2020 Memory CSD: Scope and Problem Statements

July 30th, 2020

Carolyn Duran, Frank Hady, Mark Pontarelli and Memory CSD team

Summary

Across our businesses we see key customer workloads (AI, Graphics, and multi-core) bottlenecked by limitations in DRAM bandwidth. This same DRAM is consuming an increasing portion of platform power while commanding an increasing percentage of platform cost. Our investigation across available technologies and business needs concludes that our memory hierarchy must be augmented with two new layers. The first is a Tightly Coupled Memory layer offering very high bandwidth at 1s of GB of capacities. The team concluded that Intel's Adamantine technology has the potential to be a leadership solution in this space giving XPUs a differentiating advantage. This CSD recommends a specific phased strategy to realize this advantage for Intel – this is the most important discussion of this CSD. The proposed strategy starts with a commitment for initial products and investment in multigeneration product development, expands to more products with adjusted margin and profit expectations to establish the technology, and finally scales across Intel in a manner that meets financial expectations. The second layer, In Package Memory, has emerged recently in the industry (HBM) and we see additional use-cases and volumes, which require working with the industry to improve high bandwidth DRAM variants. With a deep customer engagement and a win-win approach with memory vendors we will also explore a viable business model for inclusion of such a memory. Finally, we plan to start a memory initiative to improve business, memory technology, and system technology cohesion across Intel.

Memory CSD Scope

Increasingly data intensive applications accessing larger data sets require more from our memory and storage hierarchy. The incumbent memory technologies making up this hierarchy cannot scale rapidly enough to keep up with the increasing needs. Intel has been methodically working our way through this hierarchy, closing performance and capacity gaps. We addressed the storage performance shortfall of Hard Disk Drives starting in 2006 by innovating with NAND based SSDs and replacing those hard drives for compelling client responsiveness gains and strong DC TCO advantage. In a 2010 CSD we addressed lagging DRAM capacity scaling and increasing BOM percentage with our Optane™ strategy. With products based on this new memory technology in market we are ramping to supply the increased memory capacity and storage performance our customers need. In this CSD we focus on the next clear gap, satisfying the high bandwidth needs of new data-hungry AI applications, competitive graphics solutions, and our rapidly growing processor core counts. These applications require more bandwidth than traditional DRAM solutions can provide and more capacity than can be achieved with SRAM. DRAM performance is simply not scaling quickly enough, and with the end of Dennard scaling, increasing DRAM bandwidth requires more of

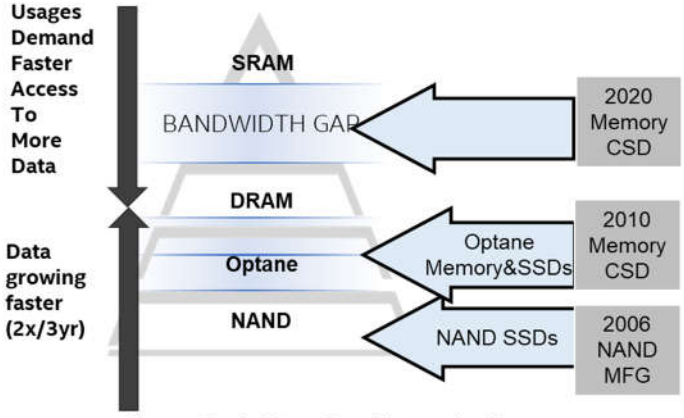

Figure 1. Addressing Hierarchy Gaps

the system's power. A successful strategy to address the memory bandwidth challenge is needed broadly across our Data Center, Client and AI/Graphics Accelerator businesses.

1 - Customer Problem Statements

In the CSD kickoff we demonstrated that Data Center, Client, and AI/Graphics Accelerator customers all feel the impact of DRAM bandwidth limitations. As a reminder, the challenges faced by each set of customers are as follows:

DC Problem Statement: *Core Count is scaling faster than DRAM bandwidth/capacity/power, creating unsustainable increases in our customer's Capex (cost) and OpEx (power), hindering our customers' ability to benefit from CPU improvements, and thus hindering CPU relevance.*

Client Problem Statement: *Client application needs are increasingly misaligned with DRAM technology, with cost-focused platforms requiring less than the minimum capacity available, and performance-focused clients suffering low graphics and multicore performance due to DRAM bandwidth shortfalls.*

Accelerator Problem Statement: for *DC AI Accelerators customers want max BW, max capacity, and min power within their large allotted systems budget, and so push the memory technology. For consumer accelerators, high graphics performance depends on high memory bandwidth to a smaller working set (1-10GB), and solutions are cost sensitive.*

As the CSD team dissected each of these problem statements we found three usages that exceed the existing platform memory hierarchy capabilities. Each featured a working-set that was a poor match for SRAM or DRAM, either requiring more bandwidth than DRAM could deliver or more capacity than SRAM allowed.

1. AI workloads feature workings sets of 10s of GB and need 100s of GB per second bandwidth with performance heavily dependent on memory capability. This workload is already challenging **AI accelerators** and will arrive to our **data center processors** as they introduce AI acceleration in 2021/22.
2. Graphics workload performance is heavily dependent on memory bandwidth as well, but capacity requirements are smaller than AI workloads with 1s of GBs sufficing. Both our **client processors** and **discrete graphics accelerators** must address these workloads.
3. Multicore application performance is rapidly increasing the bandwidth required from memory. More cores hosting more VMs in our **data center processors** is driving increased platform cost to make up for DRAM bandwidth shortfalls. Likewise, high core count **client processor** performance is throttled for applications like content creation due to DRAM bandwidth limits.

These three workloads illustrate the memory bandwidth changes our customers face, and the opportunity for Intel to solve with differentiated solutions.

2 - Intel TAM at Risk

Within Intel's forecasted 2024 $300B TAM this CSD is focused on the XPU opportunity servicing the cloud, enterprise, and client segments. Of that approximately $90B TAM, roughly half is, or will be, in a memory-constrained state based on the performance shortfalls described above. While this CSD addresses the clear and present needs of our cloud, enterprise, and client markets, it is reasonable to expect that similar constraints will arise in networking and IOT as AI and cloud native software increases in adoption in the network and edge. While each of these performance challenges are articulated as a "memory challenge", we do not recommend expanding our TAM by entering the

market for DRAM processes and fabrication. Rather, the most promising solutions to the power and bandwidth issues facing the industry today involve tighter integration of logic and memory. Potential discrete memory solutions beyond HBM are not obvious today.

3 - Technical constraints

Success in selecting a high bandwidth memory and integrating that memory effectively into a system requires staying on the right side of three key technical constraints:

1. **Memory Power** – The physics and logical construction of the memory cell contributes fundamentally to the performance and cost of that memory. It determines the energy required for bit reads and writes (pJ/b). At the memory cell level, bandwidth consumes power and generates heat.
2. **Interconnect Power** – To be useful memory must be connected to compute. Memory technologies that are process-level consistent (like SRAM) can be placed on the same die as the compute, enabling reads/writes that waste very little power. Others, like DRAM, are connected discretely and consume significantly more power moving data from die to die and package to package. For example, the >2pJ/b used to move data from LPDDR to the CPU is roughly 4x the energy of an SRAM cell itself.
3. **Bandwidth/Capacity** – In order to achieve a desired bandwidth from a memory, one must construct a minimum number of cells. This bandwidth-to-capacity ratio varies for different memory technologies – it is a characteristic of the memory technology and its logical organization. The organization can be varied within limits at increased cost. For example, DDR and HBM are both DRAM technologies but HBM is more expensive in part because of its high bandwidth organization.

