

# AMD Opteron™ Multicore Processors

Brian Waldecker | February 1, 2009

Senior Member of Technical Staff

System Optimization Engineering Group

Processor Solutions Engineering, AMD

Austin TX

Pat Conway | Presenter

Principal Member of Technical Staff

Unified North Bridge team

Processor Solutions Engineering, AMD

Sunnyvale, CA



# Outline

- AMD Roadmaps and Decoder Rings
- Hardware Overview
- Software Overview
- Questions



# Roadmap (and Decoder Rings)

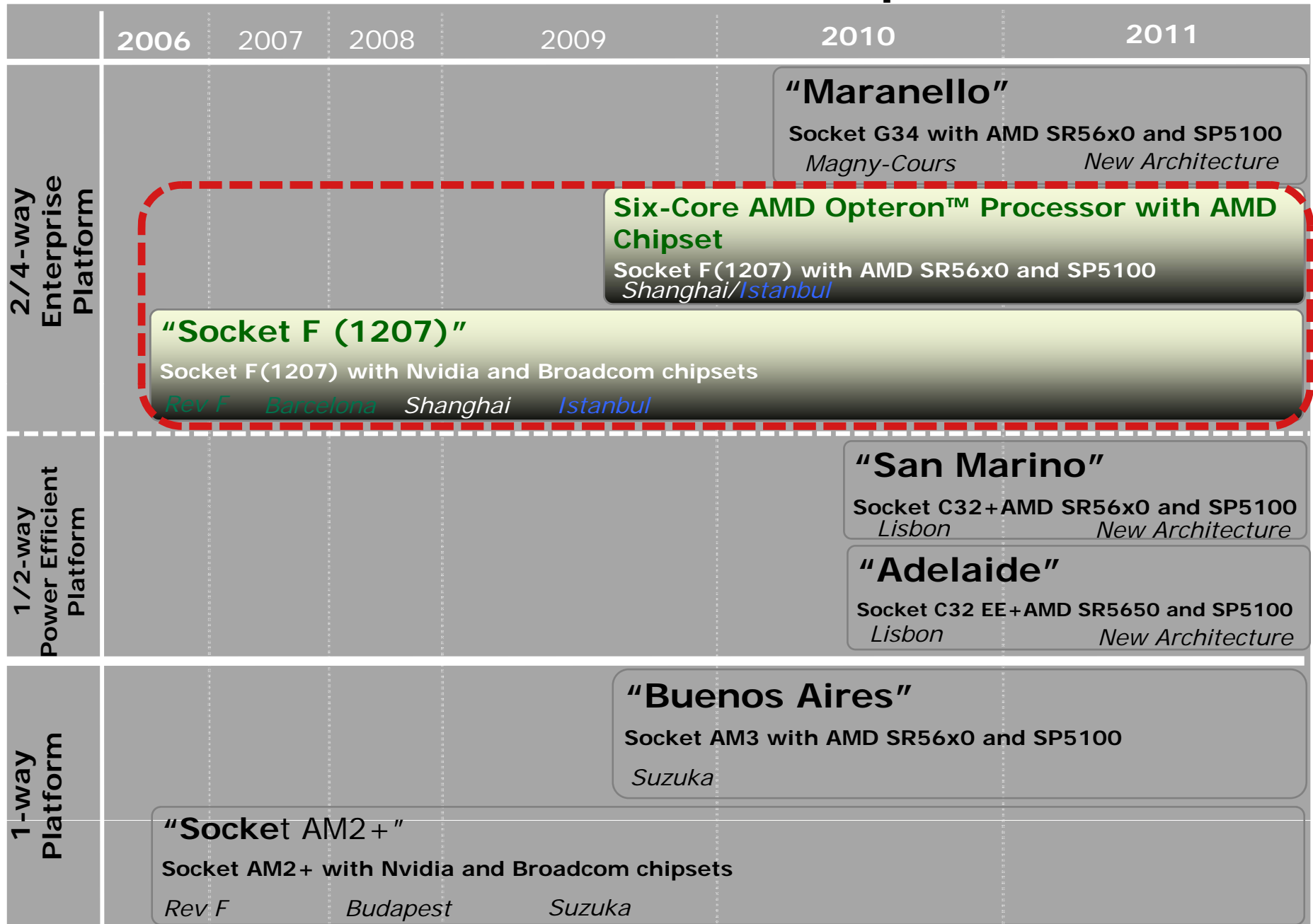


---

3 AMD Hex-Core Processors | Nersc/OLCF/NICS Cray XT5 Workshop | February 2010



# Planned Server Platform Roadmap



**"Socket F (1207)"**

Socket F(1207) with Nvidia and Broadcom chipsets

*Rev F*      *Barcelona*      *Shanghai*      *Istanbul*

**Six-Core AMD Opteron™ Processor with AMD Chipset**

Socket F(1207) with AMD SR56x0 and SP5100

*Shanghai/Istanbul*

**"Maranello"**

Socket G34 with AMD SR56x0 and SP5100

*Magny-Cours*

*New Architecture*

**"San Marino"**

Socket C32+AMD SR56x0 and SP5100

*Lisbon*

*New Architecture*

**"Adelaide"**

Socket C32 EE+AMD SR5650 and SP5100

*Lisbon*

*New Architecture*

**"Buenos Aires"**

Socket AM3 with AMD SR56x0 and SP5100

*Suzuka*

**"Socket AM2+"**

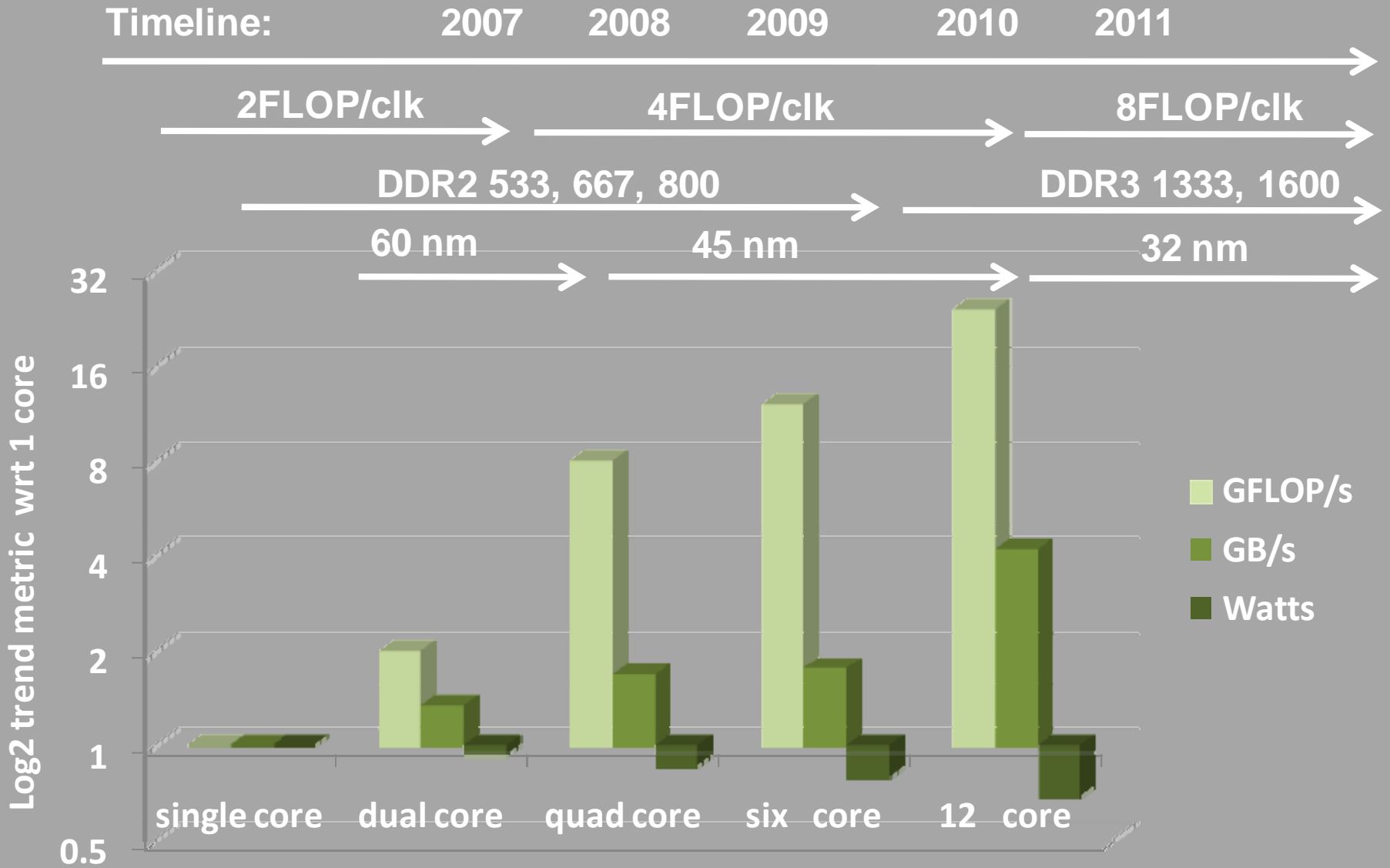
Socket AM2+ with Nvidia and Broadcom chipsets

