

Towards Energy-Proportional Datacenter Memory with Mobile DRAM

Krishna Malladi¹
Benjamin Lee²

Frank Nothaft¹
Christos Kozyrakis¹

Karthika Periyathambi
Mark Horowitz¹

Stanford University¹

Duke University²

Outline

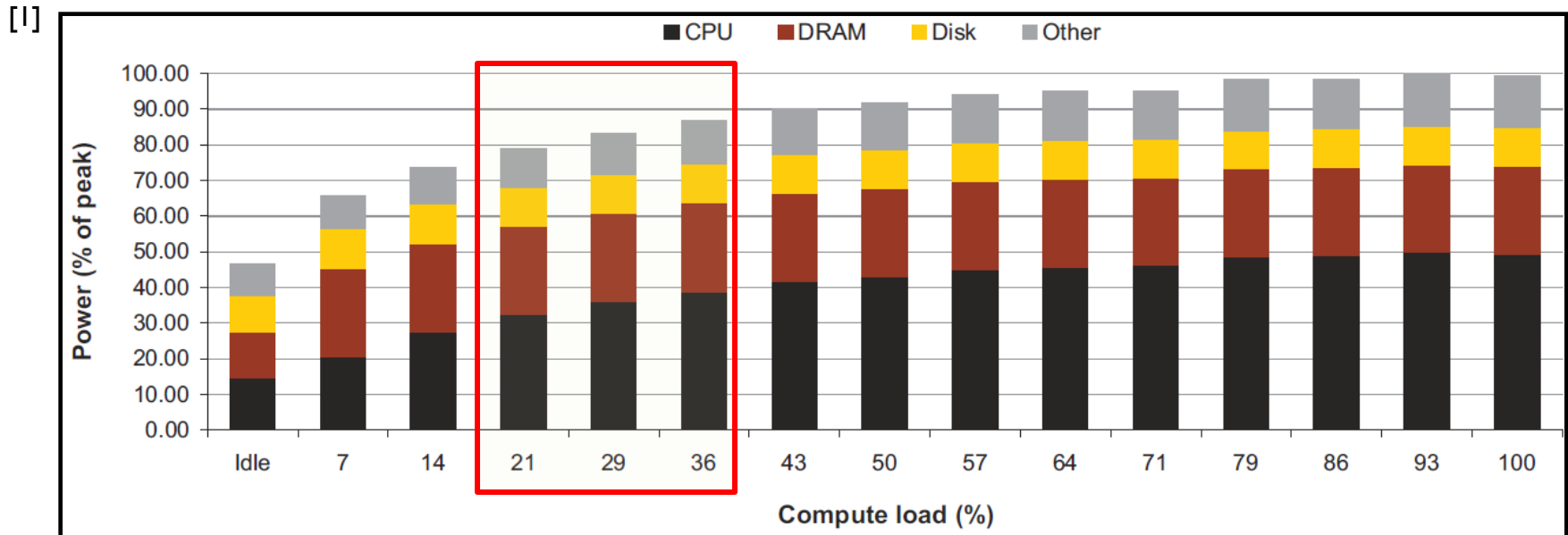
- ▶ Inefficiencies of server DRAM systems
- ▶ Mobile DRAM
- ▶ Evaluation

Outline

- ▶ Inefficiencies of server DRAM systems
- ▶ Mobile DRAM
- ▶ Evaluation

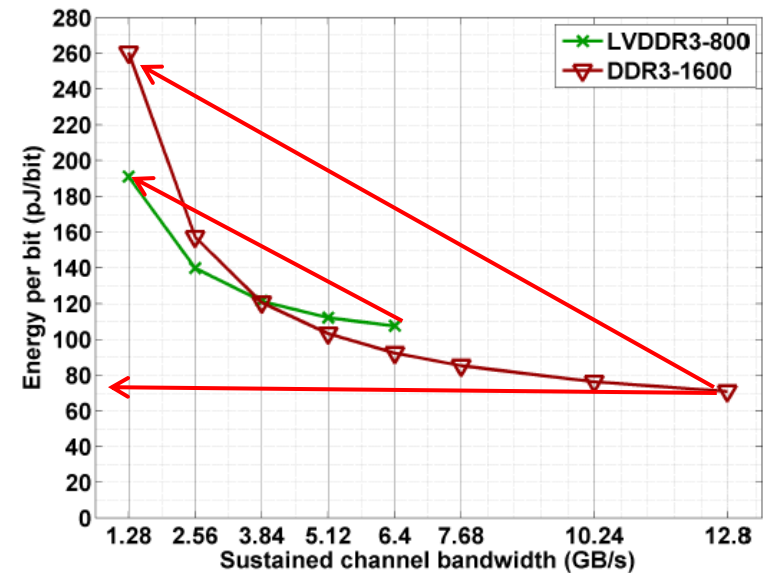
Server DRAM systems

- ▶ Server power main energy bottleneck in datacenters
 - ▶ PUE of ~ 1.1 \rightarrow the rest of the system is energy efficient
- ▶ Significant main memory (DRAM) power
 - ▶ **25-40%** of server power across all utilization points
 - ▶ Low dynamic range \rightarrow no energy proportionality
 - ▶ Power hungry active-idle and power-down states



DDR3 energy characteristics

- ▶ DDR3 optimized for high bandwidth (1.5V, 800MHz)
 - ▶ On chip DLLs, on-die-termination
 - ▶ 70pJ/bit at 100% bus utilization with 40% static cost
 - ▶ Increases to 260pJ/bit at low datarates due to static power
- ▶ LVDDR3 alternative (1.35V, 400MHz)
 - ▶ Lower voltage → Higher on-die-termination
 - ▶ Still disproportional at 190pJ/bit
- ▶ Need memory systems that consume lower energy and are proportional
 - ▶ What metric can we trade for efficiency?