4 - Memory Hierarchy Recommendation for Discussion in the CSD

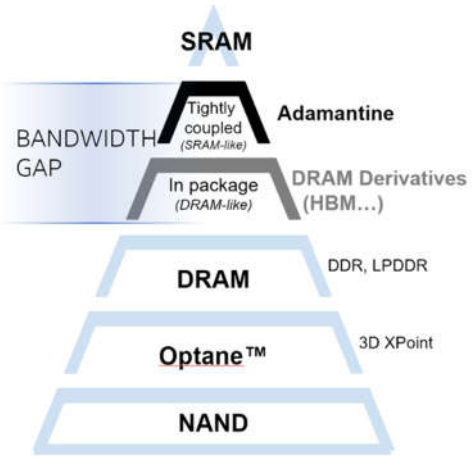

Figure 2. Memory Hierarchy

With these workload needs and memory fundamentals in mind, the CSD team first exhaustively considered the memory technologies available or likely to become available to fill the memory bandwidth gap already described. We determined that there are in fact, two distinct hierarchy layers between SRAM and DRAM, based on capacity needs. For smaller capacities, approximately 500MB to 8GB, we are defining a **Tightly Coupled** Memory layer in the memory hierarchy. Tight coupling is necessary to concurrently resolve the key bandwidth, power, and cost challenges highlighted above. Additionally, a tightly coupled memory solution allows co-design of compute and memory to achieve bandwidth and power goals. The software usage of this memory as either a hardware-managed cache, graphics resource, or perhaps in the future a full main memory, makes its introduction software feasible. Both the multicore caching and graphics use cases fit the attributes of a tightly coupled memory. Data center platforms may also benefit from a L4 cache to enable scaling from 96 to 144 cores without adding additional memory channels (in pathfinding), and both AI and Graphics accelerators may benefit from a tightly coupled memory as a replacement for Rambo SRAM (for AI) and eventually GDDR (for Graphics). For higher capacities, in the 10s of GB capacity, we use the term **In-**

Package Memory, or IPM, to refer to this layer of the hierarchy. IPM is not capable of the same high bandwidths as tightly coupled memory but allows for much larger capacities and significantly higher bandwidth than DRAM. This memory matches the requirement of AI usages in AI accelerators and our future data center processors. As an example, HBM (a DRAM based technology) is already used for our AI accelerators and will be included in a Sapphire Rapids high performance SKU.

While we will present recommendations for each of these, the bulk of the investment and hence the bulk of our discussion will focus on Intel's strategy for tightly coupled memory as an opportunity for XPU differentiation in adjacent and core markets.

4.1 – Tightly coupled memory recommendations

The CSD team assessed the competitive landscape against Adamantine (ADM), an embedded DRAM which is currently in technology development in TMG and POR for MTL-P in client. Logic vendors, including Intel, are driving larger/better SRAM caches, and several companies, including TSMC and Samsung, are investing in STT-MRAM, starting with NVM applications but potentially expanding into tightly-coupled memory applications. AMD, Apple, and TSMC are also developing the 3D stacking technologies necessary to tightly couple compute and memory, akin to Intel's Foveros solution. In this CSD, we assessed four potential solutions for the 2024 timeframe: ADM, SRAM, STT-MRAM from a logic foundry, and potential DRAM process solutions.

- SRAM: While SRAM can deliver extremely high bandwidth solutions, it simply is not capable of providing the necessary capacity at acceptable cost.
- Memory Vendor Solution: If memory vendors were to pursue a tightly coupled solution (theoretical at this point), they would be able to achieve the desired capacities for these use cases, but they are business and technology limited in their ability to deliver the necessary bandwidth/capacity ratio. Additionally, process and business differences limit their ability to tighly couple these memories with XPUs to reach both bandwidth and power targets.
- Logic solutions:
    - Adamantine(Intel): This technology is the best means to address bandwidth, power and capacity challenges. Intel conducted R&D on both ADM and STT-MRAM, but ultimately focused on ADM due to it's collectively superior performance in bandwidth, capacity and power. Note, if a logic foundry should choose to pursue an ADM-like solution, we would expect it to have similar performance characteristics. We have no indication this is happening.
    - STT-MRAM (TSMC): STT-RAM is in development at several companies, primarily as an embedded NVM solution. Should companies choose to productize STT-MRAM as a tightly coupled memory option, we would expect it to have reasonable performance across key metrics.

The team concluded that ADM is the best tighly coupled memory technology with the potential to be a differentiator for Intel in adjacent markets (Graphics, AI accelerators) and a key tool to empower compute expansion in Xeon by reducing customer's reliance on DRAM.

This bring us to the most important discussion of this CSD: Intel's resolve to commit up front to invest in a technology roadmap and the neccessary business tradeoffs to bring disruptive products with ADM to market and mature ADM in the long run as a differentiating technology at scale for Intel. We have

been here before, with eDRAM, and we are using those lessons learned to propose a different way forward. We believe Intel must make a choice: used a phased approach to ADM development and productization as shown below, or exit ADM development, decommitting it from MTL-P and suffering potential competive risks while we look to the memory ecosystem for alternatives.

Phase	What it Takes	Success
Disrupt and Commit by 2022/2023	Committed investment to ADM roadmap (100s of $M, 22nm) Targeted, small volumes (~10-20MU) Margin compromise	Design wins for MTL-P Early learning on ADM in DC applications
Innovate and Establish by 2023/2024	Continued TD roadmap BU committed resources Deep working set/use case knowledge Margin compromise	Maturing technology to drive scale Developing common requirements Design wins in key areas
Scale by 2024/2025	Ultilize ADM across businesses as a core business opportunity	Continued TD roadmap must hit cost metrics Must be competitive vs. alternatives

4.2 In Package Memory Plans

The team identified high bandwidth DRAM derivative technologies, like HBM, as the best solution for 10s of GBs of capacity needed for IA accelerator workloads. Dense DRAM fab processes enable these high capacities. At the same time the changes to memory architecture and packaging that enable higher bandwidth also significantly increase memory cost. There have been several prior attempts at enabling in package memory, with limited commercial success. In this area it is critical to apply learnings from past attempts to enable future success.

Past Experience / Learnings

- Focus on solving the customer problem first. We need a deep understanding of workloads, their use in memory, and their future trajectory.
- Given it's high capacity in package memory is visible to software, and so our plan must be software-inclusive.
- New innovations must be meet internal and memory ecosystem business models and affordability expectations.

Taking these learnings into account, this team plans to proceed as follows. First, we will strengthen our driving of memory with industry, deepening our existing collaborations to more strongly influence new-to-Intel memories such as HBM and GDDR. Secondly, we will identify and pursue an opportunity to execute a strategic partnership if and only if it is done in partnership with a definitional customer to develop a differentiated solution.

5 - Systems Plans

While much of this CSD focused on the need for multi-generational solutions to address key customer pain points around bandwidth, power, capacity, and cost, we would be remiss to overlook the significant system level efforts underway across the corporation to mitigate these pain points in the short term. System innovations including compression, memory pooling, flat 2LM, memory channel additions, etc., each provide one-time advantages. There technologies must be pursued on a product

by product basis but are not sufficient to fundamentally address the scaling limitations of DRAM and SRAM in the long term. We recommend that BU's continue to develop a portfolio of systems innovations to mitigate customer pain points. As competition is also actively innovating in this space, Intel solutions must meet standard business expectations in volumes and margin to remain competitive. As an aid to BU decision-making, we also propose the formation of a Memory Initiative, described in the next section.

6 - Memory Initiative

The CSD team has decided, with support from Rich Uhlig, to proceed with structuring a memory initiative under the Disciplined Innovation and Technology Initiatives 2020 umbrella introduced this year as part of the Culture MRC. The general initiative effort is intended to address a weakened track record of maturing and landing high-potential innovations with an associated cadence of improvements over time. Reducing splintered/siloed efforts and restoring the strong links between technology readiness and product POR intercepts are also key elements of the effort.

Through the Memory CSD process, we uncovered several examples where we believe a Memory Initiative would benefit the corporation:

1. Enable BU decisions in memory with clear, consistent data regarding the memory ecosystem.
2. Improve ability to identify patterns in customer pain points in conjunction with ecosystem intelligence to quickly align on strategies.
3. Utilize increased workload intelligence to converge memory requirements where possible and necessary to consolidate volumes aligned with Tightly Coupled and IPM innovations.
4. Provide clear direction to TMG and the memory ecosystem as applicable.
5. Quickly assess memory program and project assumptions across the company if/when business conditions change, both internally and in the ecosystem.
6. Provide early-warning mechanism for strategic corporate programs relying on product intercepts across businesses to ensure a holistic action plan.
7. Ensure robust technical readiness for new memory technologies, including understanding of the solution in the context of the whole memory hierarchy and software enabling required.

Requests/Next Steps:

The Adamantine specific elements of this initiative are dependent on the decisions made by ELT regarding the future of the technology. That notwithstanding, we are seeking agreement at the ELT level to work toward a disciplined engagement model between the technical and product communities and identify key initiative resources (Initiative lead, ADM roadmap integration lead, PM).