HPI Hardware Update – September 2016

Markus Dreseler,
markus.dreseler@hpi.de

Summary

- The new generation of Intel Xeon Phi is no longer a coprocessor, but can run existing programs as a stand-alone processor. Fast MCDRAM with 400+ GB/s is used to speed up calculations.
- Additional companies, including HPE, Samsung, IBM, and Fujitsu are working on their own Non-Volatile Memory solutions. We give an overview over existing and upcoming technologies.
- AMD gives some more information about upcoming server processor, which they hope can compete with Intel.
- IBM's new Power9 processor comes in Scale-Up and Scale-Out versions. Fast connections to external accelerators become increasingly important for IBM.

Intel releases new Xeon Phi Processor

With the new Xeon Phi "Knights Landing" processor, Intel released a new type of processor aimed at accelerating massively parallel operations, such as those occurring in the AI or HPC world. Compared to Intel's previous Xeon Phi "Knights Corner" accelerator PCIe card, the new Xeon Phi brings a number of significant changes. It can now be used as a stand-alone machine, booting a regular Linux and can, due to its x86 support, execute existing programs without any modification needed. Intel claims that executing existing binaries will already come with a performance advantage, but recompilation or optimization for the new platform are needed for full performance.

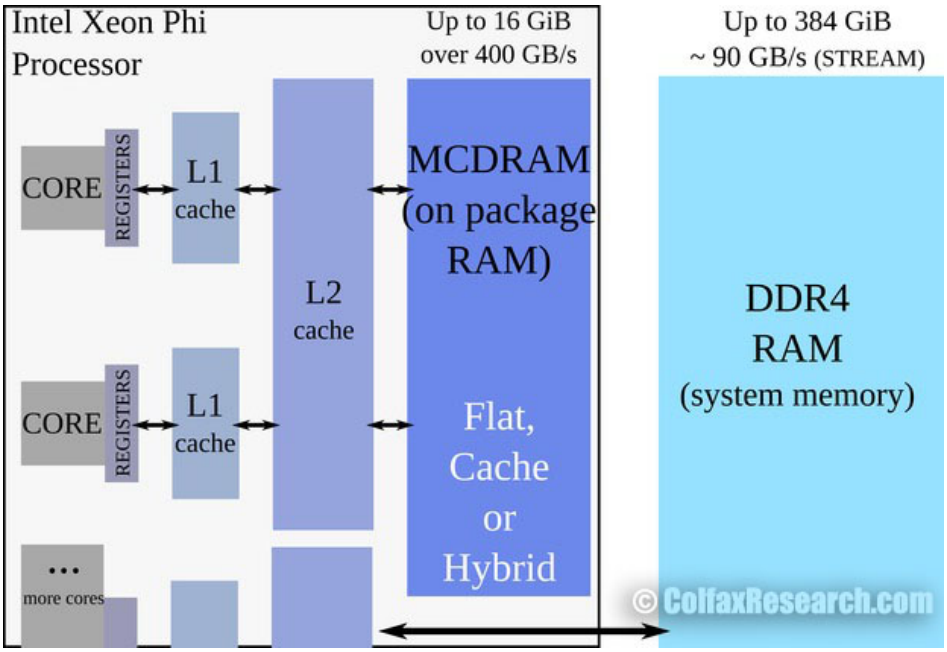

Figure 1: Memory Hierarchy in Xeon Phi [XP1]

From a hardware perspective, an interesting aspect is the inclusion of on-package High-Bandwidth Memory. The so-called MCDRAM (Multi-Channel DRAM) can deliver a performance of 400+ GB/s. Additionally, regular DRAM can be connected, but only with a throughput of 90+ GB/s [XP2]. These two types of memory can be used in three different ways: In "cache" mode, the faster MCDRAM acts as a transparent cache for the larger DRAM, in "flat" mode, it is exposed to the programmer side-by-side with DRAM so that the programmer can decide if data is to be placed on fast or slow memory, and in "hybrid" mode, some of the memory is used as a cache and the rest is exposed to the programmer.

One processor holds 72 CPU cores at 2400 MHz, 16 GB of MCDRAM, and can support up to 384 GB of DRAM. To scale beyond this, the Xeon Phi can be embedded into a network by using the integrated Omni-Path adapter. However, multi-socket setups are not possible, also because the cache coherence traffic of the fast MCDRAM could not be supported by the existing QPI interconnect [XP3].

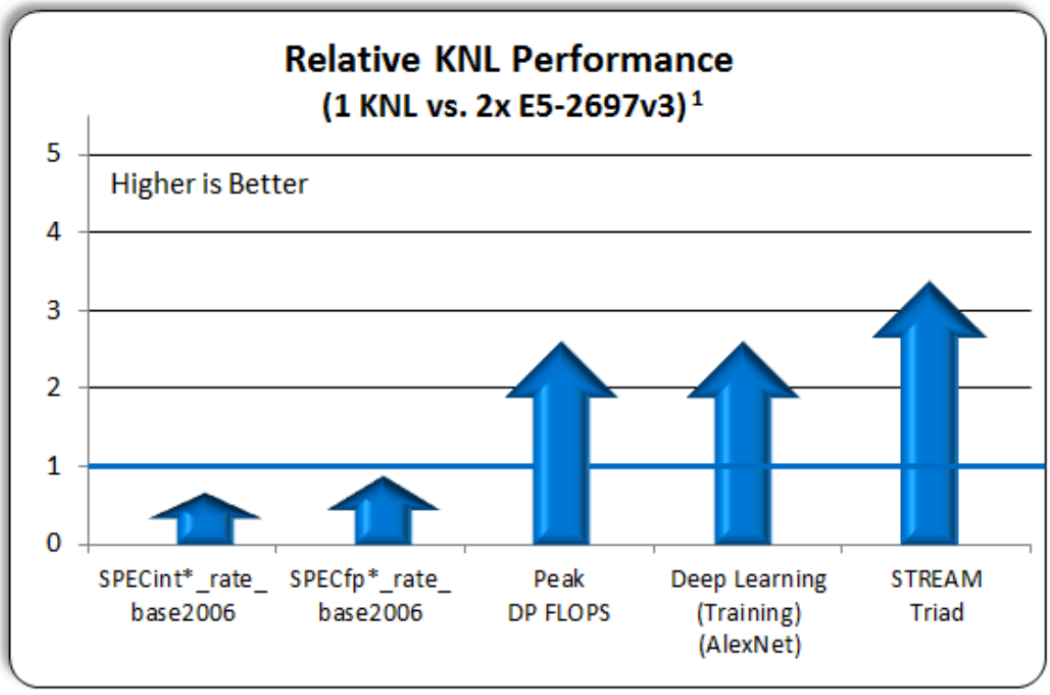

Figure 2: Preliminary Performance Comparison of Xeon Phi (KNL) compared to regular CPU [XP2]

Intel is marketing the Xeon Phi as a "server-class" processor, but appears hesitant to differentiate it from its regular Xeon line of server CPUs. An early comparison using different benchmarks is shown above. It compares the relative performance of a Xeon Phi Knights Landing (KNL) processor with two mid-range Xeon processors of the previous Haswell generation. The Xeon Phi appears to have a 2-3x advantage over regular processors when it comes to highly parallelizable workloads that heavily use a limited amount of memory. Comparing its specifications to the best available Xeon CPU also shows its advantages and its limitations.

Product Name	Intel® Xeon® Processor E7-8890 v4 (60M Cache, 2.20 GHz)	Intel® Xeon Phi™ Processor 7290F (16GB, 1.50 GHz, 72 core)
Code Name	Broadwell	Knights Landing
# of Cores	24	72
Processor Base Frequency	2.2 GHz	1.5 GHz
Max Turbo Frequency	3.4 GHz	1.7 GHz
Max Memory Size (dependent on memory type)	3072 GB	384 GB
Max Memory Bandwidth	102 GB/s	115.2 GB/s

Figure 3: Comparison of regular Xeon CPU (left) to Xeon Phi (right) [XP4]

Independent comparisons to its competitors in the GPU world, such as Nvidia's Tesla GPU are not yet available.

Non-Volatile Memories

More and more companies are entering the NVM market in one way or the other. Especially at the Flash Memory Summit in August, many new advances were presented. The following tables are meant to give an overview of the most relevant technologies that are currently discussed:

Technology	Description
NAND-backed DRAM	Not a new technology, but a combination of existing DRAM with NAND flash as used in SSDs, but attached via the memory bus. A controller transparently moves data between fast DRAM and SSD, ensuring its persistency. Batteries or capacitors are used to ensure complete write-back. Some overhead compared to DRAM, especially in write-heavy applications but faster than writing to SSD. More expensive, because data is stored twice.
MRAM	Stores information using magnetic fields. Potentially faster speeds than DRAM. Better endurance (sustainable number of writes) than other technologies.
NRAM	Uses carbon nanotubes for storage. By applying voltage to the tubes, contact between two tubes can be established or removed. This can be used to store data. Supposed to have a lower latency than DRAM, because the material changes can be performed in picoseconds – a speed at which the DRAM interface becomes the limiting factor [NV1].
Memristor / ReRAM	The Memristor is said to be a fourth fundamental circuit element, in which electric charge and magnetic flux are combined.
PCM	Phase-change memory uses an electric current to change the phase (amorphous or crystalline) of a special type of glass. This phase can be read to retrieve the data. Current challenges are the write endurance and the latency.

From a market perspective, the following companies are involved. Some smaller companies have been left out, especially if their announcements have not been substantiated.

Companies	Technology	Status ● : Product available ● : Product announced ● : Research announced
HPE "NVDIMM"	NAND-backed DRAM	● Shipped in Proliant Servers, up to 8GB, 3x more expensive than DRAM [NV2, NV3].
Viking "ArxCis"	NAND-backed DRAM	● Available in 8GB and 16GB versions, no news since 08/15.
Everspin	MRAM	● Available in 32MB, 35 ns read/write cycle (3x slower than DRAM, but claimed to be faster than 3D XPoint), very small capacity [NV4].
Netlist (+ Samsung) "HybriDIMM"	NAND-backed DRAM	● Product announced in 08/16; other than above DRAM+NAND technologies, HybriDIMM uses NAND flash also to expand capacity, transparently prefetching data to DRAM; up to 512 GB NAND + 16 GB DRAM per module. Can be used in place of regular DIMMs without a special BIOS [NV5]. Cost around 20% of regular DRAM price for same capacity [NV6].
Intel + Micron "3D XPoint"	classified	● The most-discussed product at the moment. SSD versions announced by Micron ("QuantX") and Intel ("Optane") for this year, to be 10x faster but 4x more expensive than existing SSDs [NV7,NV8]. At a later point, a version that is directly attached to the memory bus will bring a better performance, but still higher latency than DRAM.
Samsung	Z-NAND	● Competing against 3D XPoint, Samsung just announced a new "Z-SSD", which uses optimized NAND flash to reach a performance "comparable to the new Micron Quantx products" at a lower cost [NV9]. However, this only competes against the SSD versions, and is unlikely to be available as a DIMM.
Fujitsu + Nantero	NRAM	● Fujitsu has licensed NRAM technology from previously less known Nantero. It claims to be scalable even better than NAND flash with speeds better than current DRAM. A first product is planned for 2018 [NV10].
Western Digital	ReRAM	● 3D ReRAM announced in 08/16 to be used in upcoming SSDs, "universal memory" not before 2020 [NV11,NV12].
Viking + Sony	ReRAM	● Collaboration announced in 08/15, no

		news since then.
Sandisk + HPE	ReRAM	● Collaboration announced in 10/15, no news since then, Sandisk's parent is working on its own 3D ReRAM.
HPE	Memristor	● Removed from roadmap [NV13].
IBM	PCM	● Research showed performance in the range of 10x DRAM latency, no product announced yet.
IBM + Samsung	MRAM	● Research showed switching in 10ns (almost comparable to DRAM) at 11nm (allowing for higher capacity than previous MRAM technologies), no product announcement yet, "unlikely to replace DRAM soon" [NV6].

Several vendors are also working on technologies that can bring non-volatile memory to the area of storage appliances. The standard, called NVMf for "NVMe [Non-Volatile Memory express] over Fabrics", makes it possible to access SSD-like non-volatile memory via RDMA [Remote Direct Memory Access]-enabled networks, such as Infiniband. In a first experiment, using NVMf for a MySQL cluster, an improvement of 2.5x over existing flash arrays with a transactional latency of 14.3ms have been measured [NV14].

Processors

More information on AMD Zen released

AMD is slowly releasing more information about their upcoming Zen platform, which is considered to be their last hope for reentering the x86 server market. The server versions of the processor, codenamed Naples, are schedule for Q2 '17. They will feature 32 cores and, first for AMD, use SMT2 (i.e., two threads per core). Compared to previous AMD architectures, a 40% higher IPC (Instructions per Cycle) is promised. Also, AMD claims a better performance than comparable Intel Broadwell processors, but does not provide numbers or a comparison to Skylake [PR1,PR2].

Power9

IBM gave more information about their upcoming Power9 processor, which is to be released in the second half of 2017. As mentioned before, two types will be marketed: A Scale-Out (SO, optimized for two sockets) and a Scale-Up version (SU, for multi-socket systems). The SO version supports up to 4 TB directly attached memory with a bandwidth of up to 120 GB/s. The SU version can handle 8 TB, which, due to buffering, can be accessed with up to 230 GB/s at the cost of a higher latency. Also new is the information that both SMT8 (12 cores with 8 threads per core) and SMT4 (24 cores with 4 threads per core) variants will be available [PR3]. This may be due to the fact that the additional threads only brought 7% of additional performance in the Power8 architecture.

	SMT4 • 12 cores	SMT8 • 24 cores

Scale-Out
- 4 TB RAM
- 120 GB/s (direct)

Scale-Up
- 8 TB RAM
- 230 GB/s (buffered)

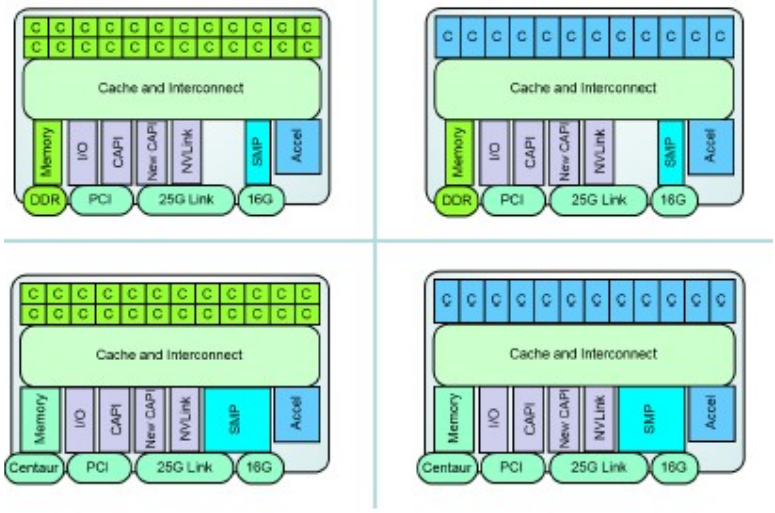

Figure 4: Power9 options [PR4]

A key feature is also the new "Blue Link" interconnect that is supposed to be a faster option for attaching GPUs and other accelerators than PCIe. These are likely to be IBM's proposal for the CCIX standard – an open standard for a cache-coherent interconnect between CPUs and other hardware [PR4].

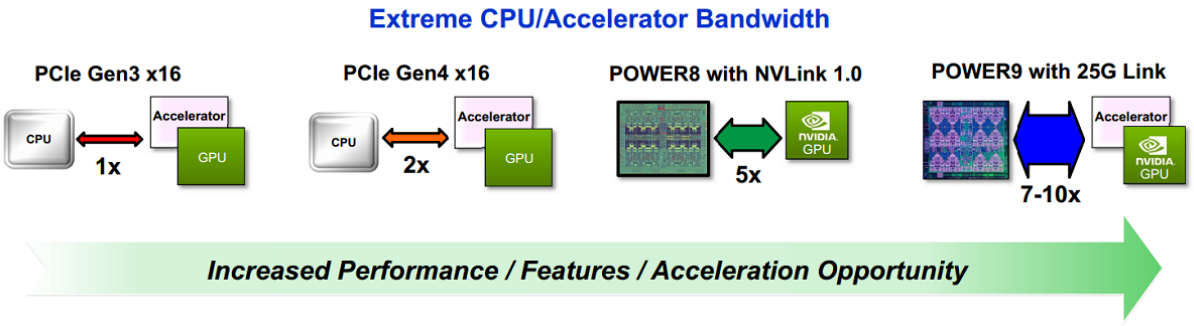

Figure 5: Connectivity from Power9 to Accelerators [PR3]

This appears to be part of a broader strategy: "As we're moving into the post-Moore's law era, you can't just […] make the general-purpose processor faster," said Bill Starke, IBM Distinguished Engineer. "It's our believe that you're going to see more and more specialized silicon. That can be in the form of on-chip acceleration, but as you can see from our approach, we tend to believe it's more flexible and deployable with off-chip acceleration. Obviously it requires extreme bandwidth, low-latency, and tight integration with your main processor complex, but that's where we see the future of computing going and you see us putting very strong investments in these directions" [PR3].

Kaby Lake

Intel released the first notebook processors from its 14nm Kaby Lake generation, successor to Skylake. Desktop CPUs are announced for January, but server CPUs are not yet officially announced. For desktop workloads, Intel claims a perfor-

mance increase between 10% and 20% [PR5], partially due to slightly higher frequencies.

Newsflash

Intel releases Silicon Photonics

Intel released the first two products using Silicon Photonics, an upcoming technology that transmits data using light in glass fibers. Unlike traditional fiber optics, however, Silicon Photonics can produce the laser signals directly on chip, without the need for additional optics. This greatly simplifies the production process and size, thus enabling light-based interfaces that are an immediate part of the CPU [NF1].

References

[XP1] http://colfaxresearch.com/knl-mcdram/
[XP2] http://www.hotchips.org/wp-content/uploads/hc_archives/hc27/HC27.25-Tuesday-Epub/HC27.25.70-Processors-Epub/HC27.25.710-Knights-Landing-Sodani-Intel.pdf
[XP3] http://wccftech.com/intel-knights-landing-detailed-16-gb-highbandwidth-ondie-memory-384-gb-ddr4-system-memory-support-8-billion-transistors/
[XP4] http://ark.intel.com/compare/95831,93790

[NV1] http://www.tomshardware.com/news/fujitsu-carbon-nanotube-memory-nram,32603.html
[NV2] http://www.computerworld.com/article/3051135/data-storage/whats-in-hpes-persistent-memory.html
[NV3] https://www.hpe.com/us/en/servers/persistent-memory.html
[NV4] https://www.everspin.com/news/everspin-readies-industry%E2%80%99s-first-256mb-perpendicular-spin-torque-mram
[NV5] http://insidehpc.com/2016/08/netlist-hybridimm-memory-unifies-dram-nand/
[NV6] http://www.theregister.co.uk/2016/08/08/samsung_and_netlist_hybridimm/
[NV7] http://www.eetimes.com/document.asp?doc_id=1330280
[NV8] http://www.theregister.co.uk/2016/08/12/xpoint_fails_to_match_intels_claims/
[NV9] http://www.eetimes.com/document.asp?doc_id=1330285
[NV10] http://www.theregister.co.uk/2016/08/31/nram_dev_nantero_signs_fujitsu/
[NV11] http://www.anandtech.com/show/10562/western-digital-to-use-3d-reram-as-storage-class-memory-for-specialpurpose-ssds
[NV12] http://www.theregister.co.uk/2016/08/16/wd_says_resistance_is_not_futile/
[NV13] https://www.hpcwire.com/2015/06/11/hp-removes-memristors-from-its-machine-roadmap-until-further-notice/
[NV14] http://www.theregister.co.uk/2016/09/01/mangstor_and_mellanox_nvmef_review_finds_it_speedy/

[PR1] http://www.theregister.co.uk/2016/08/18/amd_zen_latest/
[PR2] http://hexus.net/tech/news/cpu/95914-amd-provides-first-glance-zen-summit-ridge-performance/

[PR3] https://www.hpcwire.com/2016/08/30/ibm-unveils-power9-details/
[PR4] http://www.eetimes.com/document.asp?doc_id=1330350
[PR5] http://arstechnica.com/gadgets/2016/08/intel-unveils-kaby-lake-its-first-post-tick-tock-cpu-architecture/

[NF1] http://www.enterprisetech.com/2016/08/19/intel-launches-silicon-photonics-chip-previews-next-gen-phi-ai/