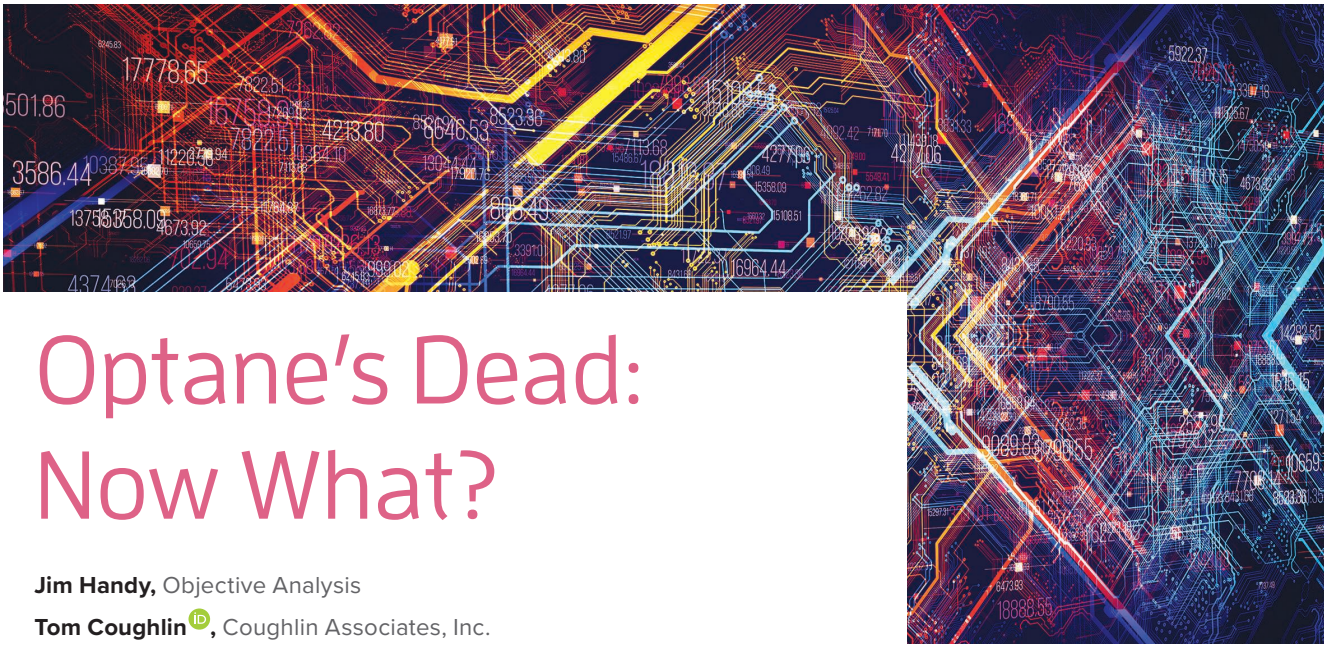
EDITOR TOM COUGHLIN
Coughlin Associates; tom@tomcoughlin.com



Optane's Dead: Now What?

Jim Handy, Objective Analysis
Tom Coughlin, Coughlin Associates, Inc.

Does Optane's demise signal the end of the use of emerging memories in mainstream computing? Not even close! Even today, magnetic random-access memory (RAM) and resistive RAM are making serious inroads into Internet of Things endpoints.

Since Intel announced in July that the company would "wind down" its Optane efforts,[1] there has been much concern over the fate of emerging memories. In this article, we explore the reason for Intel's move and the impact it is likely to have on all emerging memory technologies, based on a thorough report by Coughlin Associates and Objective Analysis covering emerging memories of all kinds.

WHAT IS OPTANE?

For those asking, "What is Optane," the simple answer is that it's an emerging memory technology that Intel introduced to improve computing's cost/performance ratio. An "Optane-persistent memory module" was to be inserted at half the cost per byte of dynamic random-access memory (DRAM) between a server's DRAM and its solid-state drive (SSD) as a new memory layer. It would perform some of the duties of the SSD at near-DRAM speeds, and it would bring down the cost of the overall system by permitting the system to use less DRAM.

This is the boldest step yet any company has taken to try to bring an emerging memory technology into direct competition with the high-volume leaders: DRAM and NAND flash. By our estimates,[2] Intel spent approximately US$10 billion trying to make this experiment a success before deciding to abandon it, as shown in Figure 1.

