Bandwidth & Latency Challenges for Multi-Core Server Performance

Fayé Briggs, PhD
Chief Performance Architect
Digital Enterprise Group
Intel Corporation

12th EMEA Academic Forum
June 12 – 14, 2007
Budapest, Hungary

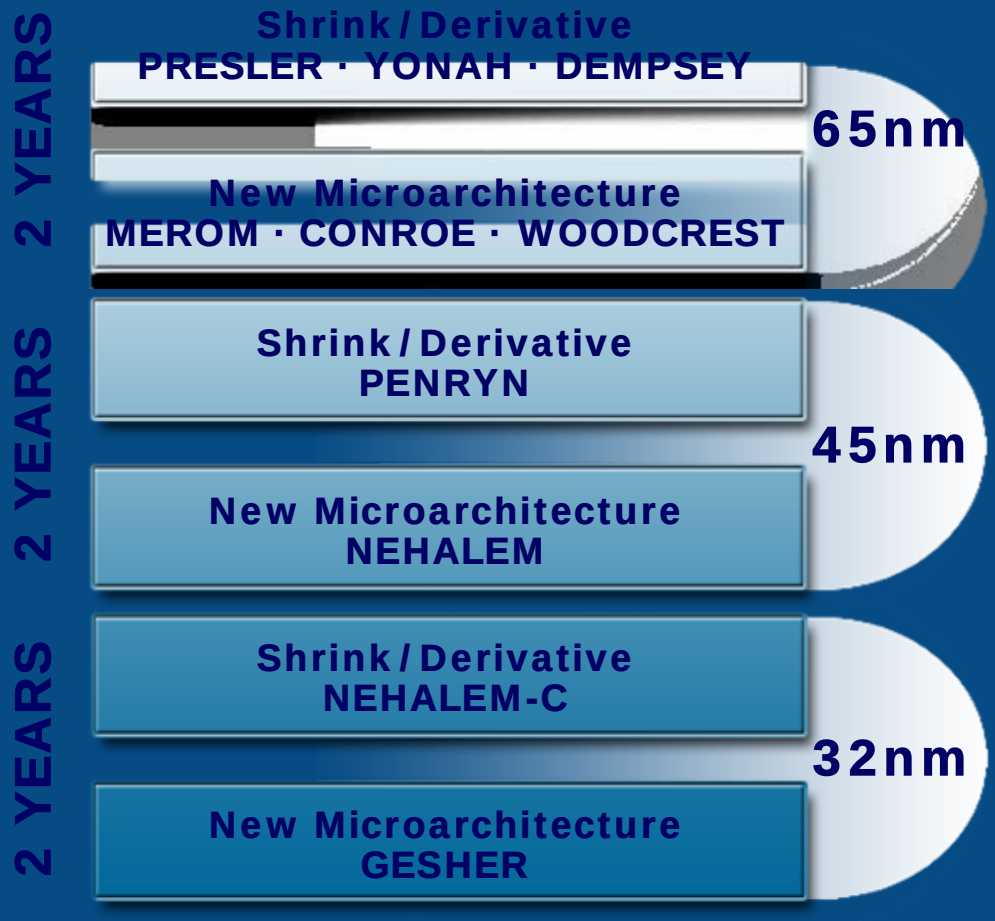
Legal Disclaimer

- LEGAL INFORMATION:
- THIS DOCUMENT AND RELATED MATERIALS AND INFORMATION ARE PROVIDED "AS IS" WITH NO WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, NON-INFRINGEMENT OF INTELLECTUAL PROPERTY RIGHTS, OR ANY WARRANTY OTHERWISE ARISING OUT OF ANY PROPOSAL, SPECIFICATION, OR SAMPLE. INTEL ASSUMES NO RESPONSIBILITY FOR ANY ERRORS CONTAINED IN THIS DOCUMENT AND HAS NO LIABILITIES OR OBLIGATIONS FOR ANY DAMAGES ARISING FROM OR IN CONNECTION WITH THE USE OF THIS DOCUMENT.
- Performance & functionality will vary depending on (i) the specific hardware & software you use & (ii) the feature enabling/system configuration by your system vendor. See www.intel.com/ for information on Intel Technology or consult your system vendor for more information.
- All dates provided are subject to change without notice.
- Intel, Pentium, Xeon, Itanium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States & other countries.
- *Other names & br&s may be claimed as the property of others.
-

Agenda

- Multi-Core Momentum
- Multi-Core Performance Challenges
- Platform Architectures & Performance
- SPEC CPU2006 Sensitivity to Bandwidth
- Bandwidth & Performance Implications of Increasing Core-Count
- Workload Based Analysis
- Summary

Microprocessor Design Model



PRINCIPLES

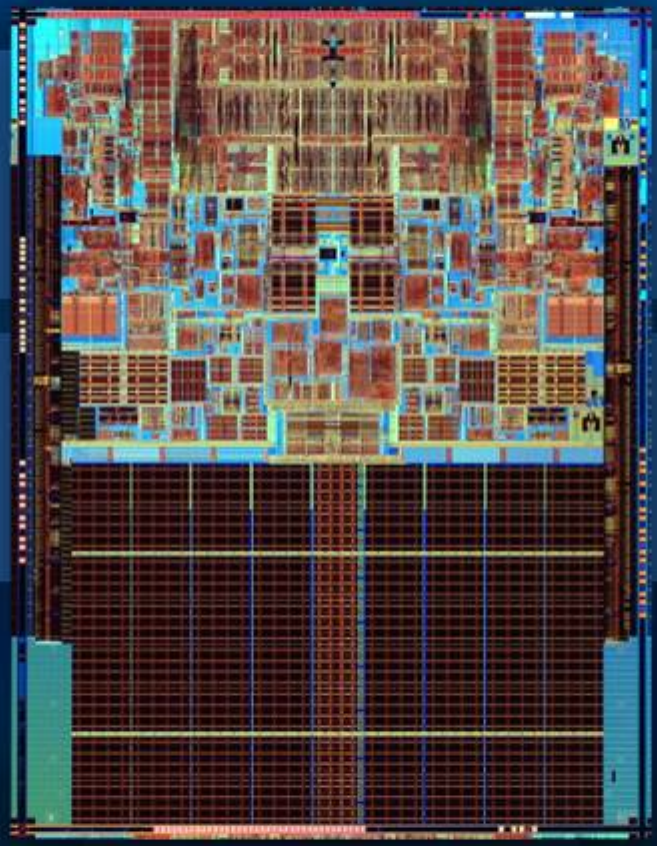
1. One micro-architecture for all high volume market segments
2. Optimized for performance/watt
3. Parallel design teams
4. No waiting on new process technology
5. Chipset cadence offset for fast ramp

OBJECTIVE: Sustained Technology Leadership

Intel Core Micro-architecture
Five Key Innovations

Intel® Wide Dynamic Execution
- 4 wide issue, retire
- macrofusion

Intel® Advanced Digital Media Boost
- 128 bit wide SSE

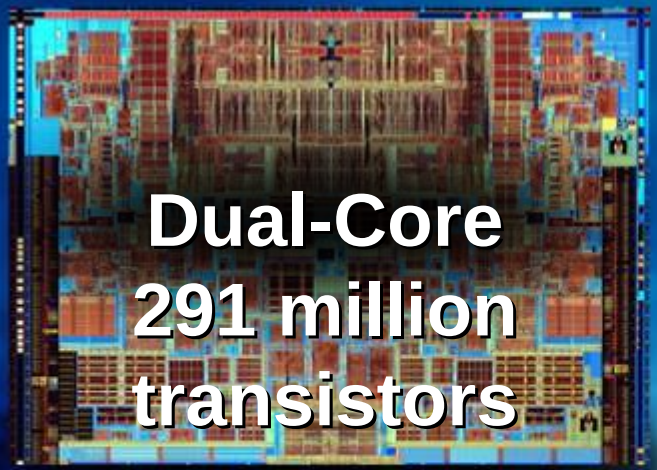


Intel® Intelligent Power Capability
- clock gating
- split buses

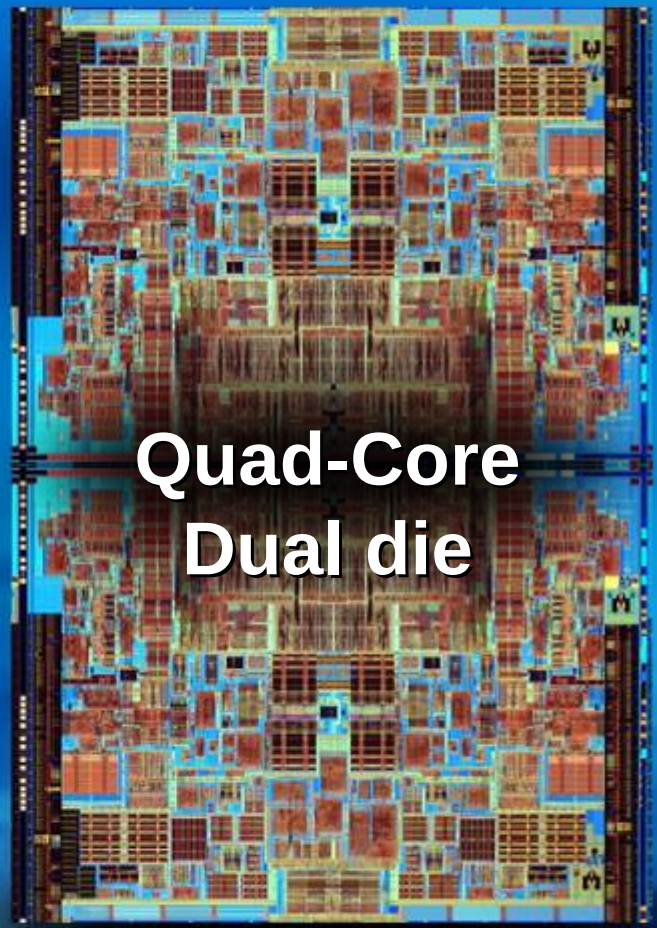
Intel® Smart Memory Access
- enhanced prefetch
- memory disambiguation

Intel® Advanced Smart Cache
- large shared cache

Multi-Core Products



2Q 2006



4Q 2006

More multi-core products expected in the future

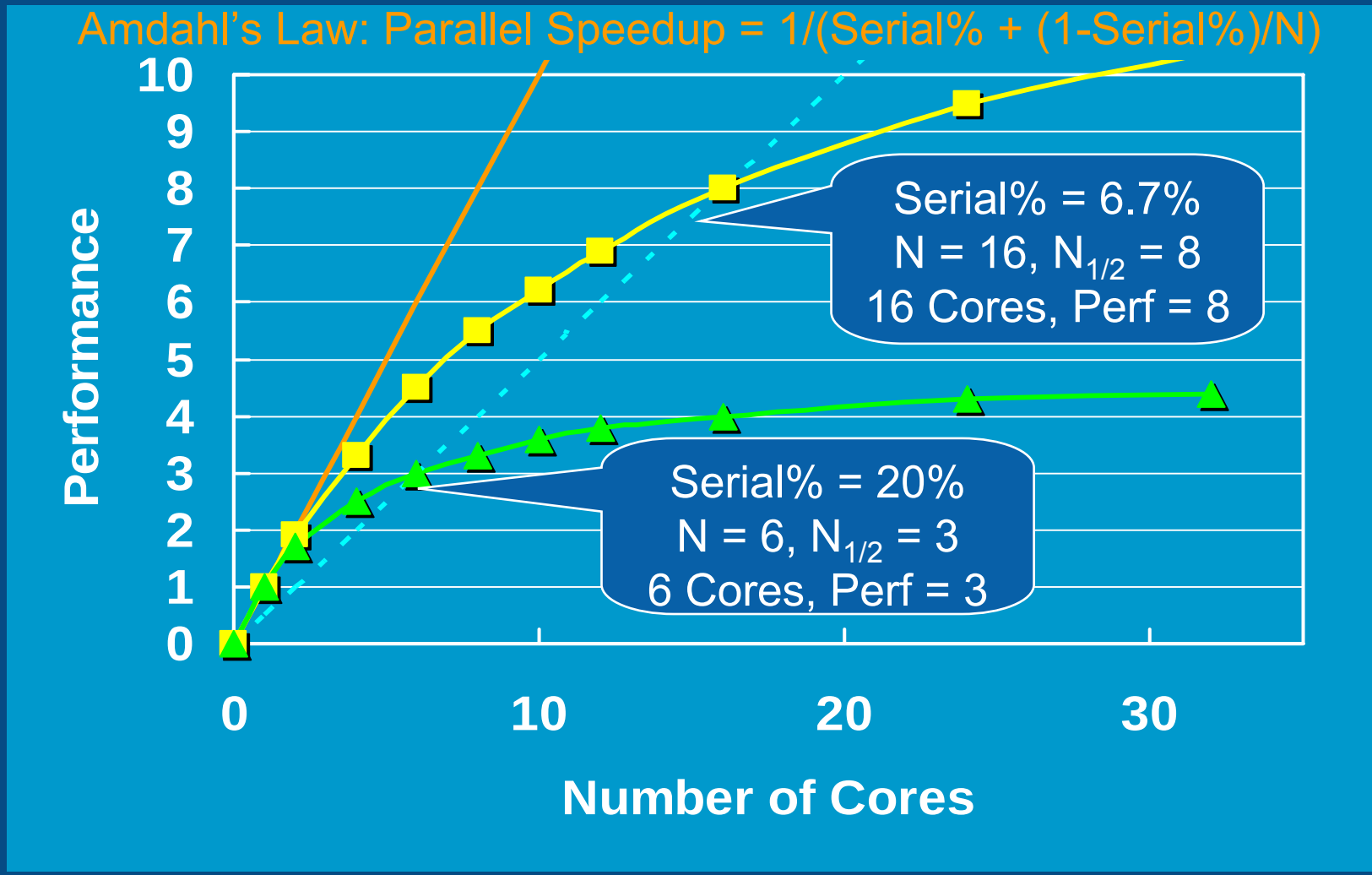
Agenda

- Multi-Core Momentum
- Multi-Core Performance Challenges
- Platform Architectures & Performance
- SPEC CPU2006 Sensitivity to Bandwidth
- Bandwidth & Performance Implications of Increasing Core-Count
- Workload Based Analysis
- Summary

Performance Challenges in Multi-Core Platforms

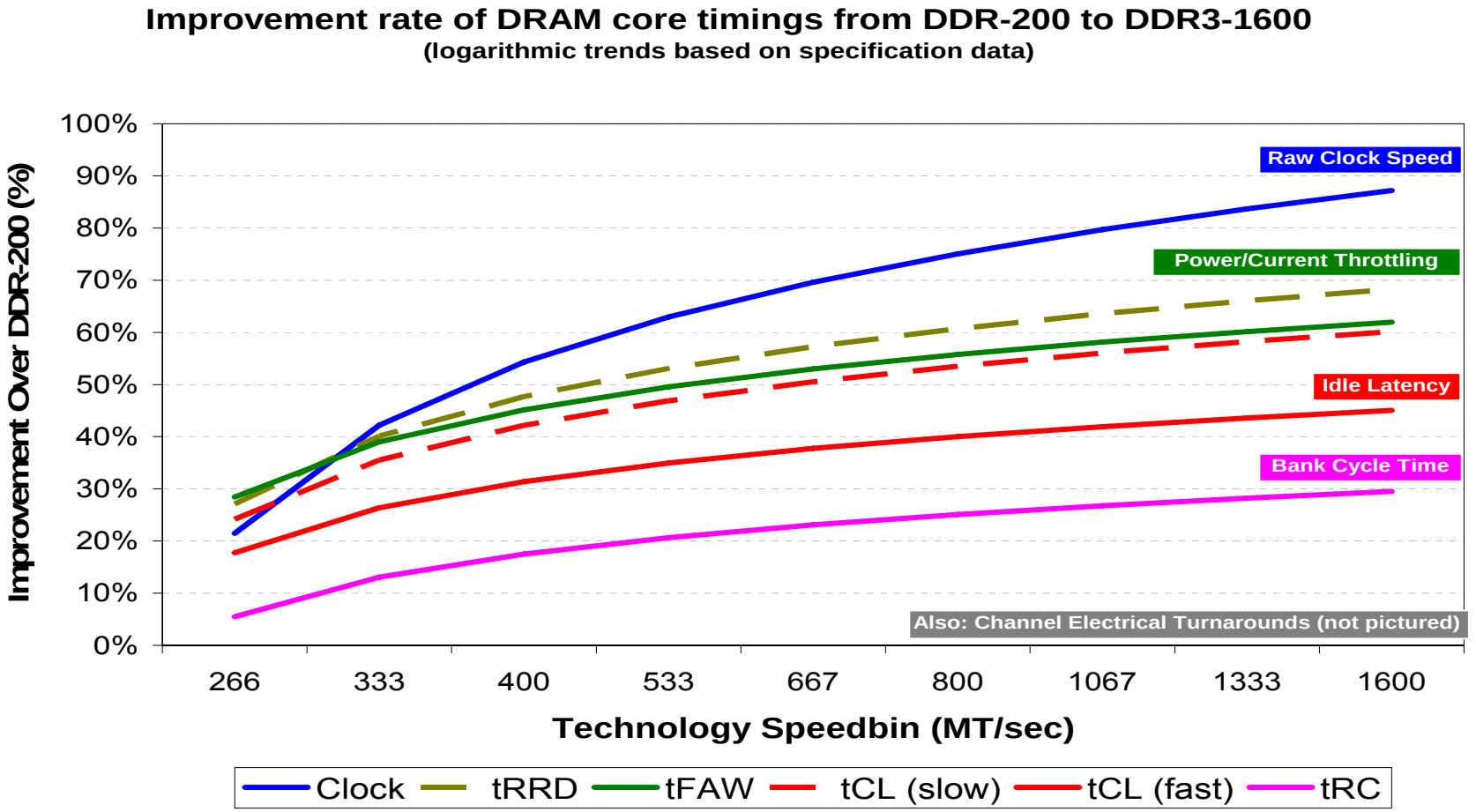
- Extracting thread level parallelism in most workloads
  - How much?
- Ability to generate code with lots of threads & performance scaling
  - New tools available
- Power limitations
- Platform latencies (idle and loaded)
- On-chip interconnect/cache infrastructure
  - Adequate on-die bandwidths & reduced miss rates
- Memory and I/O bandwidth required

Performance Scaling



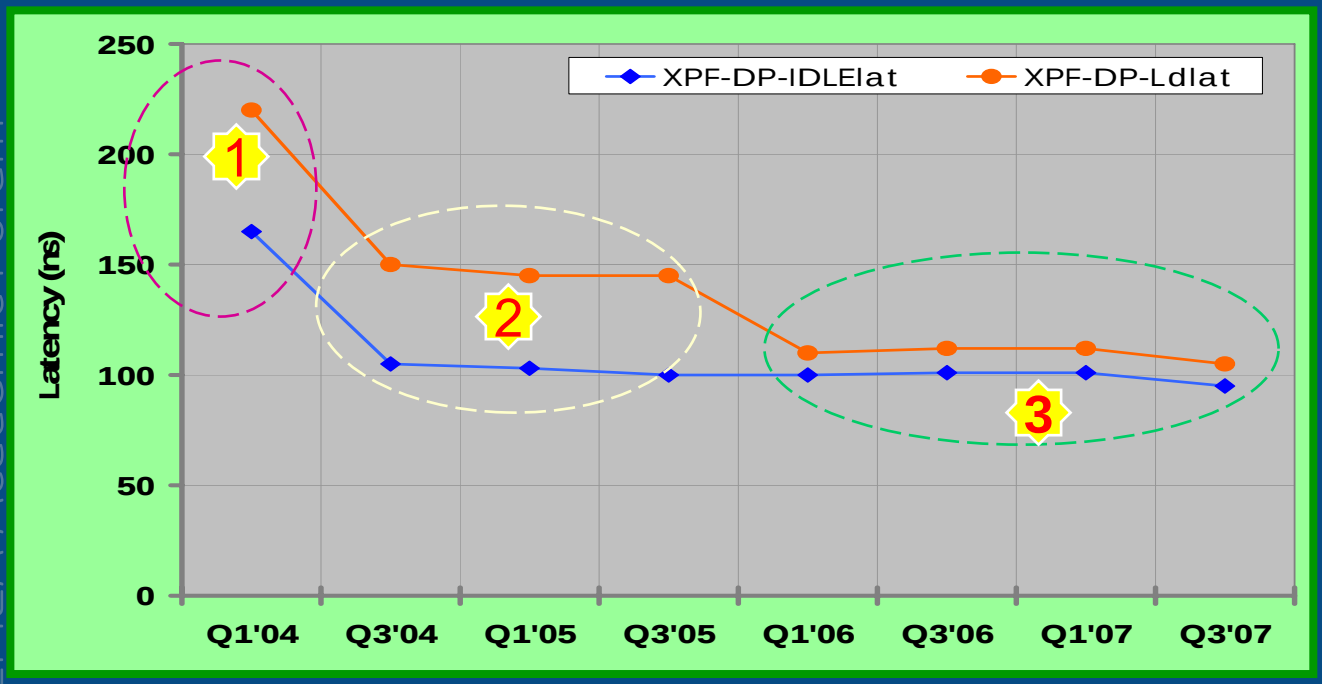
Parallel software key to Multi-core success

