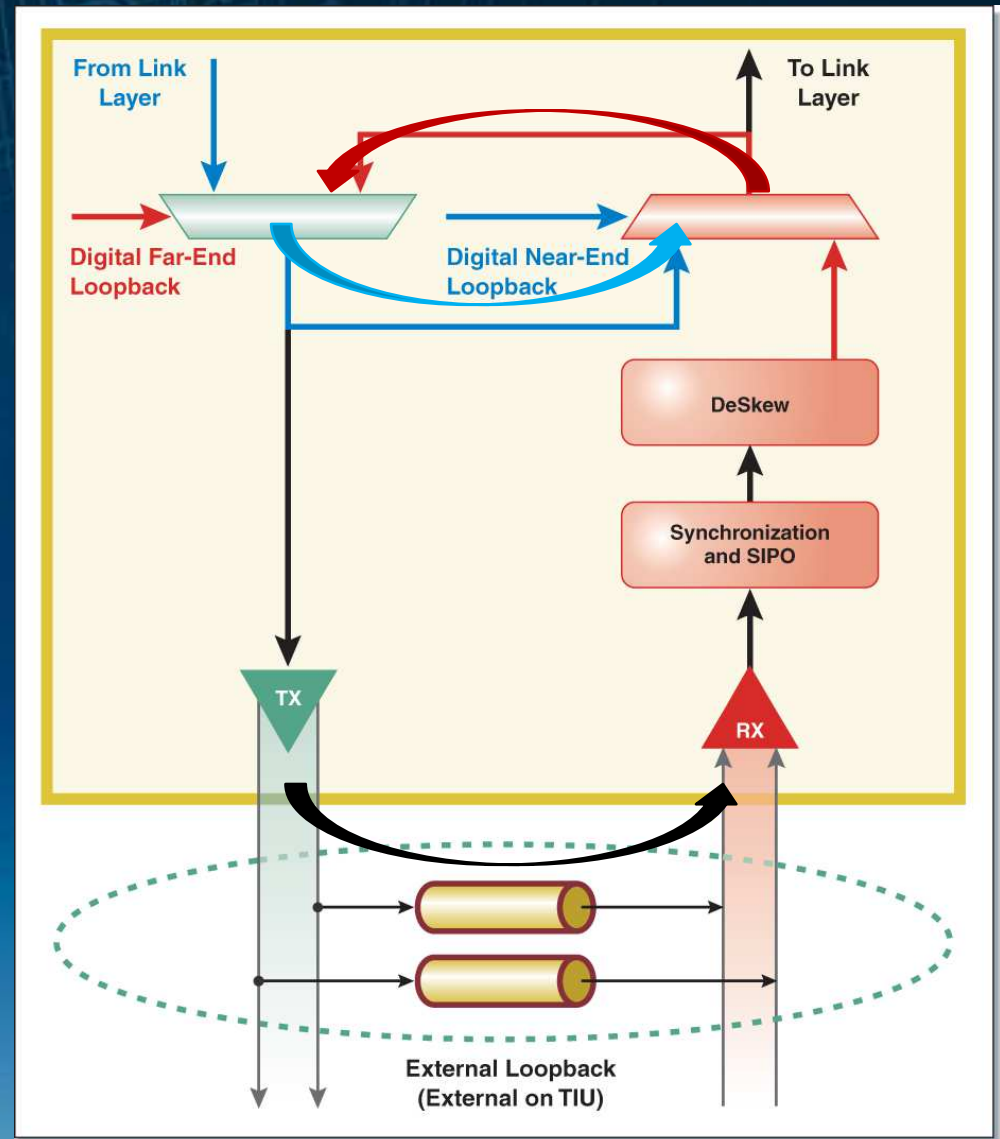
Coefficient Search

Proper selection of Tx coefficients is critical to open Data Eye at the Receiver



Support for Probe-Less Testing

- At 6.4GT/s physical probing of a link is no longer an option
- Physical layer provides additional diagnostic hooks to support several variants of loopback testing
 - Digital Near-End Loopback
 - Local Inter/Intra Link Loopback
 - Remote Loopback



Link Layer

Flow Control

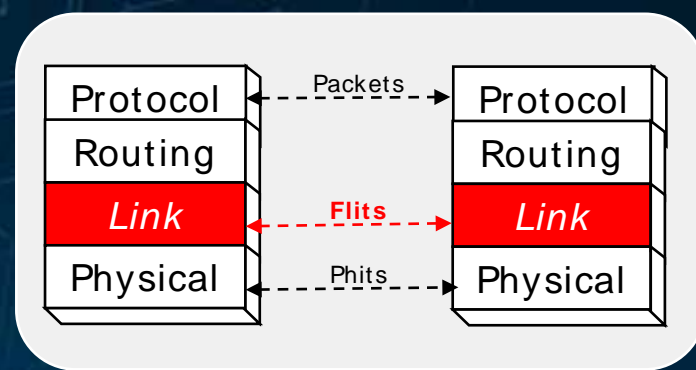
- Credit/debit scheme

Higher layer services

- Arbitrate among the multiple message classes (HOM / NCS / NCB / NDR / DRS / SNP / Special flits - link layer)
- Manage the multiple virtual networks
- Prevent protocol deadlocks
- Ensure the Link BW is realizable