WHY OPTANE DIED

What went wrong? Why was Intel unable to make this technology profitable? That's a story that takes some explaining.

When Optane was introduced in 2015,[3] it was first known as *3D XPoint memory*. This was a variation of phase-change memory (PCM or PRAM) that used a two-terminal selector, so bits could be stacked on top of one another in what is called a *Crosspoint* structure. Crosspoint arrays are the densest-possible layout of bits, and the more bits you get onto a chip the cheaper the cost per bit. If you can stack two layers

Digital Object Identifier 10.1109/MC.2023.3235096
Date of current version: 8 March 2023

you can double that efficiency, and if you can stack more layers, it's even better.

Add to this the fact that PCM can use tighter process geometries than can either DRAM or NAND flash and you get a road map for costs that reduce over time at a faster pace than established technologies.

At the 2015 introduction, Intel and its then-partner Micron claimed that 3D XPoint memory was half the die size of "Standard Memory," by which they meant DRAM. If it was half the size, then the assumption stood that it could be produced at half the cost of DRAM, but this only holds if the two technologies have the same wafer cost. The wafer costs could only be brought into line if Intel sold enough of its Optane products to drive large wafer volumes—large enough to drive down the costs.

Understanding this, Intel launched 3D XPoint-based Optane SSDs, which were intended to drive large volumes early in the technology's career since the dual inline memory module (DIMM) format required a number of changes to software architecture to become a reality.

Optane SSDs never became very popular,[4] although, as their performance was throttled by their nonvolatile memory express interface to the point that they were not all that much faster than an NAND-based SSD, while they were sold at a hefty price premium over NAND SSDs. This kept initial wafer volumes low.

By the time Intel was able to introduce Optane in the DIMM format, a task that required changes to processor architecture, operating systems, and application programs, Optane's production cost was still very high, yet it didn't make sense to buyers unless it was significantly less expensive than DRAM. The sub-DRAM price that Intel offered was important for Optane to fit into the memory/storage hierarchy shown in Figure 2. Intel struggled for roughly four years to use DIMMs to increase wafer demand, but the adoption of this relatively radical new way of structuring application programs was slow to come, and Optane continued to lose the company substantial sums of money.

When we review Intel's approach, we see nothing from an execution standpoint that we would have done differently. The failure was a strategic one, and it was based on a flawed understanding of how important the economies of scale are in the memory business.

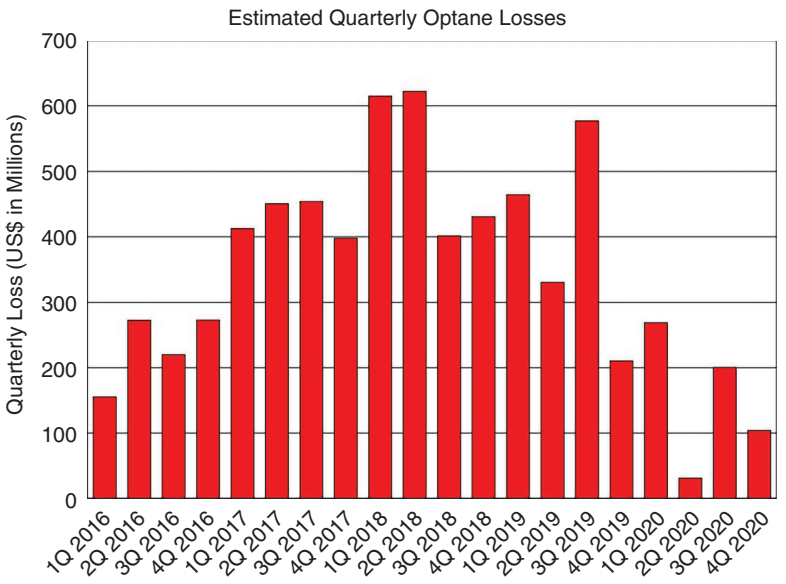


Figure 1. Estimate of Intel annual losses attributed to Optane memory. (Source: Objective Analysis.) 1Q: first quarter; 2Q: second quarter; 3Q: third quarter; 4Q: fourth quarter.