*Rev F*

*Budapest*

*Suzuka*

# AMD Multi-core Processor trends



\* More computation while using less power per core

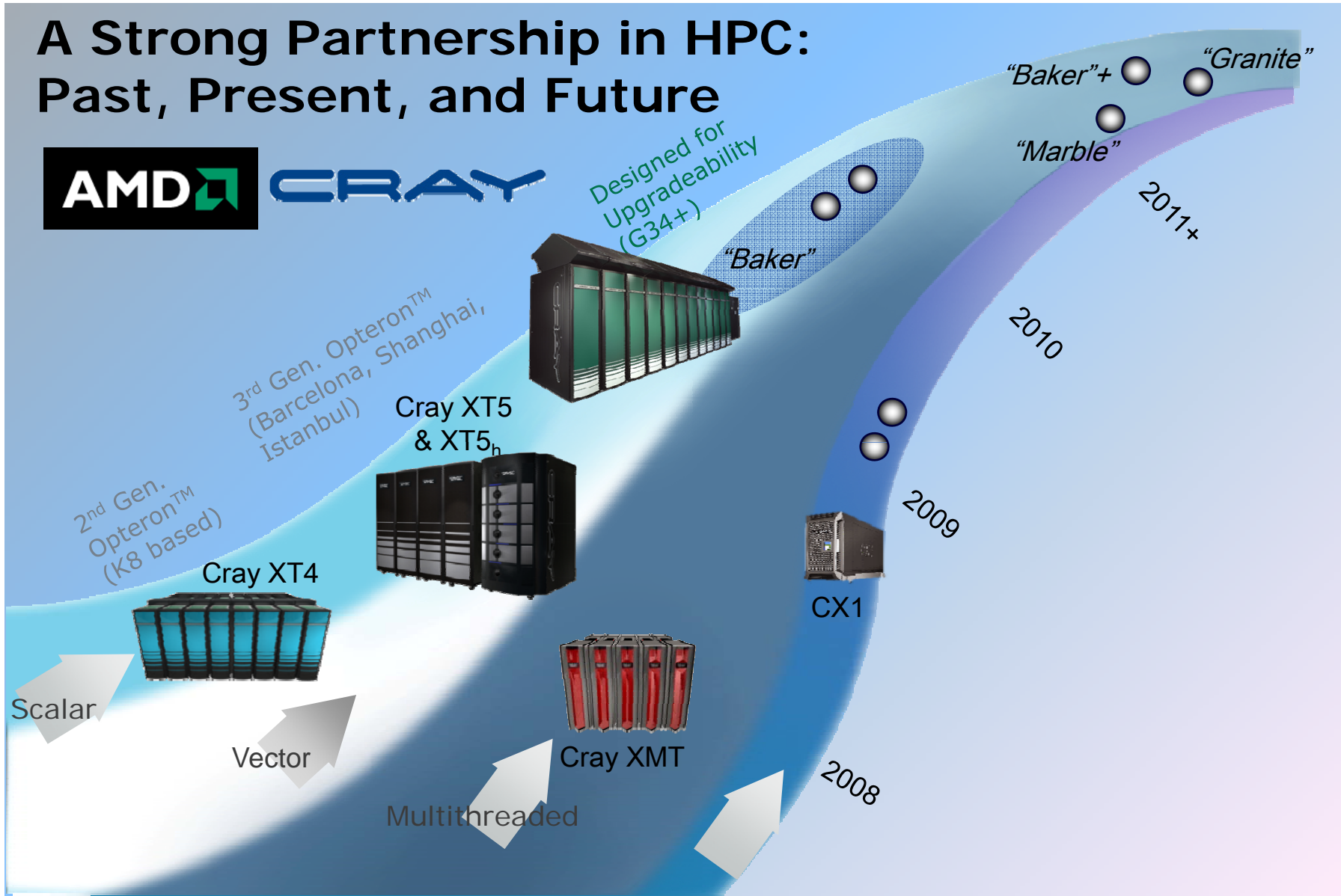
# x86 64-bit Architecture Evolution

	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**



# A Strong Partnership in HPC: Past, Present, and Future



# Hardware



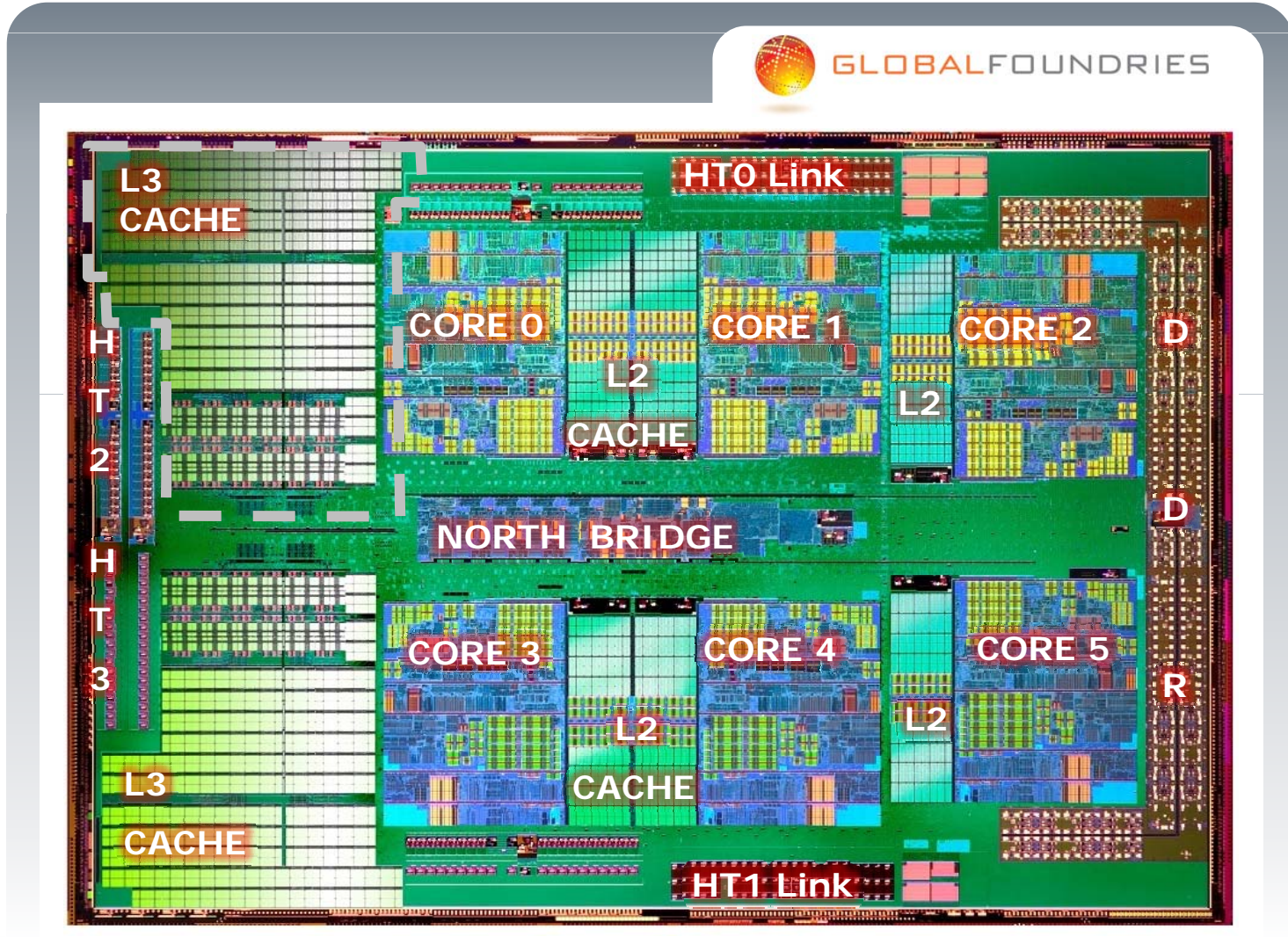
---

8 AMD Hex-Core Processors | Nersc/OLCF/NICS Cray XT5 Workshop | February 2010





# "Istanbul" Silicon



# Shanghai to Istanbul

- 6 cores (~1.5X flops)
  - Same per core L1 & L2
  - Same shared L3
  - NB & Xbar upgrades (going from 4 to 6 cores)
- HT Assist – provides 3 probe scenarios
  - No probe needed
  - Directed probe
  - Broadcast probe
- Memory BW and latency improvement
  - Amount depends on platform and configuration
- Socket Compatibility



# Cache Hierarchy

## Dedicated L1 cache

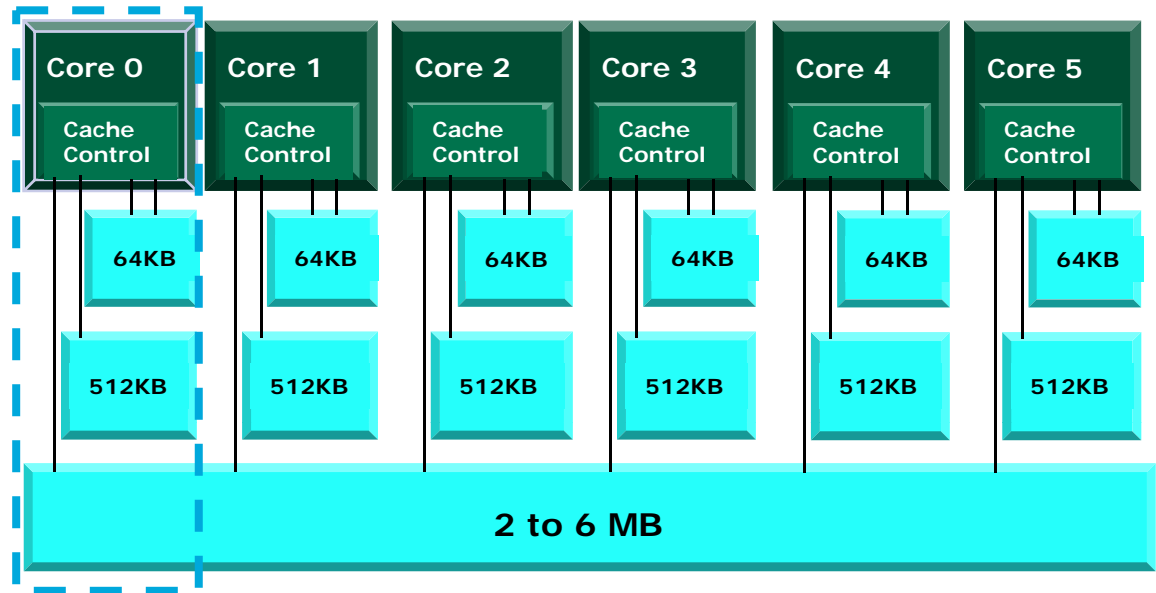
- 2 way associativity.
- 8 banks.
- 2 128-bit loads per cycle.

## Dedicated L2 cache

- 16 way associativity.

## Shared L3 cache

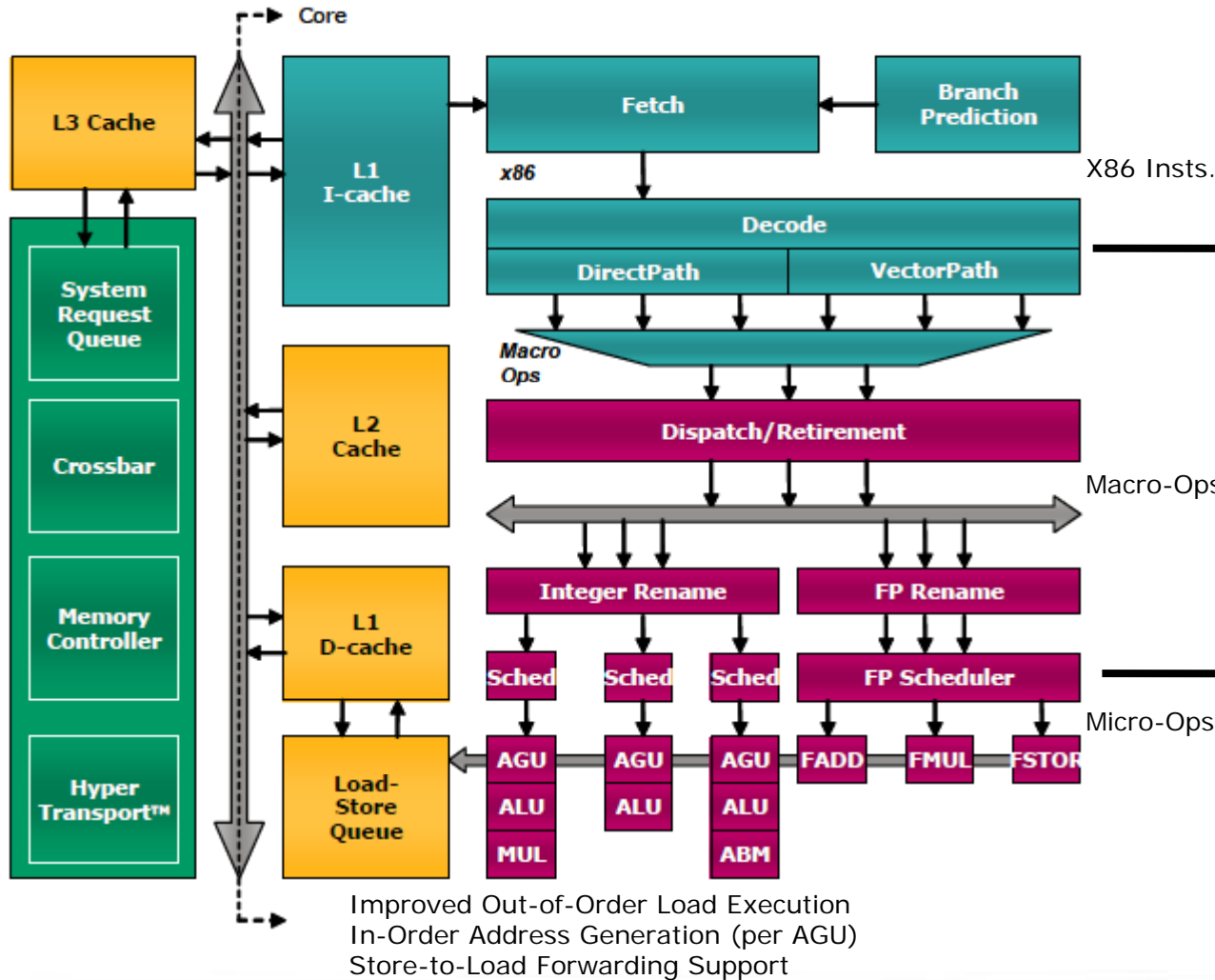
- 32 way (Barcelona), 48 way (Shanghai and Istanbul) associativity.
- fills from L3 leave likely shared lines in L3.
- sharing aware replacement policy.