Error Checking and Recovery

- CRC checking
- Link level retry recovery
- Link Self-healing for hard errors



Link Layer Quantum of Information

Physical Layer deals with Phits (20, 10 or 5 bits)

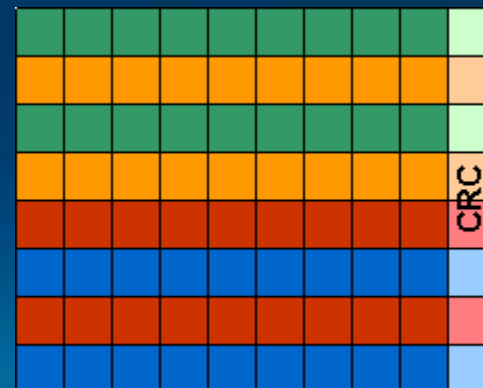
Protocol Layer deals with Messages (or Packets)

Link Layer works at a *Flit* granularity

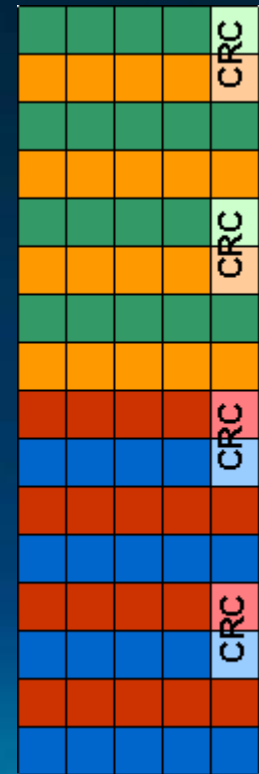
Always 72 Bits of Payload + 8 Bits CRC



Full
Width
Link



Half
Width
Link



Quarter
Width
Link



Virtual Channels Mapping Function

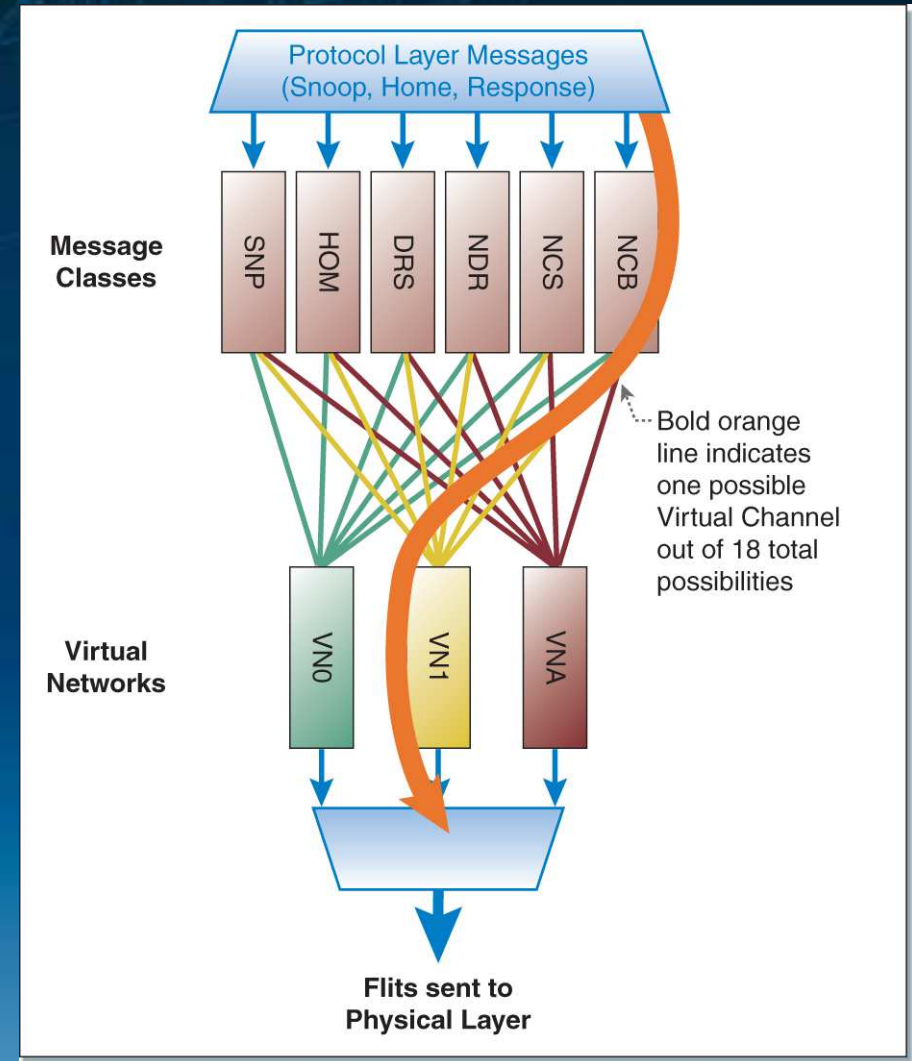
Link Layer maps Messages to VN's

- 6 Non-blocking message classes
- 3 Virtual networks defined (VN0, VN1, VNA)
- 18 Virtual channel combinations

