Evaluating the Reuse Cache for mobile processors

Lokesh Jindal, Urmish Thakker, Swapnil Haria
CS-752 Fall 2014
University of Wisconsin-Madison

Executive Summary

Problem :

Mobile SOCs – Area is money!

Cache Area = 20-30% of SOC die area

Solutions?

Technology Scaling

- Expensive, diminishing returns

Reuse Caches

- Reduced cache size, comparable performance

Results=>

50% Area reductions for 5% performance loss

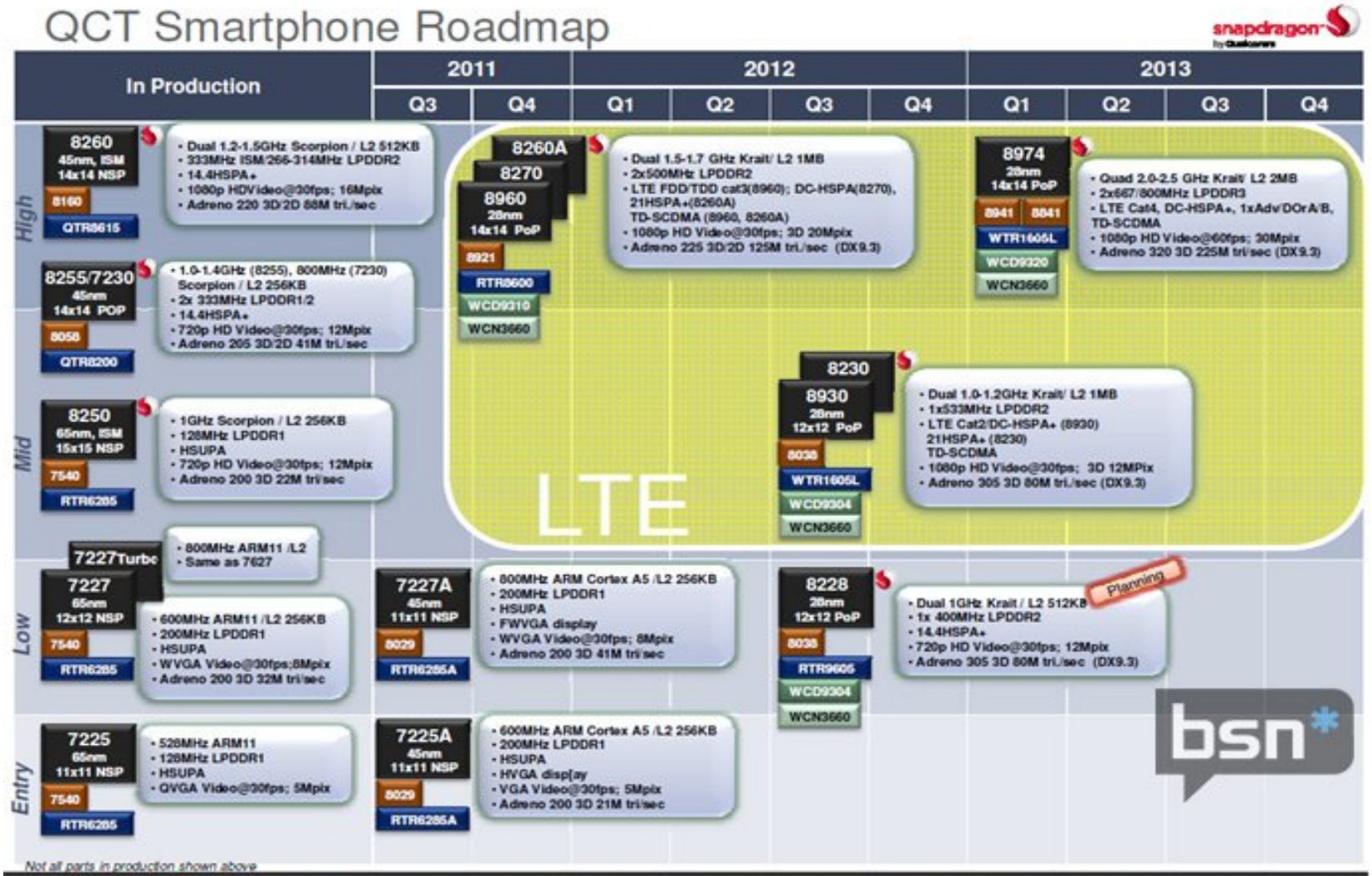
Outline

- Motivation
- Implementation
- Methodology
- Performance Evaluation
- Future Directions and Conclusion

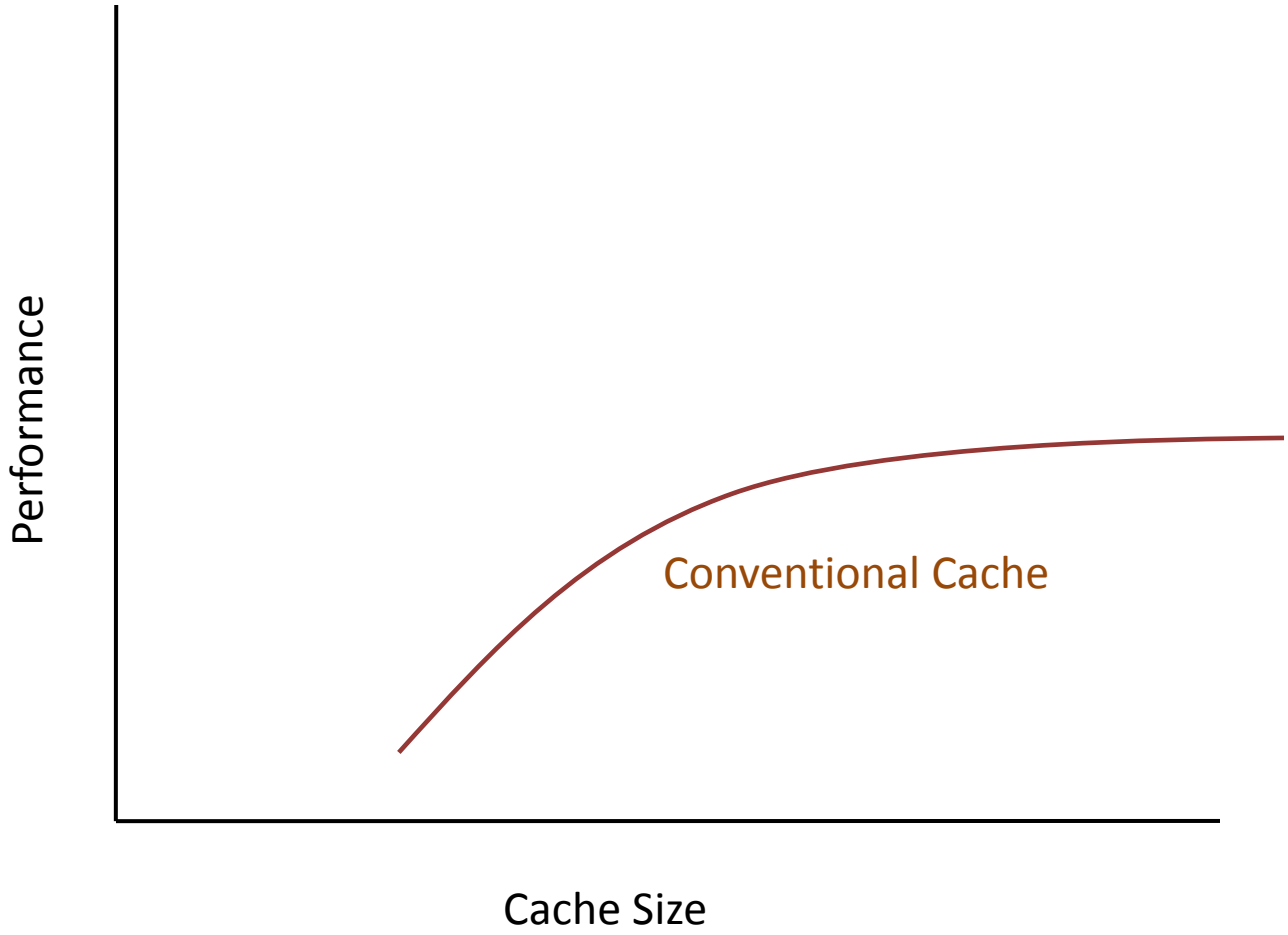
Outline

- Motivation
- Implementation
- Methodology
- Performance Evaluation
- Future Directions and Conclusion

Caches in mobile SOCs

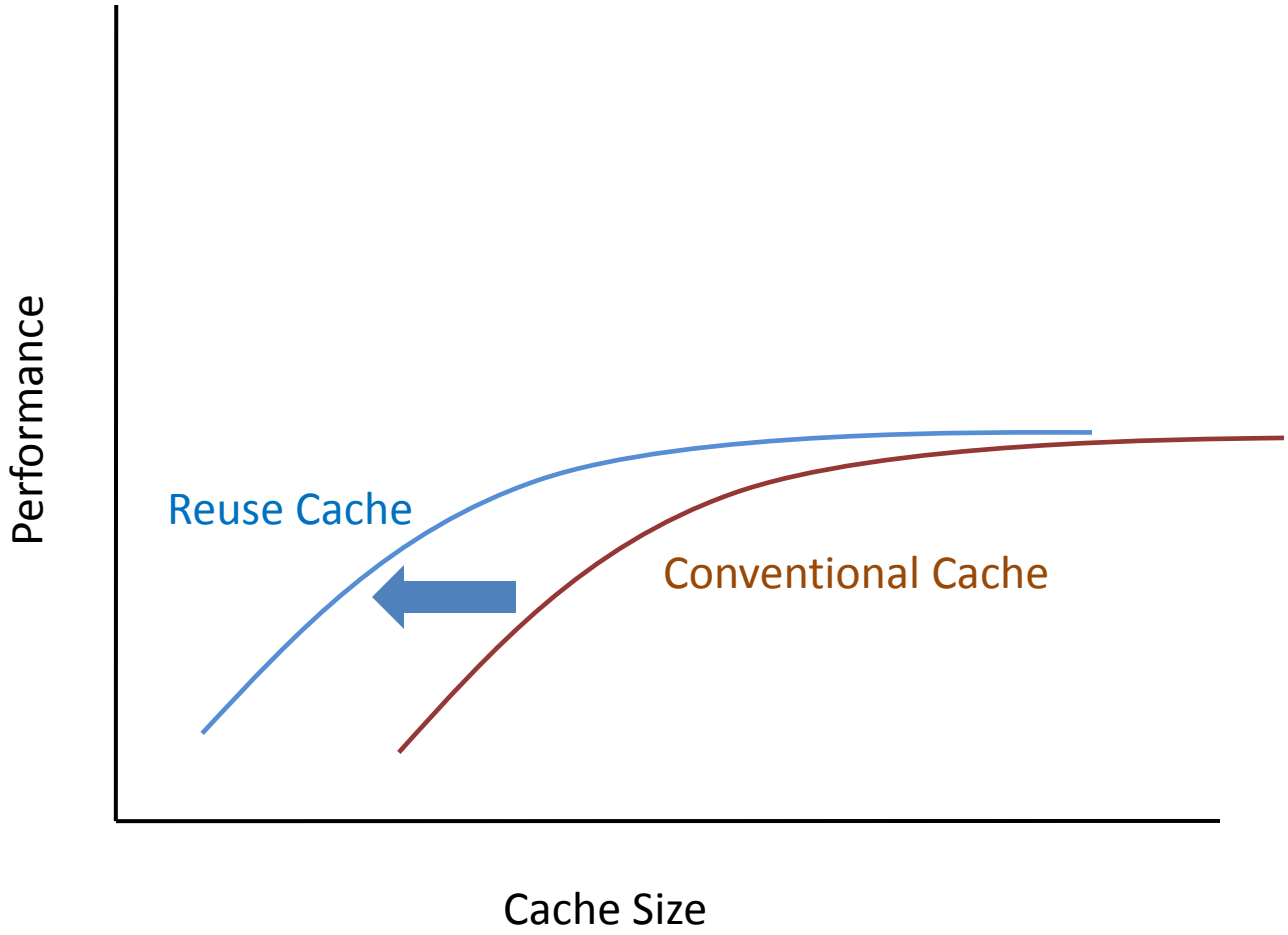


Area-Performance Tradeoffs*



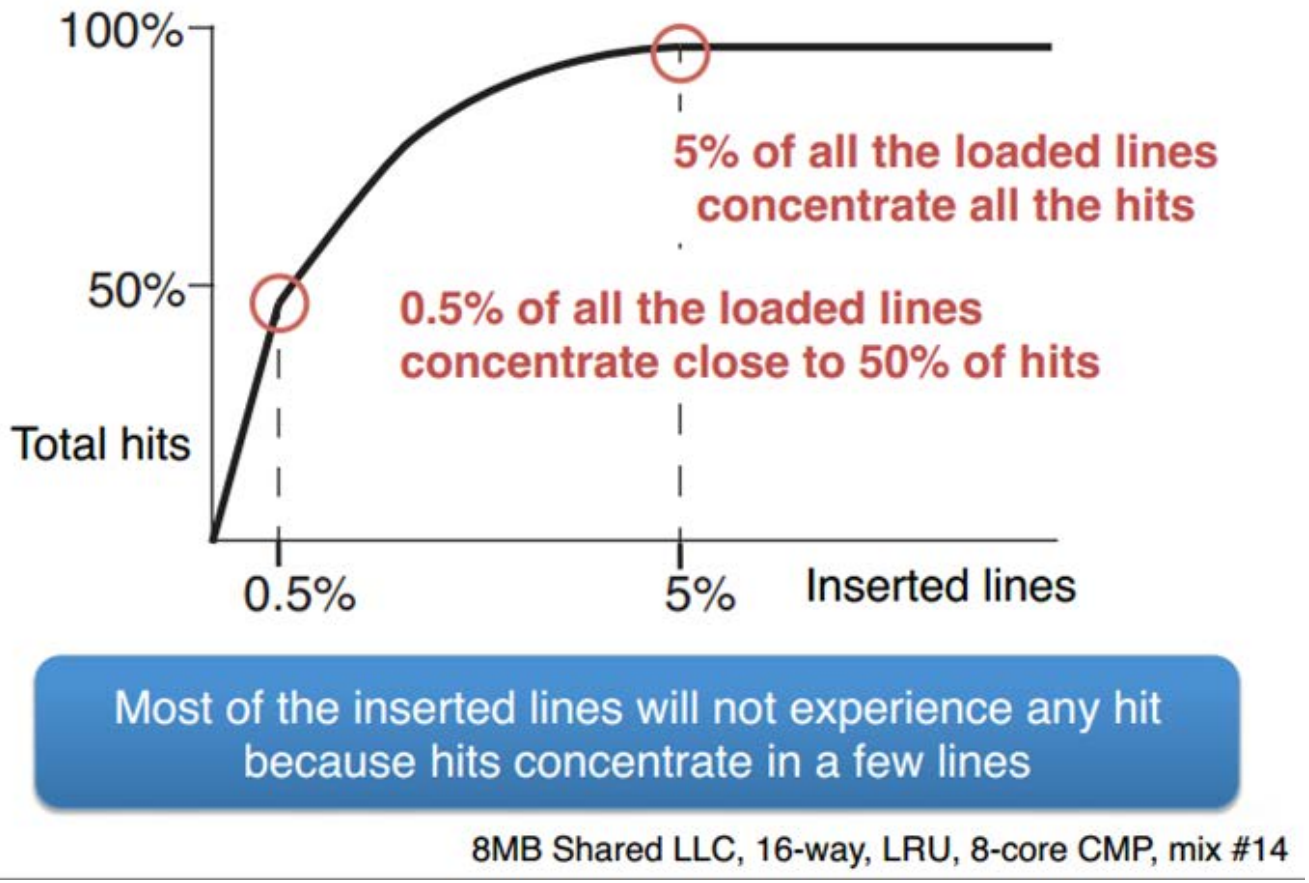
* Representative Graph

Area-Performance Tradeoffs*



* Representative Graph

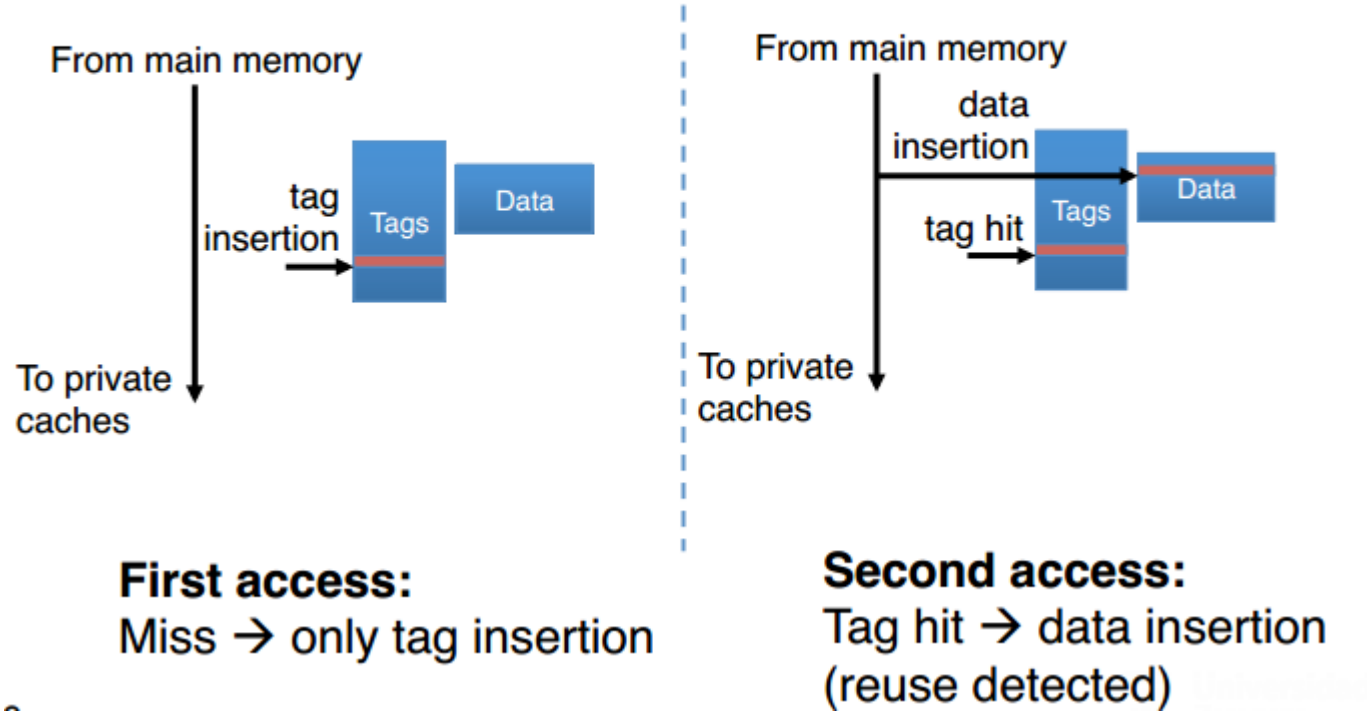
Motivation



Jorge Albericio, Pablo Ibáñez, Víctor Viñals, and José M. Llabería. 2013. The reuse cache: downsizing the shared last-level cache. In *Proceedings of the 46th Annual IEEE/ACM International Symposium on Microarchitecture* (MICRO-46).

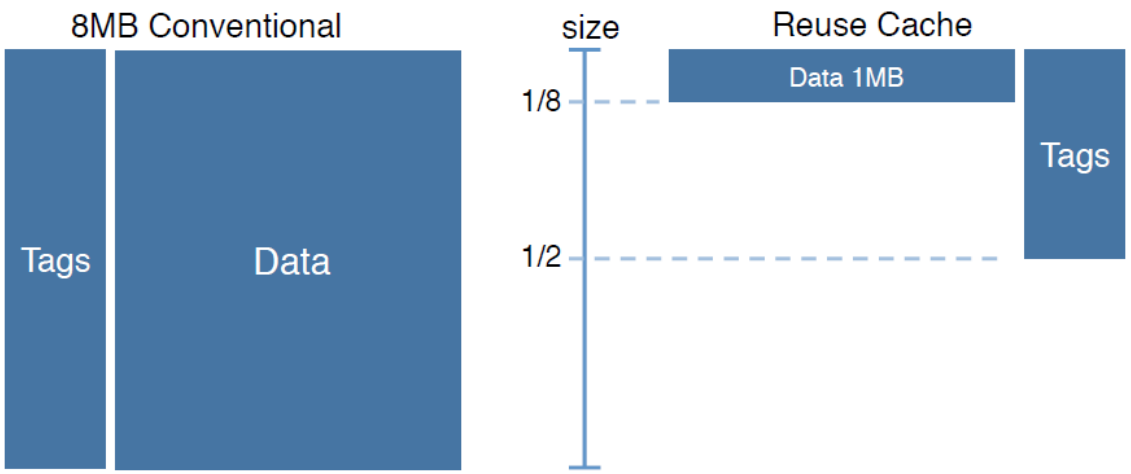
Reuse Cache : Original idea

Data is stored only when **reuse** is detected



Jorge Albericio, Pablo Ibáñez, Víctor Viñals, and José M. Llabería. 2013. The reuse cache: downsizing the shared last-level cache. In *Proceedings of the 46th Annual IEEE/ACM International Symposium on Microarchitecture* (MICRO-46).

Inclusive Shared Last-Level Cache (SLLC)



**Same average performance*
with 84% area savings**

*100 multiprogrammed SPEC CPU 2006 workloads in an 8-core CMP

Jorge Albericio, Pablo Ibáñez, Víctor Viñals, and José M. Llabería. 2013. The reuse cache: downsizing the shared last-level cache. In *Proceedings of the 46th Annual IEEE/ACM International Symposium on Microarchitecture* (MICRO-46).

Inclusive Shared Last-Level Cache (SLLC)

8MB Conventional

size

Reuse Cache

Data 1MB

1/8

Tags

Reuse Cache :
good idea for desktop processors (seemingly)
but what about for mobile processors?

