# Amdahl's Law in the Multicore Era

Explain intuitively why in the asymmetric model, the speedup actually decreases past a certain point of increasing r.

The limiting factor of these improved equations and things that are ignored by them can be discussed in lecture.
Q: How different categories of multicore chips can affect the original Amdahl's equation?

Discuss why the graphs of speedup vs. the resources used per core, given in Figure 2, generally show either a performance loss or gain for symmetric multicores as more expensive cores are used while asymmetric multicores exhibit a maximum speedup at some intermediate point.

The paper quotes when perf(r) >r , improving resources will aid both sequential and parallel execution, when perf(r) < r , improving core performance will hurt parallel execution, explain how ?

Exam Question: Explain the speedup (graphical) trends that arise from Amdahl's models for speedup by comparing to the number of r BCEs (the resources of r BCEs) for symmetric, asymmetric, and dynamic multicore systems assuming f (the fraction of software that is parallelizable) is very near 1.

In your opinion what are the two most simplistic assumption made in the paper while generating the results that can severely impact any conclusion drawn out of the results in the paper?

Consider a multicore processor system with n cores of r BCEs each. Let n=4, r=4. Explain why or why not a speed up of S(r) < r per core is cost-effective.

One possible exam question would be to show how to account for the effect of memory overhead in the different configurations.

Since dynamic multicore processors and asymmetric multicore processors can offer performance greater than the performance of symmetric processors, why are we designing symmetric multicore processors such as core 2?

In an asymmetric multicore, what happens to the optimal design point as parallelism in an application increases (shift to more or less powerful "dominant" core)?

What are the positive aspects inherited by dynamic multiprocessors from symmetric and asymmetric processors? Will the dynamic multiprocessor performance grow indefinitely with increasing number of processors and why?

A design engineer/researcher should look for methods to enhance sequential performance even if they come out to be locally inefficient. Justify.

By augmented Amdahl's law, try to explain 1) why few powerful cores on a multicore chip may show more performance improvement than many base cores on the same chip, and 2) why asymmetric and dynamic multicore chips show more performance improvement than symmetric ones.

**********

# Algorithms for Scalable Synchronization on Shared Memory Multiprocessors.

Explain how page/variable migration helps multiprocessor synchronization.

What parameters mainly result to memory and interconnect contention when synchronization constructs are used? How does architecture contribute to better performance of certain constructs?

What two negative characteristics of previous synchronization algorithms did the authors see as their biggest limitations and seek to eliminate when designing their algorithms?

Explain the Spin lock methods and more specifically MCS method. Compare the MCS method with previous Spin lock methods.

Q: There are two categories of synchronization constructs: blocking, and busy-wait. Describe each of them briefly. Also, describe when busy wait is preferred to another.

Describe the MCS list based Queueing Lock algorithm and explain how to ensure correctness if ISA does not have atomic compare_and_swap instruction.

How the MCS lock reduces network traffic and maintain fairness as well?

Explain why the compare_and_swap instruction is essential to correct functionality of the original MCS lock implementation. Further explain how the algorithm has to be modified if named instruction is not available.

A possible exam question would be to name the different spin-lock or barrier algorithms that are proposed in the paper and explain its differences. Another question would be to just explain the new algorithms that they propose and why are they different from previous barrier/spin lock existing algorithms.

Why is ticket lock preferred over MCS lock when a fetch_and_store is not available?

What are the bottleneck in today's interconnect of multiprocessors and what are the possible solution for them.
2. In modern processor which is more important single core performance, interconnects traffic or the scalability for the software people. Justify your answer.

A shared memory architecture consisting X processors, N network switches , supporting { x,y,z} atomic operations and uses distributed cache coherence. Propose a suitable synchronization mechanism and justify your choice.

What advantages does the MCS lock devised by the authors offer over the test_and_set, ticket, and array-based queuing locks?

Exam Question: The authors' new list-based queuing lock algorithm requires an atomic fetch_and_store instruction, and would also benefit from an additional instruction. What is this instruction and why would it help? What is theoretically bad about not having it?

What are the characteristics of scalable synchronization algorithms? In which systems is MCS more efficient in comparison with other synchronization methods?

If the algorithms using local-accessible flag variables can improve the performance of shared-memory multiprocessors much more, why we still need basic test_and_set spinlock or centralized barrier algorithms?