13 Credit Pools

- 6 for message classes on VN0 (Packets)
- 6 for message classes on VN1 (Packets)
- 1 for everything on VNA (Flits)

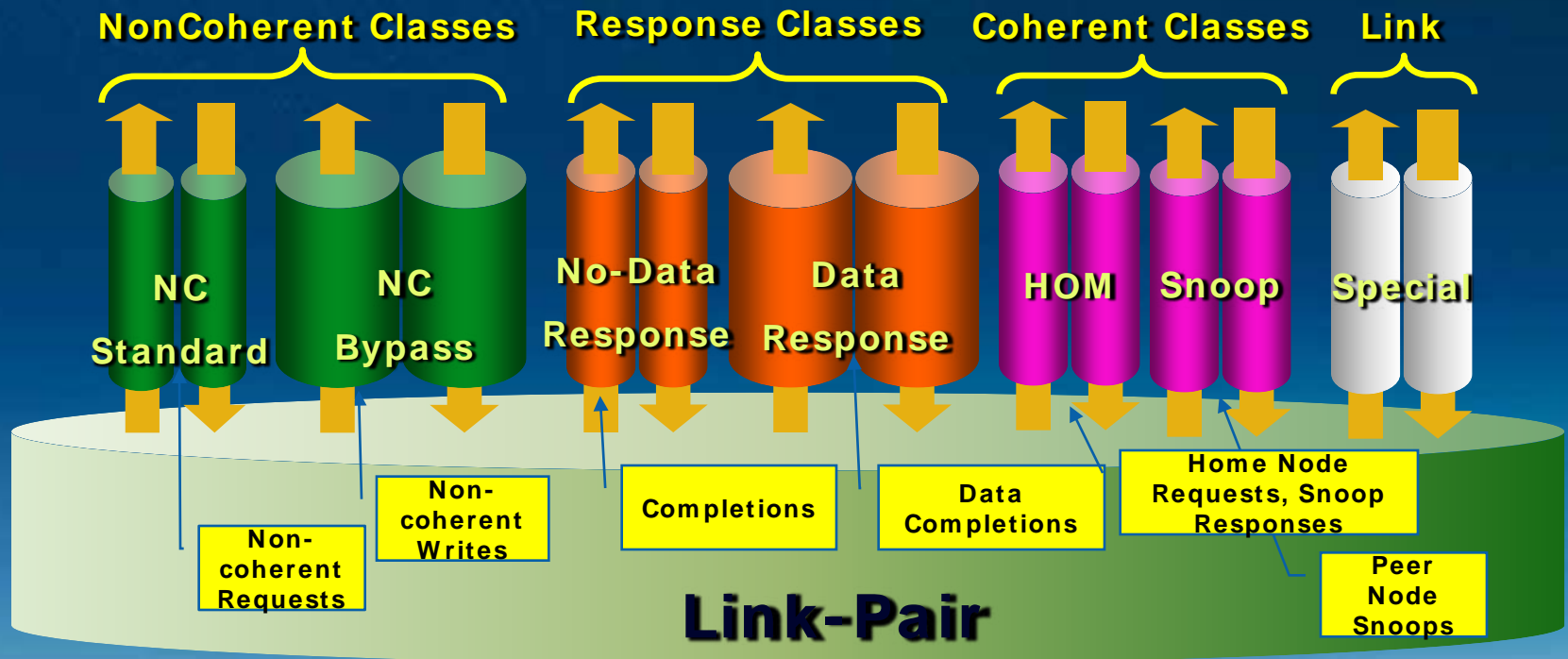
Note: Message Class and VN information are encoded in the fields of the header flit



Message Classes

Protocol events are grouped into *message classes* to prevent undesirable dependencies, i.e., situations where dependencies create deadlock or livelock scenarios

7 Total Message Classes with 6 for the Protocol Layer



Interleaving of Messages

Referred to as Command Insert Interleave (or CII)

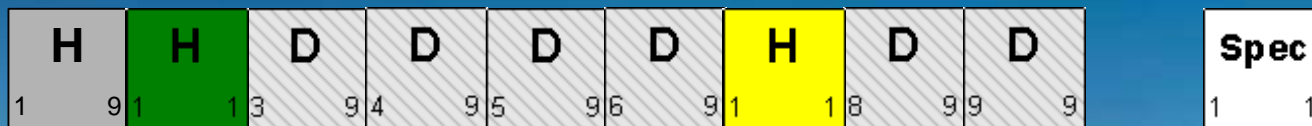
Allows for link layer to insert single (& dual) flit messages into a multi-flit messages

Benefits of CII

Eliminates any reason to Store&Forward packets in the route through case.

