Blade Computing with the

AMD Opteron™ Processor (“Magny-Cours”)

Pat Conway (Presenter)
Nathan Kalyanasundharam
Gregg Donley
Kevin Lepak
Bill Hughes

AMD
The future is fusion

Agenda

Processor Architecture

- AMD driving the x86 64-bit processor evolution
- Driving forces behind the Twelve-Core AMD Opteron™ processor codenamed “Magny-Cours”
- CPU silicon
- MCM 2.0 package, speeds and feeds

Performance and scalability

- 2P/4P blade and rack topologies
- HyperTransport™ technology HT Assist design
  - Cache coherence protocol
  - Transaction scenarios and frequencies
  - Coverage ratio
  - Memory latency and bandwidth

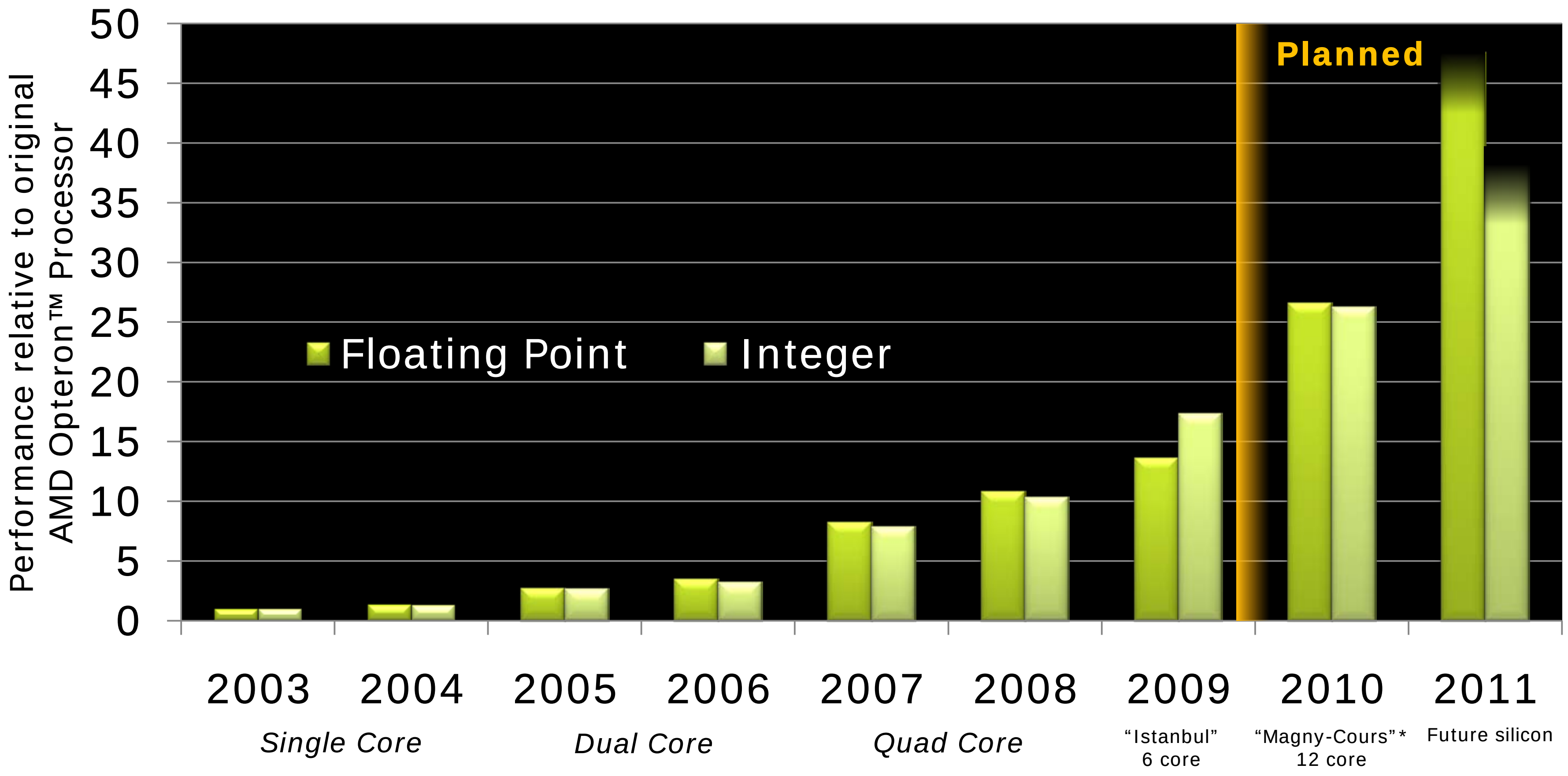
A look ahead

x86 64-bit Architecture Evolution

	2003	2005	2007	2008	2009	2010
	AMD Opteron™	**AMD Opteron™**	**“Barcelona”**	**“Shanghai”**	**“Istanbul”**	**“Magny-Cours”**