Workloads in datacenters

- ▶ **Web-search and map-reduce**
 - ▶ CPU or DRAM latency bound in stress-test and in-the-field measurements [2][3][4]
 - ▶ At peak load, need < 6% DRAM bandwidth [2]
- ▶ **Memory caching, DRAM-based storage, social media**
 - ▶ memcached and RAMCloud
 - ▶ Overall bandwidth limited by network (<10% of DRAM bandwidth)
- ▶ **Datacenter DRAM needs**
 - ✓ Low latency
 - ✓ High capacity
 - ✓ High reliability
 - ✗ High bandwidth
- ▶ **Our focus: tradeoff bandwidth for energy efficiency & proportionality**

Outline

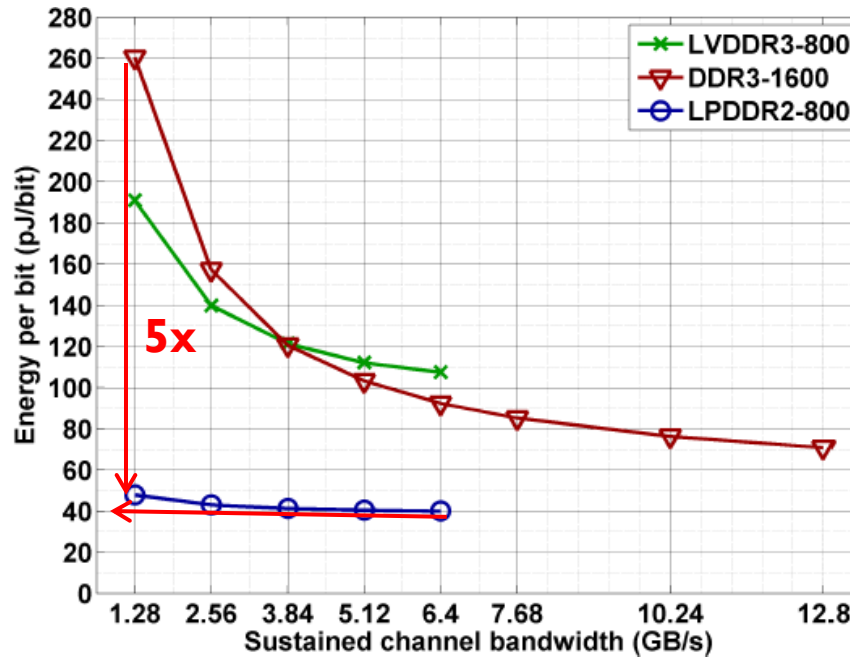
- ▶ Inefficiencies of server DRAM systems
- ▶ Mobile DRAM
- ▶ Evaluation

Mobile DRAM characteristics

Technology Parameter	DDR3	LPDDR2
Timing (tCAS, tRAS, tRC)	15, 38, 50ns	15, 42, 57ns
Active current (Read, Write)	180, 185mA	210, 175mA
Idle current (Powerdown, Standby)	35, 45mA	1.6, 23mA
Powerdown exit latency	24ns	7.5ns
Operating voltage	1.5V	1.2V
Typical operating frequency	800MHz	400MHz
Device width	8	16

- ▶ Same core as DDR3 devices
 - ▶ Same capacity per device , same access latency, same active currents
- ▶ IO interface optimized for very low static power
 - ▶ Including faster powerdown modes, no termination
- ▶ Same chip bandwidth
 - ▶ Wider interface operating at slower clock rate

LPDDR2 advantages



- ▶ Energy proportional
- ▶ Energy efficient ~ 40pJ/bit
 - ▶ 2x to 5x reduction over DDR3

LPDDR2 disadvantages

- ▶ Channel bandwidth ✓
 - ▶ Pin bandwidth is 2x lower → halves peak data rate per rank
 - ▶ Datacenter workloads require lower bandwidth
- ▶ System capacity ?
 - ▶ Not optimized for multi-chip modules or multi-rank channels
 - ▶ Inter-symbol interference (ISI) due to electrical loading
 - ▶ Datacenter workloads require high memory capacity
- ▶ Reliability ?
 - ▶ ECC works best with x4 devices
 - ▶ Complicated or expensive with x16 devices
 - ▶ See paper for details

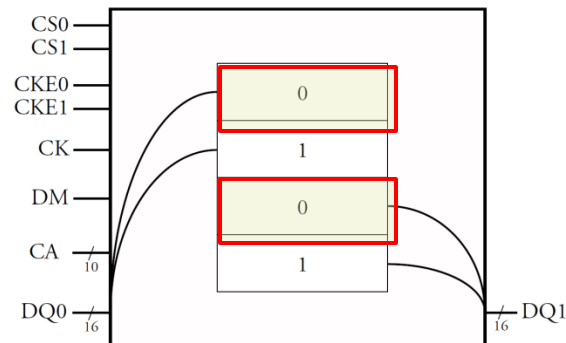
Building capacity with LPDDR2

▶ Key problems

- ▶ Wide interface → limits # devices in parallel in a 64-bit channel
- ▶ No termination → limits # devices in series due to ISI

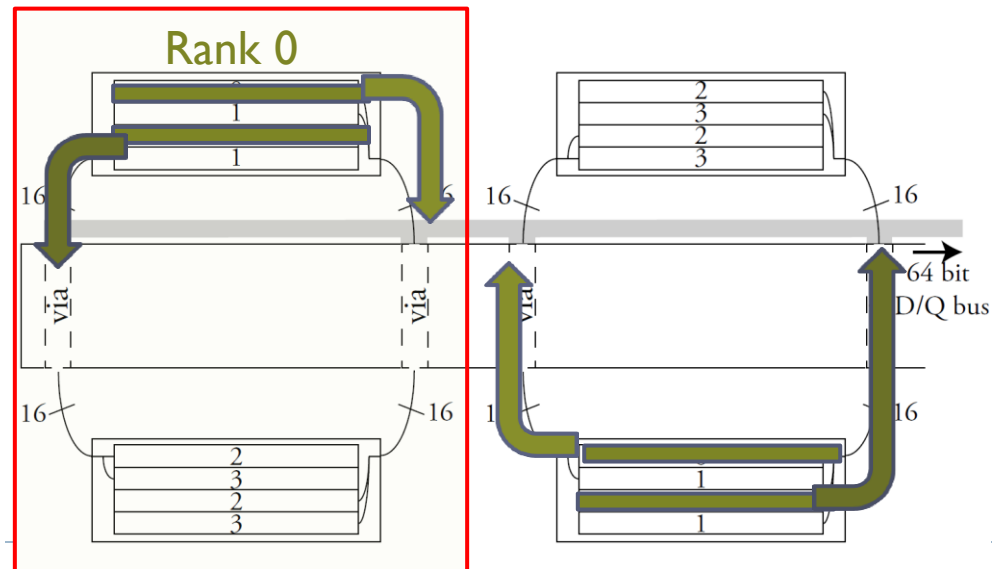
▶ Basic memory package

- ▶ Commodity LPDDR2 devices stacked (edge bonded)
- ▶ Four 2Gb x16 chips → 8Gb x32 package
- ▶ Two devices share Chip Select



High capacity LPDDR2 module

- ▶ Minimize ISI by getting stubs close together (single point load)
- ▶ Dual Line Package (DLP) module
 - ▶ Mirrored connected with on-board vias
 - ▶ Four 8Gb x32 packages → 32Gb x64 module
 - ▶ Striped ranks to minimize stub distance
- ▶ No changes needed to LPDDR2 controller



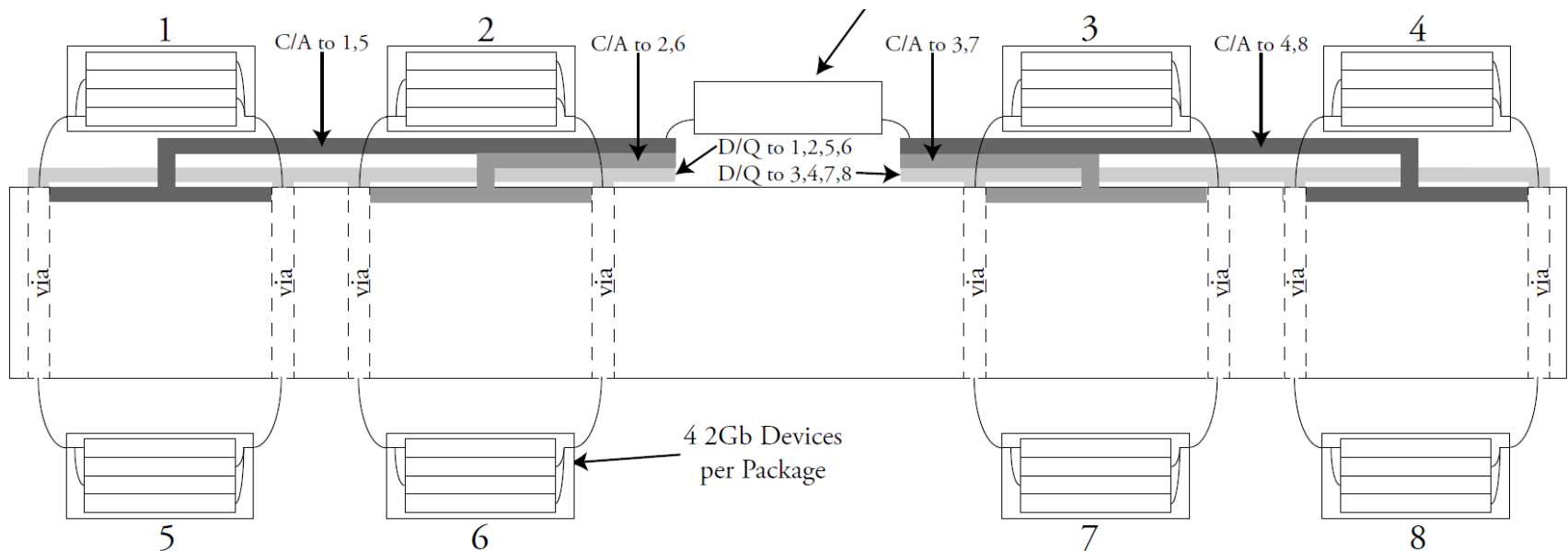
High capacity LPDDR2 channel

▶ Key problem

- ▶ More modules per channel → more ISI, degraded operation

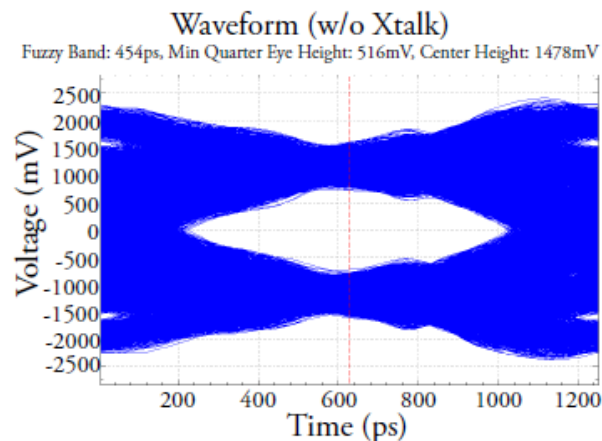
▶ Load Reduced (LR) LPDDR2 channel

- ▶ Introduce buffer to limit load on channel (similar to LRDDR3)
- ▶ 2x DQ and 4x CA lines to provide device isolation
- ▶ Two 32Gb x64 modules → 64Gb x64 channel

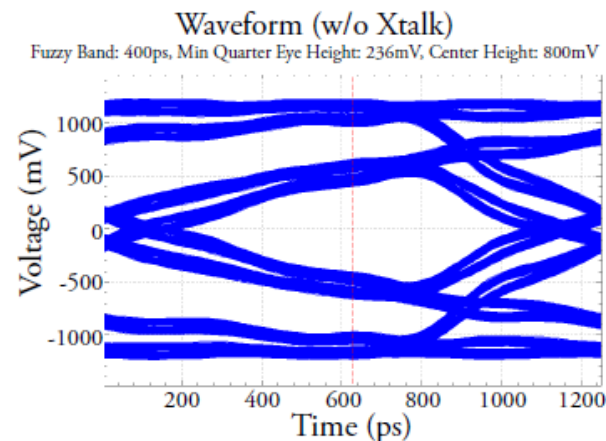


Signal integrity validation

- ▶ Signal integrity for the proposed LPDDR2 channel
 - ▶ Using SPICE and models for board traces, wire bonds, devices, connectors
 - ▶ 1.5V, 800Mbps PRBS
 - ▶ 2pF ESD cap, 2nH wire-bond
 - ▶ Industrial buffer models
- ▶ Hardest links have open eyes
 - ▶ Good time and voltage margins
 - ▶ More devices would close the eye



Device DQ to Board-buffer



Board-buffer to CA

Outline

- ▶ Inefficiencies of server DRAM systems
- ▶ Mobile DRAM
- ▶ Evaluation

Methodology

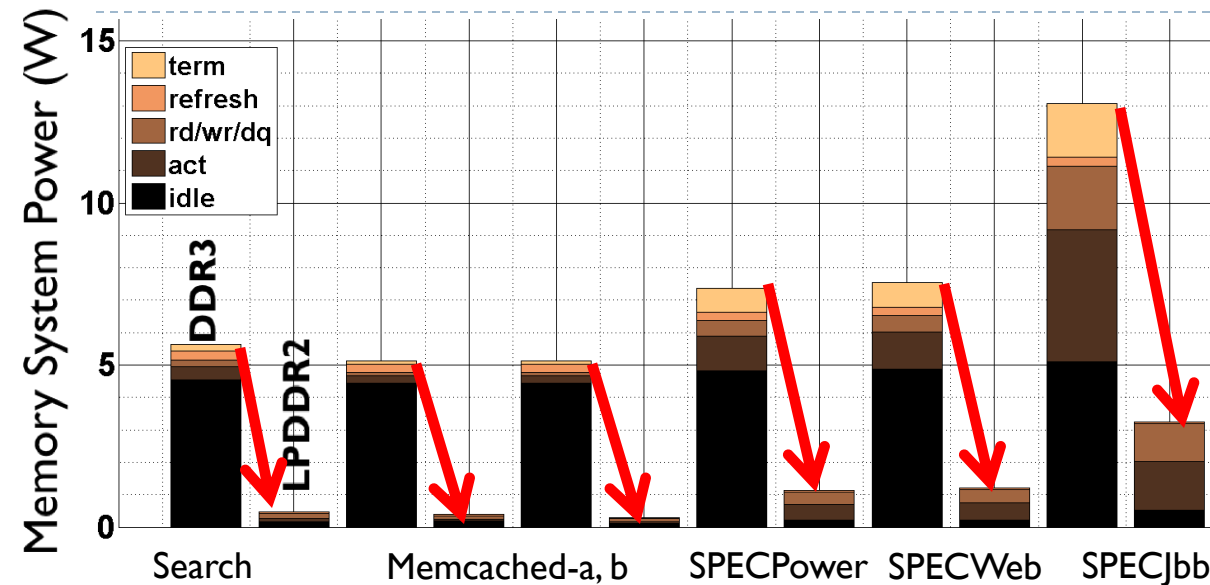
▶ Workloads

- ▶ Websearch at peak throughput
 - ▶ 30GB Wikipedia dataset , 500 top queries
- ▶ Memcached at peak throughput
 - ▶ Access to key, value pairs with 100B and 10KB values
 - ▶ Zipf popularity distribution with exponential inter-arrival times
- ▶ SPECJbb, SPECPower, SPECWeb
- ▶ Multiprogrammed SPEC CPU2006, OMP2001, PARSEC

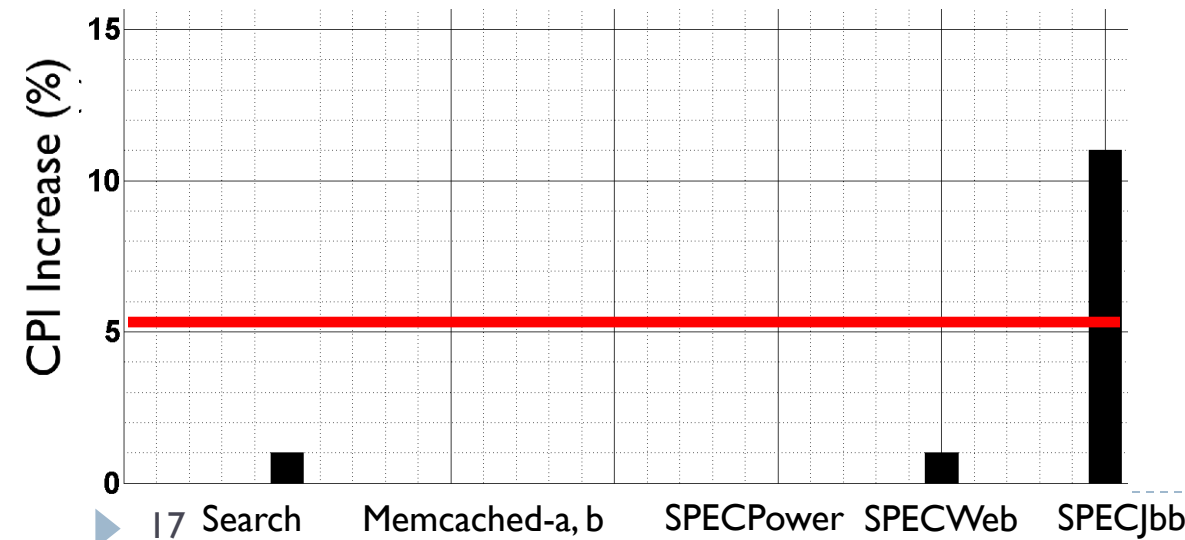
▶ System Architecture

- ▶ 8 OoO Nehalem cores at 3GHz, with 8MB shared L3 cache
- ▶ 2 memory channels: 16 GB capacity using 2Gb DDR3, LPDDR2 chips
- ▶ Validated Pin-driven simulator