DRAM Timing Improvements



Increasing gap between DRAM data clock cycle time & memory constraint timings

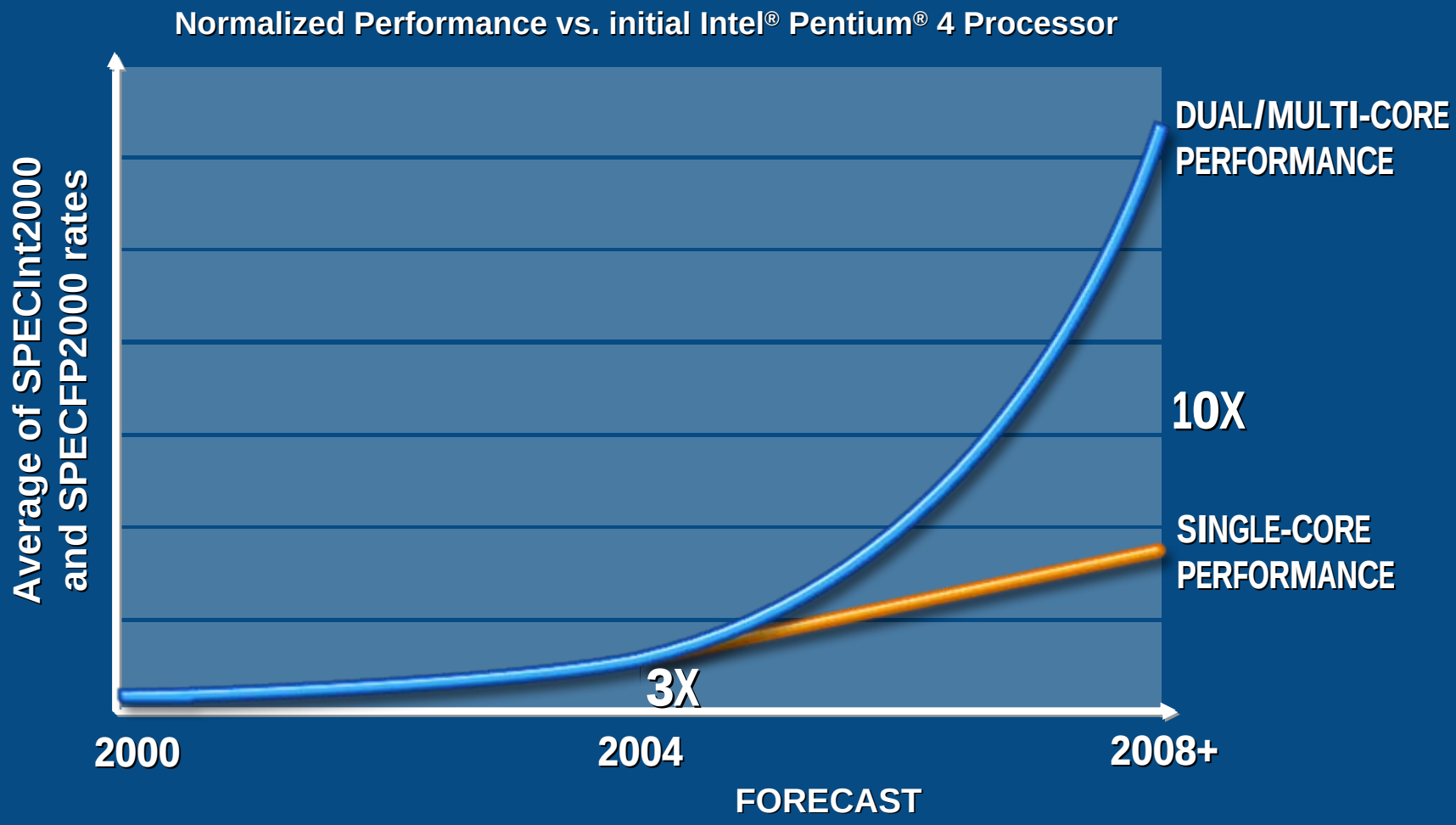
2S OLTP Average Last Level Cache (LLC) Miss Latency



1. 3-load shared
2. Faster 3-load shared
3. Dedicated Fast FSB

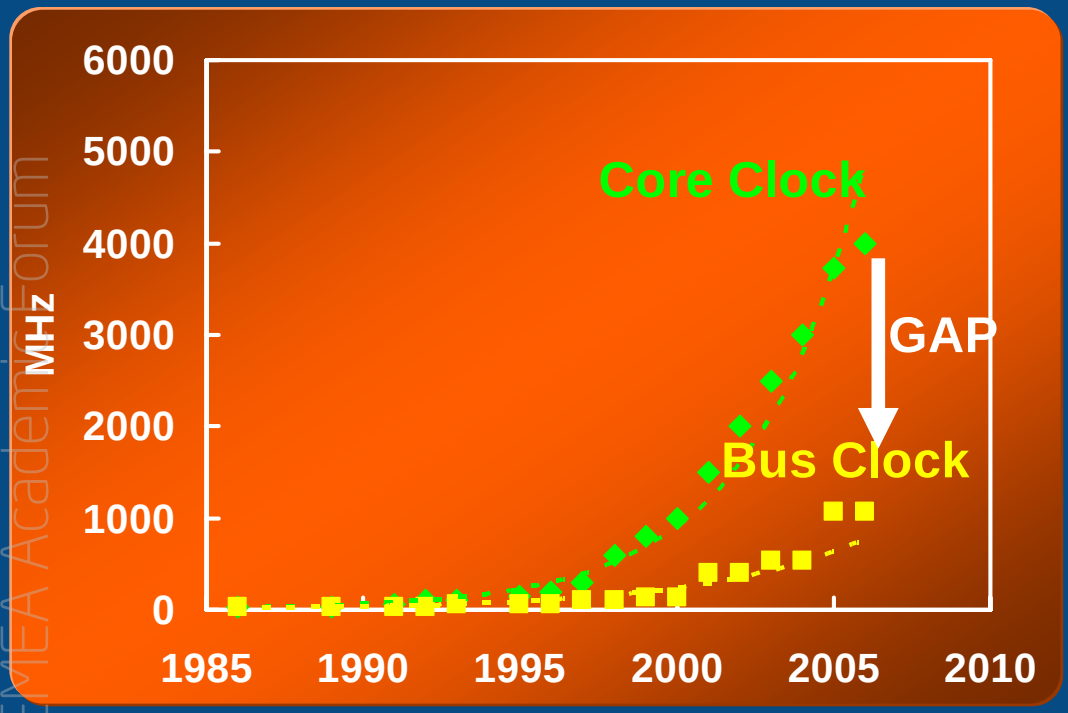
Latency reduction continues but approaching a lower bound

Bandwidth Drivers – increased parallelism



Greater parallelism drives *abrupt* increase in BW requirements

Memory BW Gap



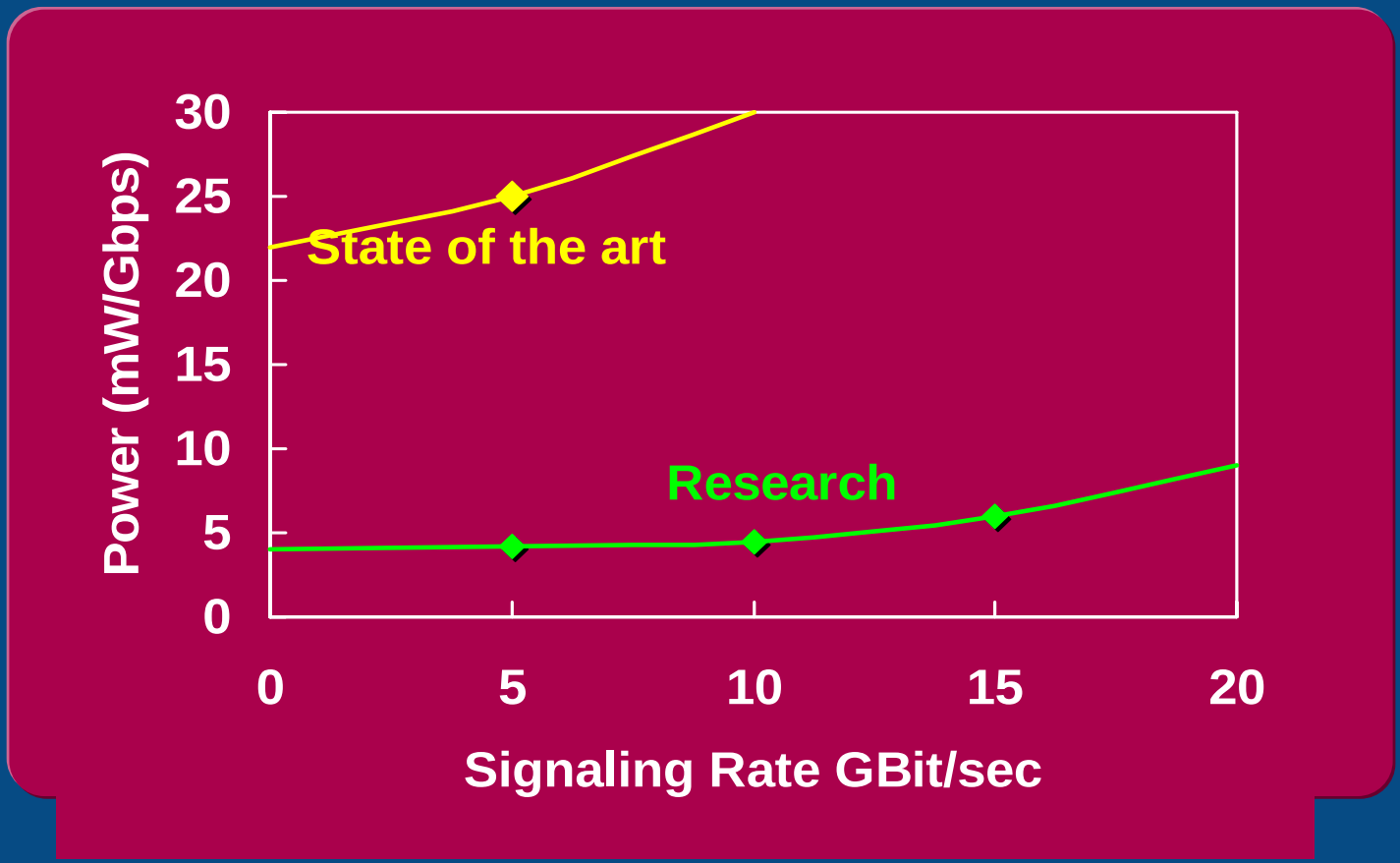
Busses have become wider to deliver necessary memory BW (10 to 30 GB/sec)

Yet, memory BW is not enough

Many Core System will demand 100 GB/sec memory BW

How do you feed the beast?