Mfg. Process	90nm SOI	90nm SOI	65nm SOI	45nm SOI	45nm SOI	45nm SOI
CPU Core	K8 ⬜	K8 🟩🟩	Greyhound 🟥🟥🟥🟥	Greyhound+ 🟦🟦🟦🟦	Greyhound+ 🟥🟥🟥🟥🟥🟥	Greyhound+ 🟨🟨🟨🟨🟨🟨 🟨🟨🟨🟨🟨🟨
L2/L3	1MB/0	1MB/0	512kB/2MB	512kB/6MB	512kB/6MB	512kB/12MB
Hyper Transport™ Technology	3x 1.6GT/.s	3x 1.6GT/.s	3x 2GT/s	3x 4.0GT/s	3x 4.8GT/s	4x 6.4GT/s
Memory	2x DDR1 300	2x DDR1 400	2x DDR2 667	2x DDR2 800	2x DDR2 1066	4x DDR3 1333

Max Power Budget Remains Consistent

Dramatic Back-to-back Gains



"Shanghai" to "Istanbul" delivers 34%
more performance in the same power envelope

**"Magny-Cours" and Future silicon data is based on AMD projections*

Driving Forces Behind “Magny-Cours”

Server Throughput

- Exploit thread level parallelism
- Leverage Directly Connected MCM 2.0

Virtualization

- Maximize compute density in 2P/4P blades and racks
- Run more VMs per server
- Provide hardware context (thread) based QOS

