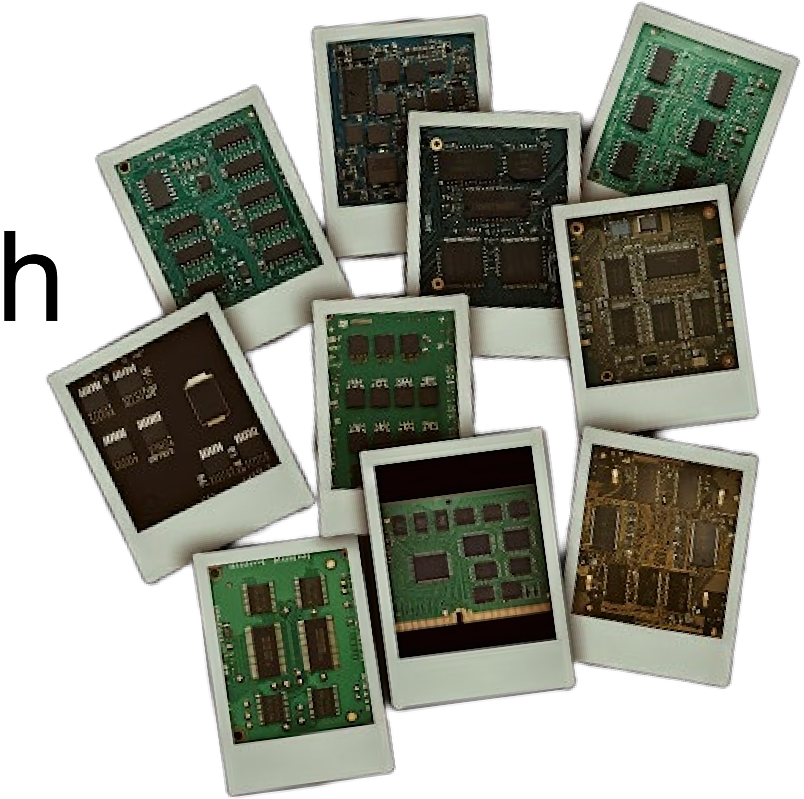
Making Memories at HyperScale with CXL

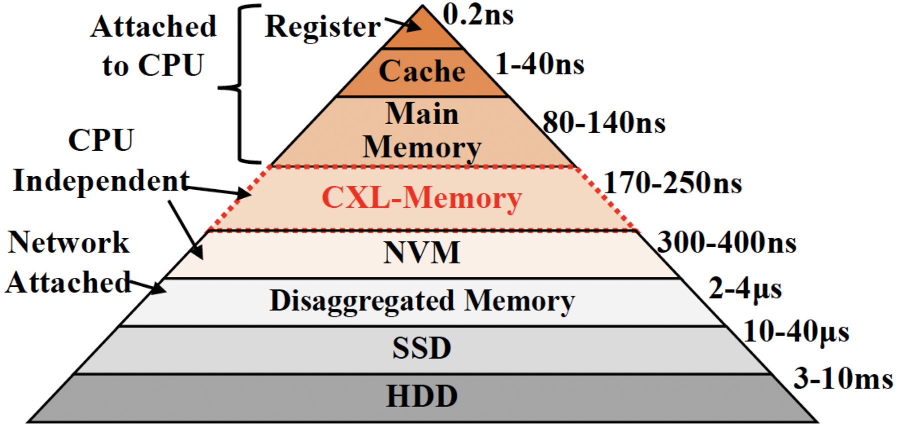
FMS 2024

Brian Morris - Google
Prakash Chauhan - Meta



The Future!

We were promised the future would bring hoverboards and CXL as a new tier in the memory hierarchy



We got our hoverboards, though maybe not as we imagined them.

How can we do the same for CXL?

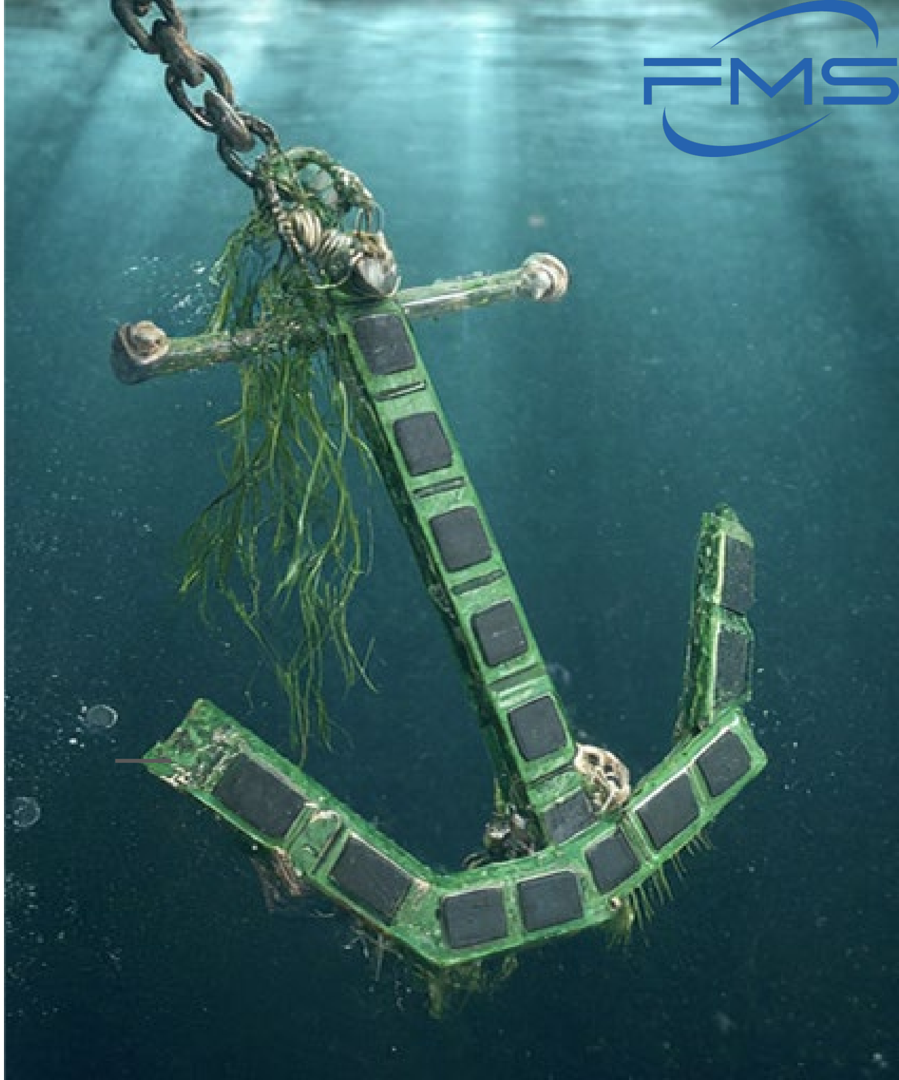
Problem:
The memory cost boat anchor

- Assume memory is ~half of platform cost
- Memory capacity scales with CPU performance
- Memory cost scales with capacity

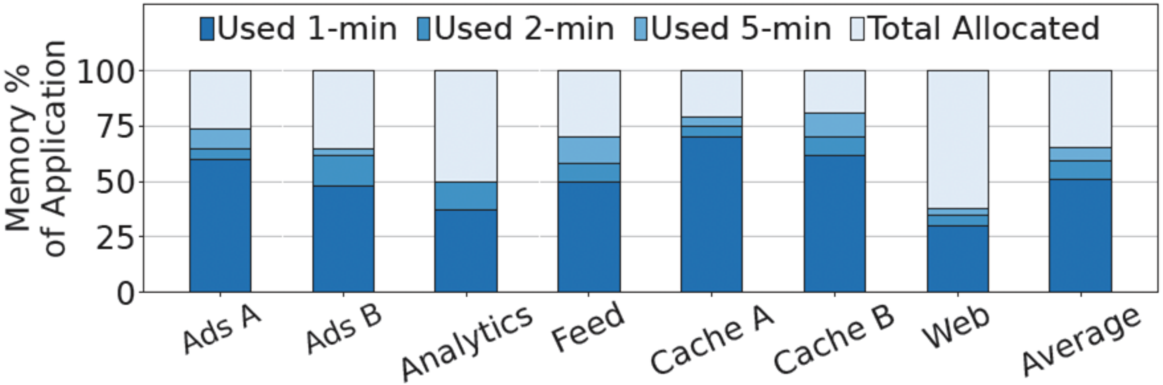
Magic* CPU with 2x performance:
2x Performance ➗ 1.5x Cost ⇒ 1.33x Perf/$