Eliminates issues that could potentially inject bubbles into the link

Minimizes lead off latency without sacrificing link utilization



CRC Protection Properties

Per Flit Calculation (not packet)

- No Additional latency incurred collecting the rest of the packet
- Protocol Flits can be consumed immediately

CRC8 Properties of a Flit

- All 1, 2, and 3 bit errors are detected
- Any odd number of bit errors is detected
- All bit errors of “burst length” 8 or less are detected
- 99 percent of all errors of burst length 9 are detected
- 99.6 percent of all errors of burst length greater than 9 are detected

Additional CRC Protection Option

- Rolling CRC across two flits
- Incur an additional flit delay in latency

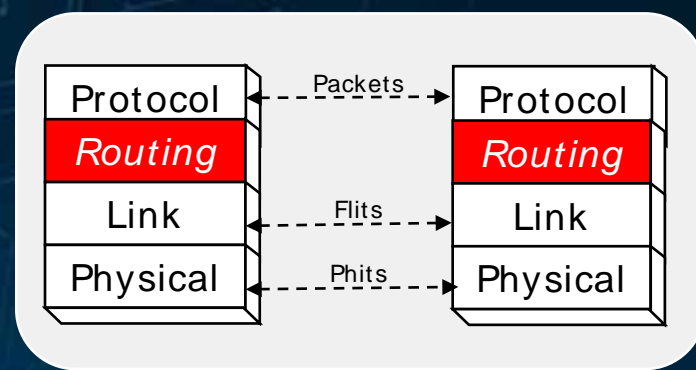
Routing Layer

Routing algorithms

- Specified through Routing Tables
- At source – based on System Address
- At intermediate points - based on destination Node ID

Programming done by firmware provides a flexible and distributed method to route Intel® QuickPath Interconnect transactions from source to destination

Programmable Routing Table is key to supporting higher level system features – OL*, Dynamic Partitioning, etc.



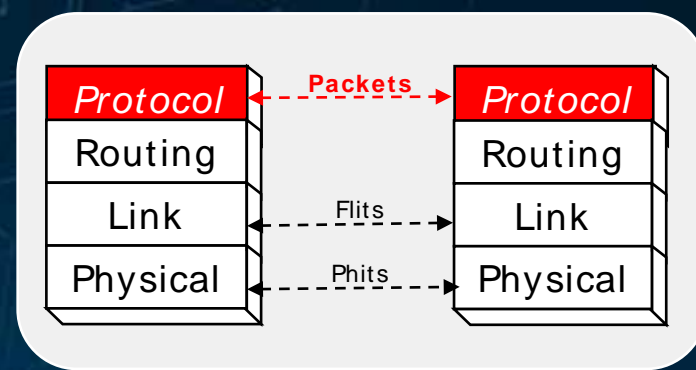
Protocol Layer

Categories of Transactions

- Coherent / Non-Coherent
- Virtual Legacy Wires
- Power Management

Flexible cache coherency protocol options

- Source Snoop option mechanism
 - Caching Agent issues snoops (or spawned by system topology) to other CA's and snoop responses to the Home Agent
 - Home Agent collects snoop responses
- Home Snoop option mechanism
 - Caching Agent issues request to Home
 - Home Agent multicasts snoop, home node collects snoop responses (directory based protocol)
- MESIF cache states (*Modified, Exclusive, Shared, Invalid, Forwarding*)
- HOM message class per source/destination pair is ordered for same addresses
 - Coherent requests and snoop responses are on HOM
- Pre-allocated resources at home agent to prevent deadlock / forward progress issues
 - No Retries / Built in conflict resolution



Agent Types

Configuration Agent (self explanatory)