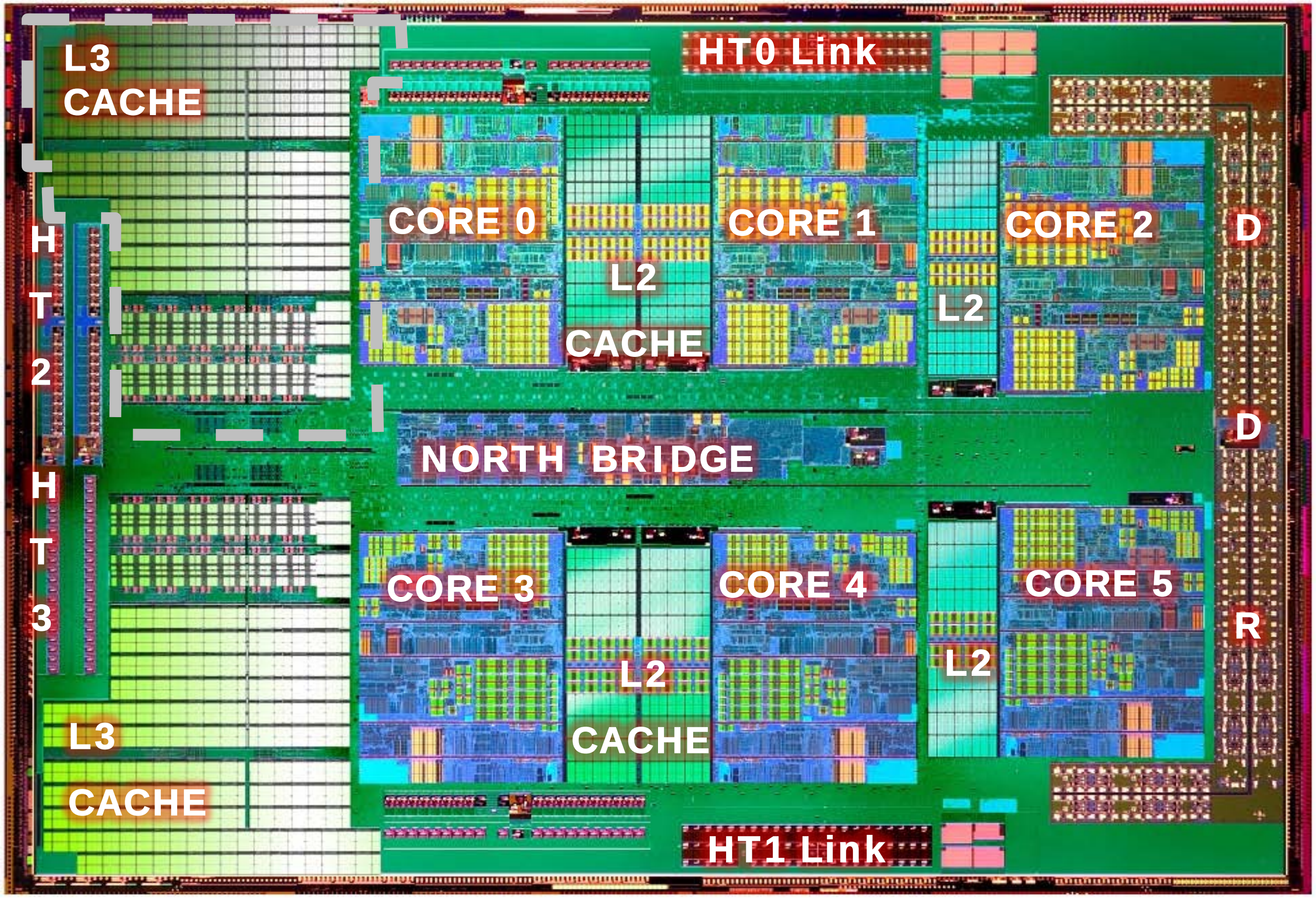
Energy Proportional Computing

- More performance, same power envelope
- Power conservation when idle

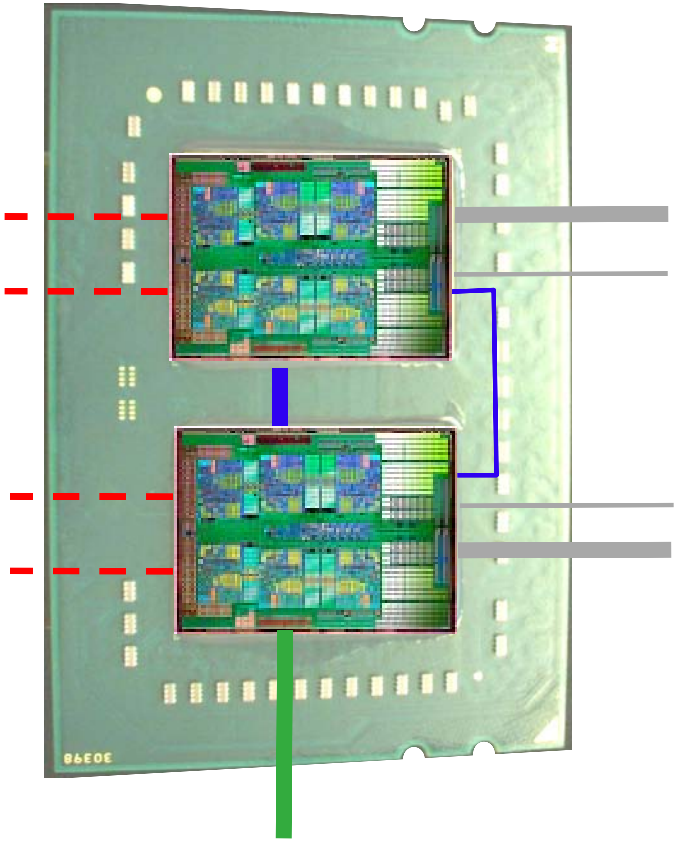
Economics

- Design efficiency – “Magny-Cours” silicon same as “Istanbul”
  - *Can help speed qualification times and customers’ time to market*
- Reasonable die size permits 2 die per reticle (Yield ⇑ Manufacturing Cost ⇓)
  - *Yield improvements can help ensure supply chain stability*
  - *Manufacturing cost savings ultimately benefit customers*