IO Pins and Power



State of the art:

100 GB/sec ~ 1 Tb/sec = 1,000 Gb/sec × 25mw/Gb/sec = 25 Watts

Bus-width = 1,000/5 = 200, about 400 pins (differential)

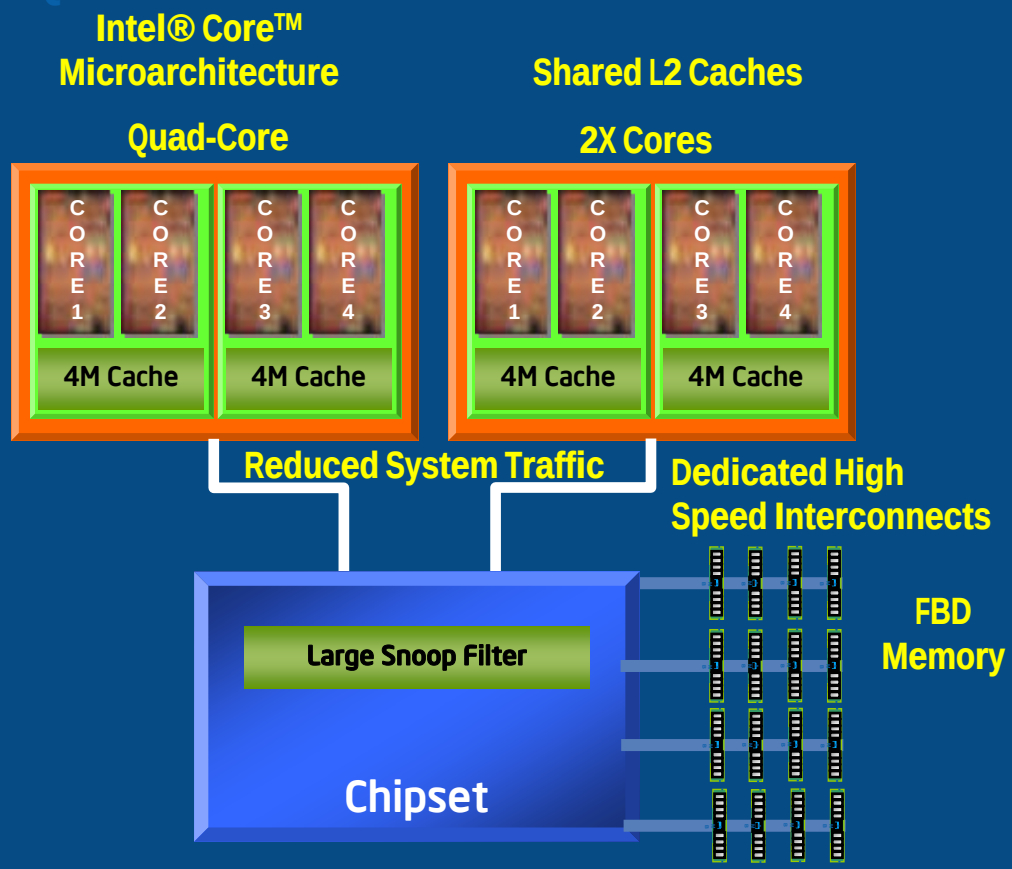
Too many signal pins, too much power

Agenda

- Multi-Core Momentum
- Multi-Core Performance Challenges
- Platform Architectures & Performance
- SPEC CPU2006 Sensitivity to Bandwidth
- Bandwidth & Performance Implications of Increasing Core-Count
- Workload Based Analysis
- Summary

2S Platform Architecture

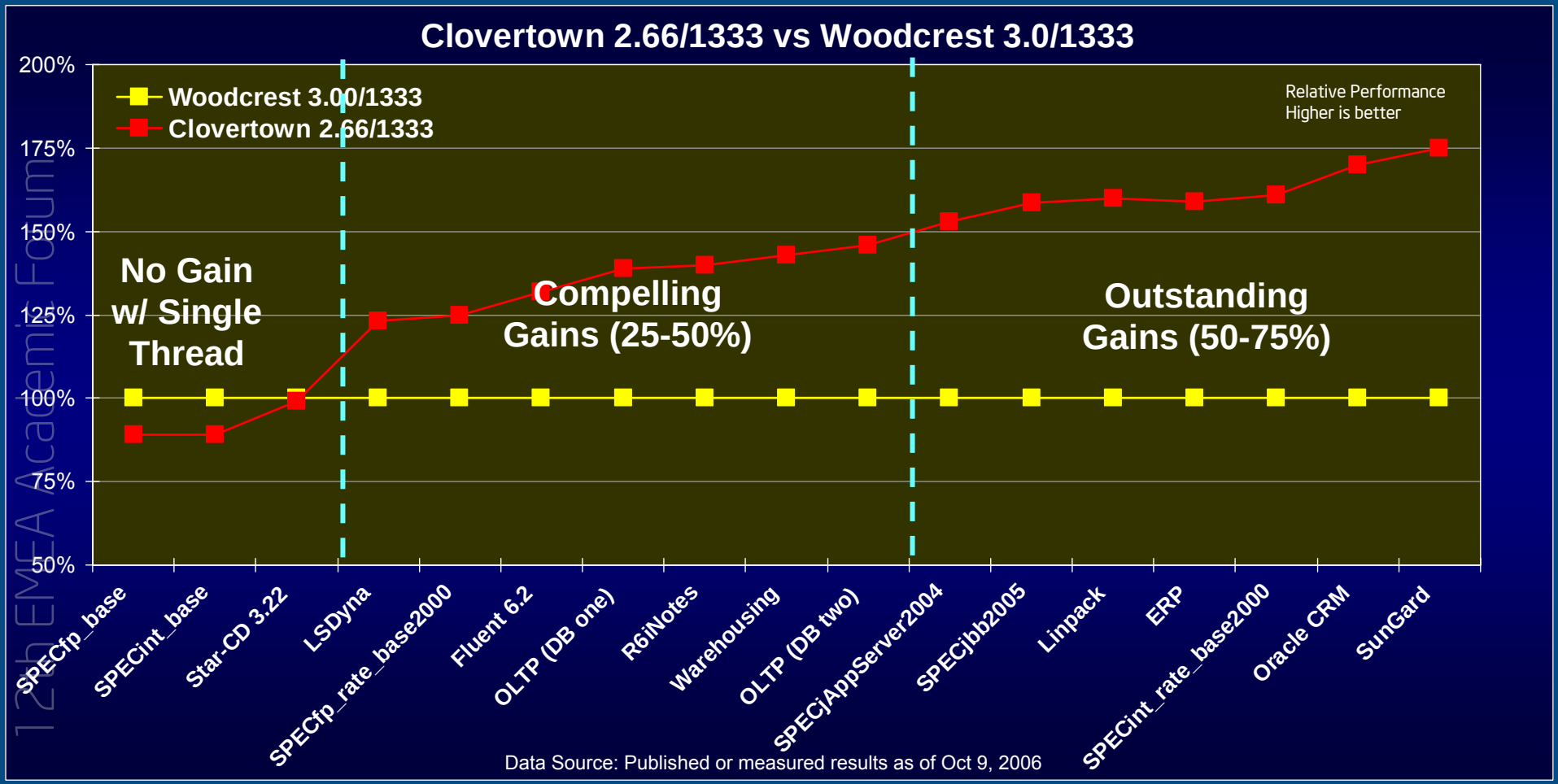
Supports Quad-Core Intel® Xeon® Processors



Enhanced Platform Capabilities Deliver Required Bandwidth for Quad-Core Performance Leadership