Conclusion:
Memory cost significantly slows platform performance per $ improvements

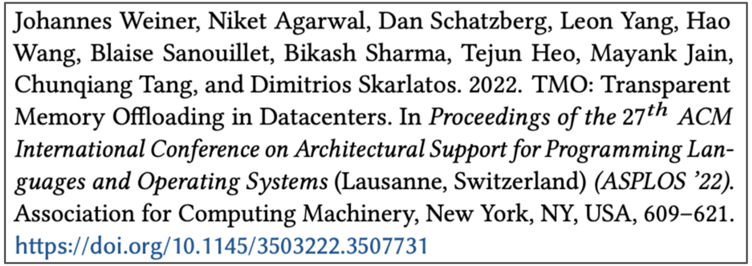
*Realistic CPU generational performance increases are nowhere close to 2x. In the End-of-Moore's Law era, this requires magic



What have we learned?
There is a lot of cold data stored in DDR!

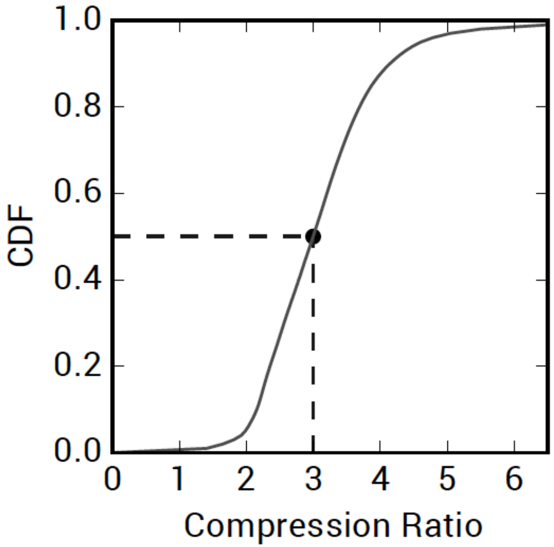


Half of memory hasn't been used in the past minute



Johannes Weiner, Niket Agarwal, Dan Schatzberg, Leon Yang, Hao Wang, Blaise Sanouillet, Bikash Sharma, Tejun Heo, Mayank Jain, Chunqiang Tang, and Dimitrios Skarlatos. 2022. TMO: Transparent Memory Offloading in Datacenters. In *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems* (Lausanne, Switzerland) *(ASPLOS '22)*. Association for Computing Machinery, New York, NY, USA, 609–621. https://doi.org/10.1145/3503222.3507731

What have we learned?
Cold data has good compressibility!

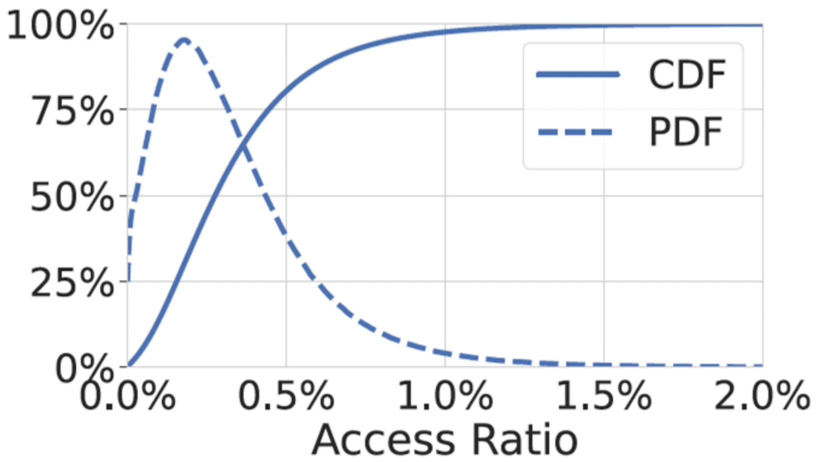


Andres Lagar-Cavilla, Junwhan Ahn, Suleiman Souhlal, Neha Agarwal, Radoslaw Burny, Shakeel Butt, Jichuan Chang, Ashwin Chaugule, Nan Deng, Junaid Shahid, Greg Thelen, Kamil Adam Yurtsever, Yu Zhao, and Parthasarathy Ranganathan. 2019. Software-Defined Far Memory in Warehouse-Scale Computers *(ASPLOS '19)*. 14 pages.