"Magny-Cours" Silicon



MCM 2.0 Logical View

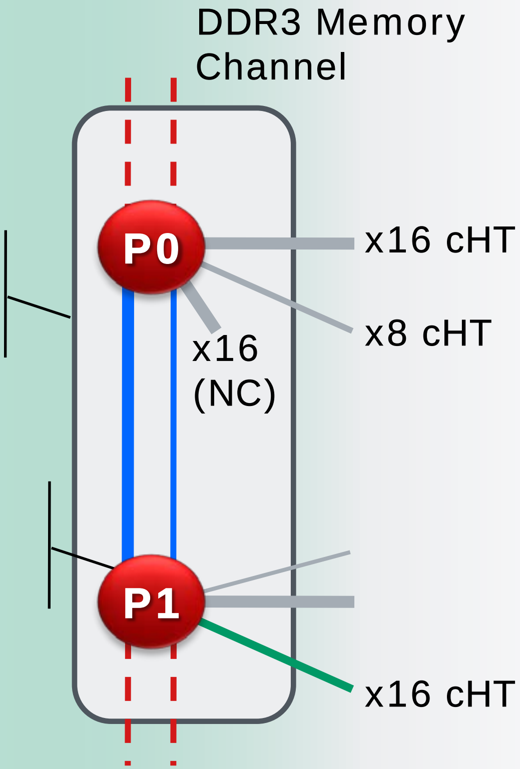


G34 Socket

“Magny-Cours” utilizes a ***Directly Connected*** MCM

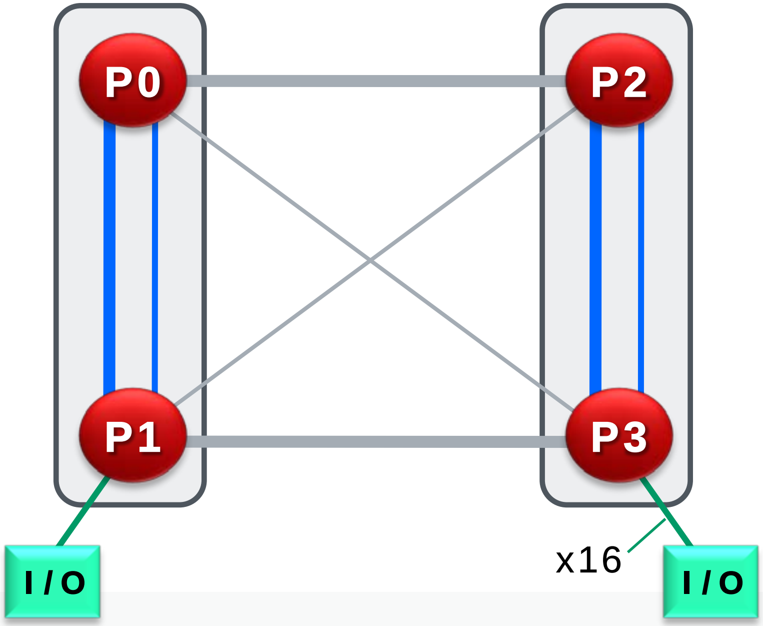
Package has 12 cores, 4 HT ports, & 4 memory channels

Die (Node) has 6 cores, 4 HT ports & 2 memory channels



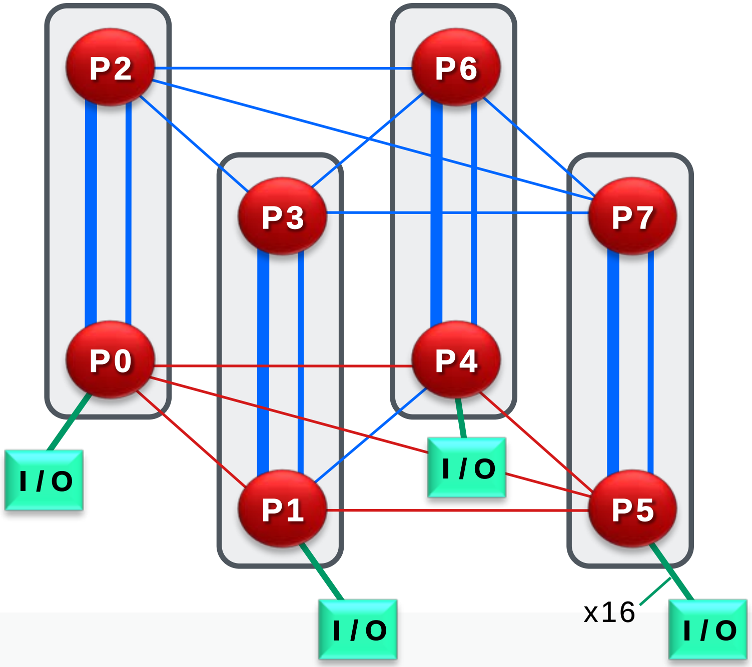
Topologies

2P



Diameter 1
Avg Diam 0.75
DRAM BW 85.6 GB/s
XFIRE BW 71.7 GB/s (*)