Datacenter workloads

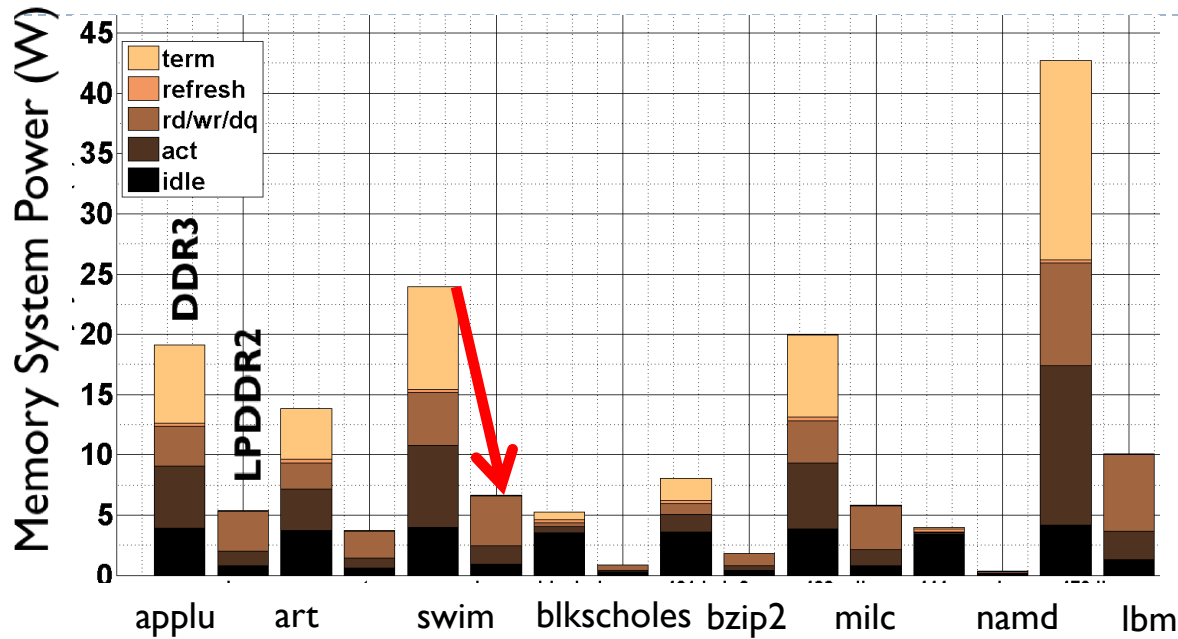


- ▶ 5-6x lower DRAM power
 - ▶ Low active-idle states
 - ▶ Very low power down state
 - ▶ No static termination
 - ▶ Fewer active devices/access

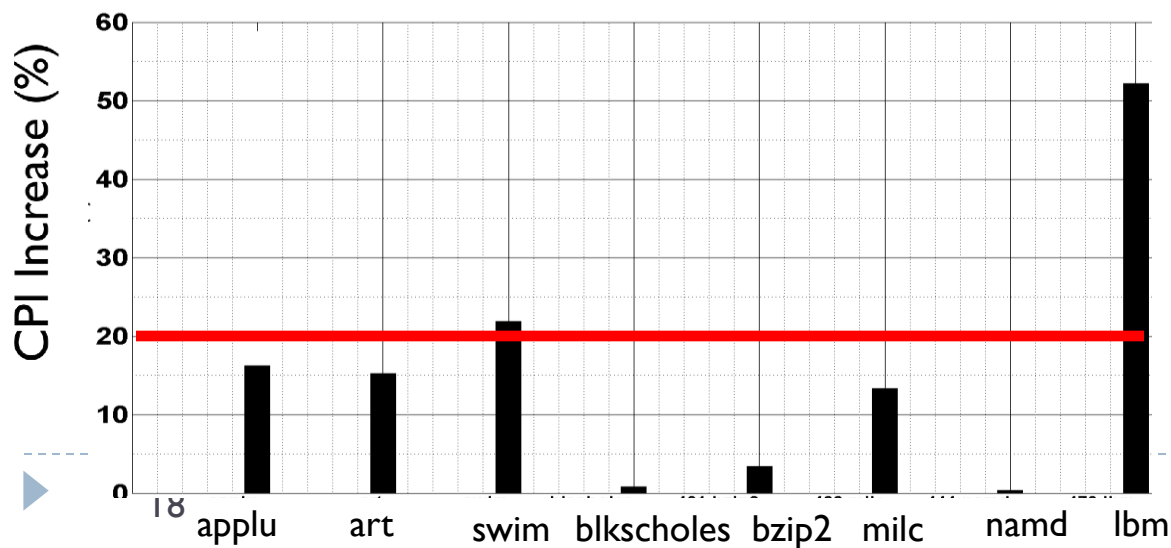


- ▶ Negligible performance impact

Other applications



- ▶ 4-5x lower DRAM power
- ▶ Similar breakdowns



- ▶ 0-55% IPC penalty
- ▶ Depends on application bandwidth requirements

TCO sensitivity to cost of LPDDR2 modules

	Xeon + DDR3 (8 cores)		Xeon + LPDDR2 (8 cores)	
	Cost (\$)	Power (W)	Cost (\$)	Power (W)
Processor (2 socket)	760	125	760	125
Motherboard	200	30	200	30
Network Interface	0	5	0	5
Memory (32GB/2-sockets)	600	40	775	10
Storage (HDD)	100	10	100	10
Total	1660	210	1835	180
No. of Servers (X 10³, in 15MW)	70		83	
TCO (in \$ per sever per month)	\$86.4		\$86.4	
Capability	1.0		1.2	

- ▶ Equal cost per server analysis
 - ▶ Based Hamilton's TCO model and Reddi et.al Bing analysis
- ▶ Can tolerate up to 30% price premium for LPDDR2 modules initially
 - ▶ Will drop as LPDDR2 modules get commoditized
- ▶ LPDDR2 improves datacenter capability by 20%

Combining energy efficient memory + processors

	Xeon + DDR3 (8 cores)		Atom + LPDDR2 (16 cores)	
	Cost (\$)	Power (W)	Cost (\$)	Power (W)
Processor (2 socket)	760	125	360	25
Motherboard	200	30	1340	3
Network Interface	0	5	0	5
Memory (32GB/2-sockets)	600	40	775	10
Storage (HDD)	100	10	100	10
Total	1660	210	2575	53
No. of Servers (X 10 ³ , in 15MW)	70		283	
TCO (in \$ per sever per month)	\$86.4		\$86.4	
Capability	1.0		4.0	

- ▶ Similar equal cost per server analysis
 - ▶ Similar results for other energy efficient processors (e.g., ARM)
- ▶ Can tolerate premiums for LPDDR2 modules, Atom boards
- ▶ LPDDR2 + Atom improves datacenter capacity and throughput by 4x
 - ▶ Note: simple cores can slowdown latency-critical queries

Other conclusions (see paper)

- ▶ Reliability

- ▶ Options for ECC with x16 devices

- Virtualized ECC

- Chipkill

- ▶ Tradeoff between parity overhead and energy efficiency

- ▶ Implications to on-chip cache hierarchy

- ▶ Improved DRAM energy efficiency magnifies LLC static power

- ▶ Question: how big should the LLC be?

- Tradeoff: reduced execution time Vs. increased static power

- ▶ Introduce AMAE metric similar to AMAT to guide analysis

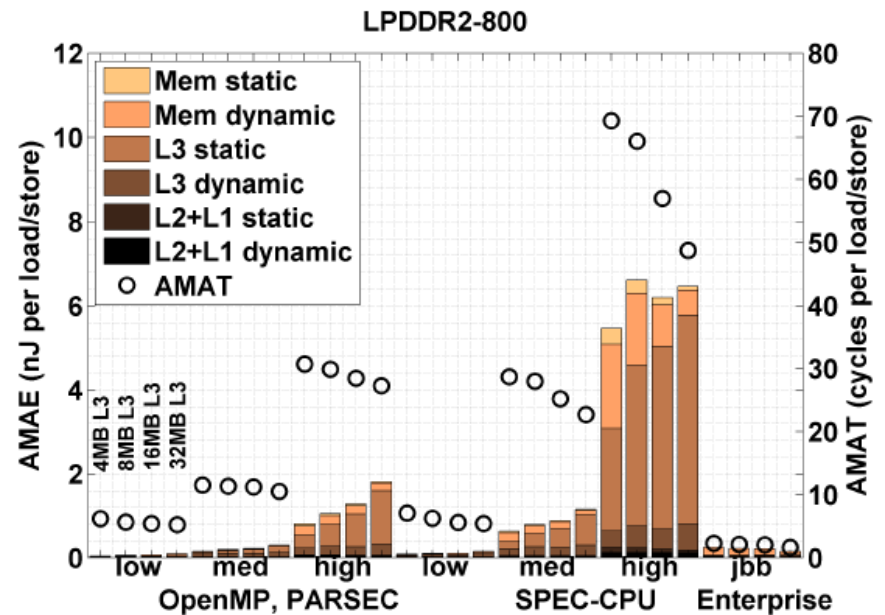
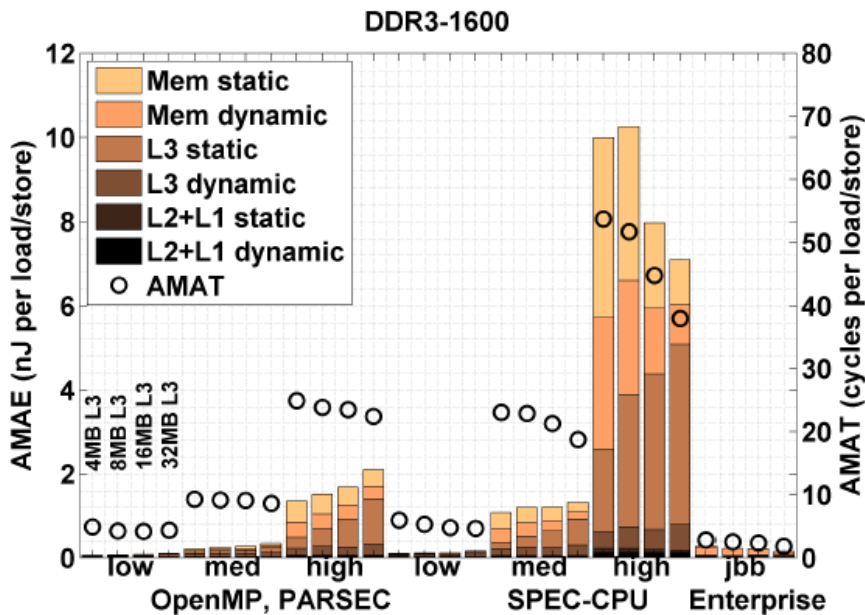
Conclusions

- ▶ **DDR3 memory systems**
 - ▶ Energy inefficient and disproportional due to high static power
- ▶ **Datacenter workloads have low memory BW requirements**
 - ▶ Low bandwidth utilization at 100% load (typical load ~30%)
 - ▶ DDR3 ill suited for these workloads
- ▶ **LPDDR2 memory systems**
 - ▶ Tradeoff peak BW for energy efficiency
 - ▶ 4-5x lower DRAM power and energy proportional
 - ▶ High capacity using die-stacking and buffered channel
- ▶ **Datacenter implications**
 - ▶ Significant capacity improvement even with higher cost modules

Thank you for your attention!
Questions?

