NTNU
Norwegian University of Science and Technology

PhD Defense Presentation

Managing Shared Resources in Chip Multiprocessor Memory Systems

12. October 2010

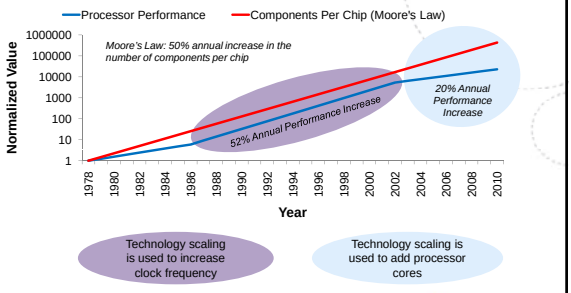
Magnus Jahre

Outline

- Chip Multiprocessors (CMPs)
- CMP Resource Management
- Miss Bandwidth Management
  - Greedy Miss Bandwidth Management
  - Interference Measurement
  - Model-Based Miss Bandwidth Management
- Off-Chip Bandwidth Management
- Conclusion

CHIP MULTIPROCESSORS

Historical Processor Performance



Technology scaling is used to increase clock frequency

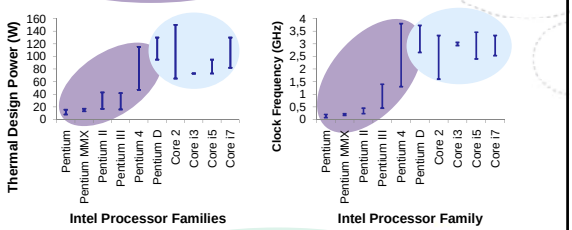
Technology scaling is used to add processor cores

Aggregate performance still follows Moore's Law

Power Dissipation Limits Practical Clock Frequency

Technology scaling increases clock frequency

Technology scaling adds processor cores

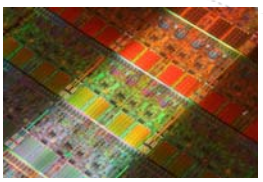


Power Wall

Source: Wikipedia, List of CPU Power Dissipation, Retrieved 02.06.10

Chip Multiprocessors (CMPs)

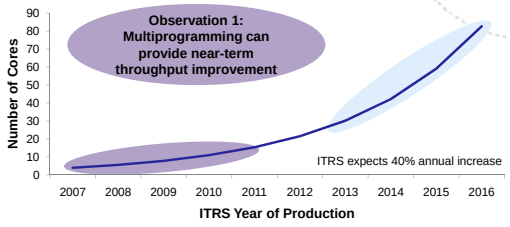
- CMPs utilize chip resources with a constant power budget
- How does technology scaling impact CMPs?



Intel Nehalem

7

Projected Number of Cores



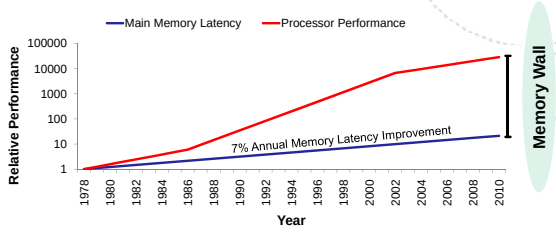
Observation 2: Software parallelism is needed in the long-term

NTNU Norwegian University of Science and Technology

www.ntnu.no

8

Processor Memory Gap



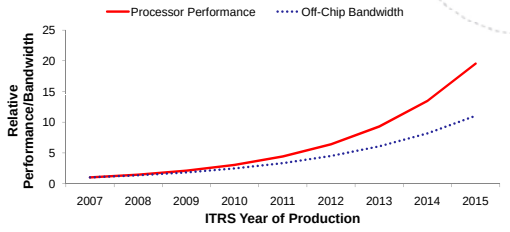
Observation 3: Latency hiding techniques are necessary

NTNU Norwegian University of Science and Technology

www.ntnu.no

9

Performance vs. Bandwidth



Observation 4: Bandwidth must be used efficiently

NTNU Norwegian University of Science and Technology

www.ntnu.no

10

Application Trends

Multi-programming

Software parallelism

Concurrent applications share hardware

Hardware Trends

Latency hiding

Bandwidth efficiency

Complex Memory Systems

Shared Resource Management

www.ntnu.no

11

CMP RESOURCE MANAGEMENT

NTNU Norwegian University of Science and Technology

www.ntnu.no

12

Why Manage Shared Resources?