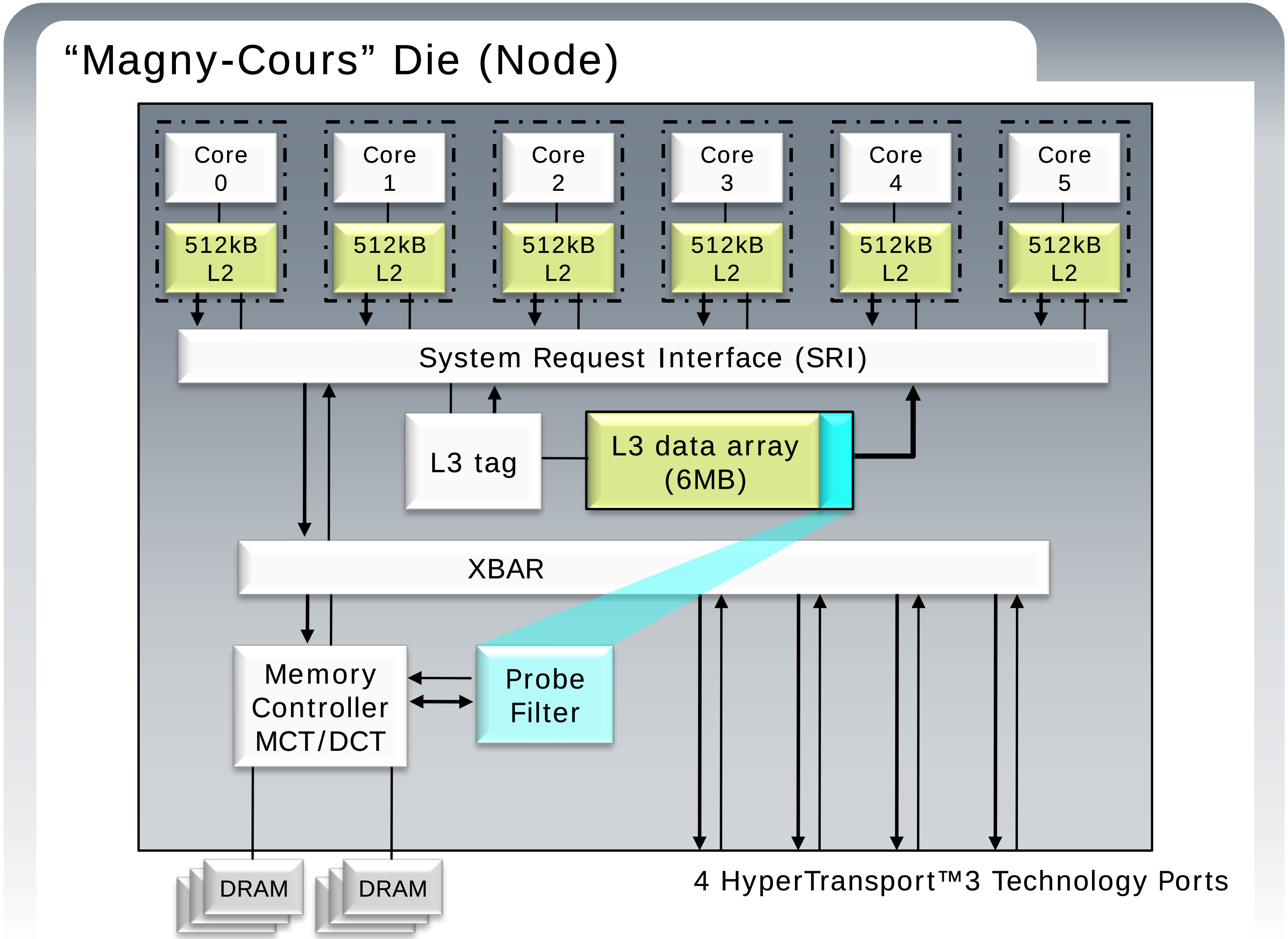
4P



Diameter 2
Avg Diam 1.25
DRAM BW 170.4 GB/s
XFIRE BW 143.4 GB/s

(*) XFIRE BW is the maximum available coherent memory bandwidth if the HT links were the only limiting factor. Each node accesses its own memory and that of every other node in an interleaved fashion.

Block Diagram



HyperTransport™ Technology HT Assist (Probe Filter)