Cold data can compress at a 3:1 ratio

More like 2:1 once we account for incompressible pages

What have we learned?
Cold data doesn't need much bandwidth!



(a) CDF and PDF of tier2 access ratio

Padmapriya Duraisamy, Wei Xu, Scott Hare, Ravi Rajwar, David Culler, Zhiyi Xu, Jianing Fan, Christopher Kennelly, Bill McCloskey, Danijela Mijailovic, Brian Morris, Chiranjit Mukherjee, Jingliang Ren, Greg Thelen, Paul Turner, Carlos Villavieja, Parthasarathy Ranganathan, and Amin Vahdat. 2023. Towards an Adaptable Systems Architecture for Memory Tiering at Warehouse-Scale. In *ASPLOS 2023* (Vancouver, BC, Canada). ACM, New York, NY, USA, 727–741. https://doi.org/10.1145/3582016.3582031

Cold data only requires ~1% of system memory bandwidth

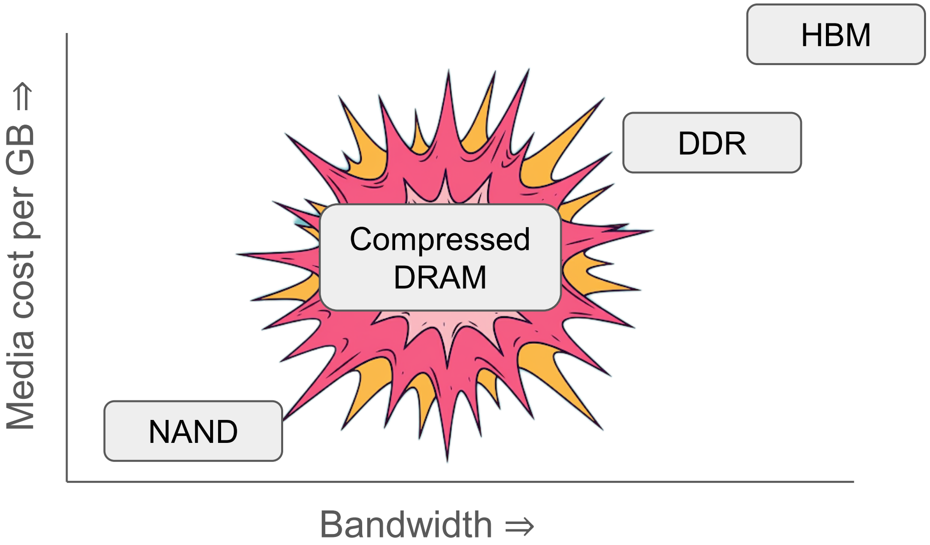
Putting it all together

Lots of cold memory that could go in a slower tier

2:1 compression is realistic across a wide range of workloads

- Halves the media cost in the slower tier
- Compression impacts bandwidth, but the slower tier only needs ~1% of platform bandwidth

Conclusion: Compressed DRAM is viable as a slower, cheaper tier of memory!



Establishing a baseline

Google and Meta introduced a base specification for a new category of CXL device at OCP focused on cost:

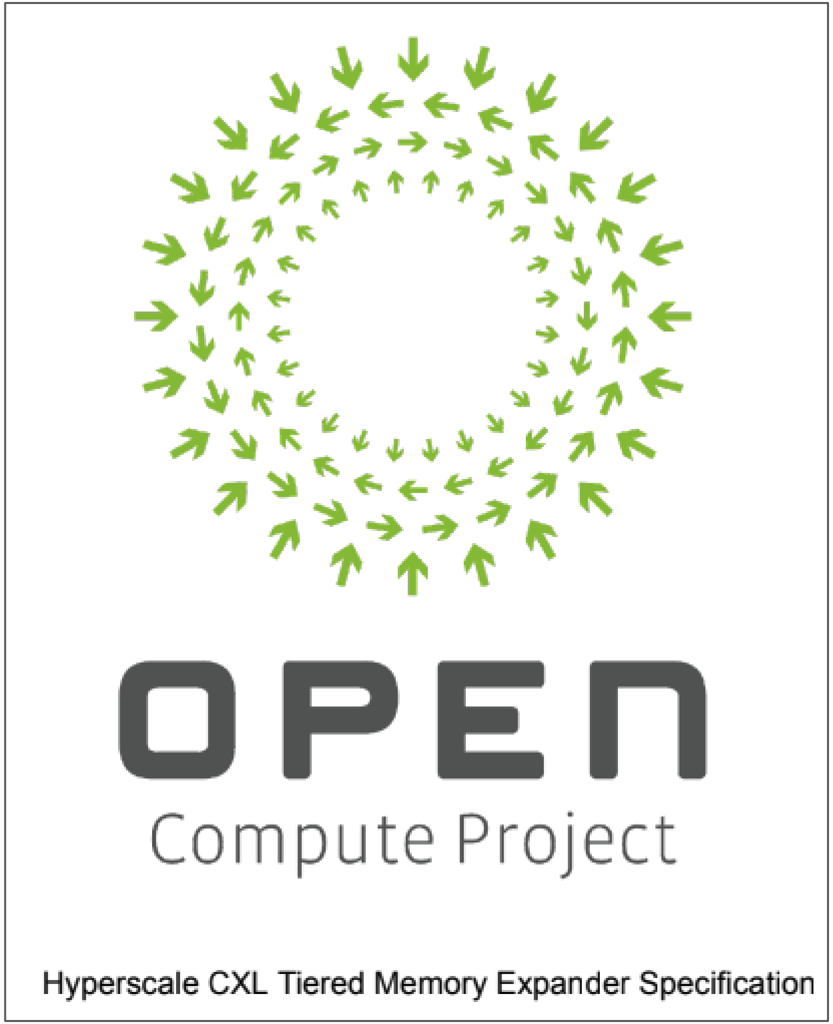
Hyperscale CXL Tiered Memory Expander

Goals:

- Can be deployed at scale
- Expands platform memory in an incremental fashion
- Improves memory cost vs commodity DIMMs

Key features:

- Support for DDR4 at 3 DIMMs per channel
- Inline memory compression at line rate
- Cache for decompressed pages



Hyperscale CXL Tiered Memory Expander Specification

Optimizing for cost

DIMMs

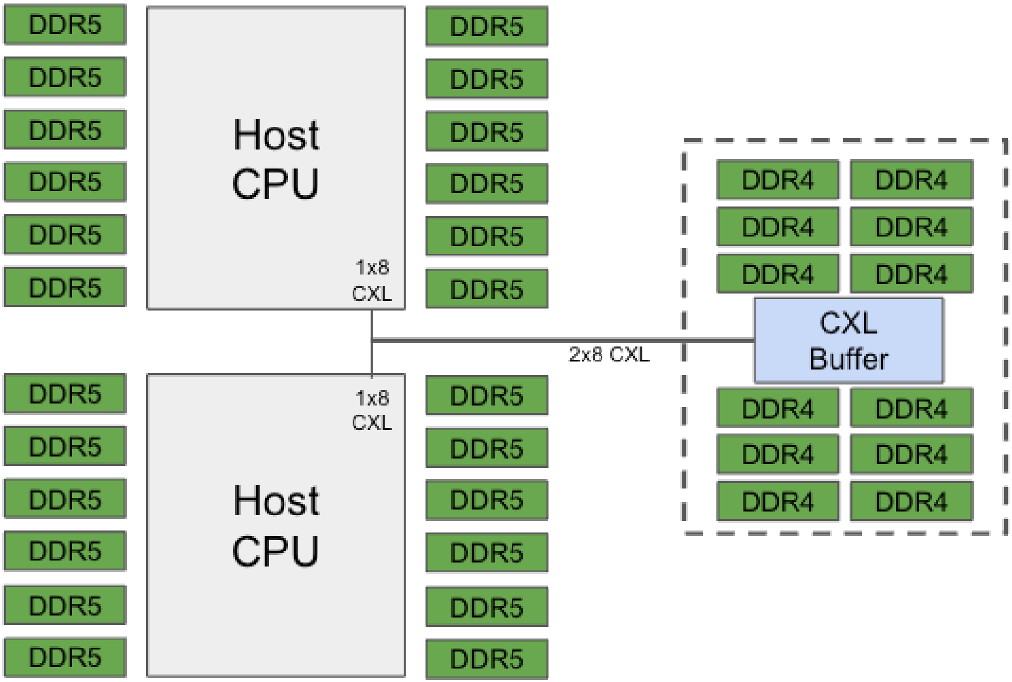
- Modular
- Serviceable
- Existing supply chain

DDR4

- Reuse & overstock
- Saves $
- Minimize embodied carbon footprint

Amortized costs over lots of DRAM!