- Provide predictable performance
- Support OS scheduler assumptions
- Cloud: Fulfill Service Level Agreement

NTNU Norwegian University of Science and Technology

www.ntnu.no

13

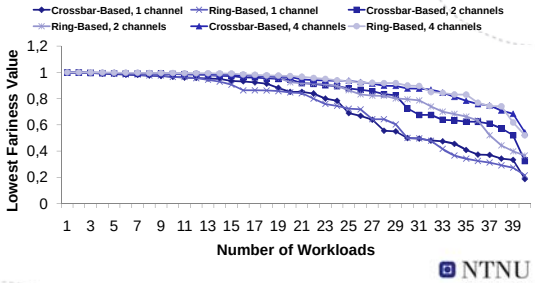
Performance Variability Metrics

- Fairness
  - The performance reduction due to interference between processes is distributed across all processes in proportion to their priorities
  - *Equal priorities:* Performance reduction from sharing affects all processes equally
- Quality of Service
  - The performance of a process is never drops below a certain limit regardless of the behavior of co-scheduled processes

NTNU Norwegian University of Science and Technology

14

Performance Variability (Fairness)



Paper B.I

NTNU Norwegian University of Science and Technology

15

Resource Management Tasks

Measurement

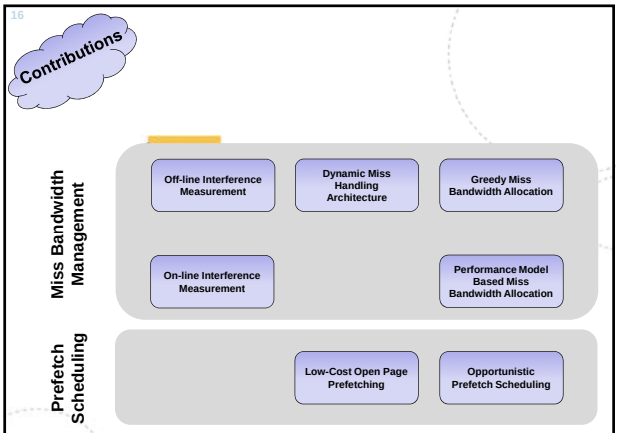
Allocation (Policy)

Enforcement (Mechanism)

NTNU Norwegian University of Science and Technology

16

Contributions



17

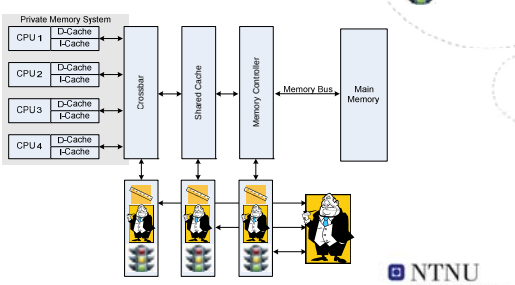
Miss Bandwidth Management

GREEDY MISS BANDWIDTH MANAGEMENT

NTNU Norwegian University of Science and Technology

18

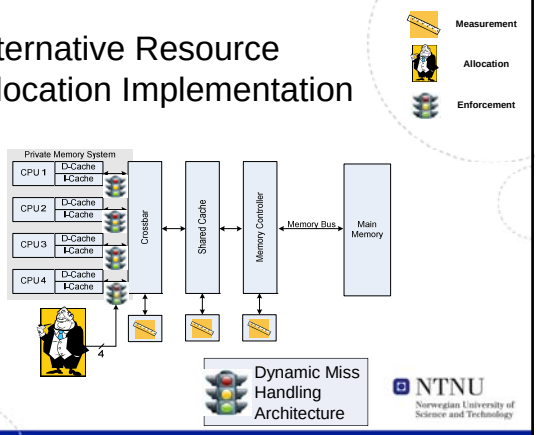
Conventional Resource Allocation Implementation