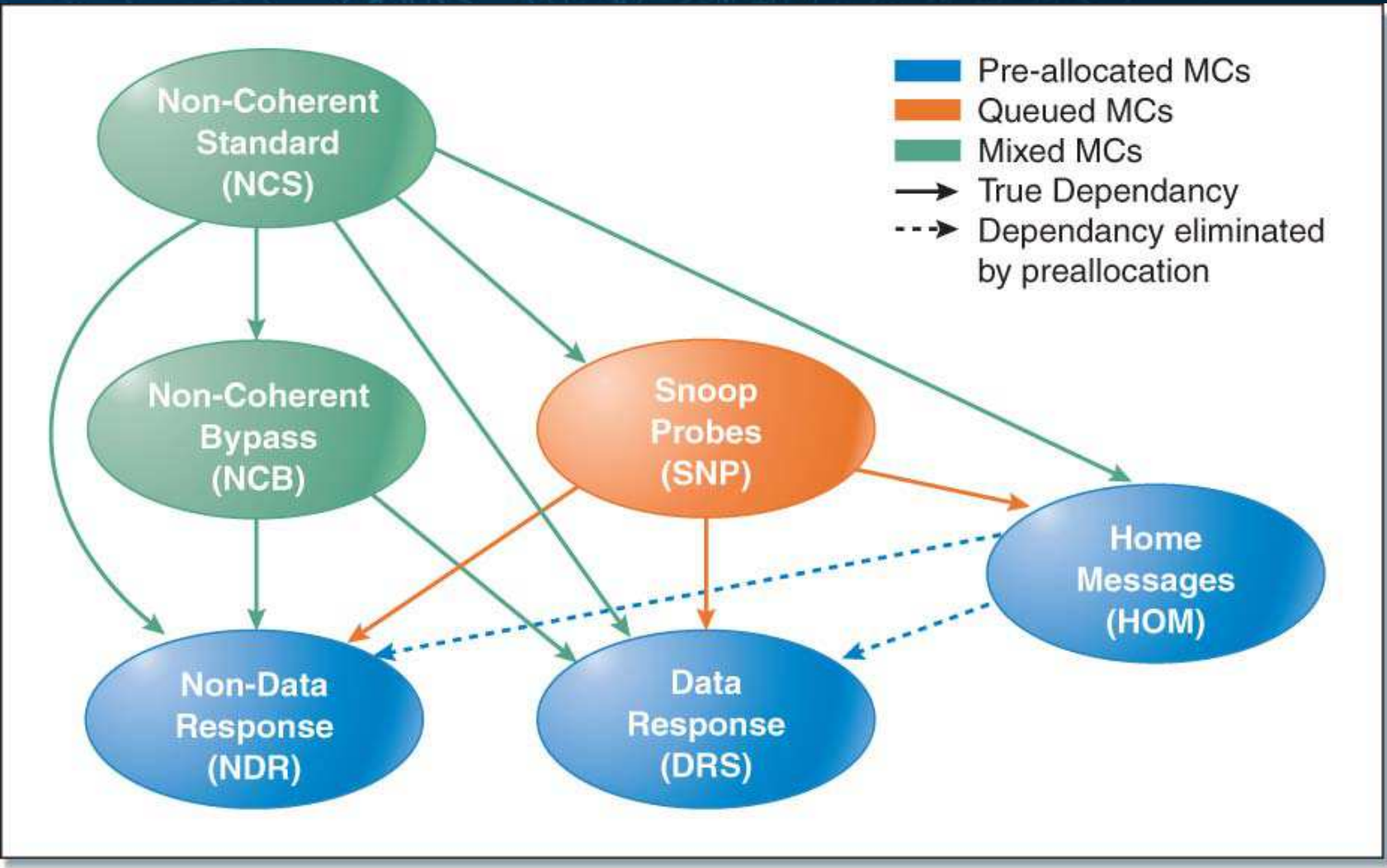
Caching Agent

- Issues request to home after cache miss
- Supplies data on snoop hits or forward requests from Home
- Detects and acknowledges conflicts to Home
- Responds with acknowledgement of conflict if snooped on an outstanding request or forced by a Home

Home Agent (coherency/order controller)

- Front end of a memory controller
- Tracks every potential outstanding request
 - Recipient of caching agents' requests
 - Recipient of all snoop responses
- Decides the next owner in address conflict cases
 - Caching agents report their observed conflicts to Home
 - Forces acknowledgement response to ambiguous conflict cases