1) Describe the changes to the MCS algorithm if compare_and_swap is not supported in the hardware.
2) Is fetch_and_store a requirement to MCS algorithm ? If not, discuss the alternatives.

Under what circumstances the FIFO based approach, for lock acquisition, employed in some queue based algorithms

may not be very effective?

Exam Question: How does the MCS lock algorithm guarantee that locks are acquired in FIFO order if atomic compare_and_swap instruction is available, but not guarantee the same if only atomic fetch_and_store is available?

Exam Question: Design a spin lock synchronization technique using a queue and describe the advantages.

---

# Memory System Characterization of Commercial Workloads

Explain why dirty miss latency increases with the number of processors/cores?

a.) What are the impacts of application scaling/simplifications for workload characterization (wrt monitoring and simulations), according to the paper?
b.) Describe the technique used by Barraso et al. to characterize the memory system of commercial workloads.

How do you think current multicore architectures (where multiple cores exist on a single chip) would perform if this study were to be performed on them?

Q: Explain the effect of larger cache on communication (coherence) misses and dirty misses? Also explain the effect of block size on false sharing?
Q: A designer claims that an SMP system with very high performance memory hierarchy can improve run-time of every commercial workload. Discuss his claim based on cost-effectiveness of using the proposed SMP?

1) What is the difference between a multiprocessor specified for commercial workloads and scientific workloads in terms of core and memory model?

2) Which specifications should be considered in design of multiprocessor servers to support OLTP database application?

3) Which specifications should be considered in design of multiprocessor servers to support DSS database application?

During design of a multiprocessor system, if you are given opportunity to modify only one parameter in the whole memory subsystem, which thing you will modify so that the designed system benefits the performance of running OLTP, most ?

Explain why or why not the local miss rate (vs. global miss rate) is important to describe the performance of a cache hierarchy. Is the characterization presented in the paper general enough to make the presented trends the basis for optimization? Why or why not?

What is hard about characterizing memory for commercial workloads? How did the authors work around the problem?

A possible exam question could be to specify what would be the challenges for the proposed workloads in multicore systems in which usually the L2 is shared instead of sharing memory at main memory level. And what other commercial workloads benefit from the use of multiprocessors nowadays.

How would a memory system optimized for OLTP differ from one optimized for DSS or Web index search?

Outline the Nature of OLTP, DSS, Web search workloads, measures to tune the performance of its memory system. Also discuss the scalability issues with these tuning measures.

1) What is the impact of cache size and number of processors on memory system performance for OLTP workloads?
2) Why is scaling of the workloads done for full system simulation? Why scaling has little or no affect on memory system characteristics according Luiz, Kourosh and Edourd?

1. What makes the commercial workloads to be so complex to characterize them.
2. What are the tradeoffs in having larger off chip cache in multicore era.

What features of these applications are important in designing a multiprocessor system?

Currently the commodity multiprocessors have been further improved compared to the time this paper is written. How these new technologies (such as cache hierarchy, processor patterns, etc) may change the memory characteristics of workloads mentioned in this paper?

Exam Question: If you had to pick one solution in order to increase the performance of a database application that receives many small transactions from many different users would you increase the cache size or decrease the latency of the cache?

Why may systems that are optimized for OLTP versus DSS lead to diverging design points?

What are the possible trade-offs a designer needs to keep in mind while designing an OLTP commercial workload? Or stated otherwise Why fraction of dirty misses increases with an improvement in memory system design for OLTP commercial workload?

Q: Which type of cache misses are the major contributors to memory performance of OLTP workloads and how can they be reduced?

**********

# A Study of Performance Impact of Memory Controller Features in Multi-Processor Server Environment.

Describe the various memory controller features presented in paper by Natarajan et al. and their suitablity to the application domain. Give examples of situations.

Do you think any of these policies can be extended to a shared bus connecting L2 caches? What changes might be needed?

1. Explain why rank interleaving reduces read latency for server benchmarks.
2. What policies show better performance in an OOO memory controller according to Chitra Natarajan?

Given features such as ooo fsb vs inorder fsb, page open vs closed access, overlapped vs non overlapped scheduling, enabling intelligent reads and writes, delayed writes for your memory controller, chose the set of features that would best suit multimedia/graphic workloads. Justify your choice

Describe the operation of and the motivation behind the intelligent read-to-write switching memory controller described in the paper. What impact does the implementation of this feature have on performance and why?

Exam Question: How can delayed write scheduling in a O-o-O memory controller reduce loaded latency and improve bandwidth utilization for server workloads in a shared memorymultiprocessor system?

Describe page-open policy, page-close policy, Non-overlapped scheduling, and overlapped scheduling in memory controller.

What are the different possible options in the design of a memory controller? What are their associated pros & cons?

Why prefetching can hurt performance if the memory controller implements open-page policy with non-overlapped schedule?

Exam Question: Explain how memory is organized into banks and the idea of pages. What is precharging? What is meant by a page hit/miss? How does this effect latency and sustained bandwidth?

What may be the impacts on performance if the memory controller with all features described in this paper is used but the memory consistency should be maintained?

Question: The authors claim that various granularities are supported. The mechanism for that is setting the mark bits for all the blocks belonging to the given object. When can this be a problem?
Answer: It is often hard to determine the boundaries of an object. For example, in order to lock an entire linked list, all the nodes must be traversed and locked appropriately. This can be a problem.

According to the paper, which microcontroller policies impact the server application performance and how?

How can the evaluations presented in this paper be extended FB-DIMMs? Unlike parallel bus architecture like DDR2 or DDR3, FB-DIMMs are serial point to point architectures. Is there going to be a significant difference in bandwidth consumption? How does the serial interconnect affect latency?

A possible exam question would be to explain the characteristics in a MP memory controller that make it more suitable for server multiprocessor type workloads.

Q: Elaborate why does CPU prefetching degrade performance in systems with page open policy?
Q: Describe the advantage of OoO FSB as compared to in order FSB.

Explain the difference between in-order and out-of-order read/write policy used in memory controller design.

What the four important method suggested by the paper to increase the memory system performance.

Exam Question: Given a certain overhead for bank conflicts and read write switching in memory design a policy for switching from reads to writes.

What features can optimize a desktop-based memory controller for MP server environments?

---

# CMP Design Space Exploration Subject to Physical Constraints

A possible exam question would be to show how to modify the different parameters of the multiprocessor (L2, pipeline length and width, number of cores) to obtain the maximum performance under different physical constraints and different types of workloads.

Exam Question: Based on the paper, compare the impact of area constraints versus the impact of thermal constraints in multi-processor systems. What will happen in terms of pipeline dimensions? Which has greater influence? What will happen in terms of the complexity of the cores, number of cores, and size of caches?

Q: Thermal constraints incur both shallower and narrower cores. Each of these two factors has a different impact on area and power. Briefly, compare the effects of a shallower core and a narrower core on power and area.

What effect does the imposition of stricter thermal constraints have on the optimal values of architectural parameters, such as pipeline width and depth, number of cores, and L2 cache size? When is voltage and frequency scaling preferable to scaling the number of cores?

1. What are the possible ways to optimize the CMP design for area and thermal constraints.
2. Given the paper constraints how you come up with a design to have good single thread performance.
Hint: Use the Amdahl law paper trick.

Exam Question: Why does imposing constraints on power density shift the optimal pipeline depth to shallower design points?

1. How do pipeline depth and width vary under thermal constraints according to Yingmin Li et.al paper? 2. What methods are suggested in the paper to meet thermal constraints for CPU bound applications?

We know that the increasing the number of cores can increase power dissipation, so in presence of thermal constraint we can lower the voltage to balance the effect. Why can't we largely decrease the voltage & increase the number of cores to enhance performance? List the consequences & remedies. What can be done to annul the effects of increase in number of cores in the presence of thermal constraints?

Exam Question: Discuss the importance of adequate cooling for microprocessors.

1) CMPs are typically optimized for throughput oriented applications like OLTP. What are the design trade offs required to accommodate single threaded performance as single threaded applications like "Microsoft Word" still continue to be important.

Do you think that an adaptive architecture would make sense? Take into account the % loss numbers presented in the paper when using a CPU-bound optimum with a memory-bound workload, and vice versa ? list any concerns that might happen.

If we want to consider more realistic cache coherence mechanism and more general workloads in CMP design space exploration, how do we modify the simulation methodology used in the paper? How the area and power constraints may impact the performance in this way?