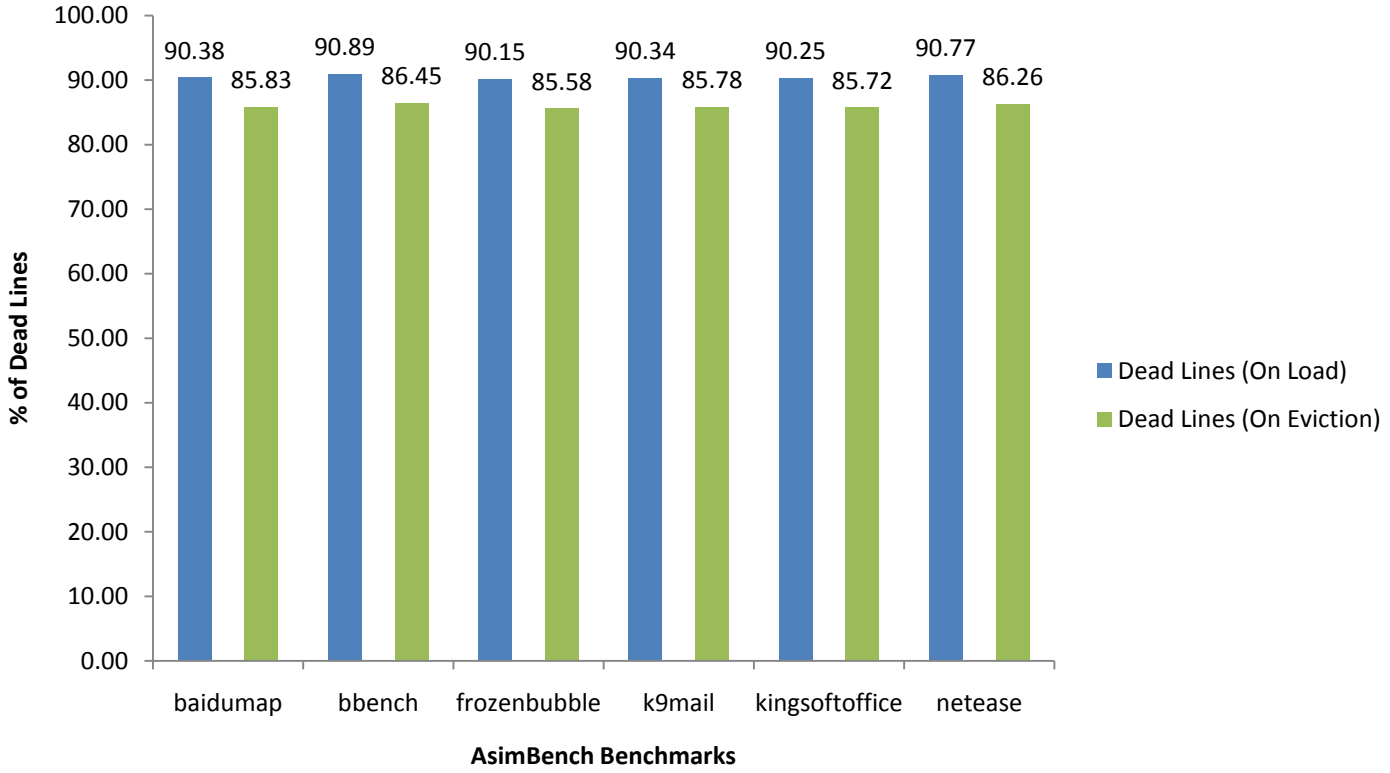
Same average performance*
with 84% ings

*100 multiprogrammed SPEC CPU 2006 wo -core CMP

Revisiting Reuse Caches for the Mobile Environment

- Questions
  - Memory characteristics of mobile workloads?
  - Spatial/Temporal/Reuse locality at L2 level?
  - Reuse Cache performance for smaller, simpler caches?

Low temporal locality

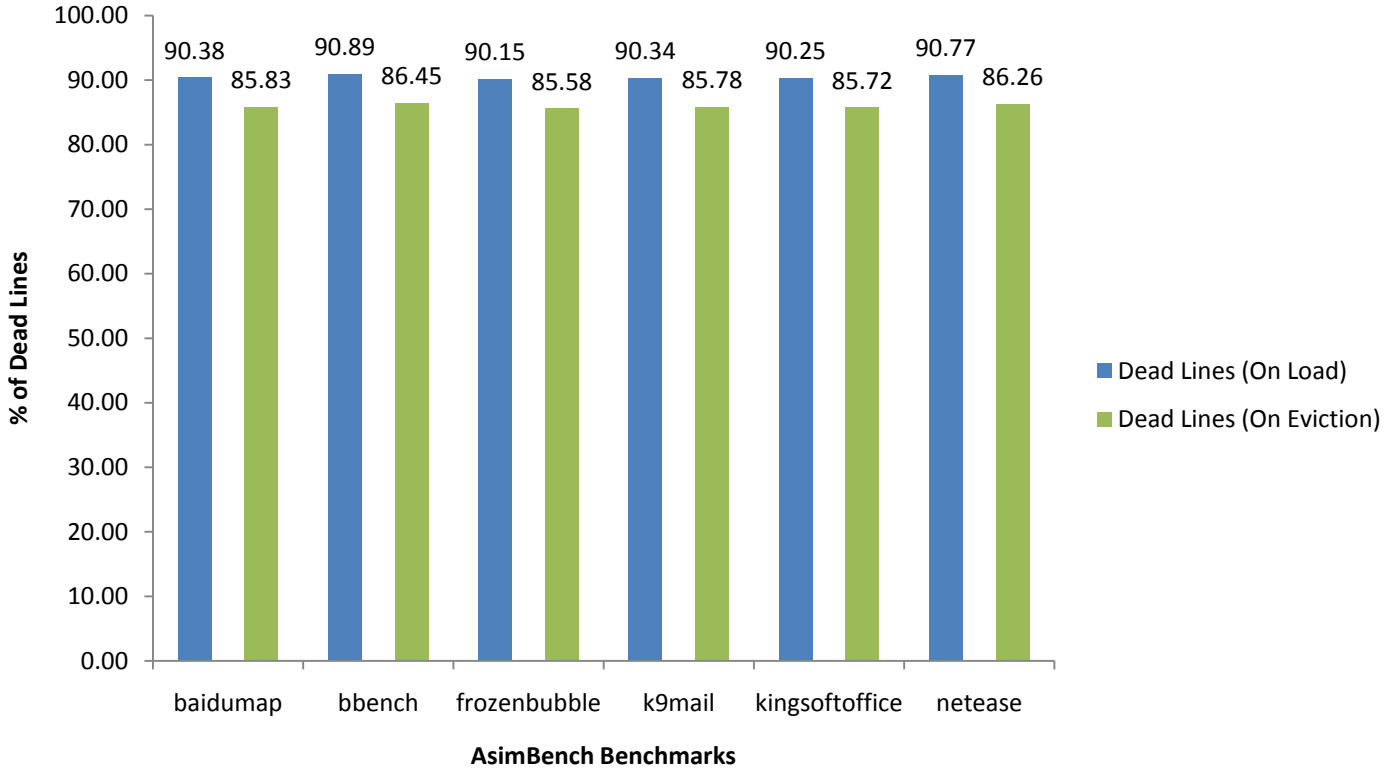


$$\%\ \text{Dead Lines (on Fetch)} = 1 - \frac{\#\ \text{Lines Reused}}{\#\ \text{Lines Fetched}}$$

$$\%\ \text{Dead Lines (on Eviction)} = \frac{\#\ \text{Unused Evicted Lines}}{\#\ \text{Lines Evicted}}$$

Low temporal locality

85-90% dead lines!



$$\%\ \text{Dead Lines (on Fetch)} = 1 - \frac{\#\ \text{Lines Reused}}{\#\ \text{Lines Fetched}}$$

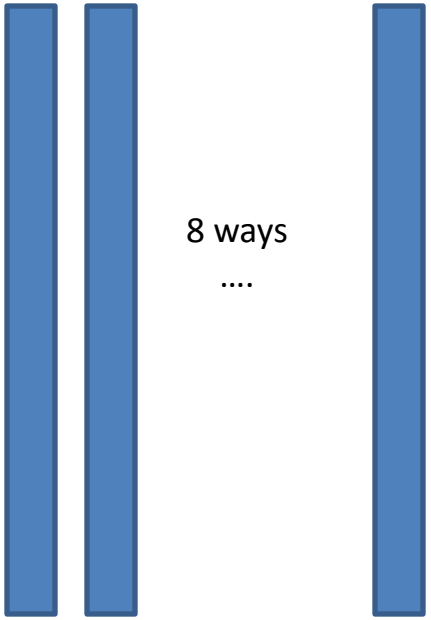
$$\%\ \text{Dead Lines (on Eviction)} = \frac{\#\ \text{Unused Evicted Lines}}{\#\ \text{Lines Evicted}}$$

Outline

- Motivation
- Implementation
- Methodology
- Performance Evaluation
- Future Directions and Conclusion

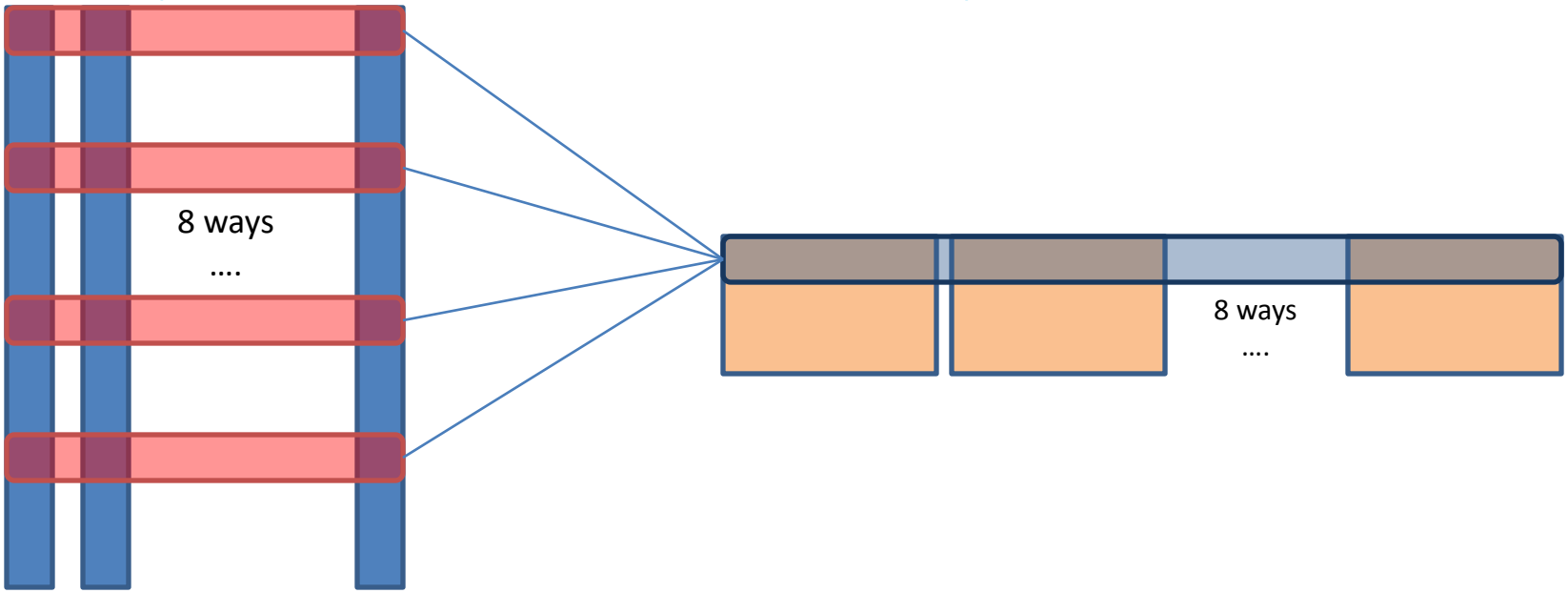
Organization

- Tag Array
  - Inclusive (aids coherence)
  - Set associative
  - (Forward) Pointer to data array entry



Organization

- Tag Array
  - Inclusive (aids coherence)
  - Set associative
  - (Forward) Pointer to data array entry
- Data Array
  - Set associative
  - Stores reused lines
  - (Reverse) Pointer to tag array entry



Coherence

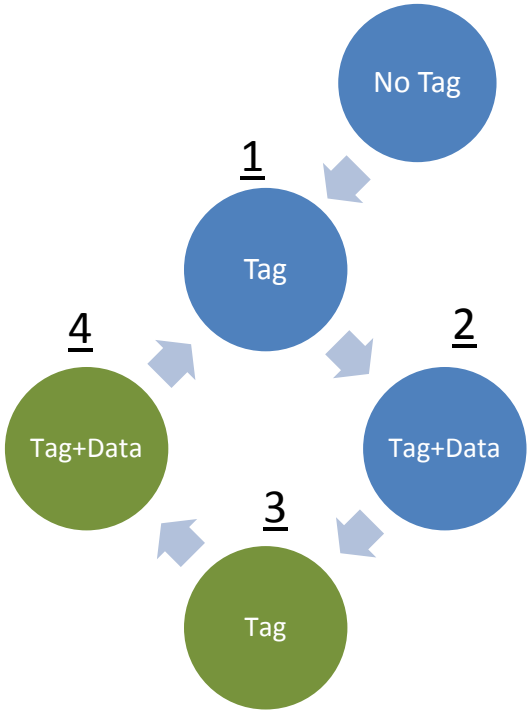
- TO-MOESI protocol
  - Small extension to the MOESI protocol
- New Tag-Only state
- Tag+Data states : Modified, Shared, Owned, Exclusive
- Transitions triggered by data insertions/evictions
- Read the paper!

