HPI Hardware Update – March 2016

Markus Dreseler,
markus.dreseler@hpi.de

Summary

- Explanation of Power8's SMT8 technology: Eight logical threads can be executed, but the scalability is limited and the last four threads only improve performance by 7%
- Samsung mass-produces HBM2 on-chip DRAM with a bandwidth of 256 GB/s
- Upcoming AMD Zen server processors will have up to 32 cores per processors and eight-channel DRAM memory
- IBM selling machines with up to 32TB of RAM

Follow-Up: Multithreading

As part of our meeting in Potsdam, the topic of hardware-level parallelism and the eight logical threads (SMT8) on Power8 CPUs was discussed. We would like to give you this information as a follow-up. Before describing SMT8, we will give an overview on hardware-level parallelism in general.

Overview on Hardware-Level Parallelism

Processor instructions are executed in four steps: Fetch, decode, execute, and write-back. With the assumption that each of these takes a clock cycle, four clock cycles are spent when executing one instruction. This is referred to as Cycles per Instruction (**CPI**). As a first improvement, these steps can be overlapped, so that four instructions are executed at the same time (see Figure 1). This is called **Pipelining**.

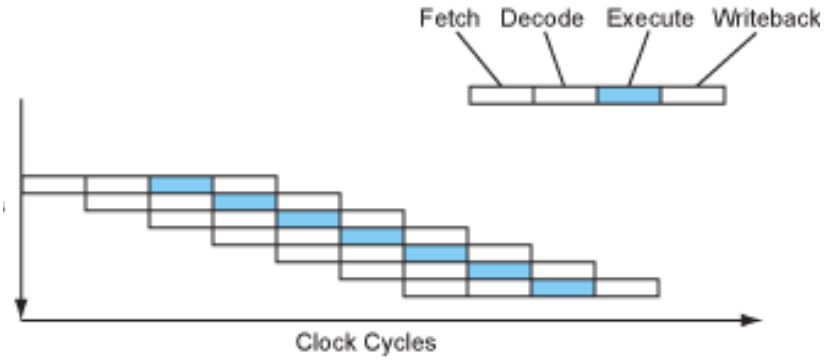

Figure 1: With pipelining, multiple instructions can be worked on at the same time [MT1]

In this first model, the clock cycle depends on the slowest step. In practice, these steps do not require the same time to complete. For example, an integer multiplication takes longer than an integer addition. To increase the clock cycle, instructions are now subdivided, so that an addition now takes one cycle and the multiplication three. This results in a deeper pipe-

line with a larger number of shorter steps and accounts for different execution latencies.

The execution step is handled by a number of **execution units**. One unit deals with integer arithmetic, another with branching, and so on. Intel's Skylake architecture has 21 units [MT2, p. 151], Power8 has 16 [MT3]. A logical next step is to issue multiple instructions at the same time and have these work on different execution units. As a result, the number of instructions per cycle (**IPC**, the inverse of CPI) increases. Intel achieves a sustainable IPC of 4, Power8 of 8 [MT4] [1]. This model is called **instruction-level parallelism (ILP)** and processors using it are called **superscalar**.

This maximum IPC can only be reached when instructions are independent of each other. Often, this is not the case – for example when an integer addition has to wait for one of the operands to be loaded from memory and the other to be calculated as the result of a multiplication. In this case, the addition has to wait for the prior instructions to return their results. This results in a **stall** of the pipeline. A "bubble" is inserted into the pipeline, which means that a time slot is wasted. To reduce the number of stalls, the CPU may reorder independent operations, so that other work is executed while the addition is waiting for its operands. This is called **Out-of-Order Execution (OOO)**.

The IPCs also explain why the clock speed of processors is an insufficient metric for comparing the performance of a processor. Processors with efficient Out-of-Order Execution can achieve more work during one clock cycle, which can offset a slower clock frequency. Literature [MT1] calls processors that spend a high effort in OOO and other ILP techniques "brainiacs" and architectures that focus on higher clock cycles "speed-demons". Figure 2 places historical and current processor architectures on these two axes. It is interesting to see how AMD developed more of a speed-demon and Intel more of a brainiac architecture. Looking back in history, this explains how Pentium has beaten Athlon in benchmarks even though Athlon had the higher CPU frequency. As the power consumption increases superlinearly with higher clock-cycles, it becomes more and more difficult to increase the CPU frequency. This is called the **power wall**. On the other hand, because ILP only works when there are unused execution units, it is not trivial to increase the IPC either.