# Core Micro Architecture

## FastPath? Macro-Ops? Micro-Ops?

Reference : Software Optimization Guide for AMD Family 10h Processors, Pub. #40546, Rev. 3.10 Feb 2009



### Notes / Considerations

X86 Insts. Avoid having more than 2 or 3 branches per 16B of instructions.

- Three Decode Categories (FastPath also called DirectPath)
- DirectPath Single - *best*
  - DirectPath Double - *better*
  - VectorPath (microcode) - *good*

Macro-Ops Macro-Ops tracked ReOrder Buffer (ROB).

- 72 entries (3 wide x 24 deep)
- In-order dispatch, retirement

Micro-Ops Micro-Ops issue from Sched to Execution Units

- "Sched" aka "Reserv. Station"
- Out-of-order issue
- FP scheduler shared across units
- INT Schedulers are "per unit"

Improved Out-of-Order Load Execution  
In-Order Address Generation (per AGU)  
Store-to-Load Forwarding Support



## TLB Review (Barcelona, Shanghai, Istanbul, Magny-Cours)

- Support for 1GB pagesize (4k, 2M, 1G)
- 48 bit physical addresses = 256TB (increase from 40bits on K8)
- Data TLB
  - L1 Data TLB
    - 48 entries, fully associative
    - all 48 entries support any pagesize
  - L2 TLB
    - 512 4k entries, and
    - 128 2M entries
- Instruction TLB
  - L1 Instruction TLB
    - fully associative
    - support for 4k or 2M pagesizes
  - L2 Instruction TLB



# Data Prefetch: Review of Options

## ■ Hardware prefetching

- DRAM prefetcher
  - tracks positive, negative, non-unit strides.
  - dedicated buffer (in NB) to hold prefetched data.
  - Aggressively use idle DRAM cycles.
- Core prefetchers
  - Does hardware prefetching into L1 Dcache.

## ■ Software prefetching instructions

- MOV (prefetch via load / store)
- prefetcht0, prefetcht1, prefetcht2 (currently all treated the same)
- prefetchw = prefetch with intent to modify
- prefetchnta = prefetch non-temporal (favor for replacement)





# Six-Core AMD Opteron™ Processor

## Performance

- Six-Core AMD Opteron™ Processor
- 6M Shared L3 Cache
- North Bridge enhancements (PF + prefetch)
- 45nm Process Technology
- DDR2-800 Memory
- HyperTransport-3 @ 4.8 GT/sec

## Reliability/Availability

- L3 Cache Index Disable
- HyperTransport Retry (HT-3 Mode)
- x8 ECC (Supports x4 Chipkill in ungangled mode)

## Virtualization

- AMD-V™ with Rapid Virtualization Indexing

## Manageability

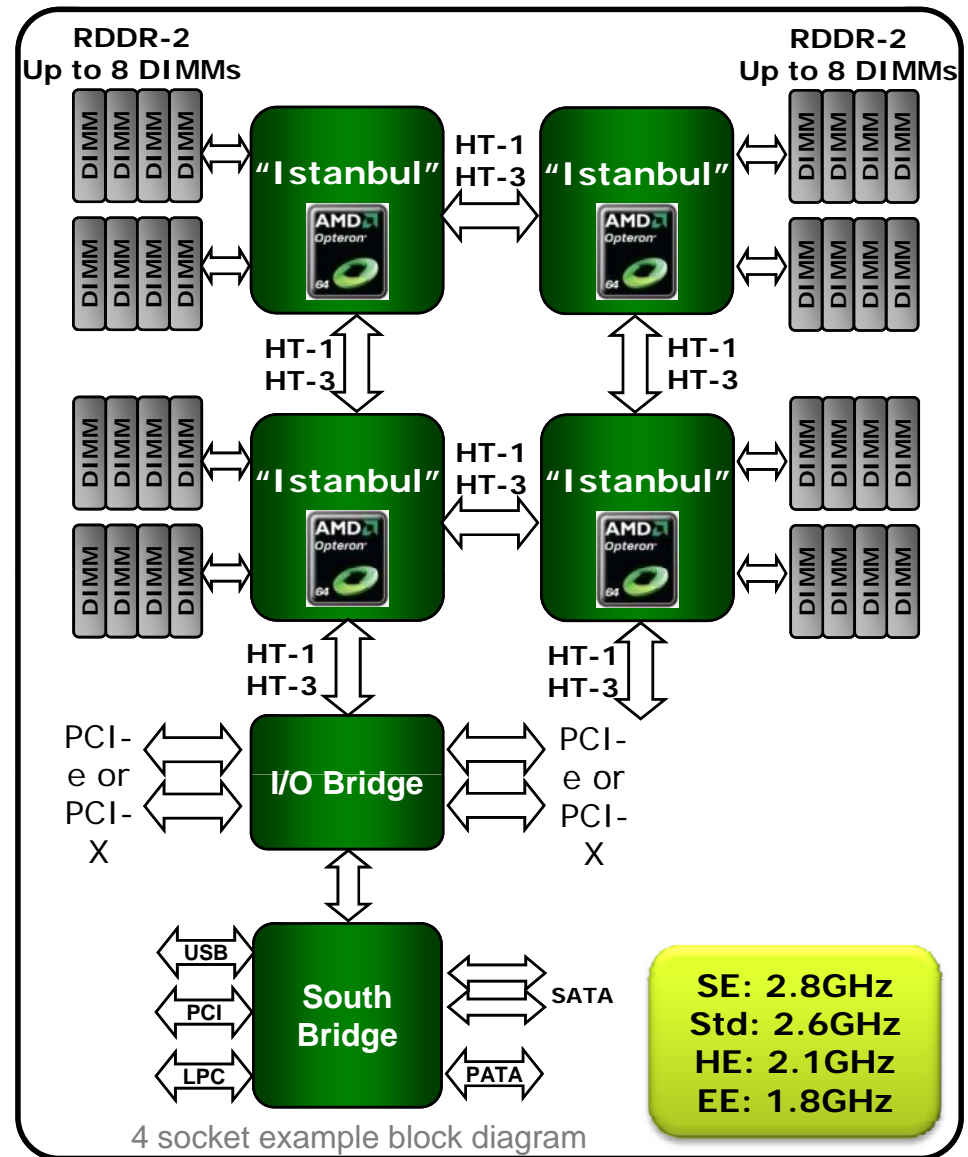
- APM Management Link\*

## Scalability

- 48-bit Physical Addressing (256TB)
- HT Assist (Cache Probe Filter)

## Continued Platform Compatibility

- Nvidia/Broadcom-based F/1207 platforms



# Six-Core AMD Opteron™ Processor

## Performance

- Six-Core AMD Opteron™ Processor  
6M Shared L3 Cache  
North Bridge enhancements (PF + prefetch)  
45nm Process Technology
- DDR2-800 Memory
- HyperTransport-3 @ 4.8 GT/sec

## Reliability/Availability

- L3 Cache Index Disable
- HyperTransport Retry (HT-3 Mode)
- x8 ECC (Supports x4 Chipkill in ungangled mode)

## Virtualization

- AMD-V™ with Rapid Virtualization Indexing

## Manageability

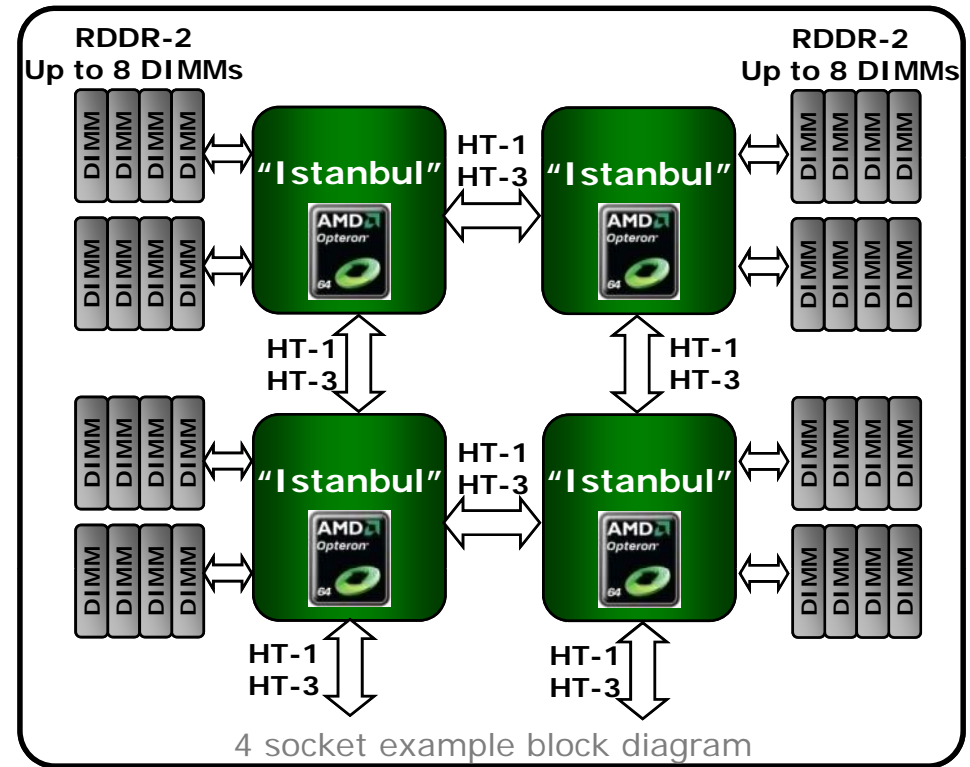
- APM Management Link\*

## Scalability

- 48-bit Physical Addressing (256TB)
- **HT Assist (Cache Probe Filter)**

## Continued Platform Compatibility

- Nvidia/Broadcom-based F/1207 platforms



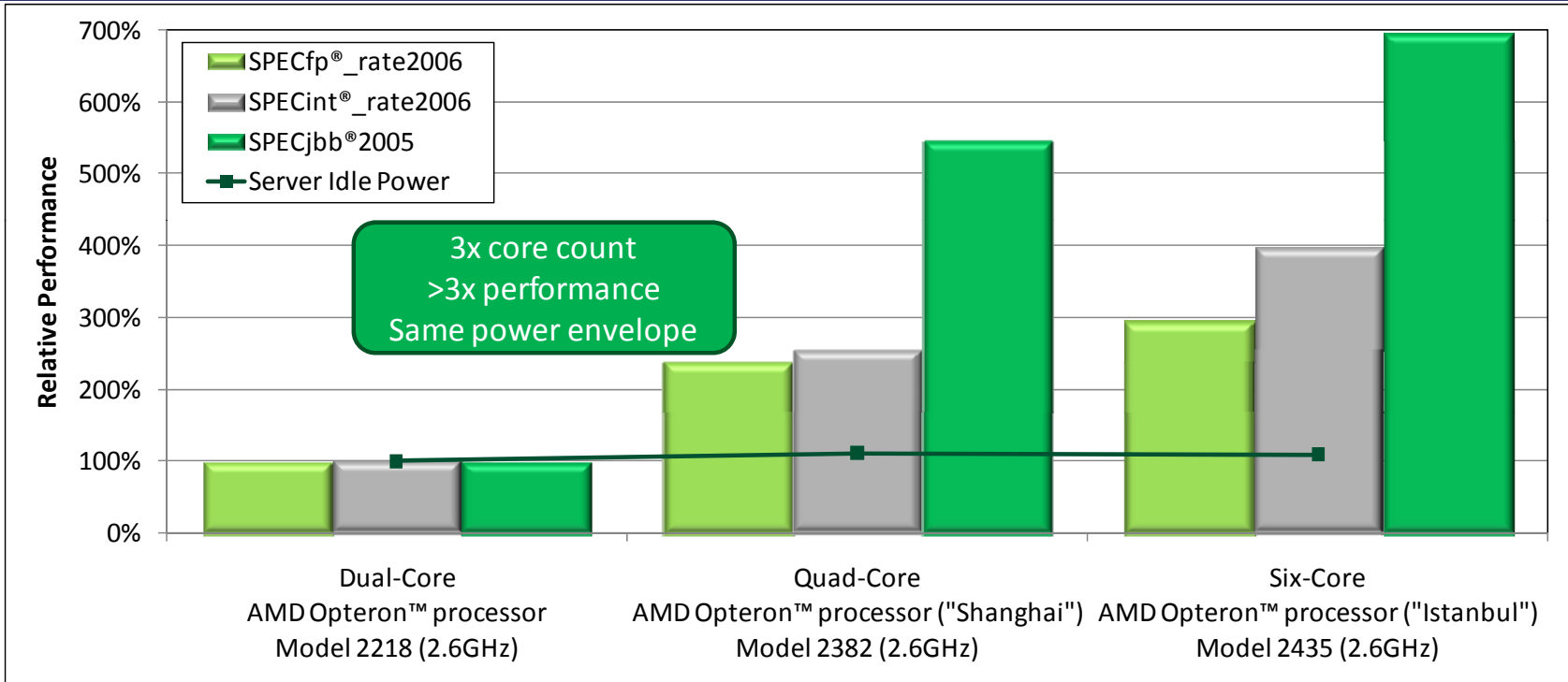
Product, Freq, Dram	STREAM Bandwidth (GB/s)*		
	2S	4S	8S
"Barcelona," 2.3/2.0, RDDR2-667	17.2	20.5	
"Shanghai," 2.7/2.2, RDDR2-800	21.4	24	
"Istanbul," 2.4/2.2, RDDR2-800	22	<b>42</b>	<b>81.5</b>





# Significantly Higher Performance in Same Power Envelope

Two-socket servers using Six-Core AMD Opteron™ processors significantly outperform two-socket servers using Dual-Core and Quad-Core AMD Opteron™ processors without consuming significantly more power



SPEC, SPECfp, SPECint, and SPECjbb are registered trademarks of the Standard Performance Evaluation Corporation. The performance results for Six-Core AMD Opteron™ processor Model 2435 and the SPECjbb result for Quad-Core AMD Opteron™ processor Model 2382 are based upon data submitted to Standard Performance Evaluation Corporation as of May 21, 2009. The other performance results stated below reflect results published on <http://www.spec.org/> as of May 21, 2009. The server idle power results are based on measurements of server active idle power for 60 seconds at AMD performance labs as of May 21, 2009. The comparison presented below is based on the best performing two-socket servers using AMD Opteron™ processor Models 2218, 2382, and 2435. For the latest results, visit <http://www.spec.org/>. Please see backup slides for configuration information.