Replacement

- Independent replacement policies for Tag and Data arrays
- Tag Array
  - Not Recently Reused (NRR)
- Data Array
  - Not Recently Used (NRU) + Shadow directory

Shadow Directory

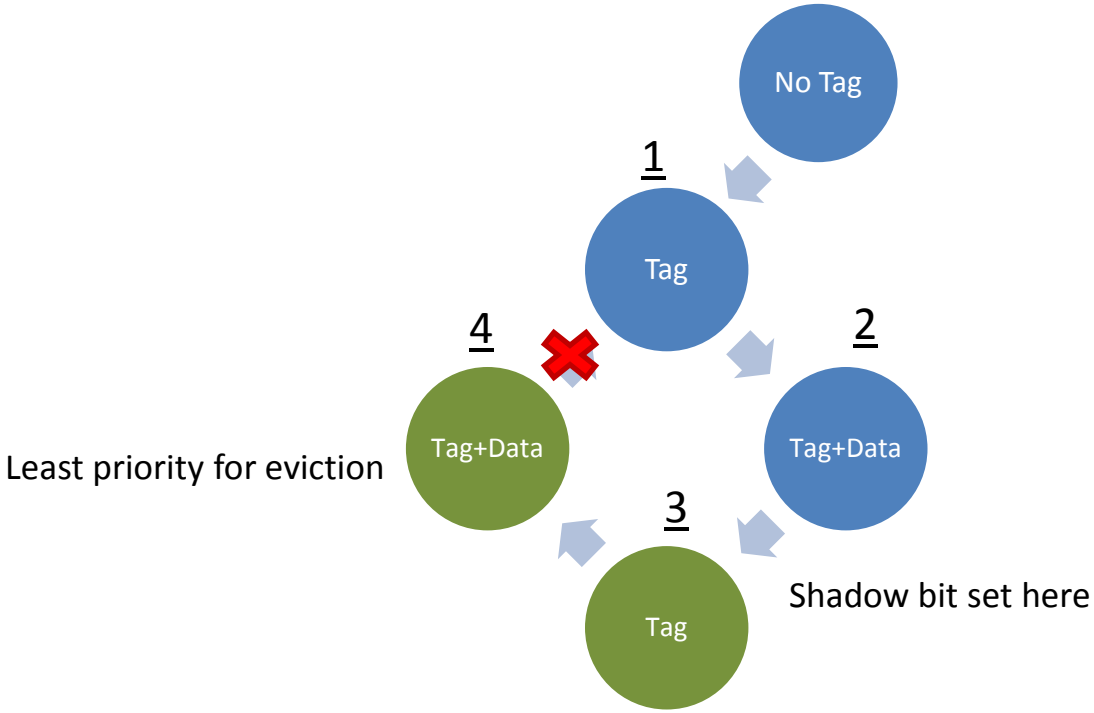
- Pomerene et al. proposed shadow directory
  - Transient lines that must be flushed quickly
  - Lines that become active after long periods of inactivity.



Prefetching mechanism for a high speed buffer store, James Herbert Pomerene, Thomas Roberts Puzak, Rudolph Nathan Rechtschaffen, Frank John Sparacio, 1984, Patent EP 0157175 A2

Shadow Directory

- Pomerene et al. proposed shadow directory
  - Transient lines that must be flushed quickly
  - Lines that become active after long periods of inactivity.
- Shadow bit added in reuse cache to break inefficient cycles



Outline

- Motivation
- Implementation
- Methodology
- Performance Evaluation
- Future Directions and Conclusion

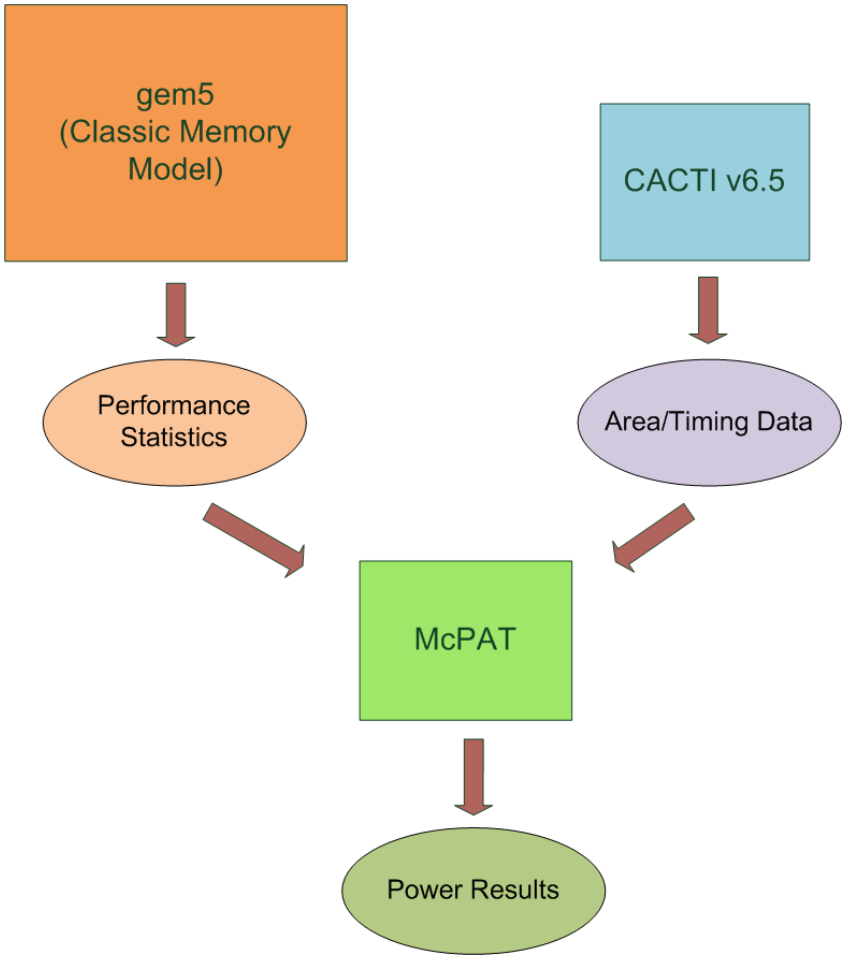
Methodology

❑Baseline
- 32KB/4way L1 I$ + 32KB/4way L1 D$
- 4MB/8way L2$ (Shared LLC)
- 8 wide OOO pipeline
- 1.4GHZ | 32nm | Low power device

(Based on ARM A15 specifications)

❑ Workloads
- AsimBench
- SPEC CPU 2006
- Self-written microbenchmarks

(functional verification)



Workload – AsimBench*

- Popular Android applications
- 11 diverse applications

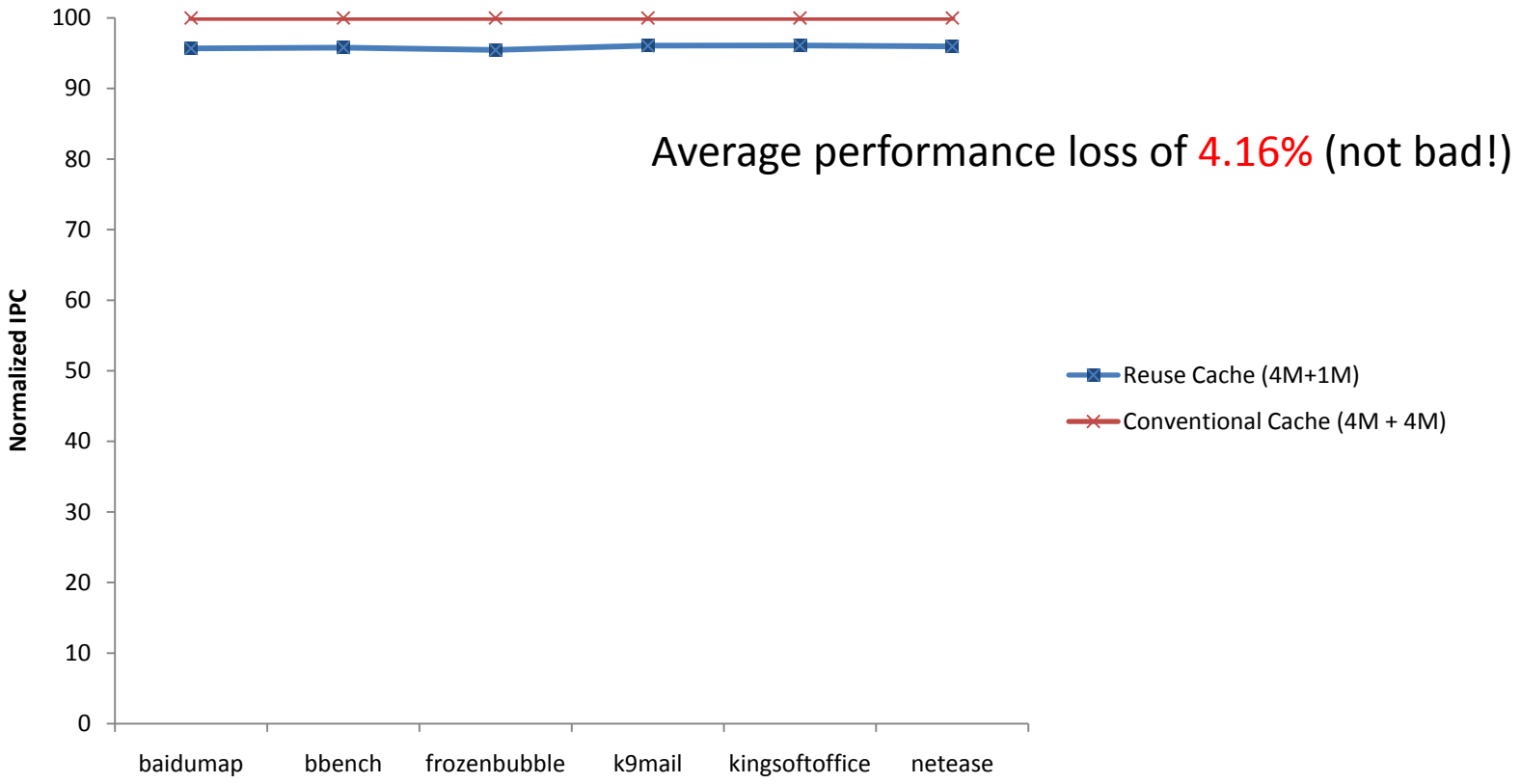
- 6 most memory-intensive apps selected for performance evaluation
  - BBench (Web Browser): Load web pages
  - K9Mail (Email): Load/Show emails
  - NeteaseNews (News): Check and load news
  - KingsoftOffice (Document Process): Open doc/xls/ppt file
  - BaiduMap (Map): Load map information of a specific area
  - FrozenBubble (Game): Load game

* Yongbing Huang, Zhongbin Zha, Mingyu Chen, Lixin Zhang. AsimBench: A Mobile Benchmark Suite for Architectural Simulators. IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), Monterey, CA, March 2014.