Message Class Dependency Diagram



Convention for Flow Diagrams

A B C

Caching Agents

H

Home Agent

MC

Memory Controller



Message traveling along ordered HOM Message Class



Message traveling along unordered snoop or response channel



Allocate requester entry or home agent tracking entry

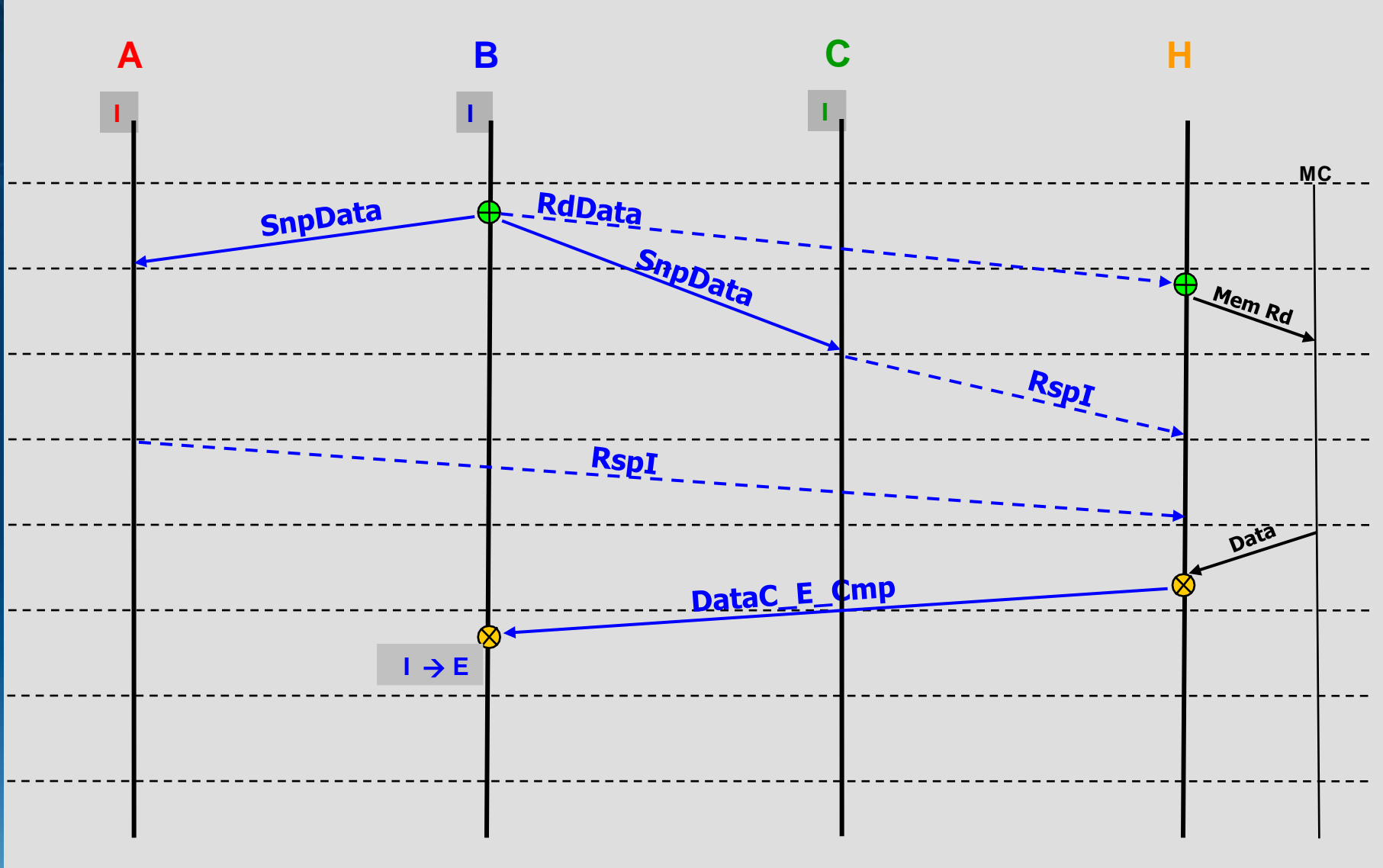


