PASTRY

Phase Change Memory Advanced universal
Switches through Thin alteRnating laYers

SEVENTH FRAMEWORK
PROGRAMME

Deliverable D3.1

"Device specifications release"

WORK PACKAGE ACTIVITY N° 3

	Name	Partner
Contributor 1	Andrea Redaelli	MYI
Contributor 2	Mattia Boniardi	MYI
Contributor 3	Matthias Wuttig	RWTH
Checked by	Matthias Wuttig and Andrea Redaelli	RWTH

Deliverable date: M03

Summary
In this deliverable, different applications for solid state memories are discussed showing that CSL-based memory can potentially satisfy most application fields. Electrical target values for the CSL based memory are thus provided to fulfil the universal memory definition.

	CONFIDENTIAL Partners	RESTRICTED	PUBLIC
Confidentiality level			X

Memory application spectrum

In order to describe the key attributes for the Chalcogenide Super Lattice (CSL) PCM it is important to consider, as a starting point, the possible applications in which the CSL PCM technology might be employed, bearing in mind that CSL PCM is designed to represent a *Universal Memory* technology since it collects all the benefits coming from the different volatile and non-volatile memory technologies.

The application spectrum of Solid state memories (SSM) covers a very broad range and different kind of memories, i.e. non-volatile and volatile. Depending on the key feature of the application, a possible list is reported below:

- Retention intensive
- Cost intensive
- Endurance intensive
- Programming intensive
- RAM read intensive
- Read intensive

	DRAM	PCM	NOR	NAND	CSL 1 layer	CSL 2 layer
Cell size (F^2)	6	5.5	10	2	4	2
Process complx	1.8	1.8	1.4	1	1.8	1.8
Endurance	1.00E+12	1.00E+07	1.00E+05	1.00E+04	1.00E+12	1.00E+12
Programming time	2.00E-08	3.00E-07	1.00E-03	2.00E-01	1.00E-07	1.00E-07
Retention time	6.40E-02	3.00E+08	3.00E+08	3.00E+08	3.00E+08	3.00E+08
Program. power	1.00E-05	2.00E-04	2.00E-04	1.00E-14	2.00E-05	2.00E-05
Average consumption	3.13E-12	2.00E-19	6.67E-16	6.67E-24	6.67E-21	6.67E-21
Norm. cost	1.08E+01	9.90E+00	1.40E+01	2.00E+00	7.20E+00	3.60E+00
Cost/operation	6.00E-12	5.50E-07	1.00E-04	2.00E-04	4.00E-12	2.00E-12
Serial through-put	5.00E+12	1.67E+10	5.00E+06	5.00E+14	5.00E+11	5.00E+11
read latency	2.00E-08	2.00E-08	6.00E-08	5.00E-05	2.00E-08	2.00E-08
read through-put	5.00E+12	1.66667E+13	1.67E+13	1.50E+15	1.66667E+13	1.66667E+13

Table 1. Metrics computed by technology parameters to map different application fields.

Table description

Table 1 describes a comparison between actual mainstream solid state memories and Chalcogenide superlattice technology. Since CSL technology should enable the use of BEOL diodes, two different options have been considered: a single layer CSL memory and a 2 layers CSL memory. Different rows in the table are related to different features:

The first one refers to physical cell size in multiple of F^2, where F is the litho capability half pitch.

The second row is the process complexity parameter that is basically related to the number of masks used to create the chip and it is normalized to the NAND one.

Third row contains the endurance that is the maximum number of programming cycles that the device can sustain before breaking.

The fourth row reports the programming time for the slower operation between "1" to "0" and viceversa in ns. NOR and NAND are very slow compared with the other ones, thus forcing to use a "page/sector" approach instead of a single bit approach for the programming operation.

The fifth row reports the retention time at the operating temperature for the consumer application that is 85C. DRAM is volatile with a retention time of 64 ms, while the other technologies are able to retain the information for 10 y.

The last row of the block represents the programming power in W. NAND is particularly low due to the programming physical mechanism that is "tunnel" effect.

The second block of rows reports new metrics developed for certain specific application and they will be described in the application paragraphs.

Retention intensive

As "retention intensive" applications, we summarize all the applications in which the stored data must be retained, without status change, for most of the life of the chip. It has to be intended as a long time storage need application. For this kind of application the most important metrics is the average power consumption to keep the data stored defined as:

$$\overline{p} = p\frac{t_{prog}}{t_{ret}},$$

where, p is the power required to write a bit, t_{prog} is the time required to write a bit and t_{ret} is the time in which the information is retained in the device. For example, at the operating temperature t_{ret} is 64 ms for a DRAM, while it is $3x10^8$ for a typical non-volatile memory.

As a consequence, a DRAM device must consume power to retain data for times longer than 64 ms. A comparison of this parameter between typical SSM and CSL-based memory is reported in table 1, where it can be seen that the average power consumption is much higher in DRAM application due to volatility of the data stored.

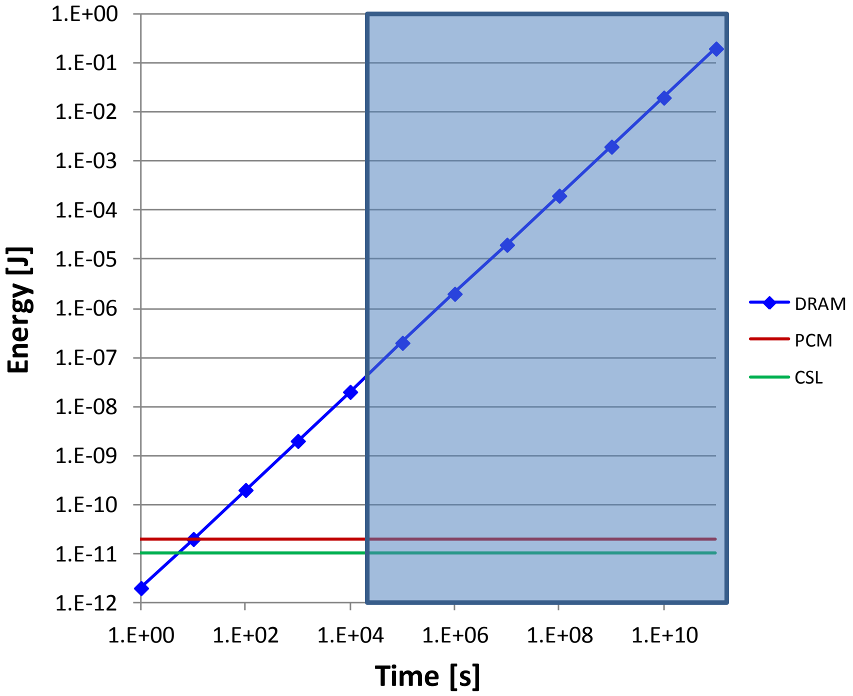

Figure 1. Consumption - Retention time trade off diagram. The blue box highlights the region in out of spec as retention time evaluated at 85C for a material able to sustain 55C for 10 y with activation energy of 2.6 eV.

To better understand the point, we reported the Energy required to keep data stored in the DRAM device as a function of the storing time needed by a certain application, compared with a PCM and a CSL device. Due to the refresh requirement of DRAM, its energy increases for times longer than 64 ms, while in PCM and CSL ones is constant due to their non-volatility feature. It is possible to see that for retention times lower than some seconds, DRAM consumption is lower than PCM and CSL. This is caused by the difference in the energy required for the single programming event that is much lower in the DRAM case. On the other hand, for application where the required retention time is higher than few seconds, CSL and PCM start to become interesting since their non-volatility ensures a lower power consumption to keep data stored. So, considering for example 1 year at 55C retention capability, with typical activation energy of 2.6 eV, that is about $2x10^4$ s at 85C, the CSL PCM technology will be a good deal in power consumption for DRAM applications since it would use up to a factor 10^3 less energy to store and retain data than a conventional DRAM.

A typical “retention intensive” application is the DRAM in certain research servers in which a lot of information must be loaded in a SSD memory to be scanned during the research operation. In these applications a lot of data must be maintained in the memory for several seconds without changing their status, but only reading them. In these kind of application a PCM, or even better, a CSL memory could be an attractive alternative.

Cost intensive

As “cost intensive” applications, we summarize all the applications in which the cost plays the major role despite of the memory performance. Typical case is the data storage, where a lot of data are written, but no direct access to the memory is performed. Usually such kinds of memories are buffered to the system through another high performance memory that loads few data at each time. In terms of cost, NAND is of course the lowest cost one. From this perspective all the other technologies cannot compete but the CSL based memory, in case of 2 layers integrated in the same chip, achieves a cost pretty close to the NAND one. It is worth noting that this is allowed by the low current required by the CSL memory (wrt conventional PCM), that enables the use of alternative switching selectors integrated in the back-end of the line, allowing the stacking of more than one single memory layer.

Endurance intensive

As “endurance intensive” applications, we summarize the applications in which endurance plays a major role. An example of “endurance intensive” application is the SSD storage in big servers where typically million cycles capability is requested. Conventional NAND SSDs have a reduced endurance, forcing frequent disks substitutions, increasing the overall cost of the memory. For this application the most appropriate metrics is the cost for operation, reported in row 3 of table 1. It is defined as the cost of the device normalized over the endurance capability. From this perspective, PCM becomes much better than NAND, and CSL results as the best performer technology. DRAM also has a low cost per operation but it is volatile so it is not suitable for SSD operation.

Programming intensive

For “programming intensive” applications, we intend the applications in which the memory is used in programming mode for most of the chip lifetime. In the memory hierarchy, DRAM and SRAM are caching intensive in systems where they are requested by the controller to execute code or process data. In this kind of applications speed plays a major role; program throughput is the right metric, reported in row 4 of table 1. The program throughput is defined as the ratio between the available power budget that the system can sustain and the energy of the single program event (the idea is that the system can work in parallel). The available power budget depends on the application, so in row 4 a normalized throughput is evaluated as:

$$\overline{T} = \frac{1}{p \times t_{prog}},$$

where p is the power for a single program event and t_{prog} the time required. Wireless application usually has a lower power budget with respect to desktop or server ones. Based on this metric NAND is the best one, taking advantage of tunnel effect for programming that is very power safe, enabling large parallel operation. Also DRAM is well positioned, while NOR and

PCM suffer from the high programming power required. CSL mitigates this issue coming closer to the DRAM performance.

RAM read intensive

As “RAM read intensive” applications, we summarize the applications in which the memory is used in random access read mode for most of the chip lifetime. In the memory hierarchy, this is the typical usage for code storage (e.g., NOR). In this case the right metric to be used is the latency time, i.e. the time required to access a single bit and it is not dependent on the power used during reading. NAND flash for example is very fast in serial read speed but it has an access time of 50 μs that completely prevents its use in RAM read intensive application (for example for execution in place of code). To use NAND memories in this kind of application, they need to be coupled to a DRAM that guarantees low access time on RAM mode but requires refresh for its operation thus being also a program intensive usage. From this perspective, PCM, DRAM, NOR and CSL are equivalent and in a good position.

Read intensive

Finally, as “serial read intensive” applications, we summarize the applications in which the memory is used in serial read mode for most of the device lifetime. A typical example is music or video streaming from an electronic support. In these cases, a lot of serial data have to be read from the support and the latency time can be neglected on the overall time budget for the operation. The read throughput in this case can thus simply be computed as the reciprocal of the energy required for the reading operation. NAND is the best performer with a large gap, while the others are almost equivalent among each other.

CSL Technology targets

Having in mind the different application fields described in the previous section, CSL based memories appear as a good candidate for a universal memory being able to find the right compromise among different application needs, as it can be seen by the table 1 where it shows the largest number of green features. Table 2 reports the target required by CSL to enable the application field listed above. In the following section, we will review each row of this table explaining their meaning.

The first row of the table reports the cell size for CSL. Conceptually, this kind of memory can be employed in a crossbar architecture thus being able to reach the $4F^2$ cell size, where F is the minimum printable geometrical feature for a given lithography. $4F^2$ is an ambitious goal since it is the minimum physical size that can be reached and it is the same as for NAND technology.

The second row reports the write-erase time target: both logic transitions have to be considered in order to give an accurate performance indication. Supposing that one of the two transitions might require a longer time to happen than its counterpart, like the reset to set transition in conventional PCM, a transition time of less than tens of ns should be indicated as the best case. While in conventional PCM the “0” to “1” logic transition time is around 100ns due to the crystallization kinetics of the whole amorphous volume, in CSL PCM such time will be likely reduced due to the fact that the logic transition is likely associated to atomic displacements; so a transition time of ten ns is expected.

Attributes	***CSL PCM targets***	***Comments***
Cell size	*4F^2*	*Cell size In a crossbar architecture. It can be further reduced in multi-layer approaches*
Write-Erase	*<10 ns*	*The limiting time is the reset to set transition time*
Retention	*85C 10 y*	*The request is related to consumer Non Volatile Memory applications. For universal memory application 55C 10 y can be sufficient*
Endurance	*>10^{12}*	*This target is the minimum to open the path for "program intensive" DRAM application*
Latency time	*20 ns*	*Same as PCM and DRAM*
Bit granularity	*Bit*	*Required for all RAM application*
Power consumption (in program)	*20 uW (1-10 pJ)*	*One order of magnitude lower than PCM*
Readiness for production	*2016*	*Tentative date to exploit project results*
Scalability	*10 nm*	*Same of PCM*

Table 2. CSL technology targets

The third row of table 2 contains the retention target. Retention specifications are strongly dependent upon the application and different specifications already exist for Non-volatile memories. In particular:

1) Consumer applications: 85C 10 years
2) Automotive applications: 125 C 10 years
3) Military applications: 150 C 10 years

For a "general purpose" universal memory the consumer specifications should be considered as a possible target of CSL PCM. Since the universal memory aims to cover also DRAM applications, a lower retention temperature is considered reasonable, i.e. 55C 10 year (85C 1 year), having in mind a possible refresh that works very often.

The fourth row reports the endurance targets. A typical target for wireless applications is 10 MCycles; for data storage applications it is much lower, i.e. 100 kCycles; for DRAM-oriented applications in which power consumption is advantageous with respect to DRAM, 300 MCycles are required; for DRAM applications, with no power constraints, 10^{12} Cycles are required. Since CSL aims to cover all these possible applications 10^{12} Cycles are required.

The fifth row shows the latency time target that is already very short in conventional PCM, about 20ns. This value has been kept constant also for CSL PCM.

The sixth row reports the bit granularity that is a peculiar properties of RAM memories such as DRAM, PCM but not of NAND and NOR. CSL must be RAM, so requires the bit granularity feature.

The seventh row reports the energy consumption target to perform a write/erase operation that should be around 20pJ; concerning DRAM applications, for a DRAM the energy required in order to retain data for a certain retention time is a function of both the retention time and the refresh rate, according to the discussion reported in “retention intensive” application.

The last two rows are less quantitative targets that simply ask for a 2X nm technology ready at 2016 and scalable for next generations.