Outline

- Motivation
- Implementation
- Methodology
- Performance Evaluation
- Future Directions and Conclusion

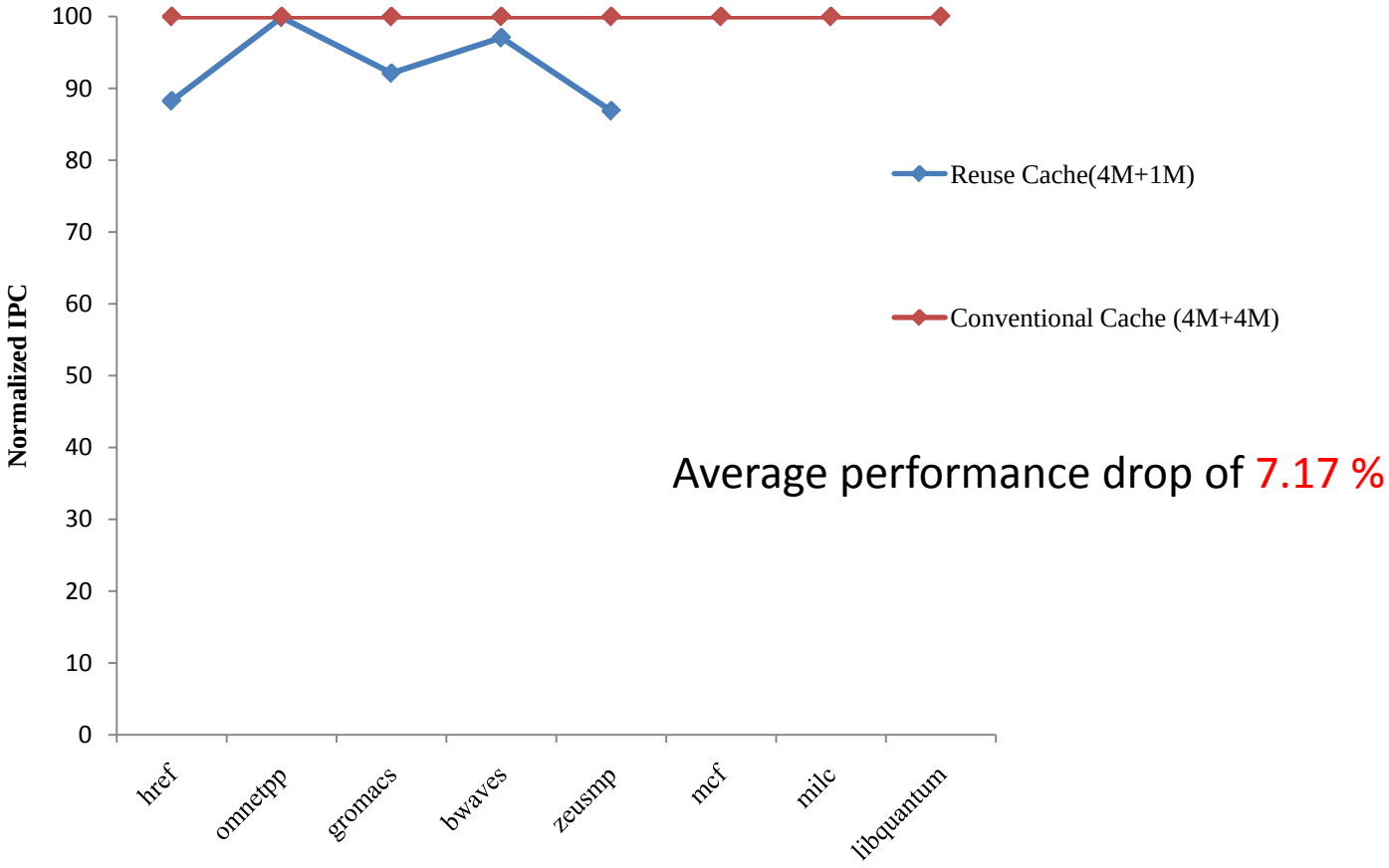
Performance - AsimBench



Average performance loss of 4.16% (not bad!)

Simulation length of 5 Billion instructions after boot in Full System mode

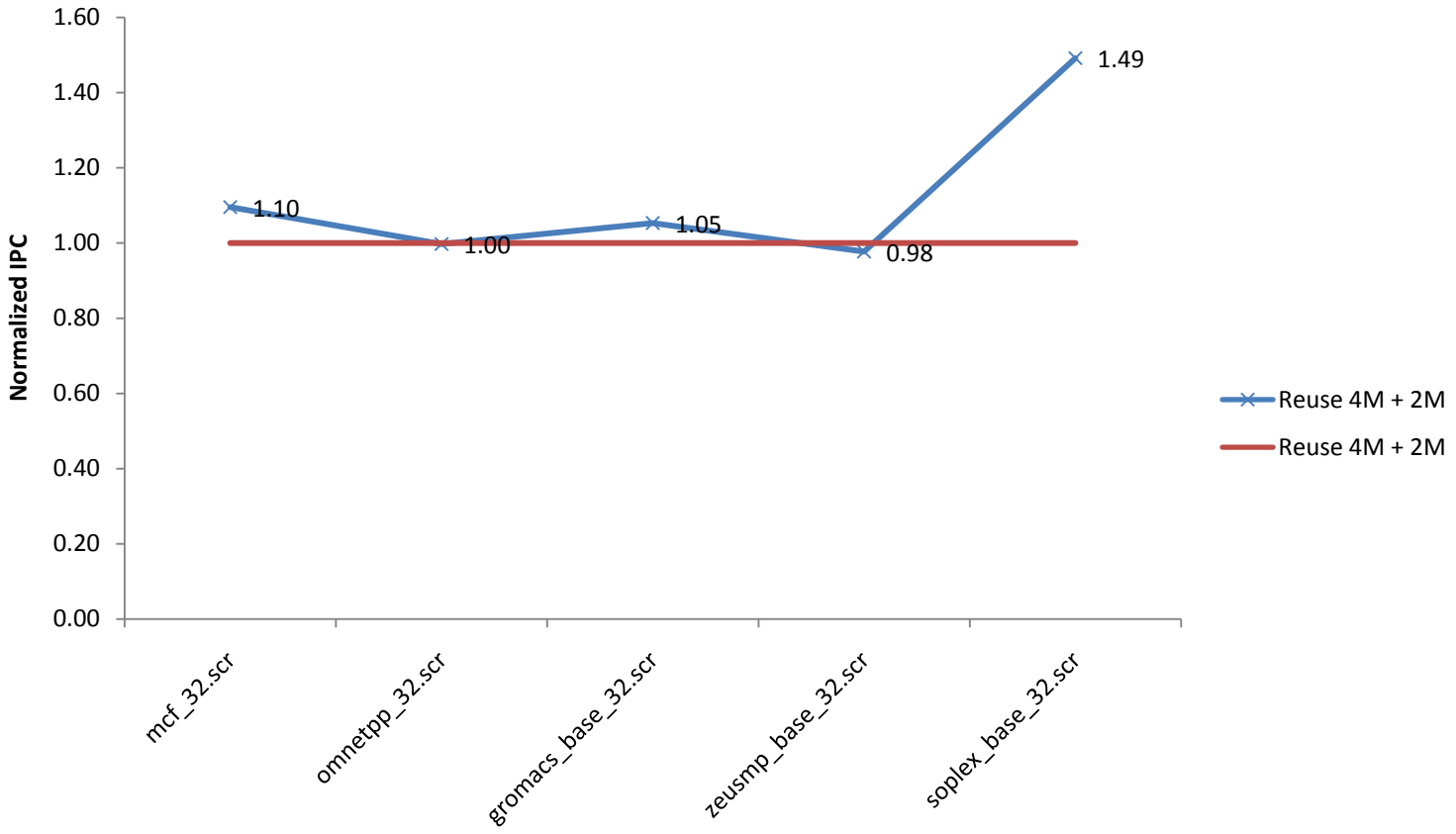
Performance – SPEC CPU2006



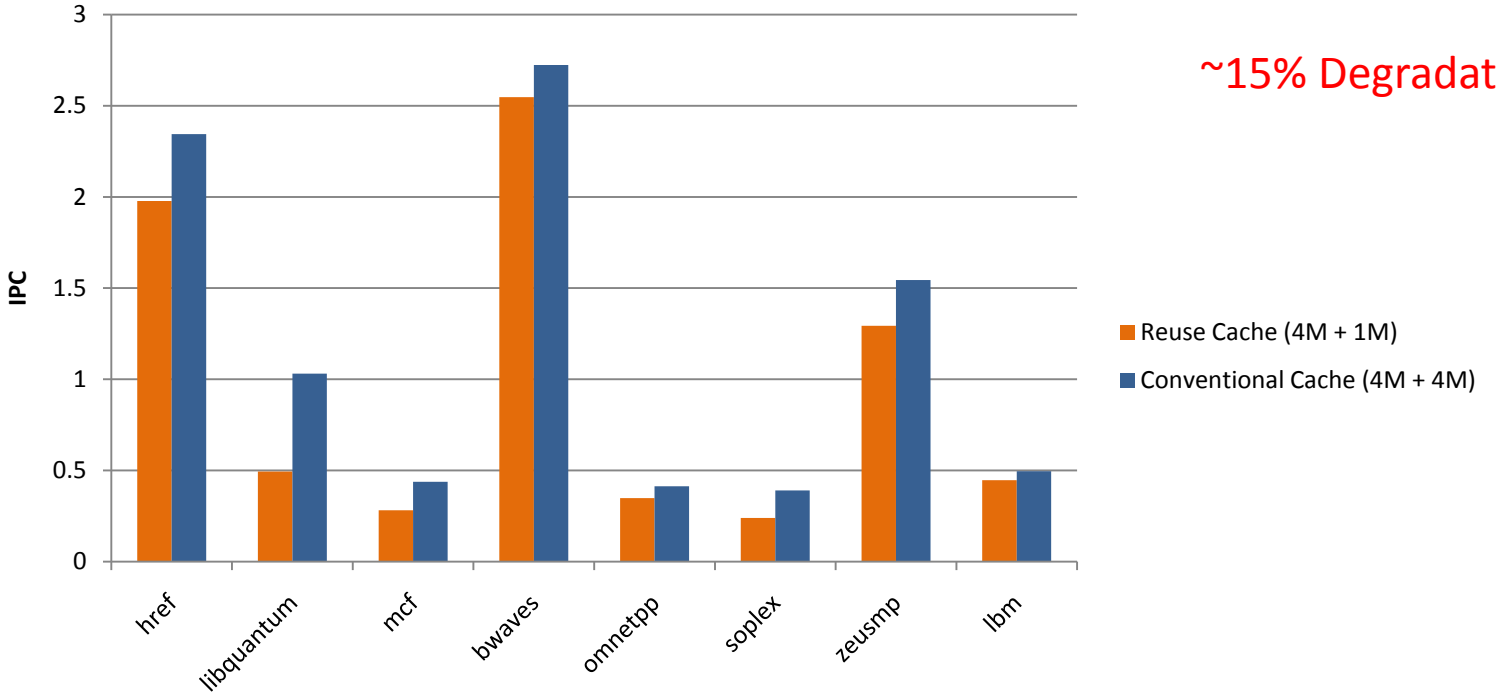
Average performance drop of 7.17 %

(Simulation length of 10B instructions in System Emulation mode)