2S Clovertown (QC) Platform Performance

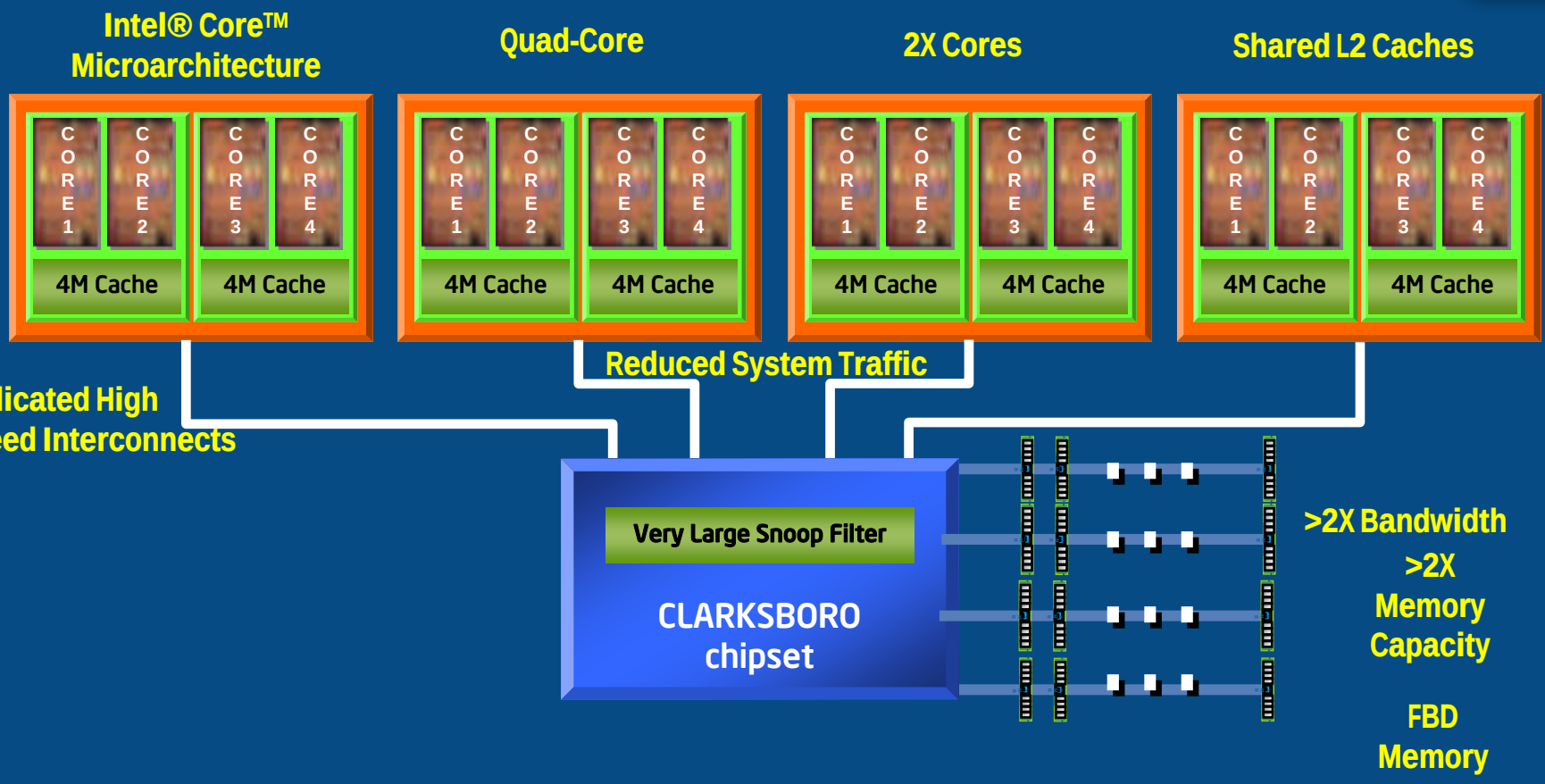
Comparison on a range of workloads



FSB-1333 enables Quad-Core Clovertown to deliver excellent gains on Multi-threaded workloads

4S Caneland Platform Overview

Quad-Core Intel® Xeon® Processor

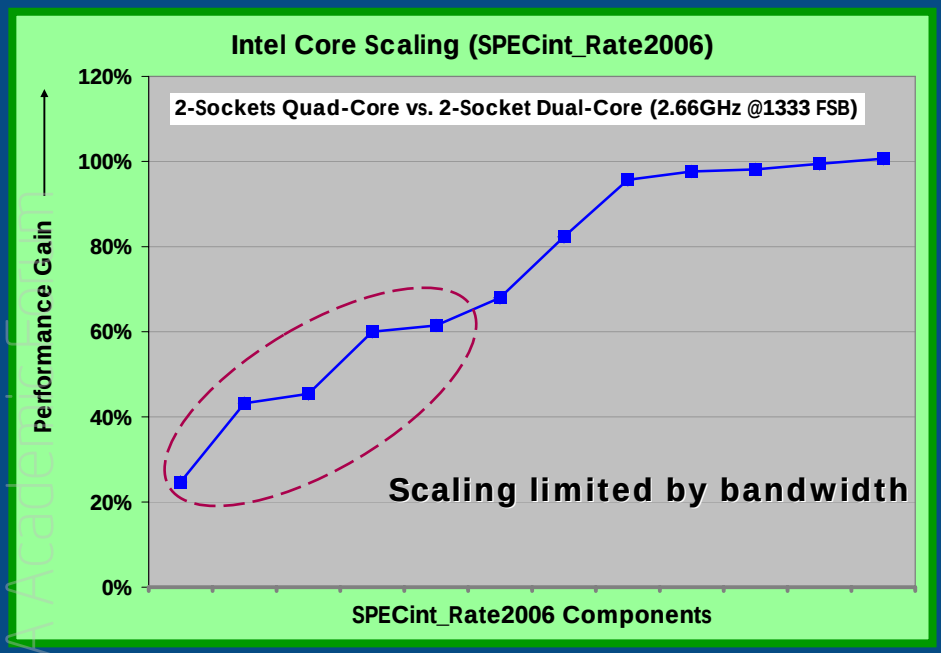


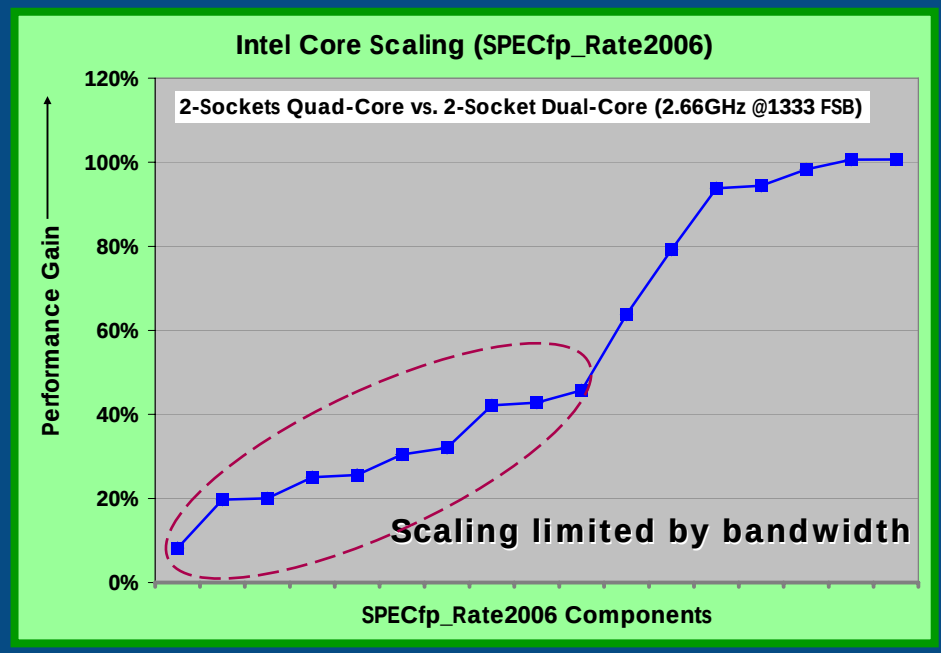
Enhanced Platform Capabilities Deliver Required Bandwidth for Quad-Core Performance Leadership

Agenda

- Multi-Core Momentum
- Multi-Core Performance Challenges
- Platform Architectures & Performance
- SPEC CPU2006 Sensitivity to Bandwidth
- Bandwidth & Performance Implications of Increasing Core-Count
- Workload Based Analysis
- Summary

2S SPEC CPU2006 – Core Sensitivity

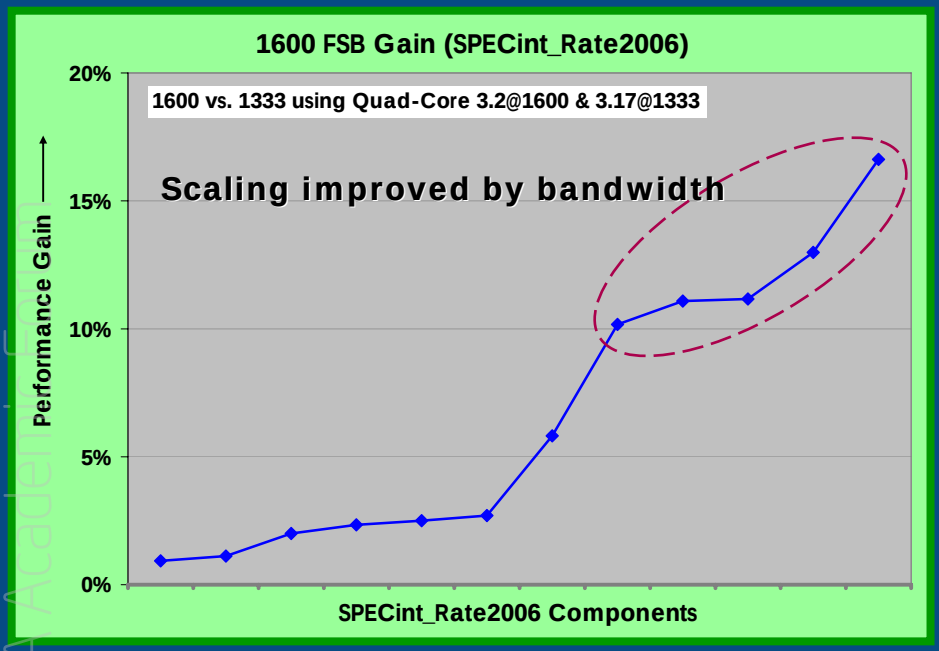


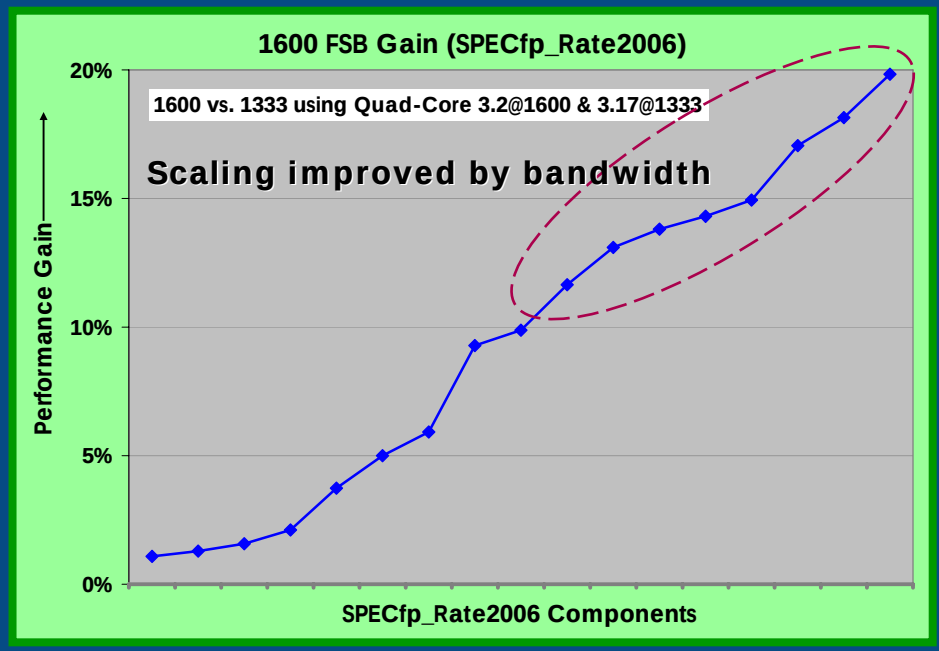


- **Most BW Sensitive SIR2006 Components**
  - **Xalancbmk, gcc, mcf, omnetpp, libquantum**
- **BW Sensitive SFR2006 Components Scaling**
  - **CactusADM, soplex, wrf, sphinx3, GemsFDTD, leslie3d, milc, bwaves**