[1] The IPC of 8 for Power does not directly relate to the number of simultaneous threads (SMT8) that we will discuss later

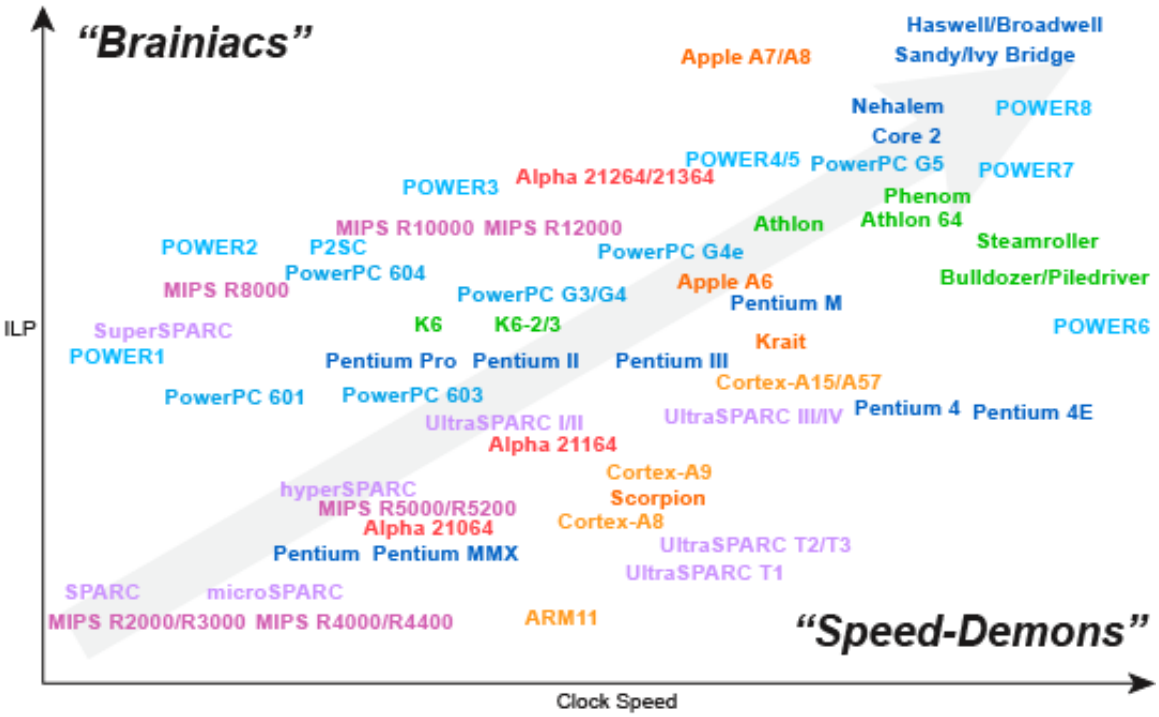

Figure 2: Some processors focus on higher CPU frequencies (speed-demons) while others focus on improved IPC (brainiacs) [MT1]

Running multiple instructions at the same time is dependent on having a high potential for parallelism in the application. Filling the pipeline bubbles with parallel tasks becomes difficult when longer stalls occur or a high level of dependency exists. This can happen when the processor has to wait on a load from DRAM and other instructions depend on the result of this load. As a result, the theoretical IPC maximum is reached only in bursts and cannot be sustained.

To fill the remaining bubbles with useful instructions, a second source of instructions is added. By running two (or more) threads on the same CPU and sharing the pipeline, stalls in the first thread can be used to execute work for the second thread. This is called **simultaneous multithreading (SMT)** or **Hyper-Threading** for Intel processors. These hardware threads differ from software threads in that scheduling is done exclusively in hardware by the CPU and does not occur the overhead known from software threading.

Each SMT thread has its own set of registers, but other resources, such as caches, execution units, and memory access are shared. As a result, two independent threads running on different processors will be faster than if they were to run on the same core.

Comparison of Power8 and Xeon

As we can see, a number of different factors determine the performance of a processor, including the clock speed, the quality of its OOO, and the number of SMT threads, but also how well the executed workload can be

parallelized by the hardware. Looking at the **SMT8 technology in the Power8**, IBM reports [MT4] that

- 2 threads delivers about 45% performance more than one
- 4 threads deliver yet another 30% boost
- the last 4 threads deliver about 7%