De-allocate requester entry or home agent tracking entry

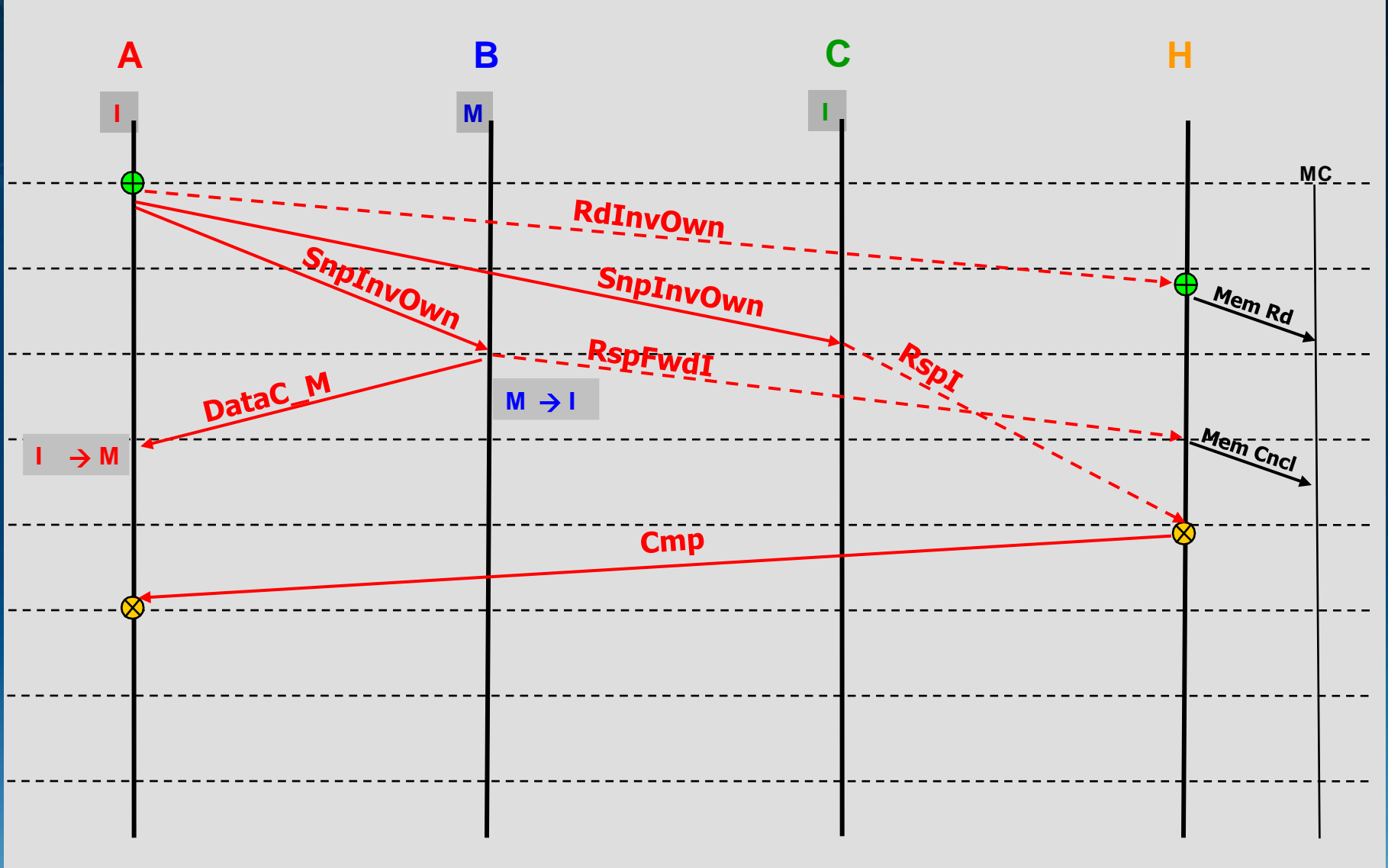


Time progresses in a down direction of the Flow Diagram

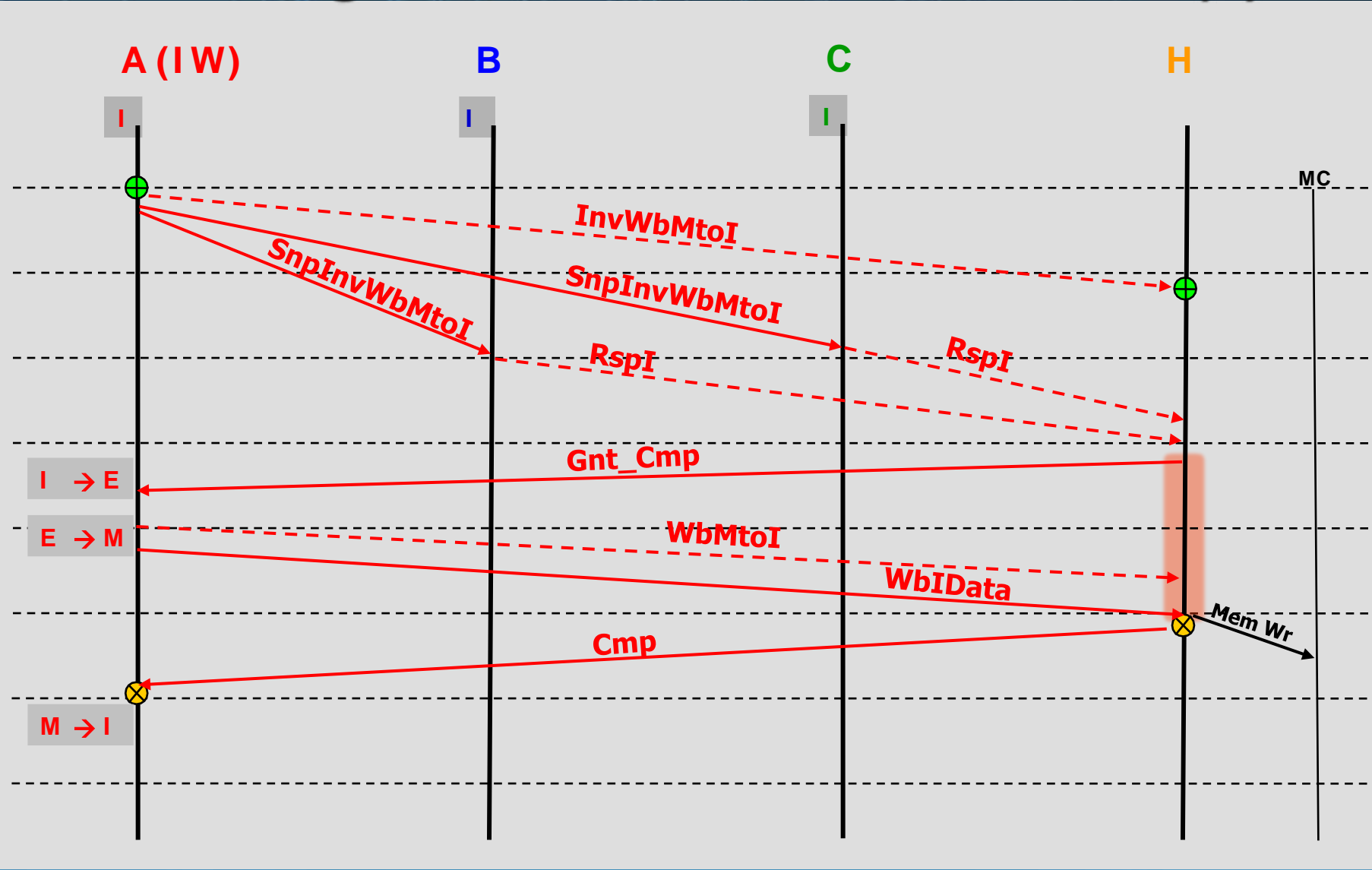
RdData request after Cache Miss



RdInvOwn with C2C Transfer



Invalidating Write with no Cached Copy



Where is this Intel® QPI Headed?