Dual-Core to Quad-Core Scaling demands adequate bandwidth

2S SPEC CPU2006 – FSB Sensitivity*





* 1600 FSB & CPU Frequencies are lab experiment environment

- Most Bandwidth Sensitive SIR2006 Components
  - **Xalancbmk, gcc, mcf, omnetpp, libquantum**
- Most Bandwidth Sensitive SFR2006 Components
  - **CactusADM, soplex, wrf, sphinx3, GemsFDTD, leslie3d, milc, bwaves**

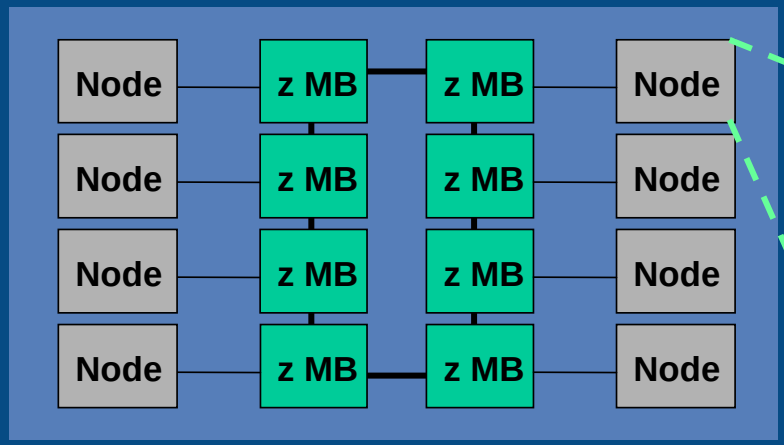
Increasing FSB bandwidth improves performance

Agenda

- Multi-Core Momentum
- Multi-Core Performance Challenges
- Platform Architectures & Performance
- SPEC CPU2006 Sensitivity to Bandwidth
- Bandwidth & Performance Implications of Increasing Core-Count
- Workload Based Analysis
- Summary

Multi-core* based on Traditional & Simple cores

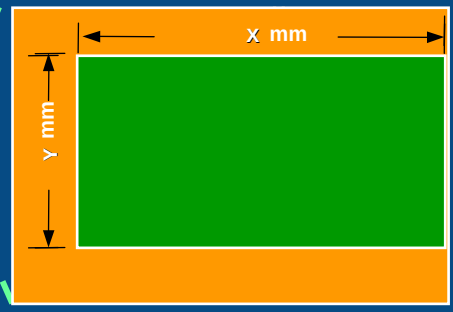
Assume Area (1 Large core) = Area (4 Small cores)



8 interconnected computing nodes, z MB Cache blocks, one per node, for 8z MB total on-die LLC

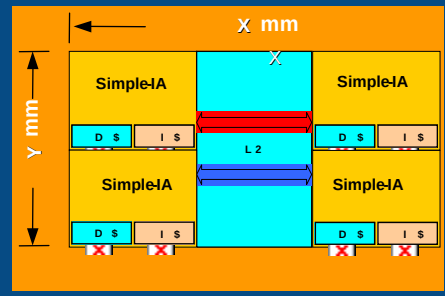
*This is Hypothetical with no Product plans

Large-core



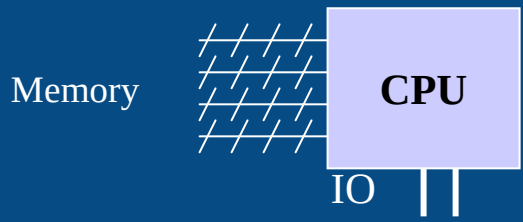
OR

Small-core



Large-core-MC: 8 Large cores (LcMC)
Small-core-MC: 8 x4 = 32 Small cores (ScMC)

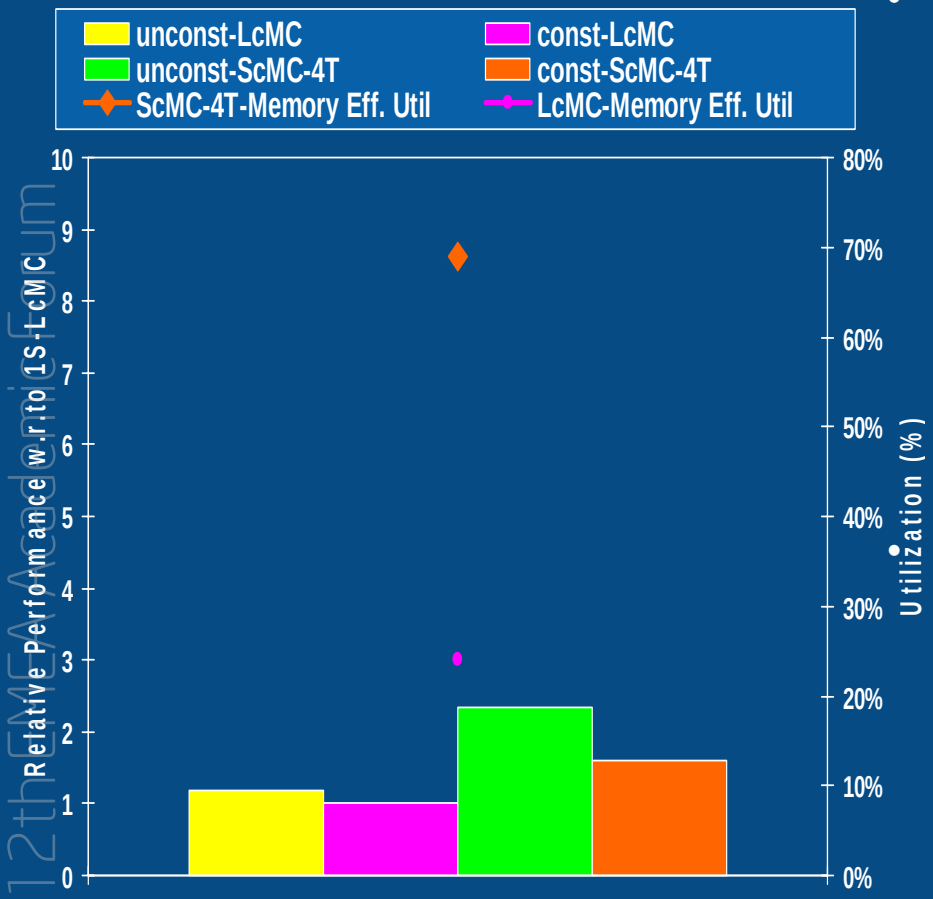
1Socket Platform Configuration



- Rest of uncore not shown
- 130W socket power envelope

- Assume no on-die bottlenecks
  - All queues, trackers etc will be sufficiently sized
  - Adequate on-die interconnect & cache BW, etc.

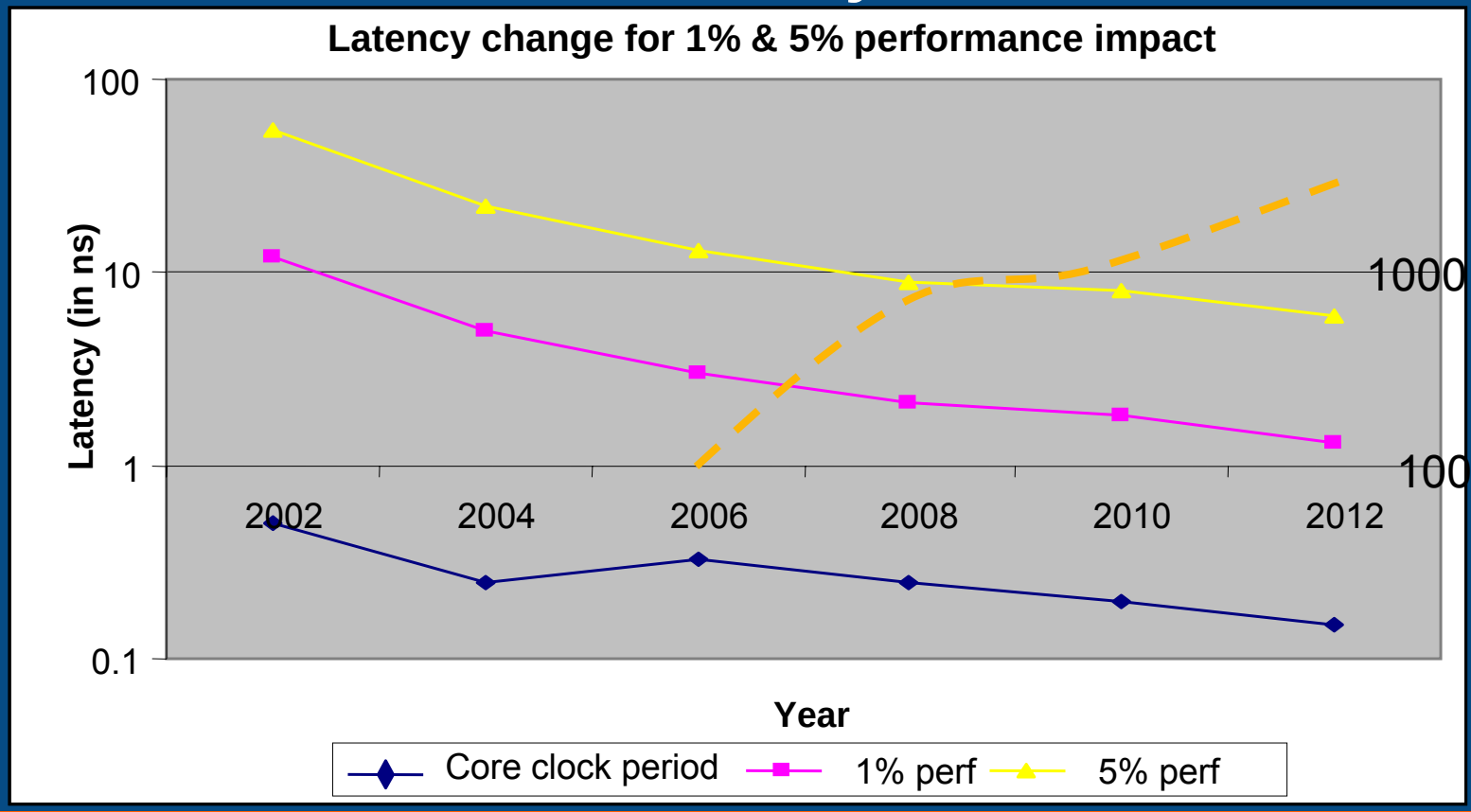
OLTP Performance: Unconstrained vs. Constrained