Optimizing caches

- ▶ Last level cache sizes & leakage
 - ▶ Need joint optimization of L3 static and DRAM powers
- ▶ Average memory access energy (AMAE)
 - ▶ Quantifies energy efficiency
 - ▶ $AMAE_{L(i)} = Ed_{L(i)} + Es_{L(i)} + MR_{L(i)} AMAE_{L(i+1)}$
 - ▶ Accounts for static and dynamic



Mobile DRAM characteristics

Technology Parameter	DDR3	LPDDR2
Static		
Idle current (power-down, standby)	35, 45mA	1.6, 23mA
Powerdown exit latency	24ns	7.5ns
Operating voltage	1.5V	1.2V
Typical operating frequency	800MHz	400MHz
Typical device width (pins)	8	16
Dynamic		
Timing (CAS, RAS, RC)	15, 38, 50ns	15, 42, 57ns
Active current (read, write)	180, 185mA	210, 175mA
Energy per bit (peak, typical)	70, 160 mW/Gbps	40, 50 mW/Gbps

▶ LPDDR2 Device

- ▶ Similar core, optimized IO interface for very-low static power
- ▶ Similar capacity (ex: 2Gb), chip bandwidth using wider parts
- ▶ Similar timing and dynamic energy currents

System capacity

▶ Key idea

- ▶ Minimize stubs by getting close to single point load
- ▶ Maximize # devices while keeping reliable link margins

▶ Basic block

- ▶ LPDDR2 commodity packages
- ▶ Edge bonded, four x16 2Gb devices

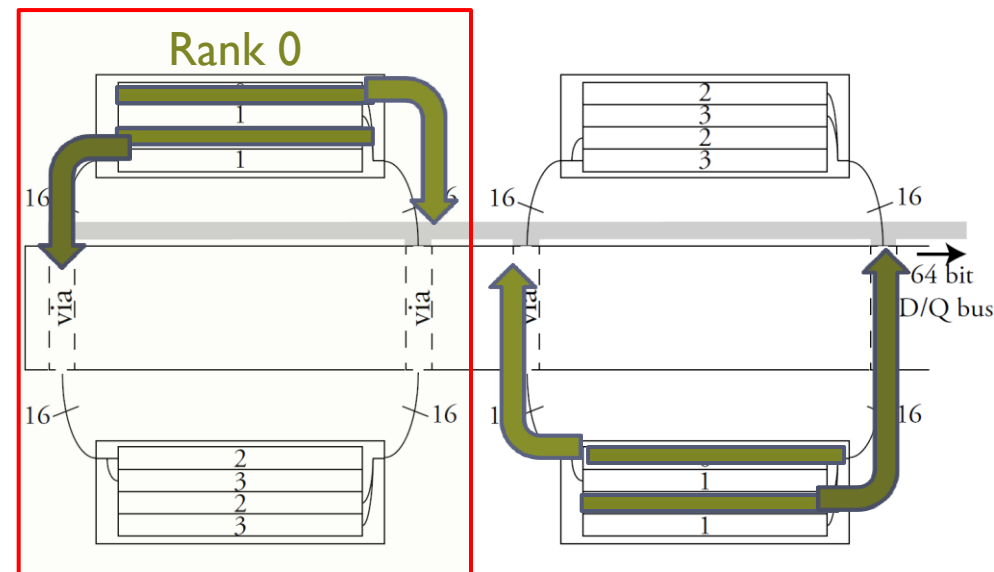
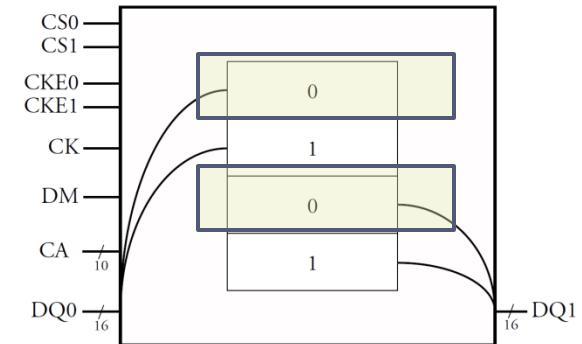
▶ Dual Line Package (DLP)

- ▶ Mirror packages with on-board vias
- ▶ 2x more load per 16-bit trace

▶ Channel interface

- ▶ Stripe ranks to minimize stubs
- ▶ Four x16 traces → x64 to controller
- ▶ No change on controller

▶ 4GB/Channel



High Capacity LPPDR2 Channel

- ▶ More packages/channel → More devices/load → Degrade SI
- ▶ Load Reduce (LR) buffer per channel
 - ▶ 2x DQ and 4x CA duplicate lines provide device isolation
 - ▶ Doubles capacity to 8GB