Configuration Exploration (WIP)



Multi-core Performance



~15% Degradation

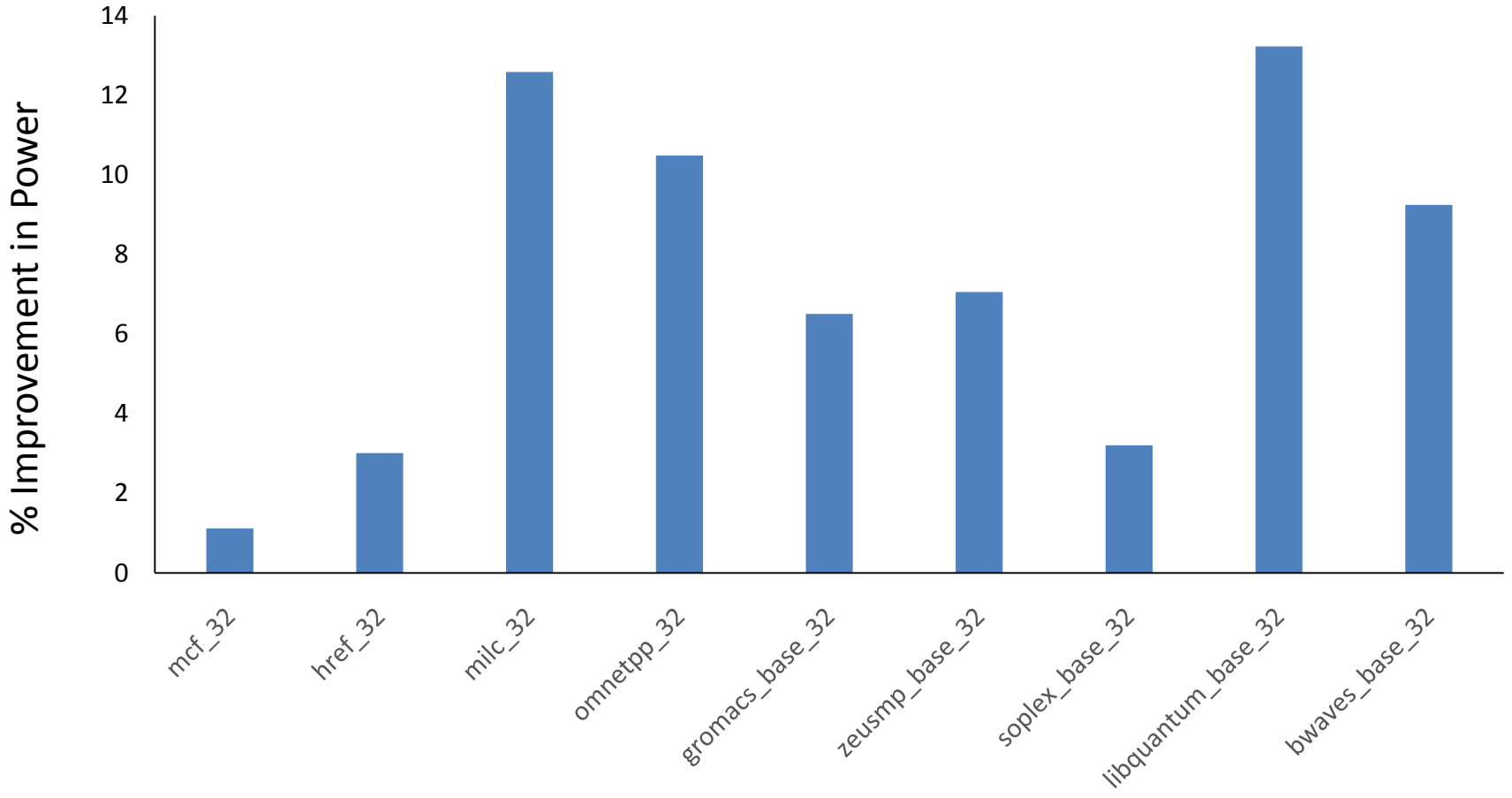
SPEC benchmarks simulated on 4 cores, running 10 Billion instructions each
(Full system + multi-core + AsimBench == Too slow)

Area Improvements

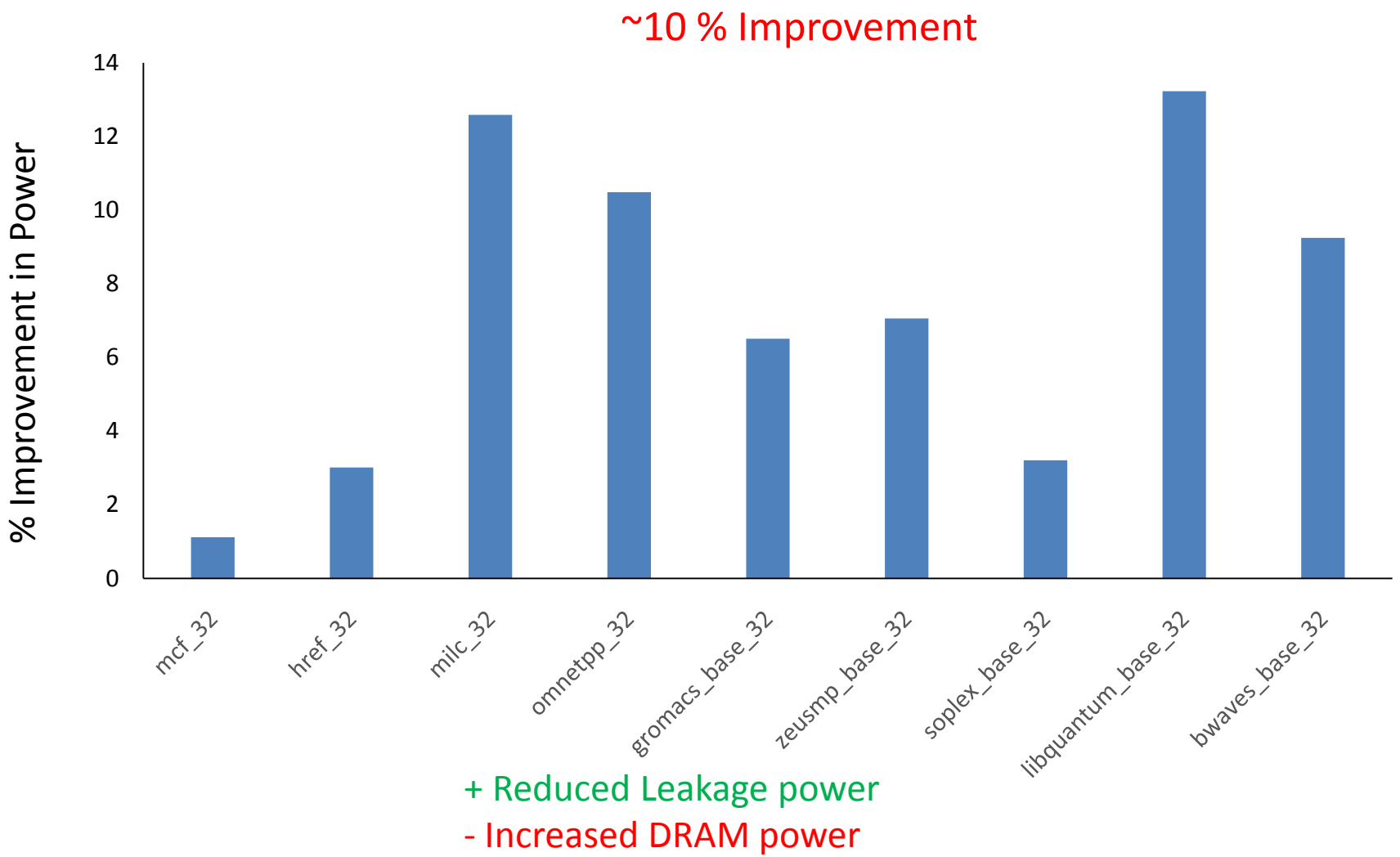
	Conventional (4M + 4M)	Reuse Cache (4M + 1M)	Reuse Cache (4M + 2M)
Area (in mm^2)	2.37	1.18	1.5

Area savings of 50% for the 4+1 reuse cache, 37% for the 4+2 cache
(Data generated using Cacti v6.5)