- 4 channels, 3 DPC ⇒ 12 DIMMs
- One CXL buffer per 432 DRAMs



Example System Config

Industry is chasing E3.s form factor, where none of this is possible 😢

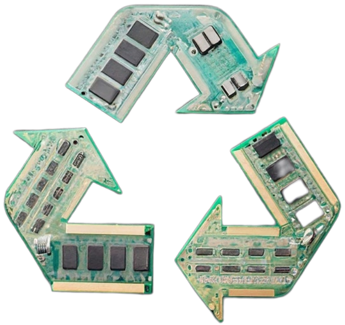
Carbon Impact

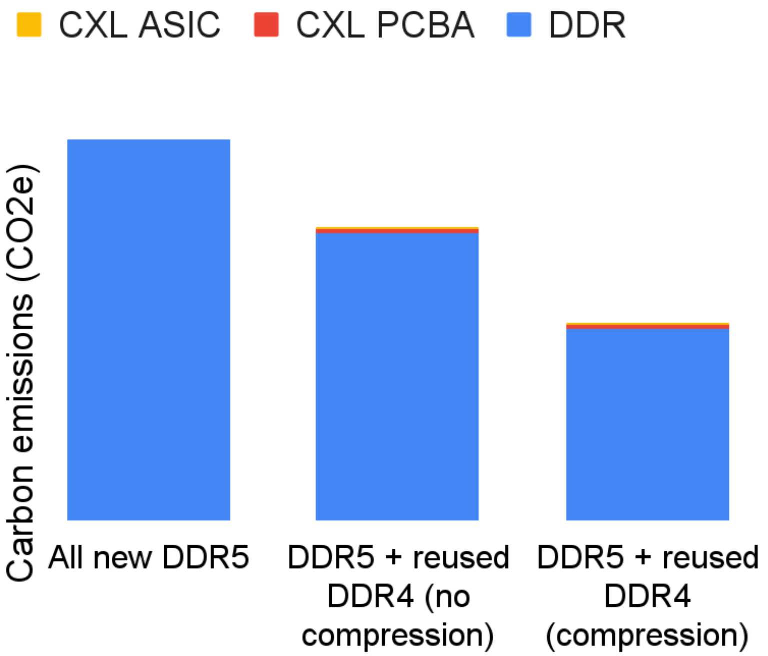
Silicon drives the embodied carbon footprint of data centers

- Majority of carbon footprint is from memory!

Analyze CXL card to accurately assess carbon footprint

- Minimizing carbon footprint of the CXL card is aligned to minimizing cost
- Compression enables even more DDR5 displacement and further carbon savings





🟨 CXL ASIC
🟥 CXL PCBA
🟦 DDR
Carbon emissions (CO2e)
All new DDR5
DDR5 + reused DDR4 (no compression)
DDR5 + reused DDR4 (compression)

Frequently Asked Questions

Question: Which workloads?
Answer: Broad range of compute workloads.

Question: What about encrypted data?
Answer: Can’t compress encrypted data.
Instead decrypt after the CXL link, compress, and re-encrypt.

Question: Why not disaggregation?
Answer: Cheaper memory is an enabler for fancy future CXL capabilities like disaggregation.
Wouldn’t you prefer a pool of cheap memory?

What's next?

OCP Spec was just the starting point:

New products should not "shoot behind the duck"

ASIC vendors:

Compression tailored to the cold memory use case:

- Lower latency decompression
- Optimizations for small block sizes
- Trade-offs in block size vs bandwidth overhead
- Bigger & smarter caches for decompressed pages

Support for multiple tiers (e.g. DRAM, compressed DRAM, NAND)

Improved reliability features for decommissioned DIMMs

Faster host link & better cold page telemetry with CXL 3.1

Standardized package interface

DRAM vendors:

There is market for lower performance, lower cost DRAM!