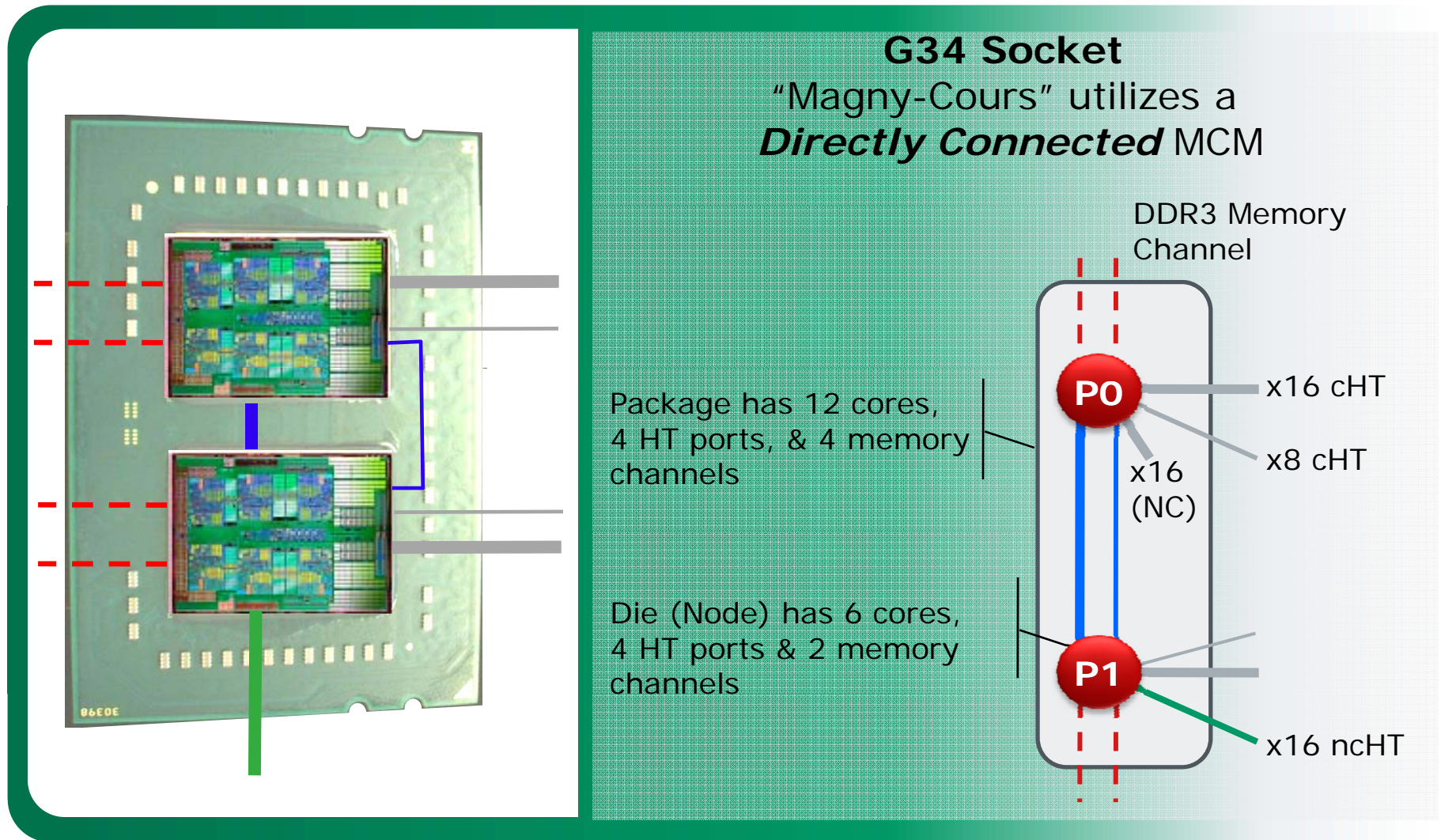


# Istanbul to Magny-Cours

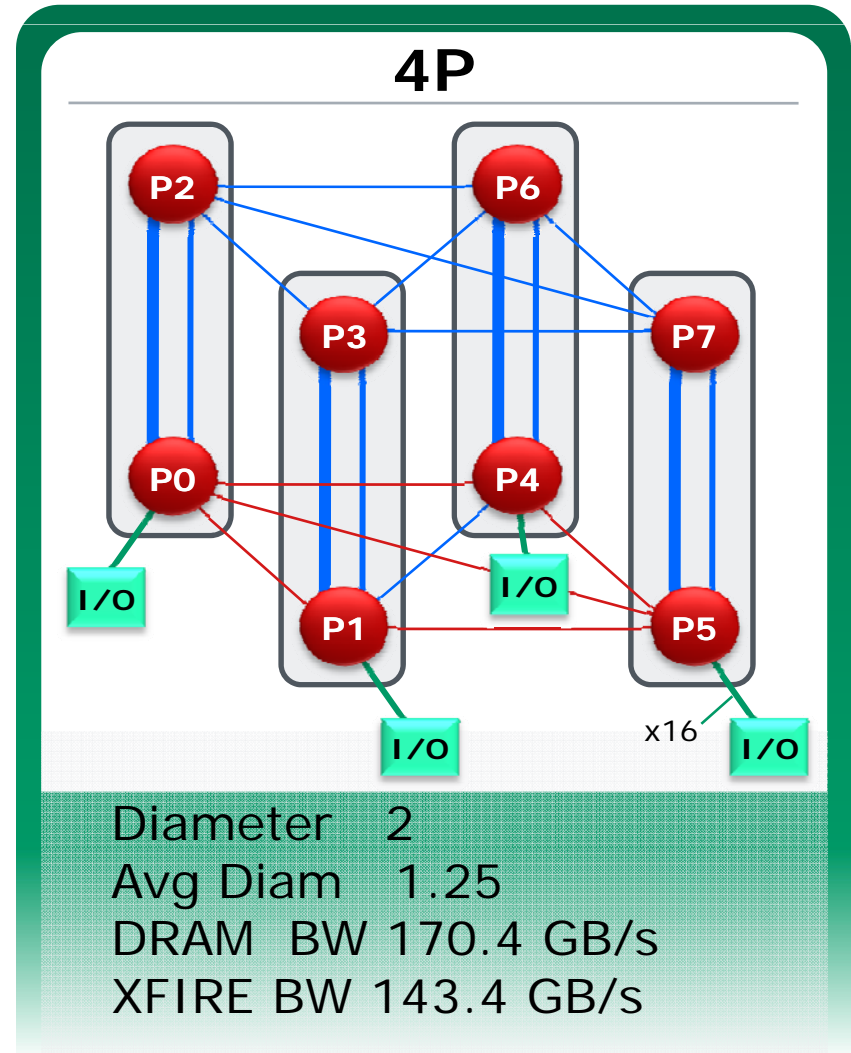
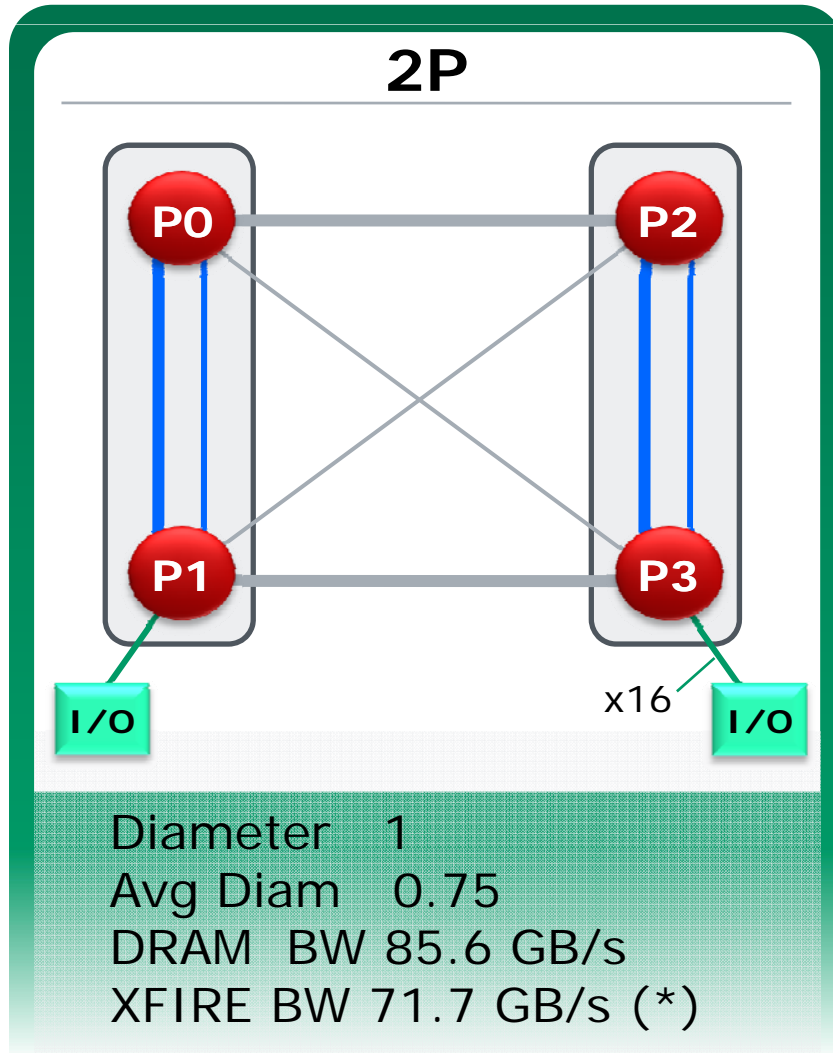
- 12 cores per socket (2 Istanbul die MCM)
  - Same per core L1 & L2
  - Same shared L3
  - NB & Xbar upgrades (going from 4 to 6 cores)
- DDR3 dimms (up to DDR3-1333)
  - 4 memory channels/socket (2 channels/die)
  - “local” memory refers to die, not socket
- Memory BW improvements
  - Same Probe filters as Istanbul
  - DDR3



# MCM 2.0 Logical View



# Topologies

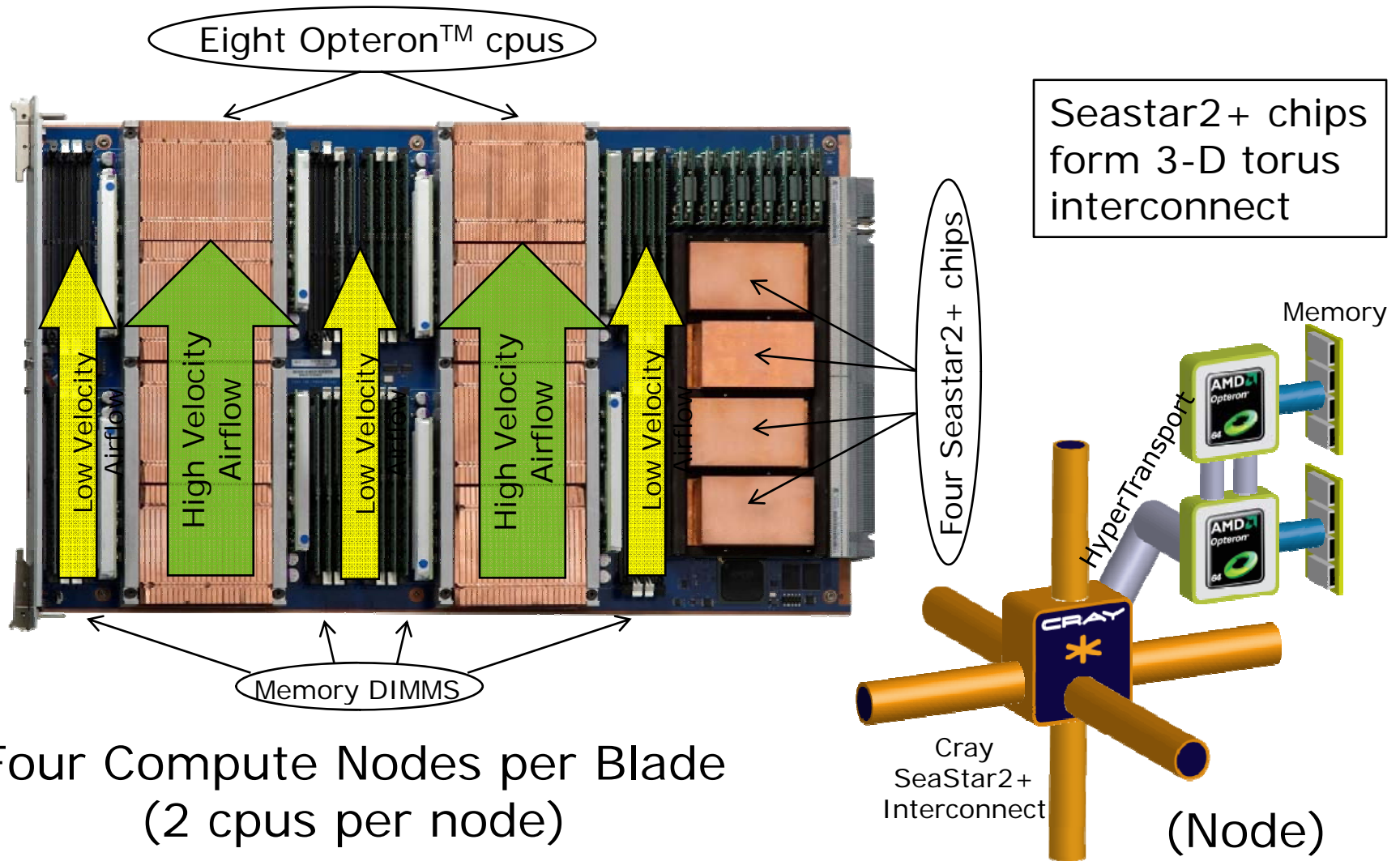


20 A (\*) XFIRE BW is the maximum available coherent memory bandwidth if the HT links were the only limiting factor. Each node accesses its own memory and that of every other node in an interleaved fashion.