% Improvement in Power



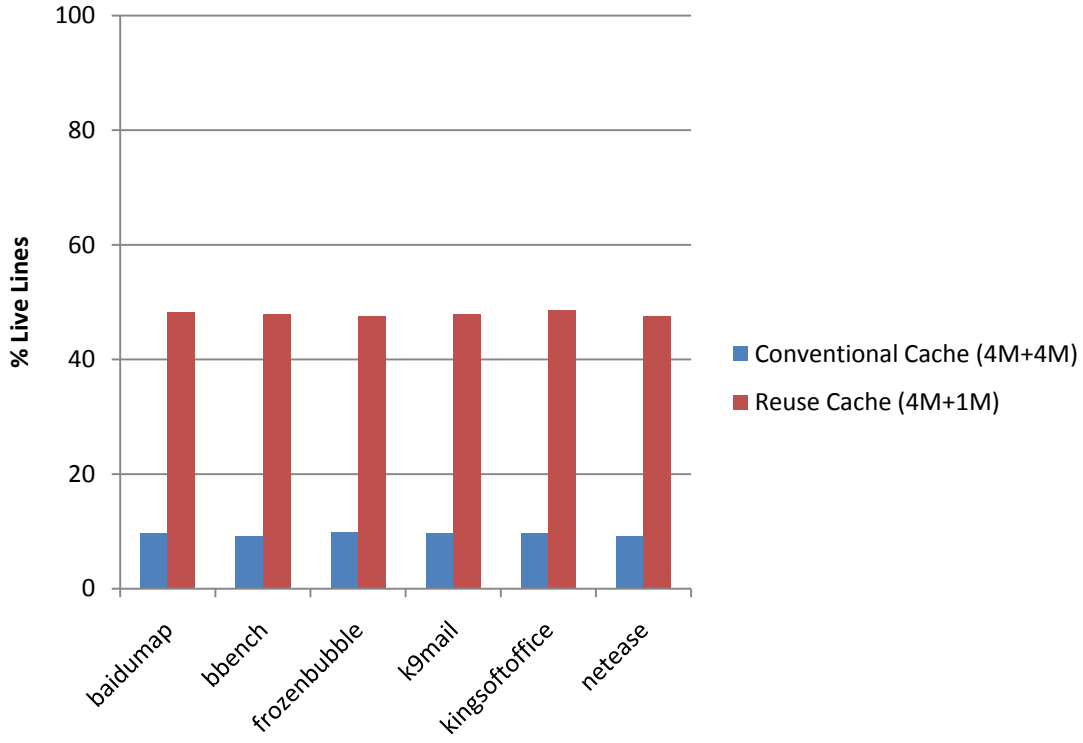
% Improvement in Power



+ Reduced Leakage power

- Increased DRAM power

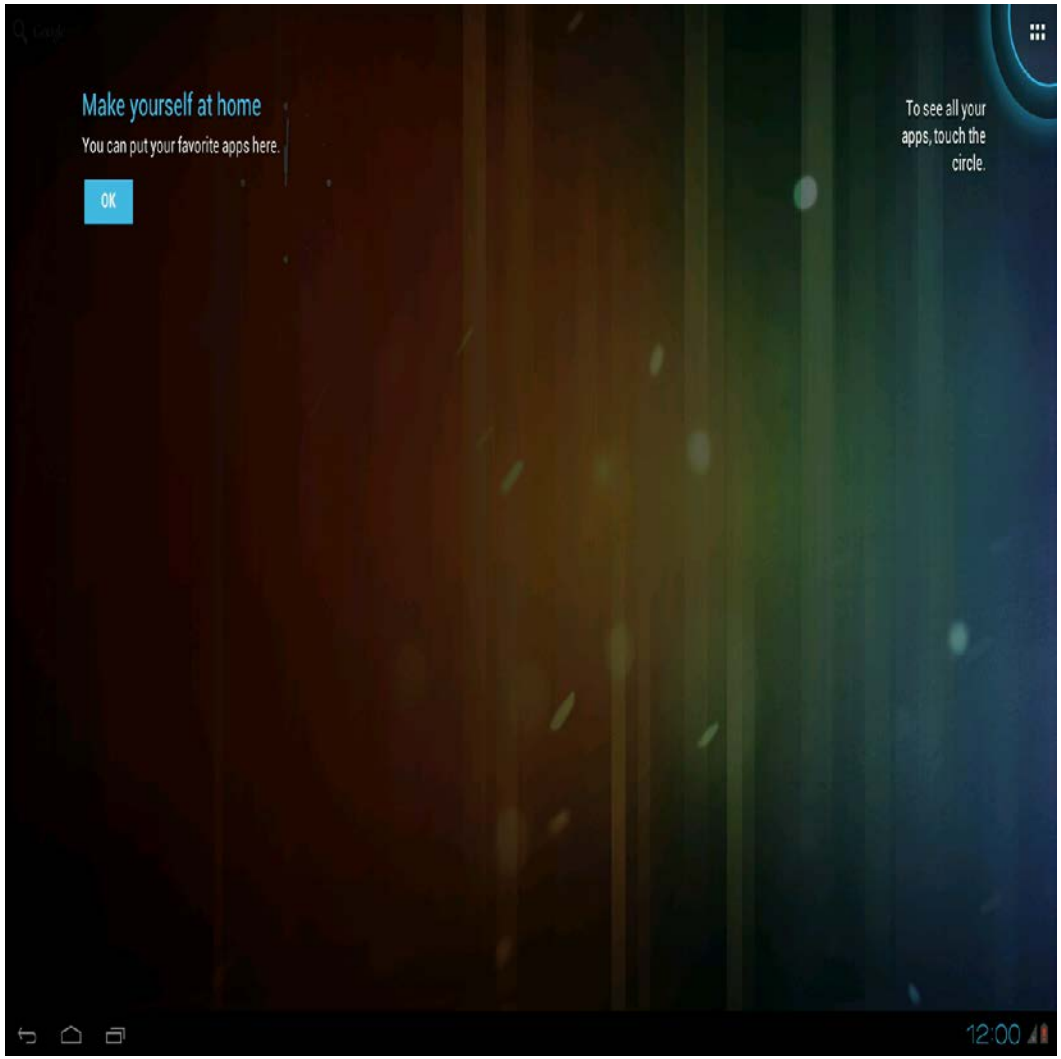
Live-ness Analysis - AsimBench

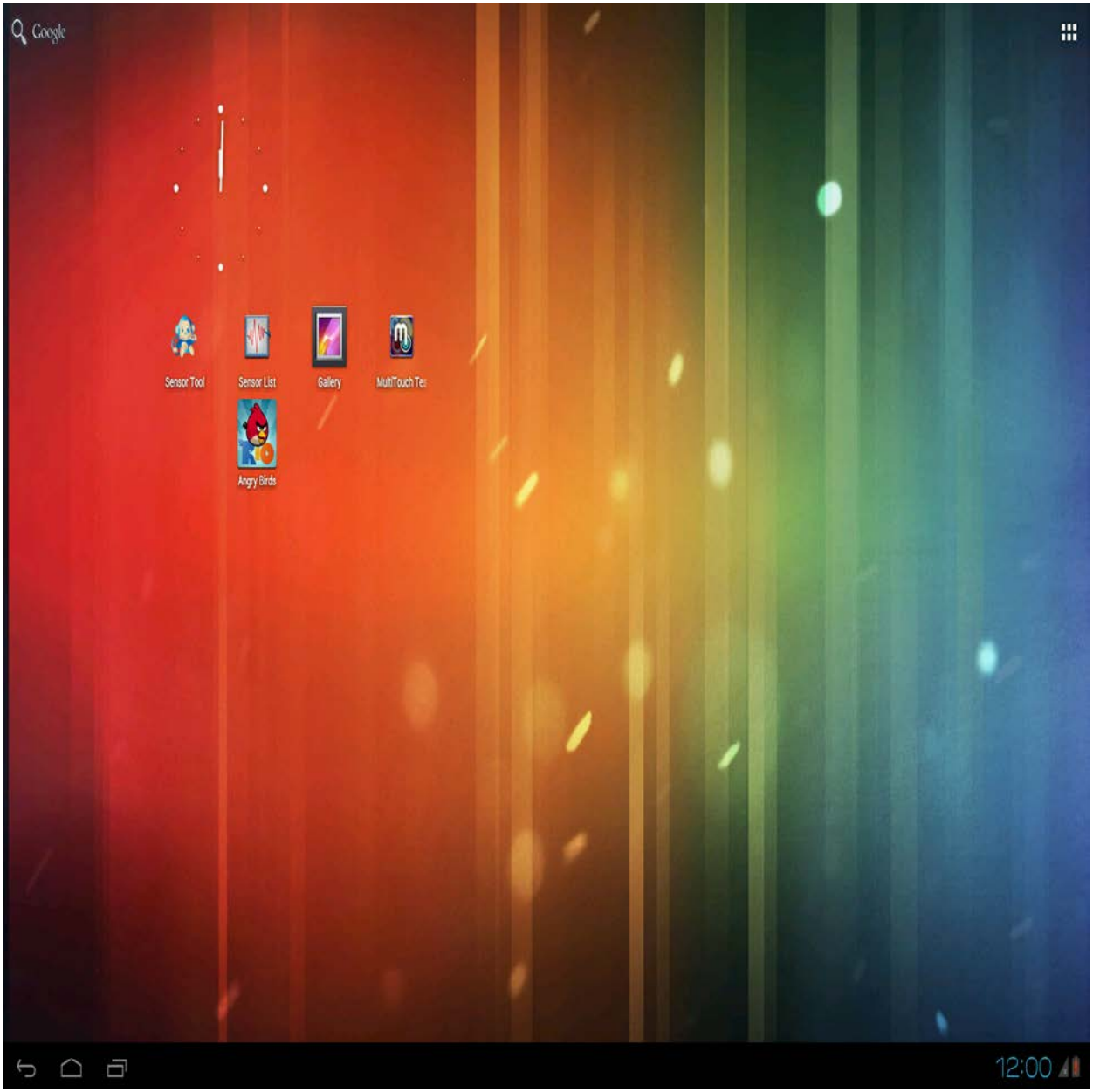


- % Live lines increased from 9.54 % to 47.95%

Pictures!

- AsimBench apps running on gem5 …


Make yourself at home
You can put your favorite apps here.
OK
To see all your apps, touch the circle.
12:00


Google
Sensor Tool
Sensor List
Gallery
MultiTouch Te
Angry Birds
12:00

1 UP
LEVEL
1
12:01

Inbox
Syncing disabled
FW: [gem5-users] fatal: SimpleDRAM system.physmem is unconnected!
4/10/2013
»huangyongbing From: gem5-users-bounces@gem5.org [mailto:gem5-users-bounces@gem5.org] On Behalf Of Andreas Hansson Sent: Thursday, March 28, 2013 5:43 PM To: gem5 users mailing list Subject: Re: [gem5-users] fatal: SimpleDRAM system.physmem is unconnected! Hi guys, As pointed out, the physmem in Ruby is "dangling" and only used to functionally have something to compare with. With recent changes in Ruby I think it can more or less be removed. Hopefully we can eventually reach a point where Ruby uses the normal gem5 ports o
FW: [gem5-users] protoc dependency
4/10/2013
»huangyongbing From: gem5-users-bounces@gem5.org [mailto:gem5-users-bounces@gem5.org] On Behalf Of Andreas Hansson Sent: Tuesday, March 26, 2013 6:08 PM To: gem5 users mailing list Subject: Re: [gem5-users] protoc dependency Hi Uday, I don't dare say what's going wrong in the last linking step, but in terms of the gem5 warning, you should be fine with the steps you mention. If you compile gem5 as an executable, does it work fine? There have been other reports of protobuf 2.5.0 causing some issues, but that could be comple
FW: [gem5-users] about CoherentBus() parameters?
4/10/2013
»huangyongbing From: gem5-users-bounces@gem5.org [mailto:gem5-users-bounces@gem5.org] On Behalf Of Ali Saidi Sent: Saturday, March 23, 2013 9:58 PM To: gem5 users mailing list Subject: Re: [gem5-users] about CoherentBus() parameters? Take a look at src/mem/Bus.py. CoherentBus inherits from Bus so all those parameters are available as well. Ali CoherentBus(clock="1GHz", width = 16). I'm wondering if there's more. I looked into the c++ code: coherent_bus.cc and coherent_bus.hh, but I really didn't find anything about "width
FW: [gem5-users] How to dump data structure symbols to trace files ?
4/10/2013
»huangyongbing From: gem5-users-bounces@gem5.org [mailto:gem5-users-bounces@gem5.org] On Behalf Of Ali Saidi Sent: Wednesday, March 06, 2013 11:52 AM To: gem5 users mailing list Subject: Re: [gem5-users] How to dump data structure symbols to trace files ? Hi Umesh, Can you try to fix the issues you mentioned and let us know if that solves the problem (or post a patch to reviews.gem5.org)? For the first issue if you include pim/full_system.hh there is a variable called FullSystem that will be true if the simulator is in FS
FW: [SPAM] Re: [gem5-users] Builded gem5 just fine but run into trouble when executing simple helloworld program...
4/10/2013
»huangyongbing From: gem5-users-bounces@gem5.org [mailto:gem5-users-bounces@gem5.org] On Behalf Of Maxime Chéramy Sent: Friday, March 01, 2013 10:27 PM To: gem5 users mailing list Subject: [SPAM] Re: [gem5-users] Builded gem5 just fine but run into trouble when executing simple helloworld program... That's weird if it works for you with swig 2.0.9. On my side, I only downgraded swig to 2.0.7 in order to have a working build. % diff /home/max/gem5_/build/X86/python/swig/event.py /home/max/gem5/build/X86/python/swig/event.p
FW: [SPAM] Re: [gem5-users] servicemanager exit on ICS
4/10/2013
»huangyongbing From: gem5-users-bounces@gem5.org [mailto:gem5-users-bounces@gem5.org] On Behalf Of huangyongbing Sent: Wednesday, April 10, 2013 1:18 PM To: gem5-users@gem5.org Cc: 'Chander Sudanthi' Subject: [SPAM] Re: [gem5-users] servicemanager exit on ICS Sounds great. Thank Chander very much. It is really helpful. Best regards, Yongbing Huang From: Chander Sudanthi [mailto:Chander.Sudanthi@arm.com] Sent: Wednesday, April 10, 2013 12:43 PM To: gem5-users@gem5.org Cc: huangyongbing@ncic.ac.cn; sungup@me.com; hamishag@
网易通行证异常登录提醒
4/8/2013
»网易通行证 邦助 亲爱的网易通行证用户： 系统检测到您的网易通行证帐号(zrx05531@163.com)于2013-04-08 10:17在北京登录了邮箱客户端(通过POP3/IMAP协议)，如不是您本人操作，请尽快修改密码，或登录帐号修复中心提交资料进行人工修复帐号（修改帐号安全信息，解绑密保）。 如您不想再收到此类提醒，可登录网易通行证进行异常登录提醒设置。 什么情况下会收到此提醒邮件？1.您的帐号已经被他人盗取，并通过其他IP地址登录；2.您正在使用的IP时常发生变动，例如您使用了移动网络；3.您的网络服务提供商有多个网络出口，并且会经常变动。 2013-04-08（本邮
魔獸世界5.2改版！多項系統全新升級！
8:28pm
»魔獸世界官方網站
开发精选技术资料免费下载
3/20/2013
»51cto.com 如果邮件无法正常显示 请点击这里!51CTO为企业级的IT朋友们精心挑选了目前比较热门，主流的网络技术、开发技术、数据库技术，以及各类程序源码、解决方案、电子书等下载资料，希望朋友们能够喜欢。当然，由于空间有限，我们此次提供的资料相对较少，如果还不能够满足你的需求，那么，请进入51CTO下载频道进行相关技术资料的下载，也可以来www.51CTO.com来浏览一下我们给各位网友们提供的策划专题和相关教程。每本资料，每个专题都经过了51CTO的精心制作，IT的朋友们赶紧行动吧，而且注册还送下载豆哦。»»精品推荐 iOS开发资料打包下载（中
网易邮箱祝你生日快乐！
3/9/2013
»网易邮件中心 尊敬的1：3月15日是您的生日，衷心祝您生日快乐！为了感谢您一直以来对网易邮箱的支持，我们特别为您准备了一份小礼物，希望您能喜欢！邮箱积分 可用于参与邮箱各种活动 网易优惠券礼包 网易邮箱生日用户专享优惠和礼券 网易红包礼券 满2减1元，兑换码：R5PHZQUWYBZ8 有效期：2013年6月1日 您的生日为注册邮箱时填写的资料，如果不对，点击这里修改»» 想换个名字或昵称？点击这里»» 网易邮件中心
Load up to 10 more
12:00