- Intel® Server Interconnect Strategy for years to come
- Intel® QPI is more than a link definition – it is an infrastructure for
 - Legacy Support for pre-existing software
 - Efficient Processing market segments features
 - Low Latency / High BW Topology
 - Invalidating Write Flow / Snoop Spawning
 - System Power Management (via PMReq Message)

Reliability/Availability/Serviceability features for the High-end and Mission Critical market segments, such as

Clock Fail-safe / Link Self-healing / Rolling CRC

Ability to re-route around a failed link

Static / Dynamic Partitioning and OL* usage of the Quiesce Message / Programmable Route Tables

Inter-socket memory mirroring / DIMM sparing / etc

For More Information

Web-based info:

- <http://www.intel.com/technology/quickpath/index.htm>
- <http://www.intel.com/technology/quickpath/whitepaper.pdf>

Book published by Intel Press:



Weaving High Performance Multiprocessor Fabric Architectural Insights to the Intel® QuickPath Interconnect

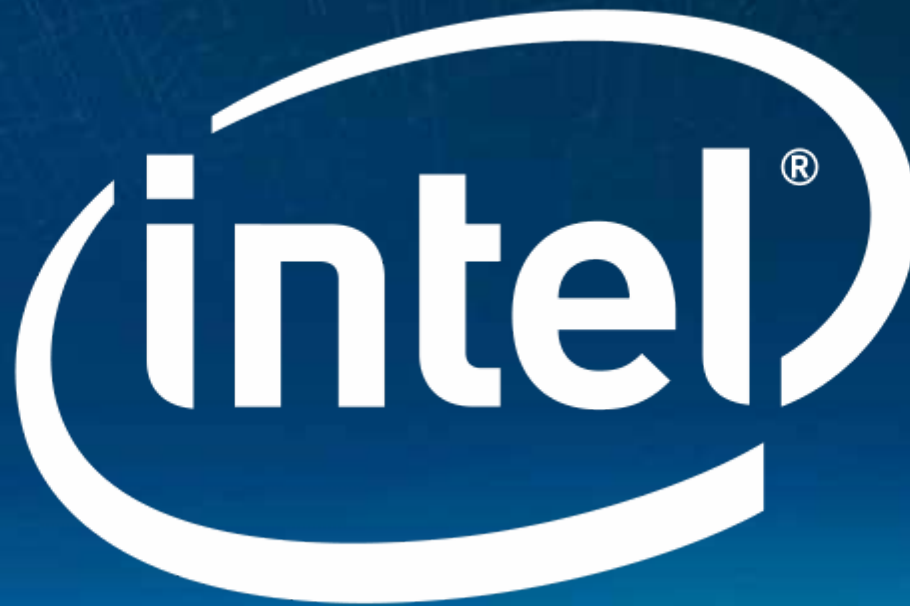
By Robert A. Maddox, Gurbir Singh and Robert J. Safranek

http://www.intel.com/intelpress/sum_qpi.htm


Training from MindShare*

- <http://www.mindshare.com>
- Contact Ravi Budruk ravi@mindshare.com





Intel® Xeon® 5500 Performance Publications

<p>SPECint*_rate_base2006</p> <p>241 score (+72%) </p> <p>IBM J9* JVM</p>	<p>SPECpower*_ssj2008 </p> <p>1977 ssj_ops/watt (+74%)</p> <p>IBM J9* JVM</p>	<p>SPECfp*_rate_base2006</p> <p>197 score (+128%) </p>
<p>SPECjAppServer*2004</p> <p>3,975 JOPS (+93%) </p> <p>Oracle WebLogic* Server</p>	<p>TPC*-C</p> <p>631,766 tpmC (+130%) </p> <p>Oracle 11g* database</p>	<p>SAP-SD* 2-Tier</p> <p>5,100 SD Users (+103%) </p> <p>SAP* ERP 6.0/IBM DB2*</p>
<p>VMmark* </p> <p>24.35 @17 tiles (+166%)</p> <p>VMware* ESX 4.0</p>	<p>TPC*-E</p> <p>800 tpsE (+152%) </p> <p>Microsoft SQL Server* 2008</p>	<p>SPECWeb*2005 </p> <p>75023 score (+150%)</p> <p>Rock Web* Server</p>
<p>Fluent* 12.0 benchmark</p> <p>Geo mean of 6 (+127%) </p> <p>ANSYS FLUENT*</p>	<p>SPECjbb*2005</p> <p>604,417 BOPS (+64%) </p> <p>IBM J9* JVM</p>	<p>SPECapc* for Maya 6.5</p> <p>7.70 score (+87%) </p> <p>Autodesk* Maya</p>

Over 30 New 2S Server and Workstation World Records!

Percentage gains shown are based on comparison to Xeon 5400 series; Performance results based on published/submitted results as of April 27, 2009. Platform configuration details are available at <http://www.intel.com/performance/server/xeon/summary.htm>. *Other names and brands may be claimed as the property of others

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, visit [Intel Performance Benchmark Limitations](#)