NTNU Norwegian University of Science and Technology

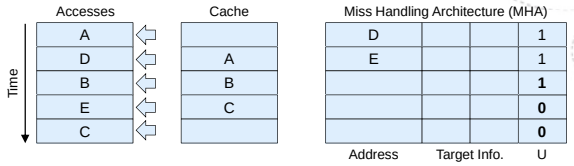
19

Alternative Resource Allocation Implementation



20

Dynamic Miss Handling Architecture



A DMHA controls the number of concurrent shared memory system requests that are allowed for each processor

21

Greedy Miss Bandwidth Management

- Idea: Reduce the number of MSHRs if a metric exceeds a certain threshold
- Metrics:
  - Paper A.II: Memory bus utilization
  - Paper A.III: Simple interference counters (Interference Points)
- Performance feedback avoids excessive performance degradations

Paper A.II and A.III

22

Miss Bandwidth Management

INTERFERENCE MEASUREMENT

23

Resource Allocation Baselines

Baseline = Interference-free configuration

Quantify performance impact from interference

Private Mode and Shared Mode

24

Interference Definition

$$I_i = L_i - \mathcal{L}_i$$

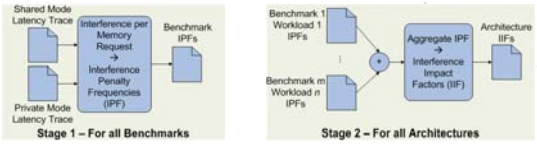
Interference = **Shared Mode Latency** − **Private Mode Latency**

$$E_i = \hat{\mathcal{L}}_i - \mathcal{L}_i$$

Estimate Error = **Private Mode Latency Estimate** − **Private Mode Latency Measurement**

25

Offline Interference Measurement



Interference Penalty Frequency (IPF) counts the number requests that experienced an interference latency of *i* cycles

Interference Impact Factor (IIF) is the interference latency times the probability of it arising, i.e. $IIF(i) = i \cdot P(i)$

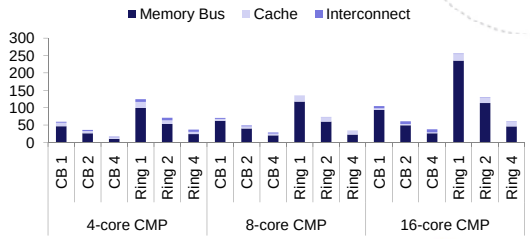
NTNU Norwegian University of Science and Technology

Paper B.I

www.ntnu.no

26

Aggregate Interference Impact



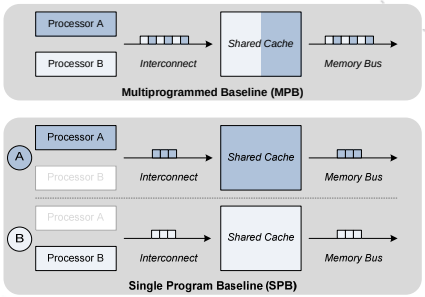
NTNU Norwegian University of Science and Technology

Paper B.I

www.ntnu.no

27

Resource Management Baselines



www.ntnu.no

28

Baseline Weaknesses

- Multiprogrammed Bas
  - Only accounts for inter ence i oned resources
  - Static and equal divisi of D M b dwidth does not give equal latency
  - Complex relationship bet urce allocation and performance
- Single Program Baseline
  - ~~Does not exist in shared mode~~

Online Interference Measurement
Dynamic Interference Estimation Framework (DIEF)

NTNU Norwegian University of Science and Technology

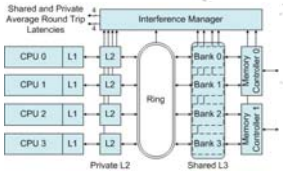
Paper B.II

www.ntnu.no

29

Online Interference Measurement

- Dynamic Interference Estimation Framework (DIEF)
- Estimates private mode average memory latency
- General, component-based framework