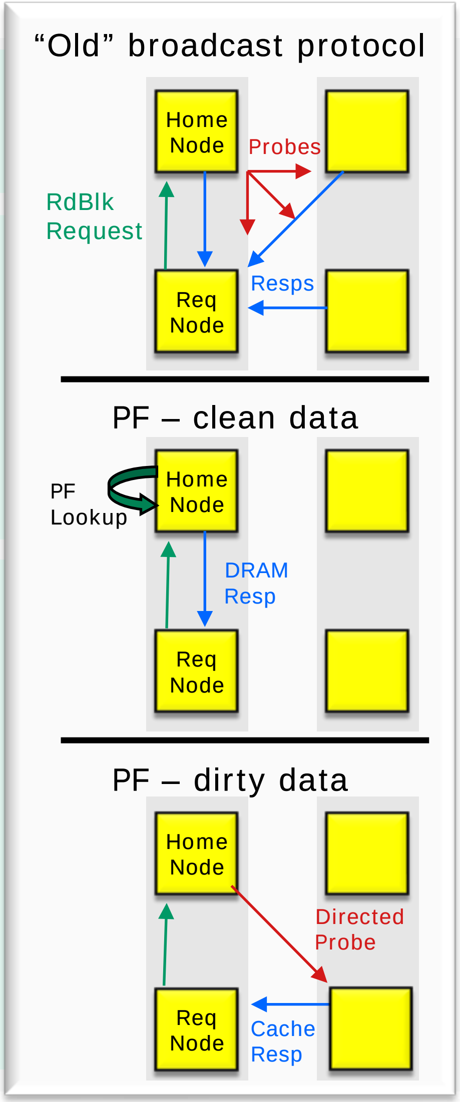
Key enabling technology on "Istanbul" and "Magny-Cours"

HT Assist is a sparse directory cache

- Associated with the memory controller on the home node
- Tracks all lines cached in the system from the home node

Eliminates most probe broadcasts (see diagram)

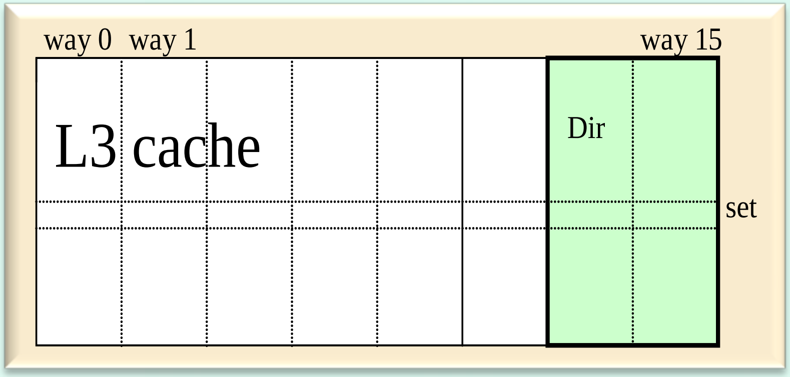
- Lowers latency
  - local accesses get local DRAM latency, no need to wait for probe responses
  - less queuing delay due to lower HT traffic overhead
- Increases system bandwidth by reducing probe traffic



Where Do We Put the HT Assist Probe Filter?

Q: Where do we store probe filter entries without adding a large on-chip probe filter RAM which is not used in a 1P desktop system?