- Unconstrained : No power/pin-count constraints
  - Cores at max design core frequency
  - Adequate interface bandwidth
    - *Effective utilizations in ~40%-50% range*
    - *Unconstrained-ScMC-4T*
      - 2.5 x memory BW as LcMC
      - 2 x IO BW as LcMC
      - **~2.5x the pin-count as LcMC**
  - Resulting in unreasonable socket power
    - *ScMC: ~300W*
    - *LcMC: ~260W*
- Constraints
  - 130W power envelope
    - *Constrains core freq (lower by 33%)*
  - Pin-count limits constrains raw peak memory BW to ~50GB/s
  - *Lowers perf by ~15% for LcMC (due to core freq lowering) & ~25% for ScMC (due to both core freq lowering & BW constraints)*

- ***ScMC-4T constrained to the same platform (pin-count, power & memory-size) as LcMC has ~1.6X the performance of LcMC.***
- ***With memory channels near saturation***

2S OLTP Performance Sensitivity to LLC Miss Latency



- As core clock speed increases, latency impact on performance increases, but latency hiding techniques can lower latency impact on performance.
- By 2012, 1.4 ns of latency could have ~1% performance impact. Note on-die interconnect throughput can increase to over 4 Terabytes/sec.

Agenda

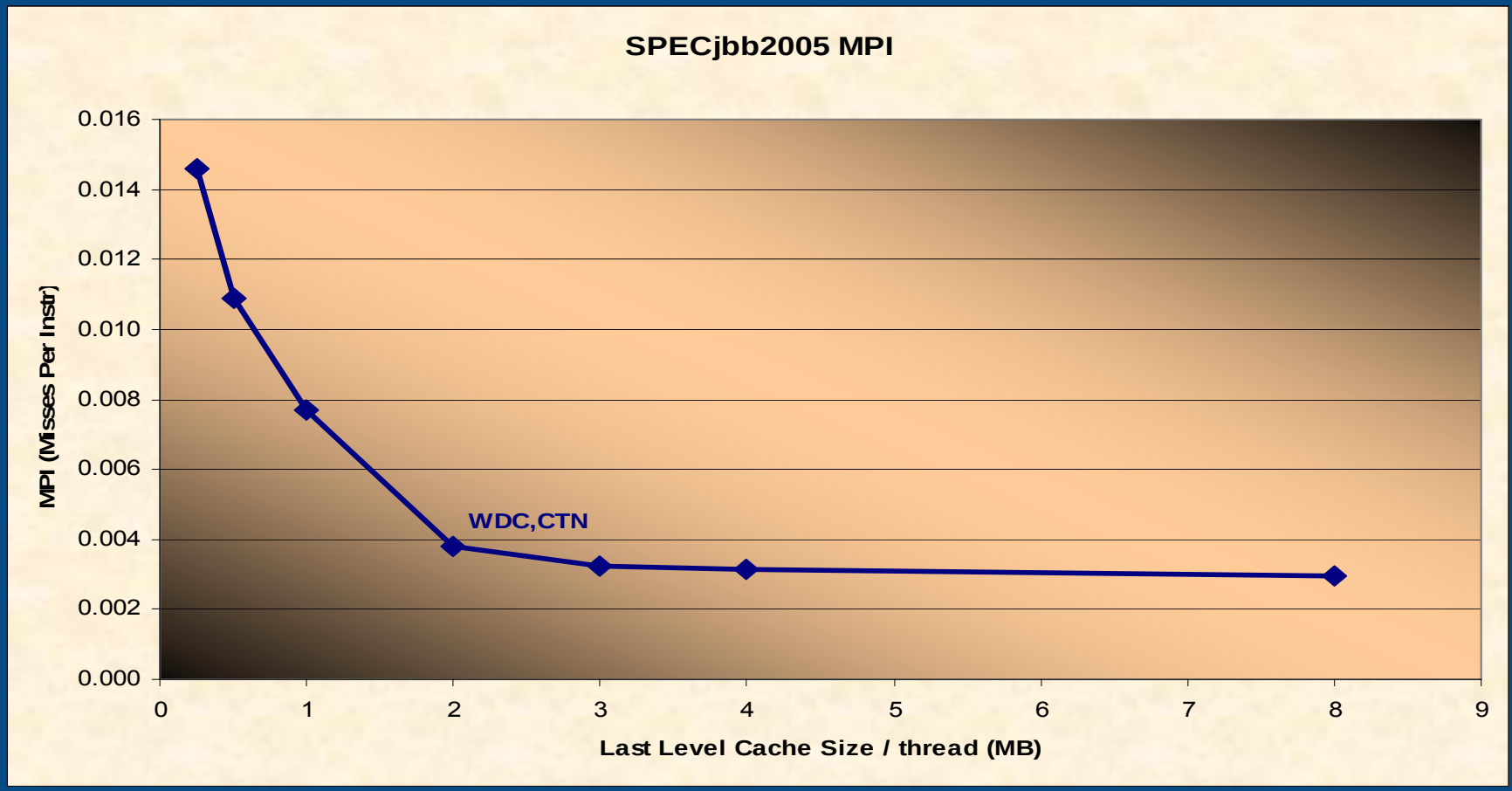
- Multi-Core Momentum
- Multi-Core Performance Challenges
- Platform Architectures & Performance
- SPEC CPU2006 Sensitivity to Bandwidth
- Bandwidth & Performance Implications of Increasing Core-Count
- Workload Based Analysis
- Summary

Workload-Based Platform Performance

- Execution Time is the product of
  - Path Length
  - Cycles Per Instruction (CPI)
  - Cycle Time
- CPI is the sum of
  - infinite-cache core cpi
  - miss rate * effective (*loaded*) memory latency
- Effective(loaded) memory latency is sum of
  - Idle latency
  - Queuing latency: driven by bandwidth
- Bad (good) news is that performance does not scale up (down) linearly with frequency

Three major components of performance drivers: core-cpi, latencies & bandwidths

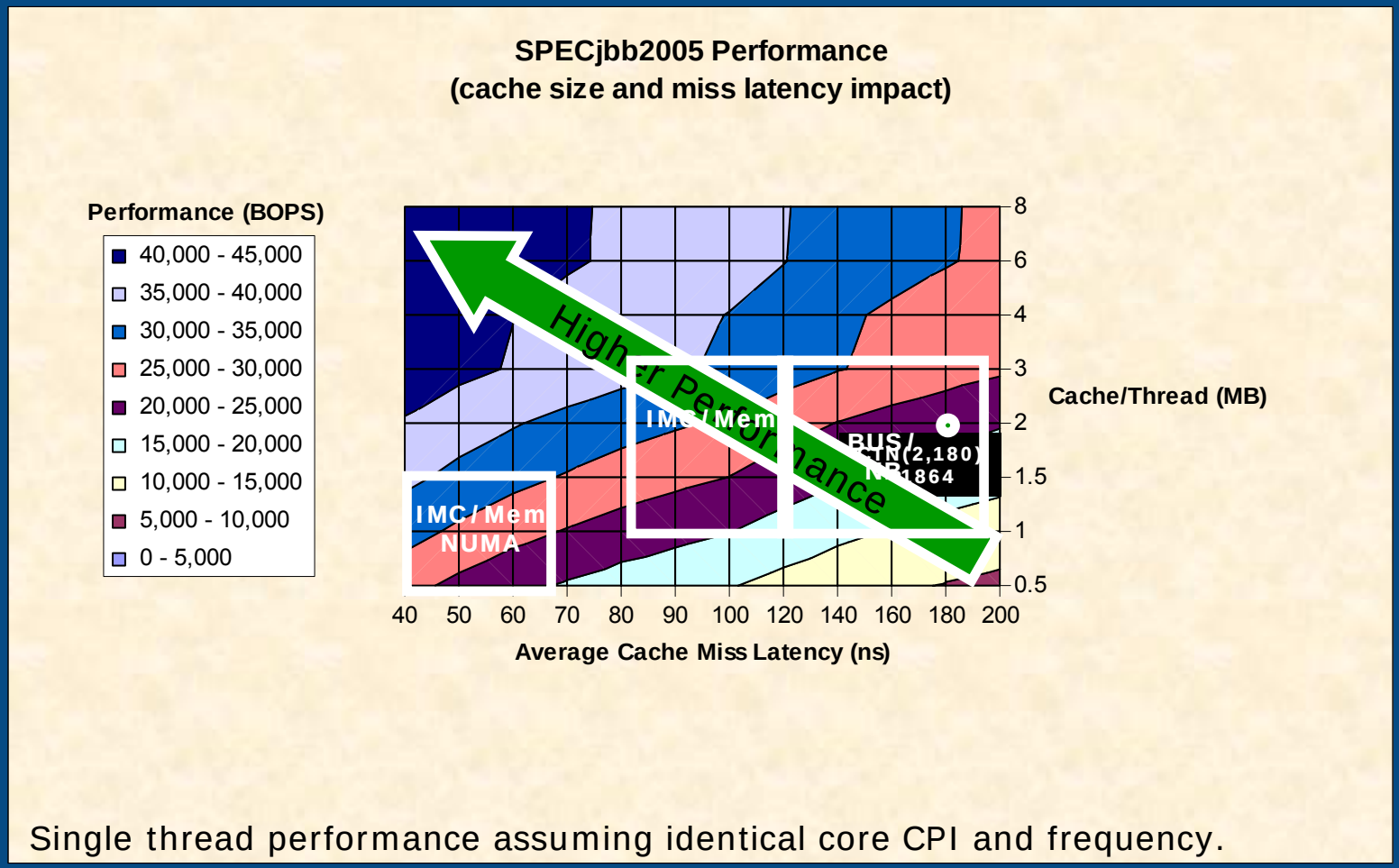
SPECjbb2005 Misses Per Instruction



SPECjbb2005 has very little sharing between threads. The shared code footprint is small (128KB) and most of LLC is partitioned between non-overlapping data sets per thread.