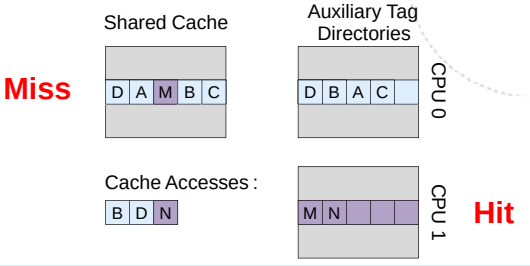
NTNU Norwegian University of Science and Technology

Paper B.II

www.ntnu.no

30

Shared Cache Interference



Eviction may not be interference

Eviction is interference

Interference latency cost = miss penalty

www.ntnu.no

31

Bus Interference Requirements

- Out-of-order memory bus scheduling
- Shared mode only cache misses and cache hits
- Shared cache writebacks

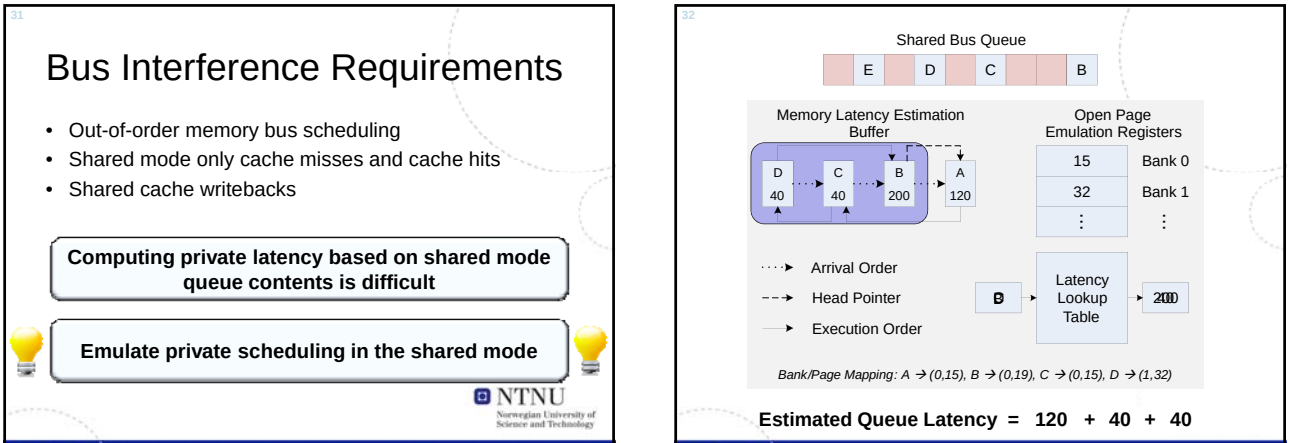
Computing private latency based on shared mode queue contents is difficult

Emulate private scheduling in the shared mode

NTNU Norwegian University of Science and Technology

www.ntnu.no

32



Estimated Queue Latency = 120 + 40 + 40

www.ntnu.no

33

Miss Bandwidth Management

MODEL-BASED MISS BANDWIDTH MANAGEMENT

NTNU Norwegian University of Science and Technology

www.ntnu.no

34

Model-Based Miss Bandwidth Allocation

DIEF provides accurate estimates of the average private mode memory latency

Can we use the estimates provided by DIEF to choose miss bandwidth allocations?

We need a model that relates **average memory latency** to **performance**

NTNU Norwegian University of Science and Technology

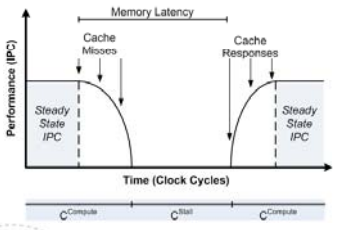
Paper A.IV

www.ntnu.no

35

Performance Model

Observation: The memory latency performance impact depends on the parallelism of memory requests



$$IPC_p = \frac{N_p}{C_p^{Compute} + C_p^{MemStall}}$$

$$C_p^{MemStall} = M_p \cdot L_p$$

Very similar in private and shared mode

Shared mode measurements can provide private mode performance estimates

Paper A.IV

www.ntnu.no

36

Bandwidth Management Flow

Measurement

Shared Mode Memory Latency

Private Mode Memory Latency

CPU Stall Time

Committed Instructions

Number of Memory Requests

Modeling

Per-CPU Models

Perf. Metric Model

Allocation

Find MSHR allocation that maximizes the chosen performance metric

Set number of MSHRs for all last-level private caches

NTNU Norwegian University of Science and Technology

Paper A.IV

www.ntnu.no

37

OFF-CHIP BANDWIDTH MANAGEMENT

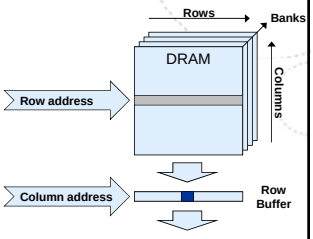
NTNU Norwegian University of Science and Technology

www.ntnu.no

38

Modern DRAM Interfaces

- Maximize bandwidth with 3D organization
- Repeated requests to the row buffer are very efficient



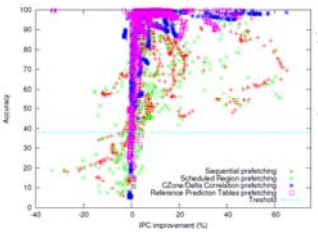
NTNU Norwegian University of Science and Technology

www.ntnu.no

39

Low-Cost Open Page Prefetching

- Idea: Piggyback prefetches to open DRAM pages on demand reads
- Performance win if prefetcher accuracy is above ~40%



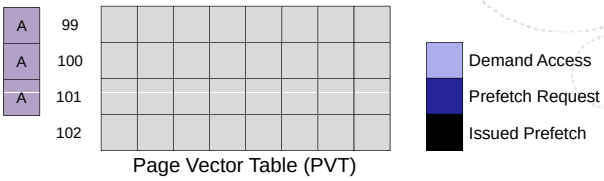
NTNU Norwegian University of Science and Technology

Paper C.I

www.ntnu.no

40

Opportunistic Prefetch Scheduling



Page Vector Table (PVT)

Idea: Issue prefetches when a page is closed

Increased efficiency: 8 transfers for 3 activations

NTNU Norwegian University of Science and Technology

Paper C.II

www.ntnu.no

41

CONCLUSION

NTNU Norwegian University of Science and Technology

www.ntnu.no

42

Conclusion

- Managing bandwidth allocations can improve CMP system performance
- Miss bandwidth management
  - Greedy allocations
  - Management guided by accurate measurements and performance models
- Off-chip bandwidth management with prefetching

NTNU Norwegian University of Science and Technology

www.ntnu.no

43

Thank You

Visit our website:
http://research.idi.ntnu.no/multicore/

NTNU Norwegian University of Science and Technology

www.ntnu.no

44

EXTRA SLIDES

NTNU Norwegian University of Science and Technology

www.ntnu.no

45

Future Work

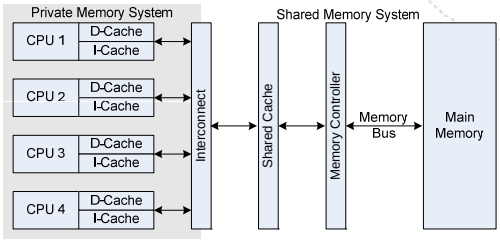
- Performance-directed management of shared caches and the memory bus
- Improving OS and system software with dynamic measurements
- Combining dynamic MHAs with prefetching to improve system performance
- Managing workloads of single-threaded and multi-threaded benchmarks

NTNU Norwegian University of Science and Technology

www.ntnu.no

46

Example Chip Multiprocessor



NTNU Norwegian University of Science and Technology

www.ntnu.no