A: Steal 1MB of 6MB L3 cache per die in "Magny-Cours" systems

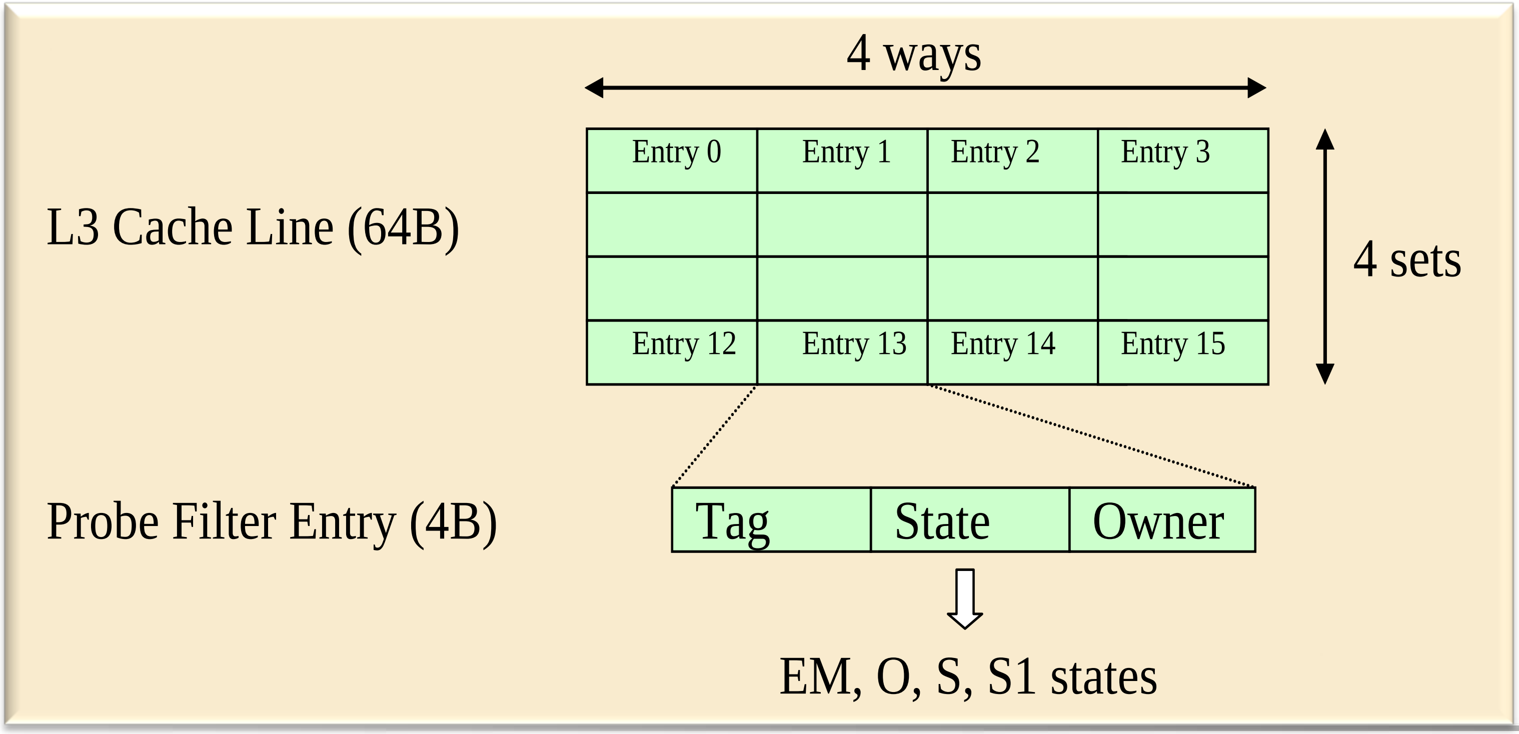


Implementation in fast SRAM (L3) minimizes

- Access latency
- Port occupancy of read-modify-write operations
- Indirection latency for cache-to-cache transfers

Format of a Probe Filter Entry

- 16 probe filter entries per L3 cache line (64B), 4B per entry, 4-way set associative
- 1MB of a 6MB L3 cache per die holds 256k probe filter entries and covers 16MB of cache



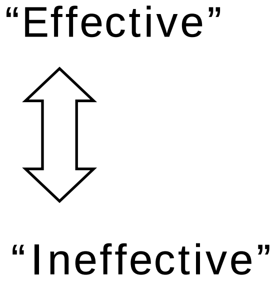
Cache Coherence Protocol

- Track lines in M, E, O or S state in probe filter
- PF is fully inclusive of all cached data in system
  - if a line is cached, then a PF entry must exist.
- Presence of probe filter entry says line in M, E, O or S state
- Absence of probe filter entry says line is uncached
- New messages
  - Directed probe on probe filter hit
  - Replacement notification E ->I (clean VicBlk)

Probe Filter Transaction Scenarios

	PF Hit					PF Miss (*)				
	I	O	S	S1	EM	I	O	S	S1	EM
FETCH	-	D	-	-	D	-	B	B	DI	DI
LOAD	-	D	-	-	D	-	B	B	DI	DI
STORE	-	B	B	B	DI	-	B	B	DI	DI