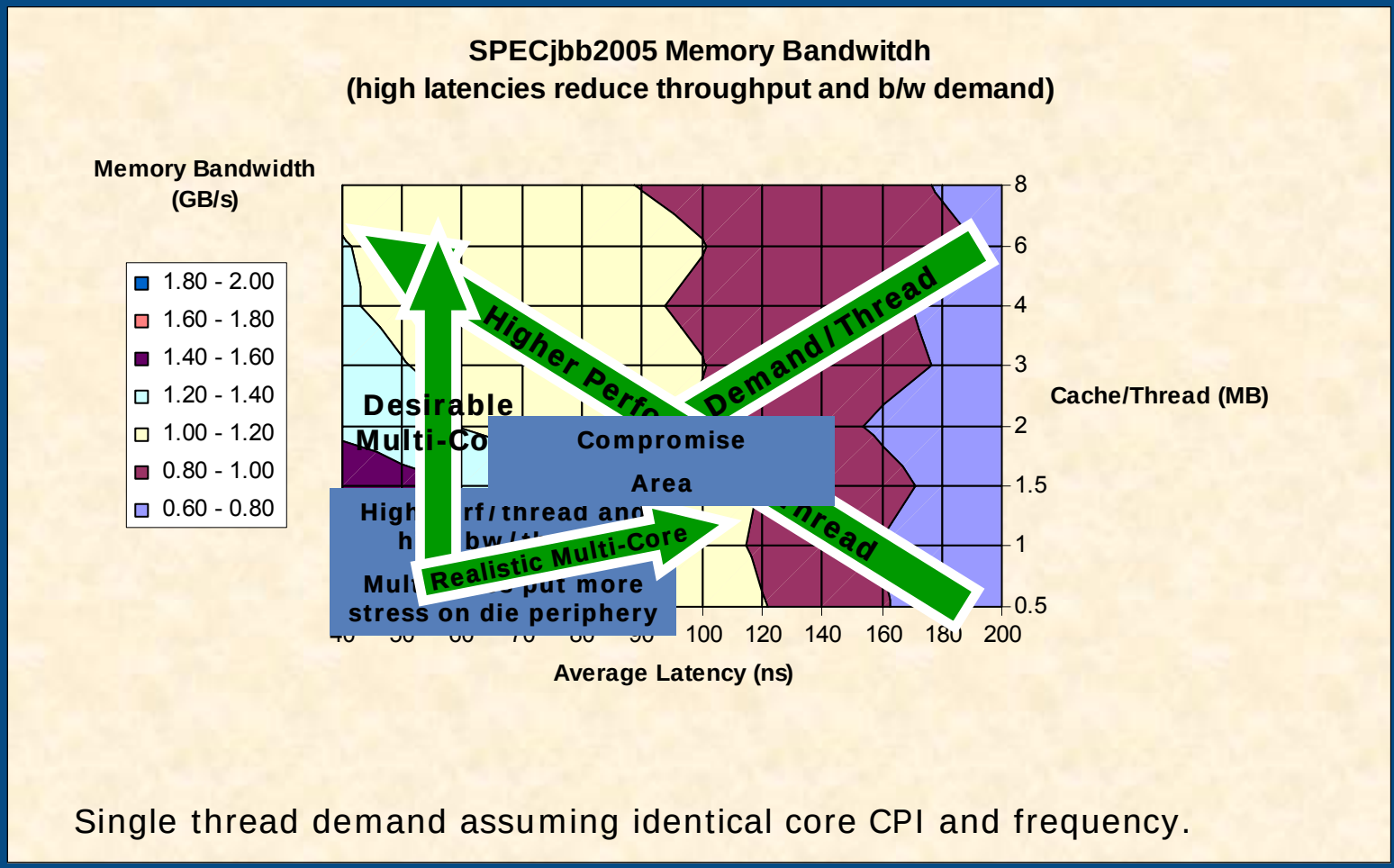
MPI sensitivity to cache size is very high up to 2-3MB. A segment of compulsory misses persists even for very large caches. The performance effect of these compulsory misses can be very damaging at high latencies..

Effect of Cache Size and Miss Latency on SPECjbb2005 Performance



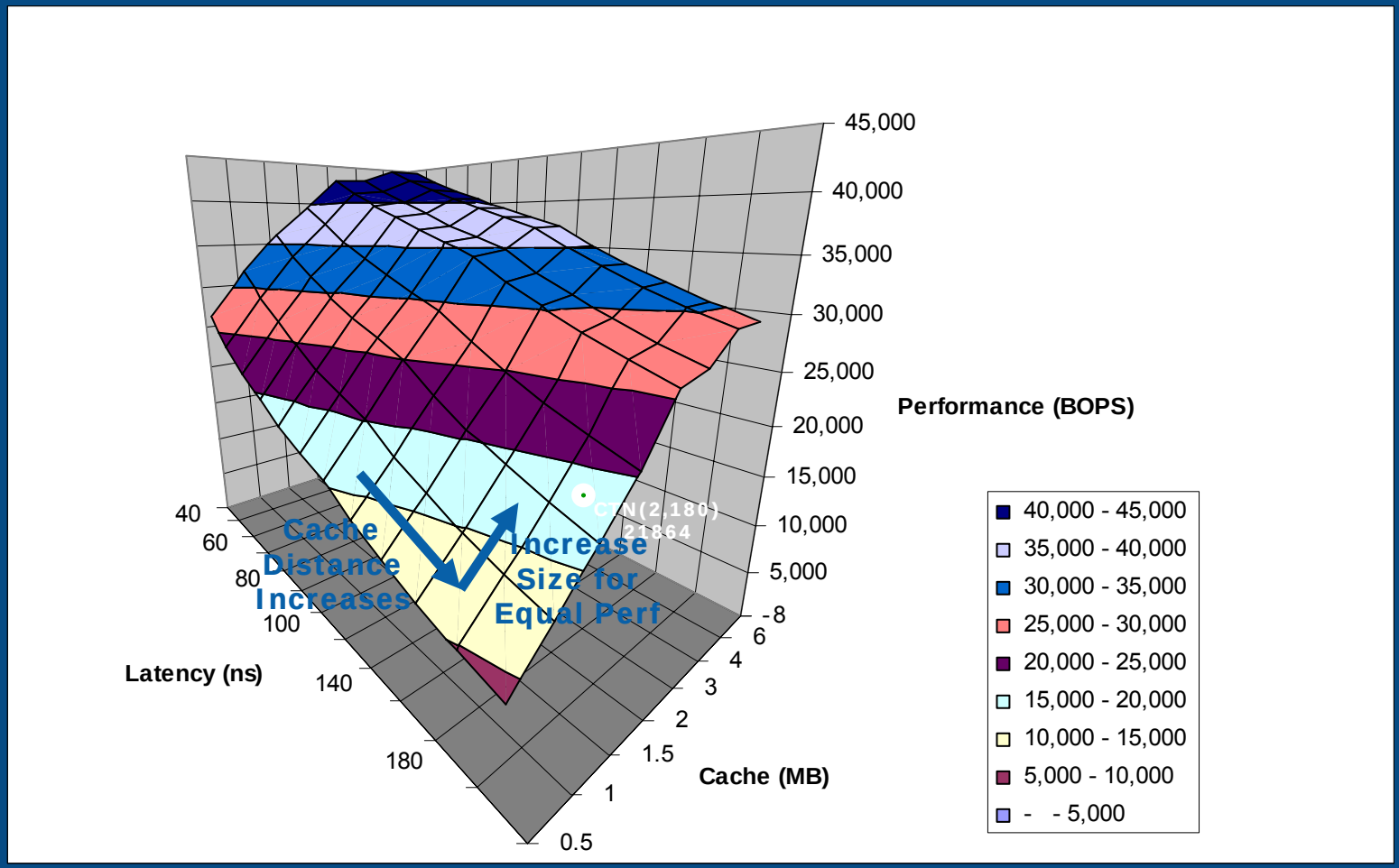
Single thread performance assuming identical core CPI and frequency.

Effect of Cache Size and Miss Latency on SPECjbb2005 Bandwidth Demand



Single thread demand assuming identical core CPI and frequency.

SPECjbb2005 Performance Drivers



Assumes a single-thread per cache

Agenda

- Multi-Core Momentum
- Multi-Core Performance Challenges
- Platform Architectures & Performance
- SPEC CPU2006 Sensitivity to Bandwidth
- Bandwidth & Performance Implications of Increasing Core-Count
- Workload Based Analysis
- Summary

Summary

- Performance growth will be driven by more multi-core products
  - Supported by great software tools to enable better application parallelization.
- Power will continue to be a major challenge for more performance delivery
  - I/O and on-die interconnects may dominate socket power.
  - Power reduction techniques research is critical
- On-die interconnects must scale to support the BW growth with min latency.
  - Latency sensitivity critical as ~1.4ns in latency will have a 1% perf impact in '12
- Platform bandwidth demand will continue to grow as more cores are added to the platform.
  - Many multi-threaded workloads demand higher bandwidth with multi-cores
  - May need to increase socket pin counts to mitigate slow BW/pin growth.

Acknowledgements

Shameem Akhter, Shekhar Borkar, Bruce Christenson,
Kathy Debnath, Ravi Iyer, Adrian Moga, Chitra Natarajan