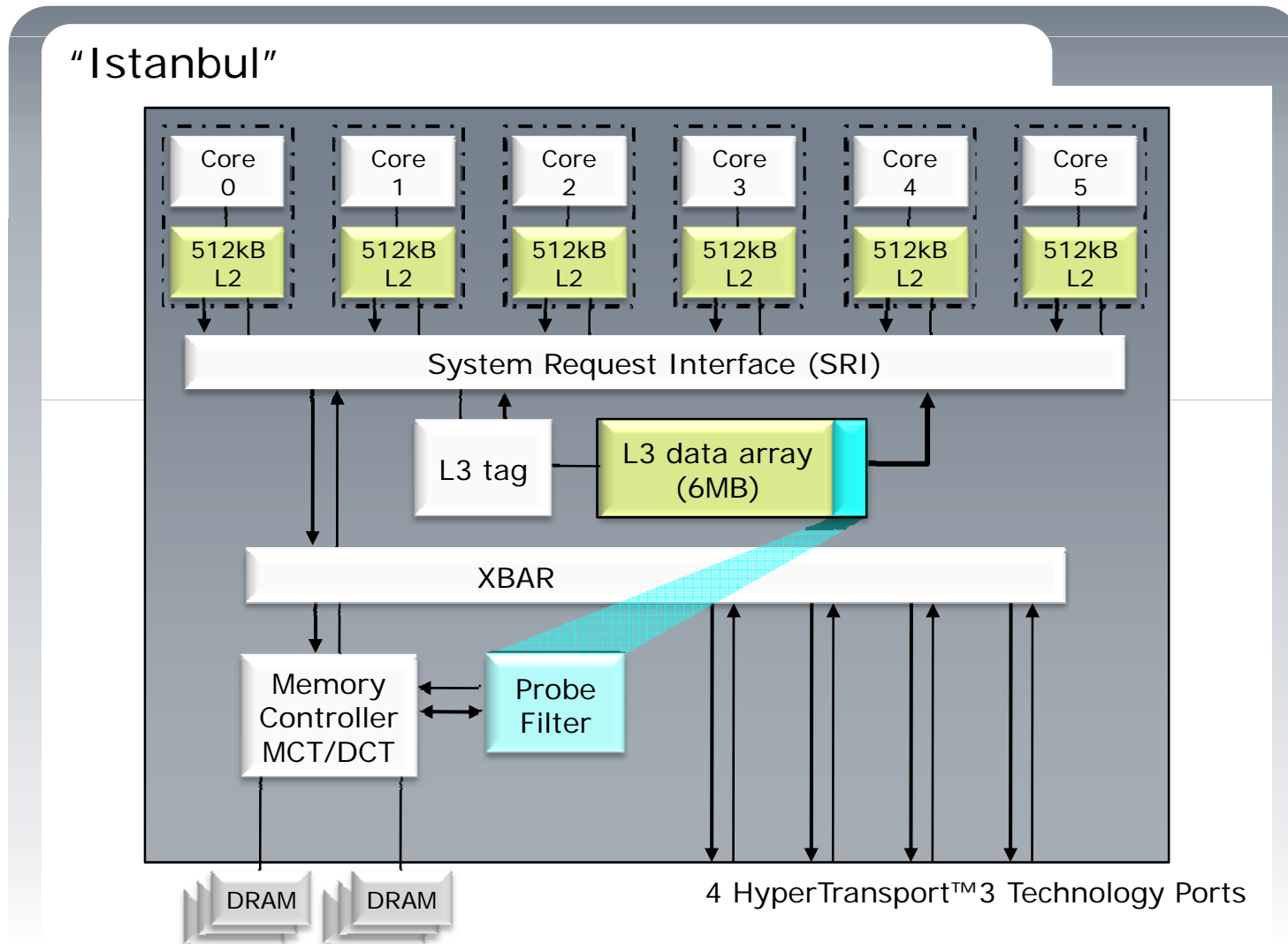
# Cray XT5 Blade and Compute Node



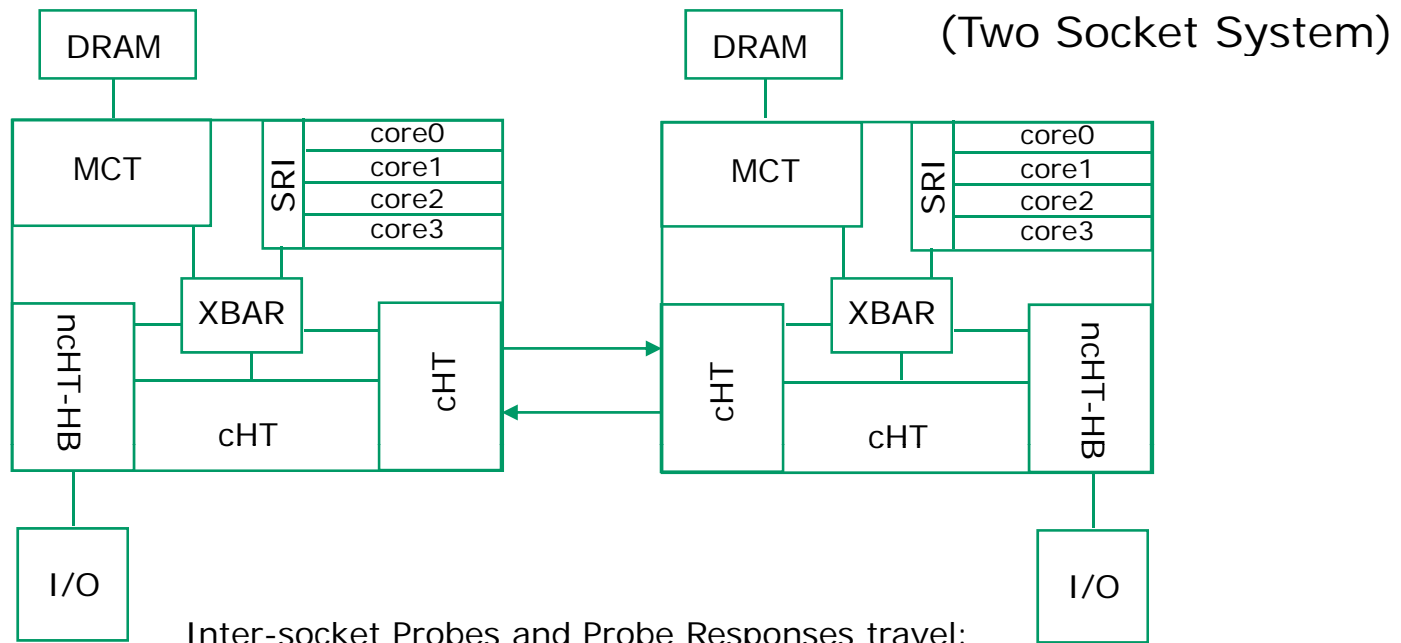
# HT Assist Feature (Probe Filters)



# Istanbul Block Diagram



# Multi-Socket System Overview



Inter-socket Probes and Probe Responses travel:  
SRI -> XBAR -> cHT -> cHT -> XBAR -> SRI

Probes Requests initiate at home memory node, but return directly to node making initial memory request.

key:

- cHT = coherent HyperTransport
- nCHT = non-coherent HyperTransport
- XBAR = crossbar switch
- SRI = system request interface (memory access, cache probes, etc.)
- MCT = memory controller
- HB = host bridge (e.g. HT to PCI, SeaStar, etc.)





# 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®
- VMARK™ 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.



# HyperTransport™ Technology HT Assist (Probe Filter)

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

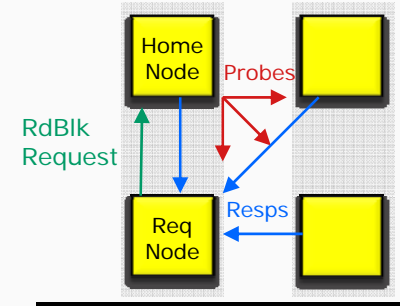
HT Assist is a sparse directory cache

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

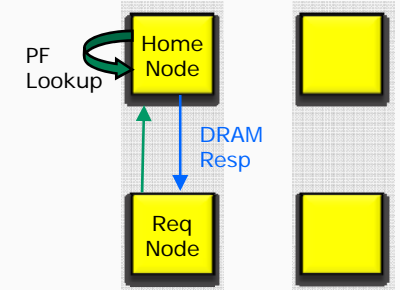
Eliminates most probe broadcasts (see diagram)

- 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

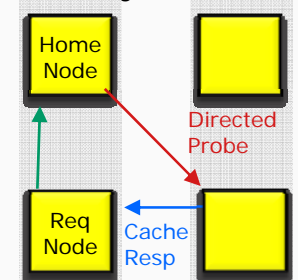
"Old" broadcast protocol




PF – clean data



PF – dirty data



# 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)

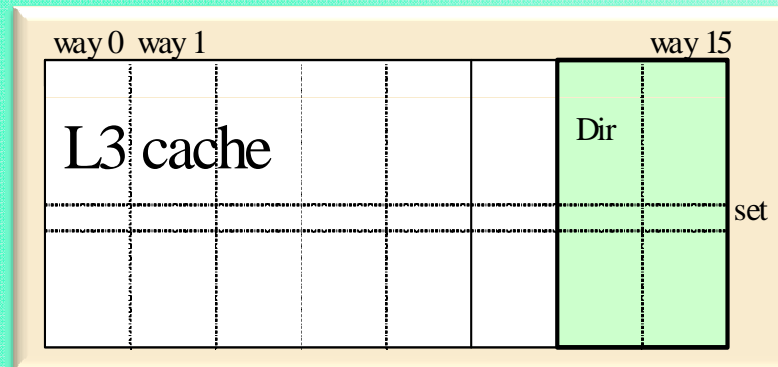




# Where Do We Put the HT Assist Probe Filter?

**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



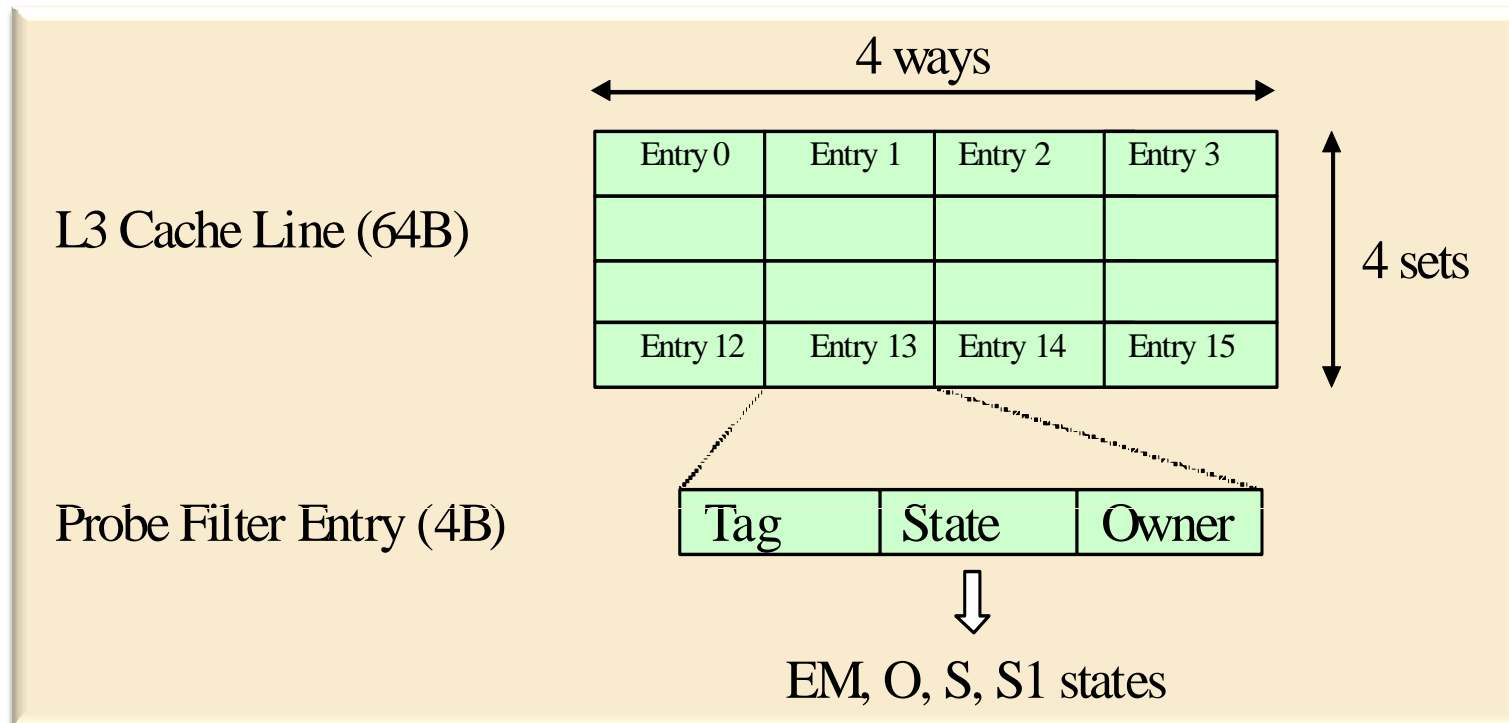
Implementation in fast SRAM (L3) minimizes

- 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



# Software



# Compiler Options

- The Portland Group (PGI) family of compilers
  - Support for Linux and Windows.
  - Debuggers and Profilers for OpenMP and MPI.
- GCC and GNU Tools for AMD (gcc, glibc, binutils/gdb)
  - AMD actively contributes improvements targeting AMD cpus.
  - More information available at <http://developer.amd.com/cpu/gnu/Pages/default.aspx>
- C/C++/Fortran compilers based on Open64 technology
  - Available for download on AMD Developer Central at <http://developer.amd.com/cpu/open64> in source and binary forms.

■



# Some Notable PGI Flags

## A Good Base Set of Flags

C/C++

-fastsse -Mipa=fast -Mipa=inline -tp shanghai-64

Fortran

-fastsse -Mvect=short -Mipa=fast -Mipa=inline -tp shanghai-64

(note: "-tp shanghai-64" is also appropriate for Istanbul)

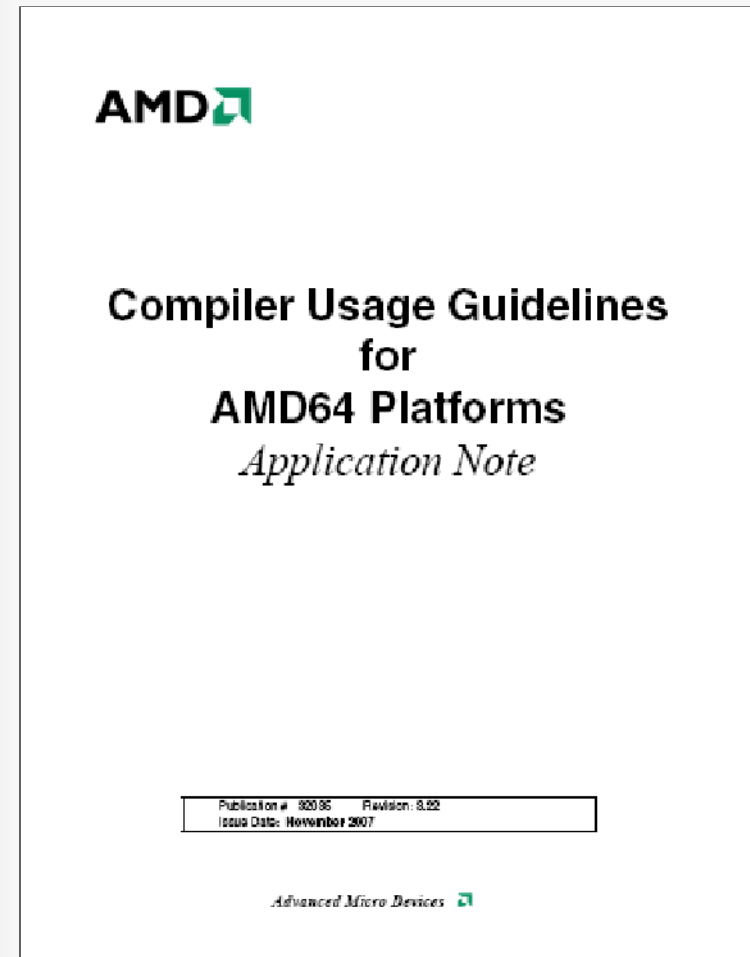
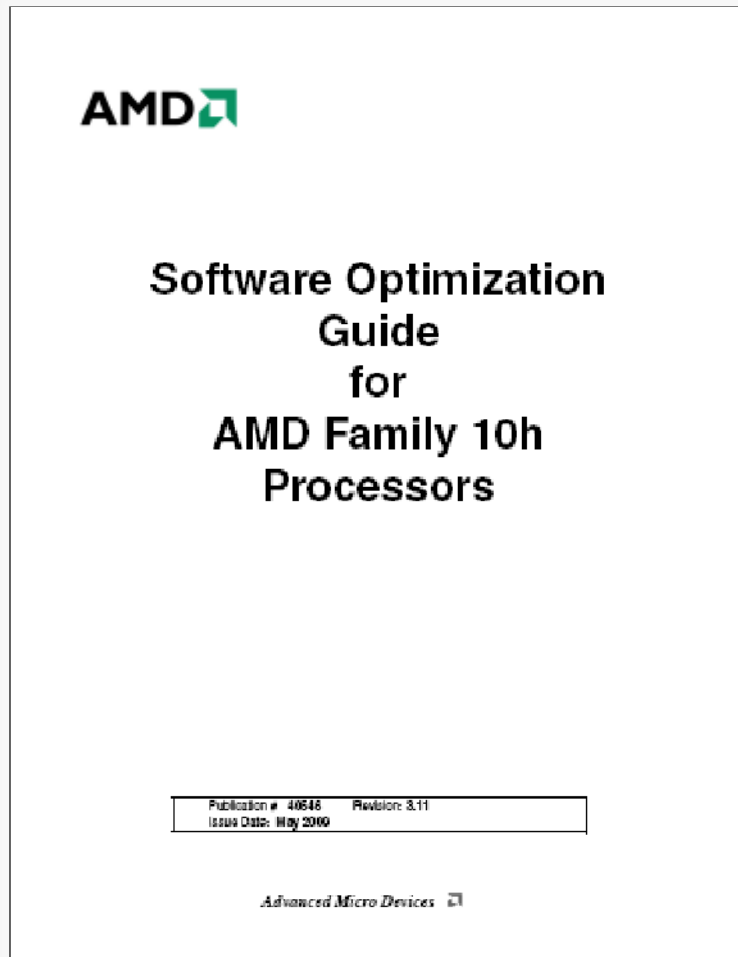
## And Some More to Usually Consider

Flag	Purpose
-Msmartalloc	Use smart malloc routines.
-Msmartalloc=huge	Use smart malloc routines with large pages (depends on amount of OS allocated large pages and the PGI_HUGE_PAGES environment variable).
-Mvect=fuse	Enables vectorizer to fuse loops.
-Mpfi=indirect (first pass) -Mpfo=indirect (second pass)	Use profile feedback optimizations. (requires compiling, doing training run(s), and recompiling).
-Msafeptr (C/C++ only)	Says arrays don't overlap and different pointers point to distinct locations. (has many sub-options).
-Mfprelaxed	Allows relaxed precision for certain Intrinsic functions (sqrt, rsqrt, order, div).





# Software Optimization & Compiler Guidelines



<http://developer.amd.com/documentation/guides/Pages/default.aspx>

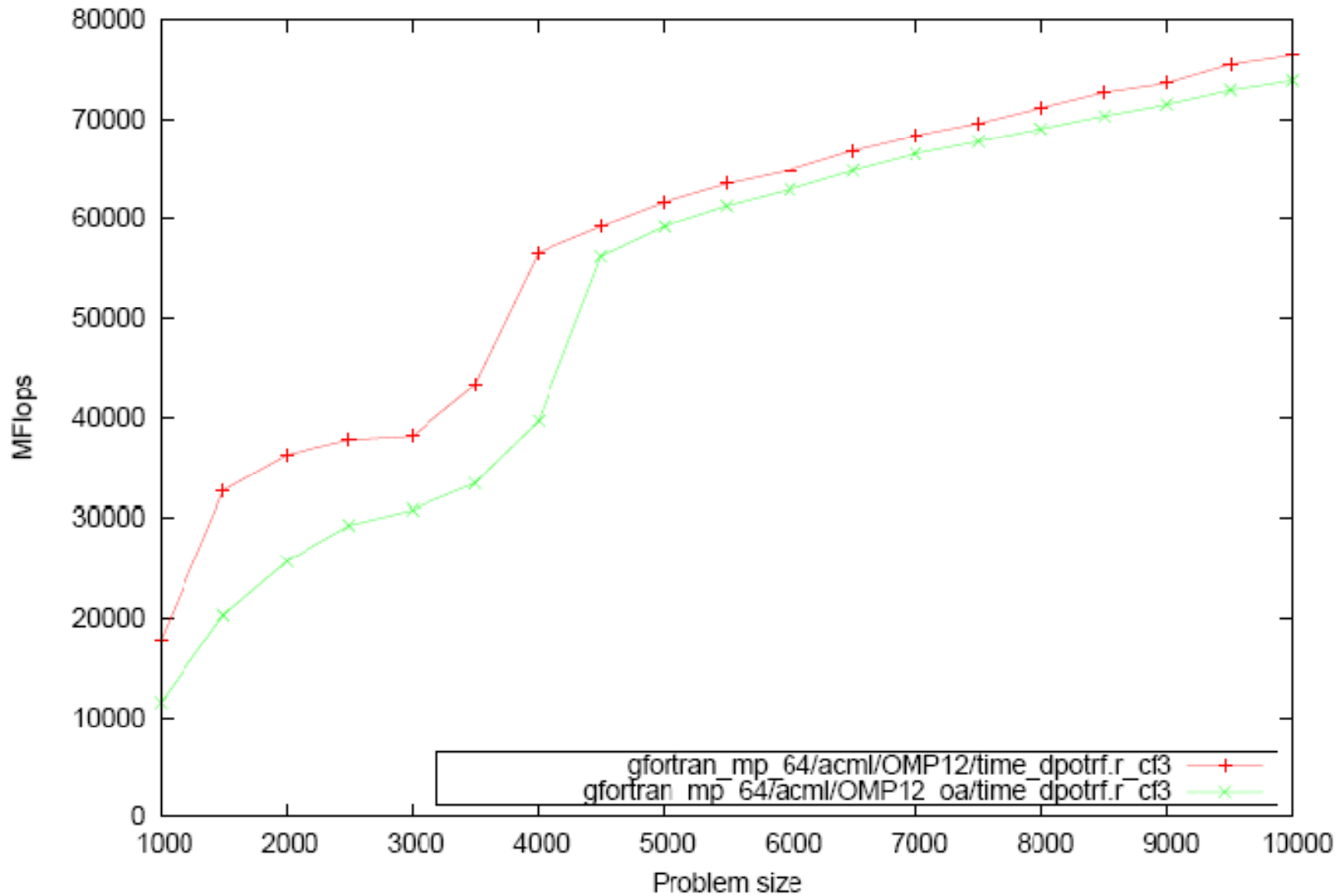
# ACML 4.3.0

- L3 BLAS improvements
  - SGEMM for Six-Core AMD Opteron™ Processor
  - New Intel DGEMM and SGEMM kernels
    - Supporting Woodcrest, Penryn, Nehalem
    - Competitive with MKL
  - New DGEMM "fast memory allocation" scheme
    - allows improved performance of other routines (such as LAPACK) which make heavy use of DGEMM
- Six-Core AMD Opteron™ processor tuning for Level 1 BLAS
  - xDOT, xCOPY, xAXPY, and xSCAL
- 3DFFT performance improvements
  - Now outperforming MKL
- AMD Family 10h tuning for real-complex FFTs
  - csfft, dzfft, scfft and zdfft have been re-tuned for FP128

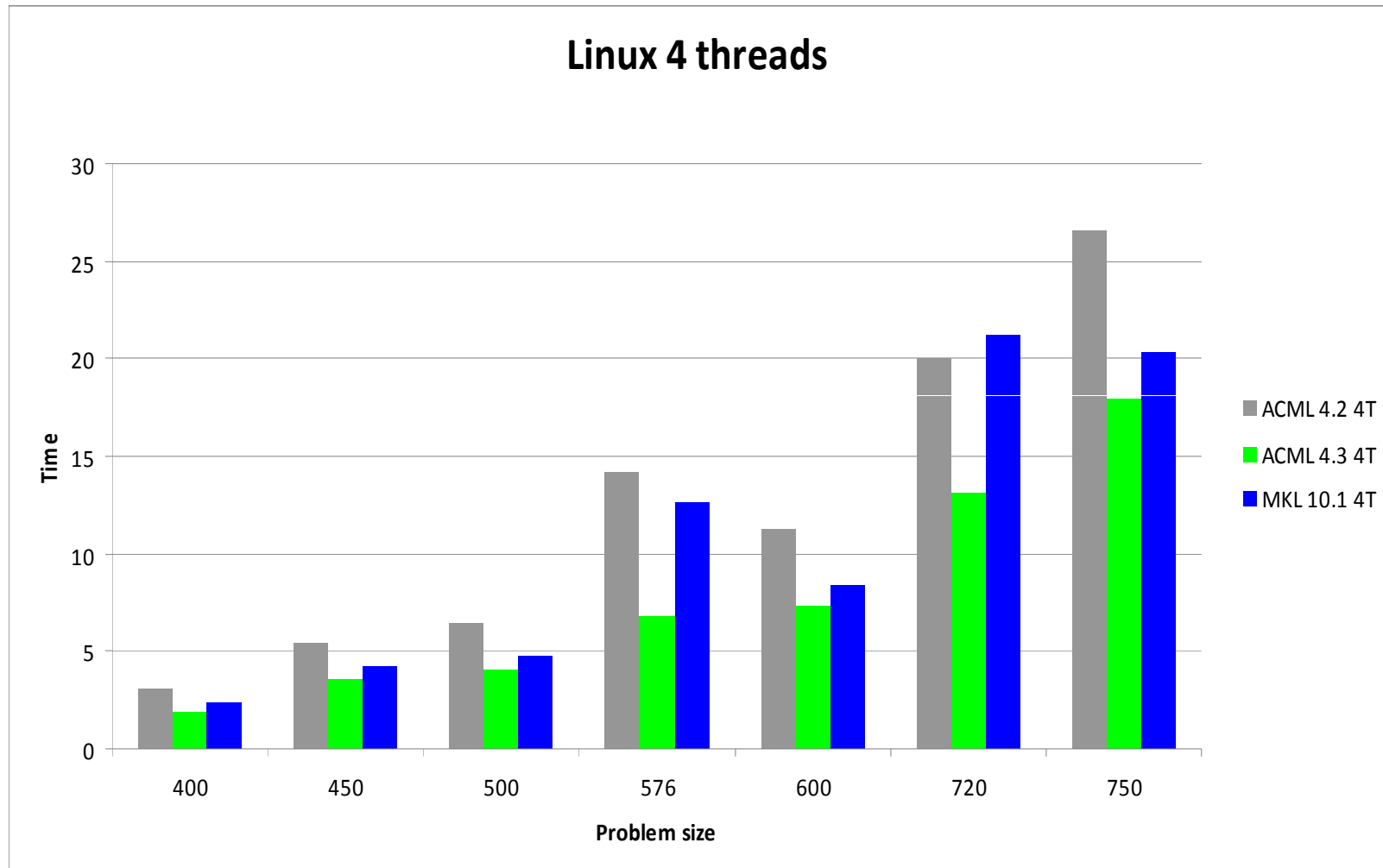


# New Memory Allocation Scheme

Timing dpotrf



# 3DFFT Performance Improvement



*Lower is better*



# ACML 4.3.0 Supported Compilers

- PGI
  - 8.0-6
- GCC/GFORTRAN
  - 4.3.2
  - Now backwards compatible with GCC 4.1.2 and 4.2
- Open64
  - 4.2.1
- Intel Fortran 11.0
- Microsoft Visual Studio 2008



# ACML 4.4 (Just Out)

- Ensure proper scaling with Istanbul/Magny-Cours
- Resolve scaling issues with small problems
- Family 10h optimized ZGEMM, associated L3 routines
- Double Complex 2D/3D-FFT improvements



# AMD Developer Central

(for more info and downloads)

## Downloads

- ACML and AMD LIBM are offered as free downloads to registered members of AMD Developer Central.

## Forums

- The ACML forum is an excellent place to find answers.

## Blogs

- Watch for blog entries by members of the ACML and LIBM team.

**<http://developer.amd.com>**



## (A bit about Stream Computing)

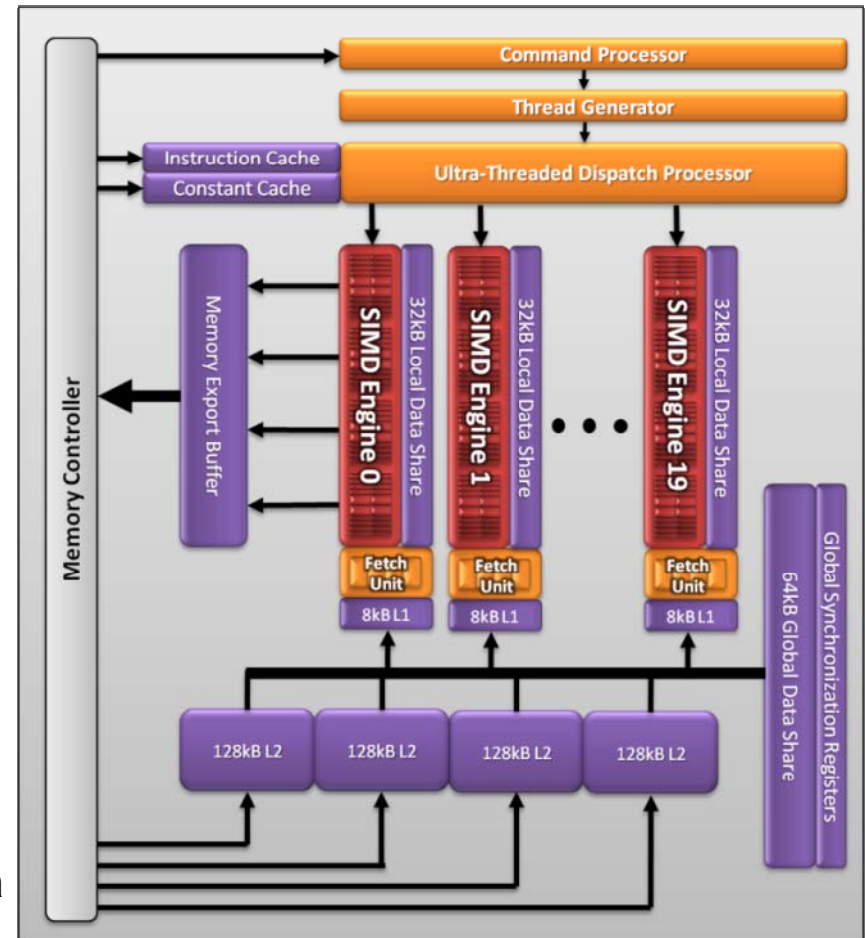




# ATI Radeon™ HD 5870 Graphics Architecture

2.72 Teraflops Single Precision  
544 Gigaflops Double Precision

- Full Hardware Implementation of DirectCompute 11 and OpenCL™ 1.0
- IEEE754-2008 Compliance Enhancements
- Additional Compute Features:
  - 32-bit Atomic Operations
  - Flexible 32kB Local Data Shares
  - 64kB Global Data Share
  - Global synchronization
  - Append/consume buffers
- ATI Stream SDK beta
  - <http://developer.amd.com/streambeta>



# Some GPGPU thoughts

- Massively Parallel Compute capability
  - FLOPS per watt and per mm<sup>2</sup> are truly impressive.
- Programmability
  - Improving but not painless.
- Moving Data onto and off of the GPGPU
  - Must do enough computation to amortize this.
- RAS
  - Not the same level as general purpose CPUs yet.

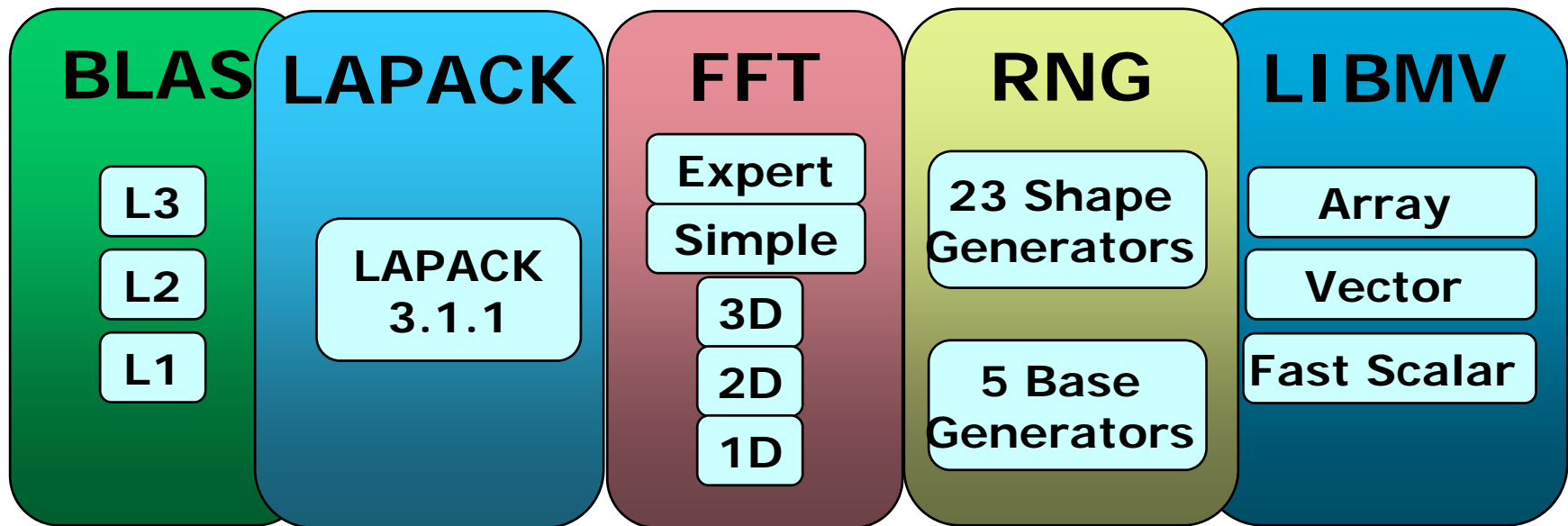


# ACML-GPU 1.0

## Released March 2009

Selected BLAS routines enabled for GPU

- DGEMM, SGEMM will run on GPU if present
- Small problems ( $N, M, K < 200$ ) run on CPU
- Supports Quad- and Six-Core AMD Opteron™ processors



## Trademark Attribution

AMD, the AMD Arrow logo and combinations thereof are trademarks of Advanced Micro Devices, Inc. in the United States and/or other jurisdictions. Other names used in this presentation are for identification purposes only and may be trademarks of their respective owners.

©2009 Advanced Micro Devices, Inc. All rights reserved.



# Questions



# Backup



# MOESI Cache Coherency Protocol