Legend	-	Filtered
	D	Directed
	DI	Directed Invalidate
	B	Broadcast Invalidate

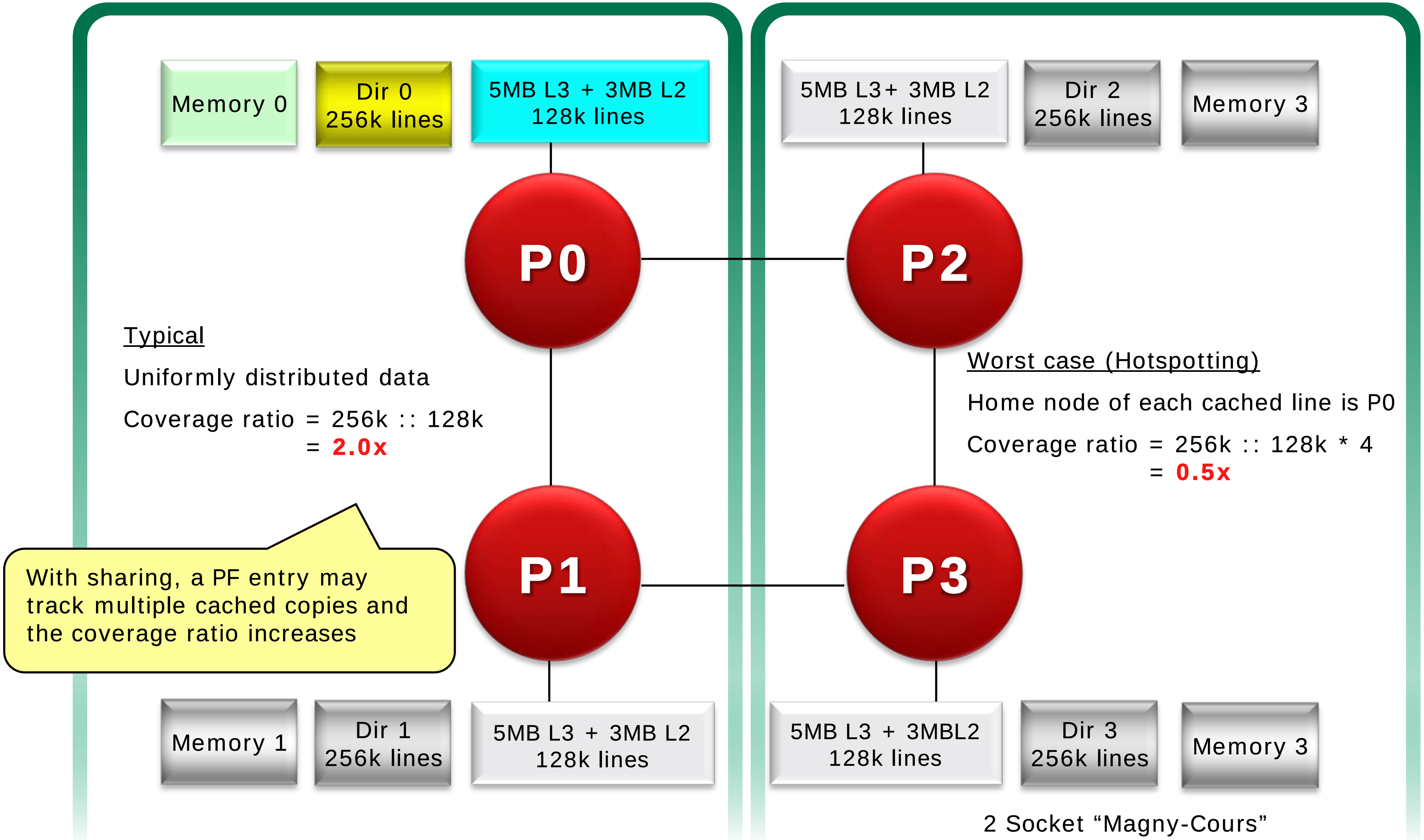

“Effective”

“Ineffective”

(*) PF miss implies line is Uncached (no broadcast necessary). State refers to the state of the line to be replaced upon allocation of new PF entry.

Traditional “Cache Hit Ratio” does not measure effectiveness of probe filter

Probe Filter Coverage Ratio



Typical

Uniformly distributed data

Coverage ratio = 256k :: 128k
= **2.0x**

Worst case (Hotspotting)

Home node of each cached line is P0

Coverage ratio = 256k :: 128k * 4
= **0.5x**

With sharing, a PF entry may track multiple cached copies and the coverage ratio increases

2 Socket “Magny-Cours”

HT Assist and Memory Latency

With “old” broadcast coherence protocol, the latency of a memory access is the longer of 2 paths:

- time it takes to return data from DRAM and
- the time it takes to probe all caches

With HT Assist, local memory latency is significantly reduced as it is not necessary to probe caches on other nodes.

Several server workloads naturally have ~100% local accesses

- SPECint®, SPECfp®
- VMMARK^{TM} typically run with 1 VM per core
- SPECpower_ssj® with 1 JVM per core
- STREAM

Probe Filter amplifies benefit of any NUMA optimizations in OS/application which make memory accesses local

SPEC, SPECint, SPECfp, and SPECpower_ssj are trademarks or registered trademarks of the Standard Performance Evaluation Corporation.

A Look Ahead

Socket compatible upgrade to “Magny-Cours” is planned with

- More cores for additional thread-level paralleism
- More cache to maintain cache-per-core balance
- Same power envelope
- Finer grain power management

New processor core (“Bulldozer”)

- Planned brand new x86 64-bit microarchitecture
- 32nm design
- Instruction set extensions
- Higher memory level parallelism

Thank you!

Disclaimer & Attribution

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION