Software/Hardware Co-design for Big Data and the impact of Future Hardware Trends

1. Software/Hardware Design

The looming end of Moore´s Law in the horizon is forcing a rethink of the classical hardware and software design of big data computing substrates. On one side, this issue is forcing system designers to leverage unconventional new hardware technologies; and this in turn is forcing the software developers to overhaul their designs for these new technologies. This requires the software and hardware to be co-designed for big data.

Hardware experts in general and computer architects in particular have disregarded software issues in the past with painful results. Consequently, many (supposedly great) changes in HW architecture did not survive. A prime example is the Cell processor (Playstation 3 processor). This was a master-Slave processor model programmed using DMAs which was extremely difficult for programmers to develop code for. Another example of HW designers disregarding software issues is the Itanium processor from Intel. There the Very Long Instruction Word (VLIW) processo explicitly aim to harness instruction level parallelism through the Compiler. However, compilers could not extract required parallelism.

The opposite is also true for recent big data implementations, i.e. software needs to be hardware conscious. Consider what happens if this is not the case: The Terasort contest lists the top platforms for sorting 100TB data. The Number 1 platform runs on vanilla Hadoop (doesn't care which HW it runs on) with 2100 nodes, 12 cores per node, 64 Gb per node, 24.000 cores and 134 Tb memory taking 4300 seconds. Now the Number 2 platform (Tritonsort) is optimized for HW, and is written in C with 52 nodes, 8 cores per node, 24 Gb, 416 cores and 1,2 Tb memory taking 8300 seconds. This shows the importance of hw/sw co-design for big data: Vanilla Hadoop may be easy to program, but needs 57X more cores, 100X more memory, and only gets 2X performance.

In the Rethink Big project (of which we are the coordinator), we take a hw/sw co-design approach towards the goal of producing a European roadmap for hardware and networking for Big Data.

2. Future Hardware Trends:

At BSC, we are also doing research in future Big Data Hardware including processor design and storage system architectures. The hardware that we foresee in future Big Data platforms are a good companion emerging distributed/dataflow programming models. On one side, we foresee much more frequent use of 3D stacking to help solve the bandwidth and capacity issues. Stacking memory on top of logic will bring us closer to a processor-in-memory (PIM) type of architecture, at least in philosophy, and PIM architectures are a good fit for programming models which exploit locality and runtime-managed movement of data as well as migrating computation to data. This additional available bandwidth through 3D stacking will usher a new period where memory I/O bounds will cease to be a major preoccupation. Now, the question will be how to exploit this bandwidth, and one solution that we think is promising to revisit vector processors; which are known to leverage bandwidth very efficiently. On the other side, we foresee that emerging non-volatile memory technologies, such as memristors

and STT-MRAM will become more prominent due to their attractive properties of no-leakage, density and persistence. These densities are higher than current magnetic disk technology, and it implies that in the future all storage including disk and main memory might be composed of these non-volatile memories. This makes complicated file systems unnecessary and increases the relevance of data-driven task based programming models. A particular limitation of current 3D stacking implementations is thermal dissipation due to increased leakage. One of the attractive properties of emerging non-volatile technologies is their lack of leakage. This makes it prudent, in our view, to combine 3D stacking with emerging non-volatile memories to solve the thermal problem; while increasing stacking depth and memory capacity. This non-volatile 3D stacked multi-core processor-in-memory will facilitate a new architecture where object-based storage is a first-class citizen. The methods of the objects stored in the 3D stacked memory will be processed in the cores that are connected through 3D vias.

3. Big data memory design as a use case:

In the volatile memories the designs are moving to 3d stacking with 3 new proposal standards: Hybrid Memory Cube (HMC), High Bandwidth Memory (HBM) and Wide I/O 2, each of these technologies are targeting a different market.