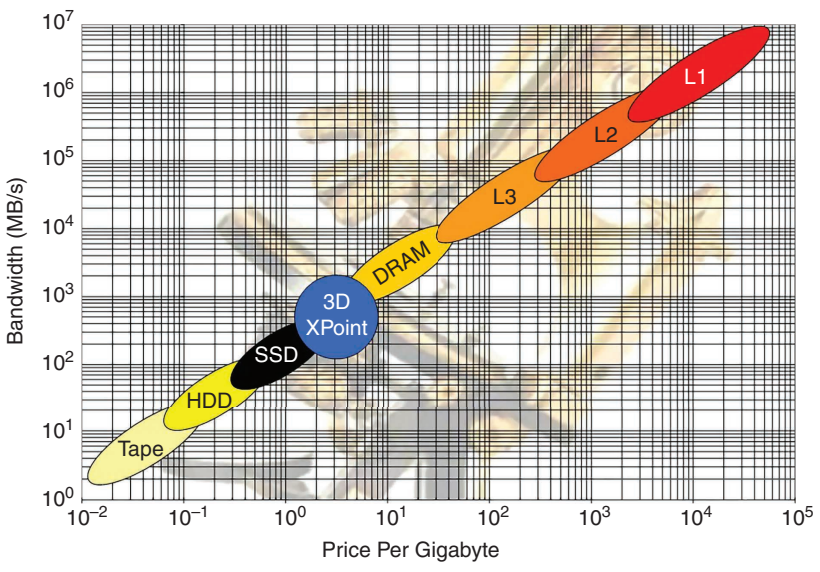


Figure 2. A representation of the memory/storage hierarchy, including 3D XPoint memory. (Source: Objective Analysis.) HDD: hard disk drive.

ECONOMIES OF SCALE: LESSONS LEARNED FROM NAND FLASH

An example makes it easier to understand just what went wrong. Since its introduction, NAND flash has had an apparent cost advantage over DRAM. An NAND flash die with the same number of bits as a DRAM, that is, produced using the same process geometry as DRAM, will be half as large as the DRAM die.[1]

(For those with a deeper understanding of NAND flash, we add that this is based on single-level cell NAND, the largest die size for that number of bits.)

You would think that the cost to produce a byte of NAND flash, then, should have always been roughly half the cost of a byte of DRAM. But that was wrong.

Figure 3, from the emerging memories report, compares the sales price of NAND flash with that of DRAM. Memories often sell at production cost, so the data in this chart can be used to approximate cost.

NAND didn't cross below DRAM until 2004, the year that NAND wafer volume reached approximately one-third that of DRAM. At this point, the cost to produce an NAND wafer must have been about twice the cost of a DRAM wafer. Until that point, NAND wafers were even more expensive. This underscores the importance of wafer production volume, or "the economies of scale," to overall costs.

3D XPoint memory's wafer volume was never even 1/10th as high as that of DRAM. At its historic growth rate, it might never reach a high-enough volume to have become profitable.

Could Intel have managed things differently to get the wafer volume up? Indeed, it could have. As an extreme example, let's say that the company aggressively sold its high-performance Optane SSDs at prices that matched standard NAND flash SSDs. That action would certainly have ramped volume high enough to drive costs to the necessary level, but it would have multiplied Intel's financial losses to many times our US$10 billion estimate.

The question remains: What does this mean to other emerging memory technologies, like magnetic RAM (MRAM), resistive RAM (ReRAM), and ferroelectric RAM (FRAM)?

WHAT OTHER MARKETS EXIST TODAY?

A number of companies have seen some level of financial success in their emerging memory efforts. Everspin and Renesas sell MRAMs, Adesto sells ReRAMs, and Infineon has a successful FRAM business. Other companies use these same technologies as embedded memories in microcontrollers, including Fujitsu, STMicroelectronics, and Texas Instruments, and some foundries support MRAM and ReRAM, including Taiwan Semiconductor Manufacturing Company, Samsung, and GlobalFoundries.

These technologies are usually favored in applications that either need a nonvolatile memory that consumes very low energy or that is exposed to high levels of radiation. High radiation is often a problem in space, but it's also an issue in surgical instruments that are sterilized with high X-ray doses. These are very low-volume applications. Low energy is key to wearable devices, from wearable fitness devices, to pacemakers and Internet of Things endpoints. These are fast-growing markets.

But none of these markets consume a very large share of semiconductor wafers, and that share is likely to remain relatively low for the immediate future, preventing these technologies from reaching important-enough wafer volumes to drive down costs. The quandary is figuring out a way to prevent emerging memory technologies from being relegated to their current niche markets.

HOW EMERGING MEMORIES CAN BREAK OUT OF THE LOW-VOLUME BUSINESS

There are two obvious paths to drive emerging memory volume. The first is to become relatively widely used as the embedded memory in microcontrollers and other systems on chip (SoCs). This will drive increased wafer volume, which will reduce production costs. The other is to carefully target other higher-priced memory technologies and displace them based on lower cost. This one would seem relatively

Low energy is key to wearable devices, from wearable fitness devices, to pacemakers and Internet of Things endpoints.

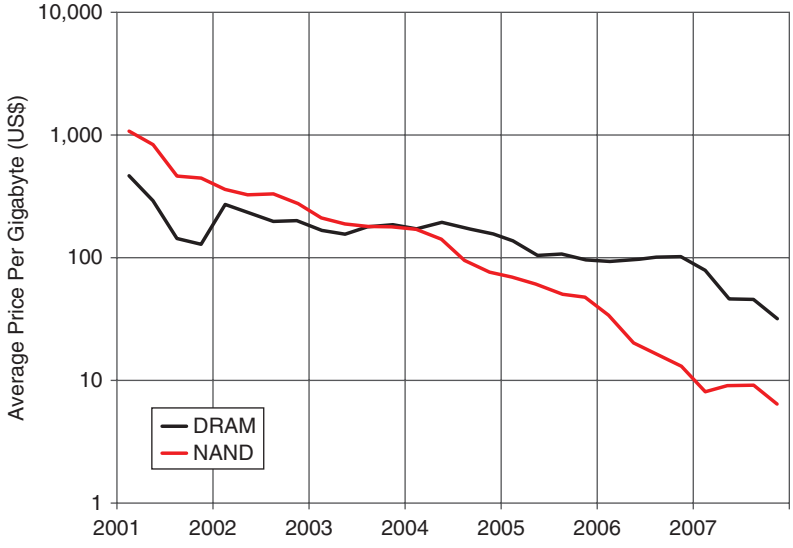


Figure 3. History of the sale price per capacity of DRAM and NAND flash. (Source: Objective Analysis.)

simple as technologies like electrically erasable programmable read only memory and SRAM have enormous die sizes compared to any emerging memory technology, but a smaller fraction of these chips' cost is the die, while package and test consume a more important share of costs, and system designers need a lot of motivation to convert from a widely sourced technology that they currently use to a new technology that may be sole sourced. A very strong sales push and some good financial backing would be required to make this work.

Of these two, the most likely path to higher volumes is through embedded memories because CMOS logic processes need to find an alternative to their current memory mainstays: NOR flash and SRAM. They have been driven to this point by the industry's constant pursuit of Moore's law.

HOW MOORE'S LAW SCALING DRIVES A NEED FOR EMERGING MEMORIES

Emerging memories' greatest opportunity has come about because of Moore's law scaling.[5] The semiconductor industry's relentless pursuit of lower costs through transistor shrinks has reached a point where it no longer works to simply perform lithographic shrinks of conventional transistor configurations; new transistor types have had to be developed, including fin field-effect transistors (FETs), ribbon FETs, and gate-all-around transistors.

This great change in transistor structure is being accompanied by a great change in the types of embedded memory available within SoCs. NOR flash and SRAM are poised to be displaced by emerging technologies. Let's look at each in turn.

> CXL allows memory to be allocated to the server with the most pressing current need, and then to be reallocated to other servers after that need has been fulfilled.

28 nm: The End of the Road for NOR Flash

NOR flash, the most common, nonvolatile memory in SoCs, has run into a brick wall. There is no NOR flash technology that works with FinFETs, the transistor that the semiconductor industry must use to produce CMOS logic chips on processes smaller than 28 nm.

At the moment, this is not that great of an issue as the SoCs that are built on sub-28-nm processes are largely high-end processors for servers, PCs, and cell phones. These don't typically incorporate any nonvolatile memory.