Exam Question:
a.) Describe the method of decoupled core and cache-interconnect simualtion introduced by the authors and explain the method invovled to improve accuracy of the simualtion results.
b.) In the paper by Li et al, the authors emphasize the importance joint optimization of parameters under given physical constraints. Briefly explain the parameters considered and the effect of physical constraints on them.

---

# Heterogeneous Chip Multiprocessor

A possible exam question would be to identify the benefits that heterogeneous CMP present for the common desktop user and the existing design challenges for those chips.

Why heterogeneous cores are expected to perform better than homogenous cores for the application that shows widely varying run time phase behavior?

Exam Question: Compare the throughput-performance characteristics of homogeneous chip multiprocessors to heterogeneous multiprocessors. When is it good to use one or the other? What makes more sense in a "realistic"

application setting? Why?

Exam question:
1. What are the different software and hardware challenges in designing the heterogeneous architecture.
2. How they mitigate the Amdahl's law.

Justify the need for heterogeneous cmps and compare them with homogenous cmps in terms of power and performance. Also Discuss the challenges involved in a heterogeneous cmp designs.

How does heterogeneous chip multiprocessors design mitigate Amdahl's law?

What kind of application may be less benefited from heterogeneous CMP than from a large and complex processor? How the shared-memory workloads complicate the design of heterogeneous CMP?

Q: Explain how the single ISA heterogeneous CMP's offer good power efficiency according to the Kumar et.al: paper?

Explain why heterogeneous CMP design is multicore aware architecture?

(From the side panel) Outline the main points the authors make about the importance of interconnect in CMP design.

Exam Question: How do heterogeneous CMPs exploit Amdahl's Law to provide power-and area-efficient performance?

*Exam Question*
What technique can be applied to multicore architectures to increase power efficiency that gives it a great advantage over the clock-gating and F/V scaling of uniprocessors?

1) What are the software overheads of heterogeneous CMPs?
2) Compare single-ISA heterogeneous CMPs with other multi-core models in power and performance aspects.

For heterogeneous CMP's, how the statically & dynamically scheduled cores be treated by the compiler?

In general, CMP design space is a function of what parameters that need careful consideration before design choices are made?

Q: Describe the throughput advantages of heterogeneous cores as compared to homogenous cores. Under which circumstances, homogenous cores show better performance and throughout?

In the paper by Kumar et al., what are the changes required in the software to exploit the benefits of a heterogenous design.

---

# Architectural Support for Software Transactional Memory

Exam Question: Compare/contrast the differences between HASTM and HyTM. In which cases would HyTM be able to perform better than HASTM and why?

Q: What are the advantages offered by the HASTM technique over conventional STM described in the paper? Explain in brief.

Would single thread performance be that important of a factor in bandwidth, throughput-focused server systems? Can the overhead for STM be overlooked?

Explain how HASTM's mark bits allow for aggressive mode, where read-set logging is skipped.

Q: Describe the main disadvantage of STM as compared to HTM and briefly discuss how HASTM can provide an answer to this issue.

What is the key idea behind HASTM? 2. Compare and contrast between HTM, STM and HASTM.

What aspects of HASTM usage give performance and flexibility advantage? Briefly describe each.

What are mark bits and mark counters? Explain how they are used to implement barriers in HASTM.

What does HASTM add to architectural state of a thread? How does it use them for commiting or aborting a transaction?

Exam Question: Explain the advantages and disadvantages of both hardware and software approaches to transactional memory.

If we modify HASTM so that we can keep marked bits of L1 cache lines after they are evicted into L2, how do the performance and overheads change in multiprocessor environment?

Compare and contrast STM, HASTM and Hybrid TM. Illustrate cases when each of them performs the best.

Why can aggressive mode HASTM optimize read barriers across atomic blocks but STM cannot?

Discuss the ISA extensions and demonstrate their usage to optimize the overheads in STMs for the common case according to the paper by Saha et al.

Even with the long running transactions( overflowing L1 cache), how HASTM can yield better performance than HyTM ( Hybrid TM) ?

Discuss the modes under which HASTM can operate. Elaborate why Aggressive mode is not always successful.

A possible exam question would be to mention the differences and benefits and disadvantages of STM, HTM, Hybrid TM and the proposed HASTM.

If you had significant locality which bits would you use for interleaving?