Hybrid Memory Cube: This technology was create by a consortium lead by, Altera, Arm, IBM, Micron, Open Silicon, Samsung, SK hynix and Xilinx. The HMC is a 3d stacking memory on top of a die of high speed logic, connected through-silicon-via (TSV). The HMC can be connected to the cpu as near memory mounted adjacent to the cpu, or as scalable modules form factor. It is possible to connect up to 8 modules to work together and also support atomic operations. The consumption is 70% lower than DDR3, the performance is up to 15X, also can use 90% less space due a bit density and the form factor.

High Bandwidth Memory: This proposal is defined as JEDEC standard. The main use seems to be as replacement for the GDDR5 in graphics cards, but it is not restricted to this use. It has up to 8 independent channels each one of 128 data bits plus capability for ECC. The bandwidth will be up to 256GB/s and the capacity will be up to 8GB of memory and can use a self refresh mode.

Wide I/O 2: This is also a JEDEC standard, in this case it is oriented to mobile devices. The main goal here is low power and high bandwidth. The connection with the cpu is done by an interposer. It has 8 64-bit wide channels.

There are also other technologies as the produced by Tezzaron (DiRAM4) that is comparable in performance and capacity with the previous ones but is constructed using a very thin wafer, this allows to have more vertical connections that are used mainly to fix errors in the wafer. Other technology that is used today but as last level cache is eDRAM (embedded DRAM) , that allows to integrate a large cache in the same die, however the cost per bit is larger than the usual DRAM chips.

The non-volatile memories can be used for two different aims, one as storage and the other as replacement for main memory. The characteristics will be different in both cases. The main characteristics are: the size of the memory cell, the retention time (the time that a bit will be in

the same state without being corrupted), the endurance (the number of write/erase cycles that the cell allows without losing its properties to store), the time to write or erase the cell, the voltage need to write or erase the cell and the capability to store more than one bit per cell. Today the use of non-volatile cells are principally used as storage, not as replacement for main memory, but in the near future there is a large probability that this will change, for few reasons, many of the non-volatile cells can be constructed with a smaller feature size and can hold more than one bit in each one, so the density and the price will be cheaper; other reason is the energy consumption, the non-volatile cell only dissipates power when they do an operation, but the RAM cells, consume always, by means of leakage or due refresh operations; the non-volatile ram will have less thermal problems to be stacked and there are many technical problems to scale down the ram cells.

The NAND flash memory is today the most use one, but due to endurance problems and erase latency it can be used only as storage, not as replacement for DRAM. Most of them are 2D but this year start to be in the market 3D chips with 32 layers that has better capabilities in term of access time and endurance.

There are many non-volatile technologies that can replace both RAM cell and NAND cells, but no one fulfill all the desired characteristics or the cost per bit is too high.

Phase-change memory (PCM) are faster reading and writing than NAND, it can be constructed in such way that can replace DRAM due the addressing capabilities. The endurance is orders of magnitude better that NAND, but in any case is not enough to replace directly the DRAM, but is a very good candidate to construct hybrid memories which combines a small volatile and non-volatile RAM. The price is expected to be as low as DRAM cells. Today there are some chips in the market constructed with this technology, but recently Micron withdraw to use it.

Ferroelectric RAM (FeRAM) are faster reading and writing than NAND, they has very good endurance near DRAMS cells, and the work is similar in the sense that the reading is destructive. The problem to be a replacement for DRAM or for NAND is that the cells are not too dense and cannot store multiple bits in one cell.

Magnetoresistive random-access memory (MRAM) is one of the technologies that has many characteristics that allows to be used as replacement for DRAM, except for the size, that is comparable with the SRAM cells.

Spin-transfer torque RAM (STT RAM) this is one of the emerging technologies that has more opportunities to be converted in a universal memory, for the density, possible feature size, speed and endurance. Today still not competitive with DRAM, but the projections of the evolution are optimistic.

Resistive RAM (ReRAM) or/and memristor These are a various types of cells that has a common principle. This is also a technology that can replace the DRAM and also de NAND, the projections are very promising with a very good endurance, very low operational voltage, multi level cells, some of them CMOS compatible and allow 3D Stacking and vertical cells. Also these type cells can be used to store a weight in artificial neural networks.