As Moore's law progresses, and as microcontrollers and other high-volume products begin to migrate below the 28-nm node, a replacement for NOR will be required, and this technology will be FRAM, ReRAM, MRAM, or perhaps some other emerging memory technology that is less prominent today.

SRAM Scaling Issues and Their Role in This

NOR flash isn't the only problem. As processes shrink, the SRAM used for embedded, rewritable memory will not shrink as aggressively as will the logic.

The impact of this is that most performance processors for servers, PCs, and cell phones will not shrink in proportion to the process as a large part of the die area (often 50% or more) consists of SRAM cache memories.

Figure 4, also from the emerging memories report, illustrates the desired shrink path based on an image of a typical processor chip. The top half of the chip is the logic, while the bottom half is the SRAM. As processes shrink, the die size shrinks along with it to produce a proportionally smaller, and less costly, chip.

But SRAM in advanced process nodes no longer scales in proportion to logic, leading to a scenario more like that shown in Figure 5: the red line represents the desired chip size, and the black line represents what is happening today, thanks to the fact that SRAM no longer scales in proportion to the process.

As SRAM is already significantly larger than any emerging memory technology, there is already motivation to replace the SRAM with an emerging memory technology. The only thing standing in the way is that SRAM writes are a few-times faster than writes into an emerging memory technology, and that those writes cause wear that leads to bit failures.

Workarounds already exist for these problems, and processor designers are already interested in the fact that a megabyte of SRAM might be replaced by

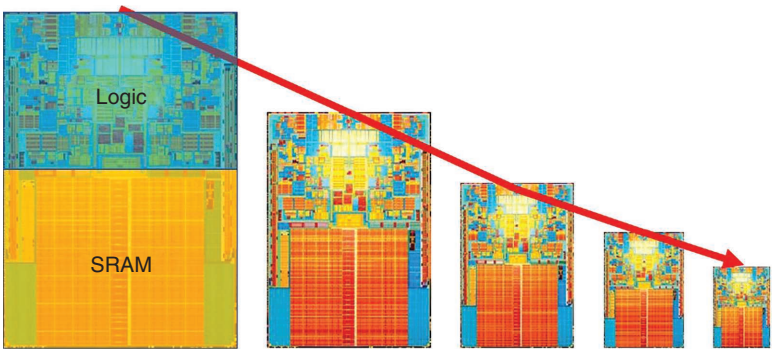


Figure 4. Chip-size reduction where logic and memory shrink at the same rate. (Source: Objective Analysis.)

6 MB of an emerging technology in the same die space at a small expense in write speed and wear. In the near future, we are likely to see SRAM caches diminish in size as slower/larger emerging memory caches are used to augment them.

The Role of Chiplets

Meanwhile, a radical new approach to packaging is likely to be brought into play. Microprocessors have already started to be produced using a few "chiplets" in a single package instead of a single monolithic die. This allows the processor to be designed as if it was using a die much larger than anything that can be built today. It also opens the door to the possibility that L2 or L3 caches, the slower, more distant ones, might be moved to a different chip than the processor, with very wide buses running at extreme speeds between the two.

This would allow the cache chip to be built using a different process technology than the processor, with the processor using high-performance CMOS while the cache chip is built using an emerging memory process. Both chips would be less costly as a result while providing higher performance.

The Universal Chiplet Interconnect Express (UCIe)[6] interface standard has been developed to help accelerate this effort. The UCI is not burdened with the legacy of standard input–output (I/O) pins, and this helps reduce pin capacitance, shorten interconnect lengths, reduce I/O power, and make interchip communication speeds significantly faster. We view UCI as a path to improved price/performance that can help continue the pace of processor improvements despite today's scaling challenges.

WILL COMPUTE EXPRESS LINK PLAY A ROLE?

Compute Express Link (CXL)[7] is a new interconnect technology designed to accelerate communication between large pools of memory and the host processor to faster rates than are supported by NVMe. As Intel originally indicated that CXL might be a good way to attach Optane to processors, is it a vehicle to promote other non-Optane emerging memory technologies like MRAM and ReRAM?

In a word, no. CXL is a way to improve system price/performance in those systems that occasionally need very large memories, but more often require much smaller memory. It is the last step in disaggregation after server virtualization and composable infrastructures. CXL allows memory to be allocated to the server with the most pressing current need, and then to be reallocated to other servers after that need has been fulfilled.

In this environment, there are only two types of memory that make sense in a CXL system: DRAM and something cheaper than DRAM. Today, all the emerging memory technologies are more costly than DRAM, so they are a poor choice for CXL memory pools.

> In the near future, we are likely to see SRAM caches diminish in size as slower/larger emerging memory caches are used to augment them.

HOW THIS IS LIKELY TO PAN OUT

When we weave this all together, we see the following likely direction for the success of emerging memory technologies:

- Discrete or standalone emerging memories will serve niches but will not drive sufficient volume to become mainstream.
- Embedded emerging memories will ramp to high-volume wafer production that will drive out the costs, making discrete emerging memories more affordable.
- Much of this volume will stem from CMOS logic migrating past 28-nm processes
- After an emerging memory replaces embedded NOR flash, it will move on to replace a substantial share of embedded SRAM.
- The broadening use of chiplets and UCI will also help to increase emerging memory wafer volume by using a dedicated emerging memory process to unburden memory from the processor while reducing costs.

This will all take a long time, but in the end, it will allow at least one emerging memory technology to move out of the shadows of market niches into the mainstream.

These arguments are spelled out in significantly greater detail in a research report jointly published by

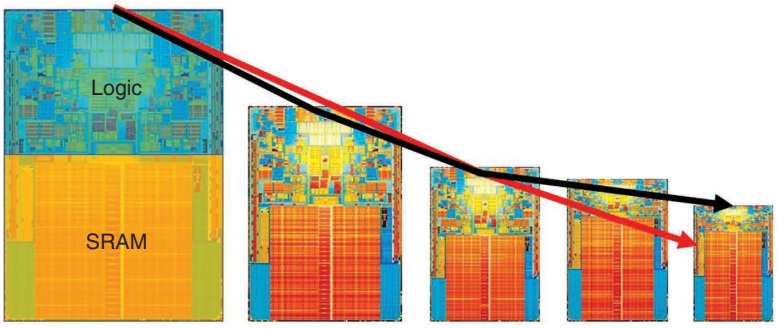


Figure 5. Chip size reduction where logic and memory do not shrink at the same rate. (Source: Objective Analysis.)

Objective Analysis and Coughlin Associates, from which the bulk of this article has been extracted. For more information visit http://Objective-Analysis.com/reports/#Emerging, or http://TomCoughlin.com/product/emerging-memory-report/.

ACKNOWLEDGMENT

Tom Coughlin is the corresponding author.

REFERENCES

1. P. Gelsinger, private communication, Jul. 28, 2022.
2. "Emerging memories enter the next phase," Objective Analysis, London, U.K., 2022. [Online]. Available: https://objective-analysis.com/reports/#Emerging
3. "Press release: Intel and micron produce breakthrough memory technology." Intel. Accessed: Jul. 28, 2015. [Online]. Available: https://www.intc.com/news-events/press-releases/detail/324/intel-and-micron-produce-breakthrough-memory-technology
4. B. Tallis. "The Intel Optane SSD 800p review." AnandTech. Accessed: Dec. 28, 2022. [Online]. Available: https://www.anandtech.com/show/12512/the-intel-optane-ssd-800p-review
5. C. A. Mack, "Fifty Years of Moore's Law," *IEEE Trans. Semicond. Manuf.*, vol. 24, no. 2, pp. 202–207, May 2011, doi: 10.1109/TSM.2010.2096437.
6. Universal Chiplet Interconnect Express, Beaverton, OR, USA. [Online]. Available: https://www.uciexpress.org/
7. "Compute Express Link™: The breakthrough CPU-to-device interconnect," Compute Express Link Consortium, Beaverton, OR, USA. Accessed: Dec. 28, 2022. [Online]. Available: https://www.computeexpresslink.org/

JIM HANDY is the general director of Objective Analysis, Los Gatos, CA 95032 USA. Contact him at jim.handy@objective-analysis.com.

TOM COUGHLIN is president of Coughlin Associates, Inc., San Jose, CA 95124 USA. He is a Fellow of IEEE. Contact him at tom@tomcoughlin.com.

Digital Object Identifier 10.1109/MC.2023.3241504