Outline

- Motivation
- Implementation
- Methodology
- Performance Evaluation
- Future Directions and Conclusion

Future Directions

- Replacement Policies (NRR + Shadow) for Fully Associative data array
- Explore allotment policies
- Better uses for extra tag entries
- Longer/Better simulations using simpoints
- Multi-core simulations using AsimBench to validate coherence protocols
- Comprehensive energy analysis

Summary

For mobile workloads, reuse cache promises

+ Significant area reduction (~50%)

+ Decent power savings (~10%)

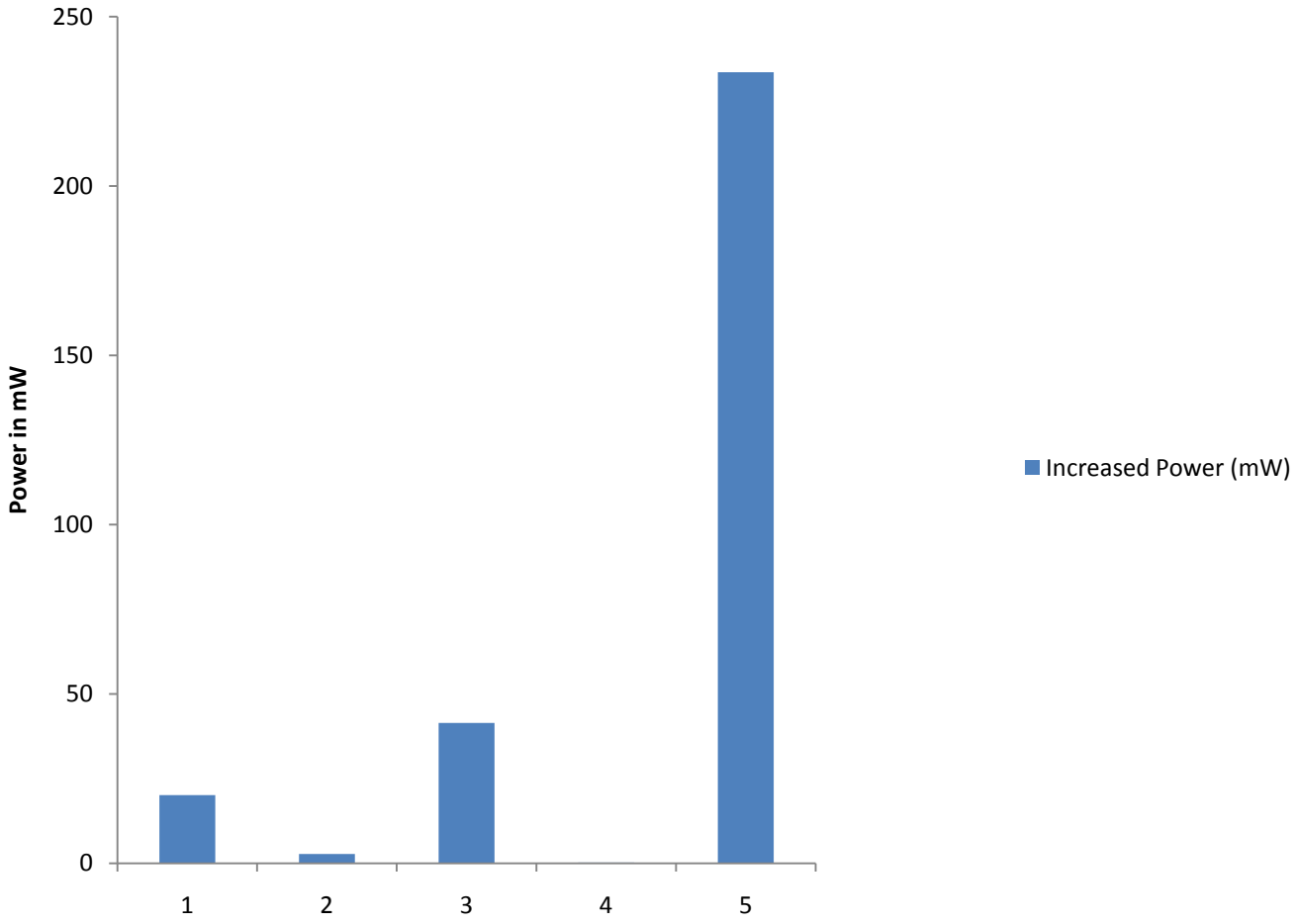
- Marginal performance loss (~6%)

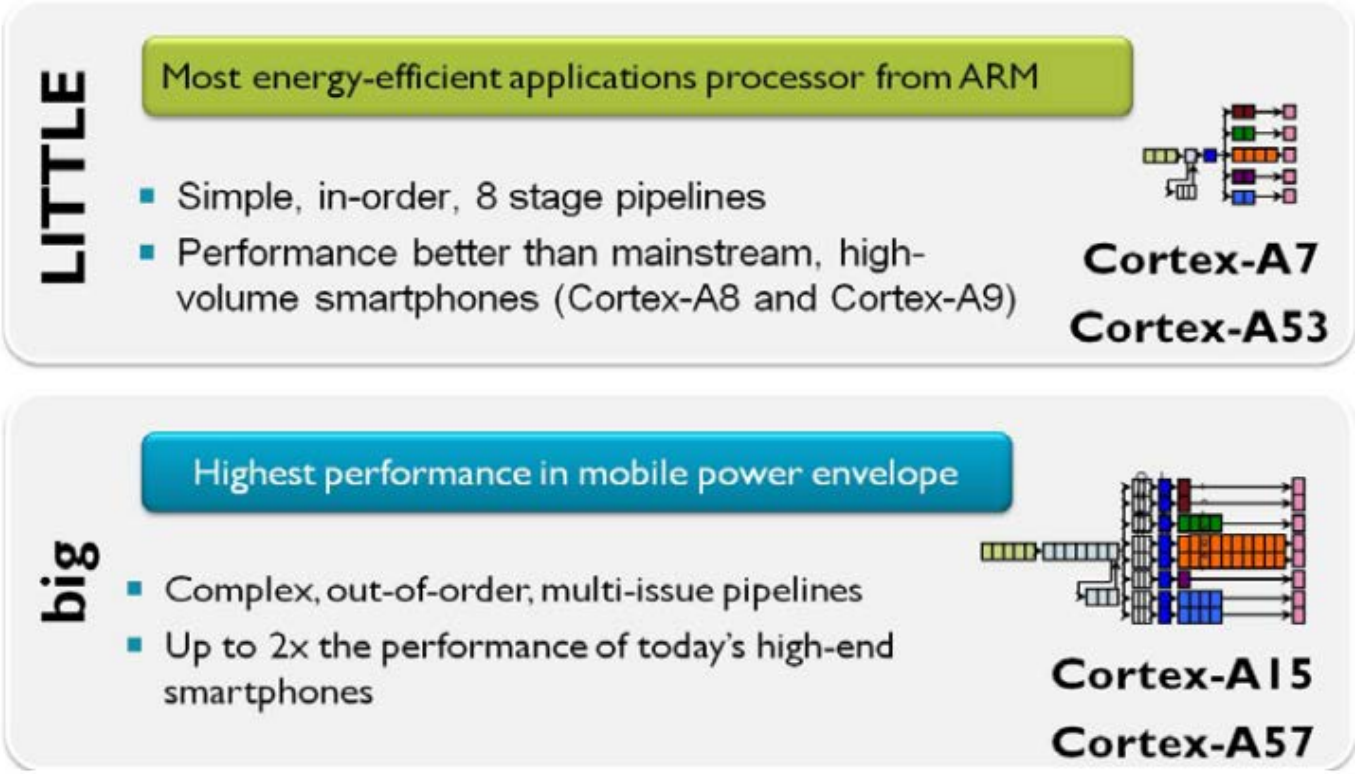
Questions?

Thank You!!

BACKUP

Configuration Exploration





L2 => 1 MB Cache
Animate this to cross out 1 MB and say 50% of data area

L2 => 4 MB Cache
Cross out and say Performance of 16 MB Cache

Results

- Mapping exploration -> IPC , ASIM + SPEC
- 4+4 vs (4+1) spec , asimbench
- Shadow vs NRR
- Boot times
- Bigger is better
- CACti graph
- Energy graph
- Dead lines

Baseline

- 32KB/4way L1 I$ + 32KB/4way L1 D$
- 4MB/8way L2$ (Shared LLC)
- 8 wide OOO pipeline
- 1.4GHZ | 32nm | Low power device

- Based on ARM A15 specifications

Challenges

- ARM aarch32 incompatible with gem5's Ruby memory model
- Simulation infrastructure for ARM in full system mode

Contributions