Xeon E7v3/POWER8 Comparison		
Feature	Intel Haswell-EX Xeon E7	IBM POWER8
Process tech.	22nm FinFET	22nm SOI
Max clock	2.5-3.6 GHz	3.5-4.35 GHz
Max. core count **Max. thread count**	18@2.5 GHz 36 SMT	12@4.2 GHz 96 SMT
Max. sustained IPC	6 (4)	8
L1-I / L1-D Cache	32 KB/32 KB	32 KB/64 KB
L2 Cache	256 KB SRAM per core	512 KB SRAM per core
L3 Cache	2.5 MB SRAM per core	8 MB eDRAM per core
L4 Cache	None	16 MB eDRAM per MBC (64/128 MB total)
Memory	1.5 TB per socket (64 GB per DIMM)	1-2 TB per socket (64 GB per DIMM)
Theoretical Memory Bandwidth	102 GB/s (independent mode)	204 GB/s
PCIe 3.0 Lanes	40 Lanes	32 Lanes

Figure 3: Specifications of Haswell and Power8 processors

Looking at the other specifications of Power8, shown in Figure 3, it appears that the Power8 outperforms the Haswell processors in most aspects. However, the devil is in the details. For example, the advantage of larger L2 cache in the Power8 is offset by its lower speed. In the end, only benchmarks with the actual workload can give information on which architecture is better for the use case at hand.

The description of hardware-level parallelism is largely based on [MT1].

High Bandwidth Memory 2

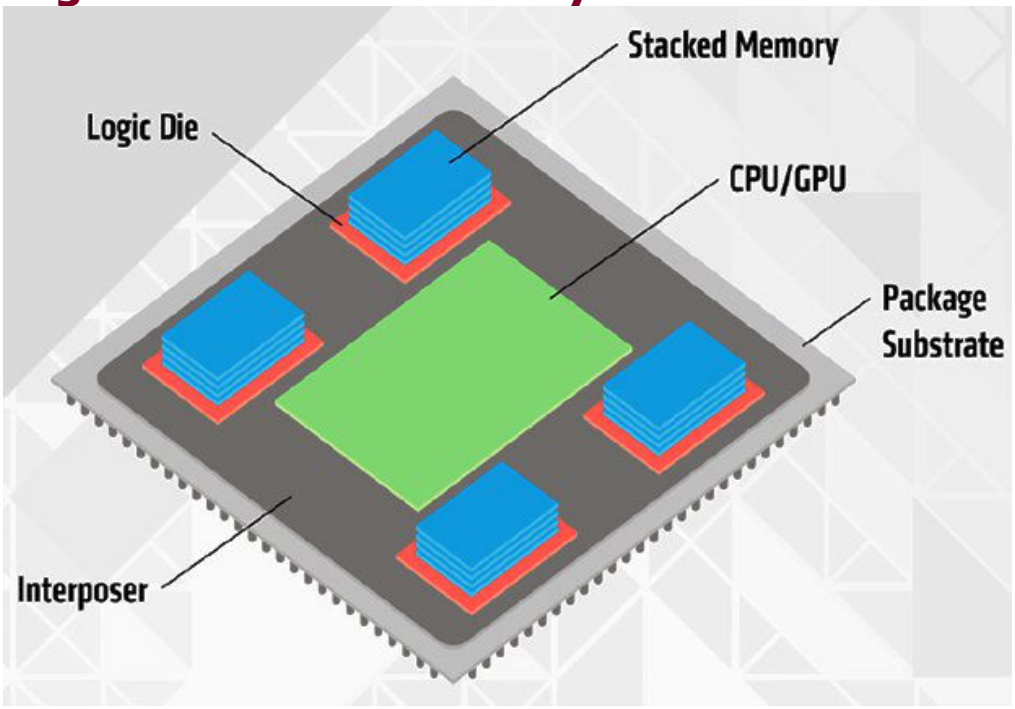

Figure 4: Basic Layout of HBM [HB4]

Samsung has announced mass production of HBM2 DRAM. This new type of RAM comes with a bandwidth of 256GB/s [HB1]. In other words, their HBM2 is said to be ten times faster than current DDR4 RAM with 25.6 GB/s per channel. Additionally, Samsung's HBM2 is said to double the power efficiency (measured as bandwidth per watt) compared to GDDR5 [HB2]. This is due to an increased bus width that allows for a reduced clock speed [HB3].

A key difference to current DRAM is that the memory is connected to the processor using a so-called interposer. This interposer is located on the chip, allowing for a closer integration between processor and memory, thus reducing the latency [HB4].

With package sizes of 4GB and later this year 8GB, this type of memory will, at first, be more important for graphics cards and accelerators. In the near future, it will not be seen in the server market, mainly due to cost reasons. In the long term, however, Samsung also plans to use HBM2 to strengthen its position in the HPC market.

For programmers, this will likely mean a system in which HBM and DDR memory are used side-by-side. HBM might here act as another cache level for DRAM contents or be an independent memory area for highly used data structures. Also, hybrid modes are a possibility. Intel uses a similar model for their MCDRAM in the upcoming Xeon Phi, codenamed Knights Landing [HB5].

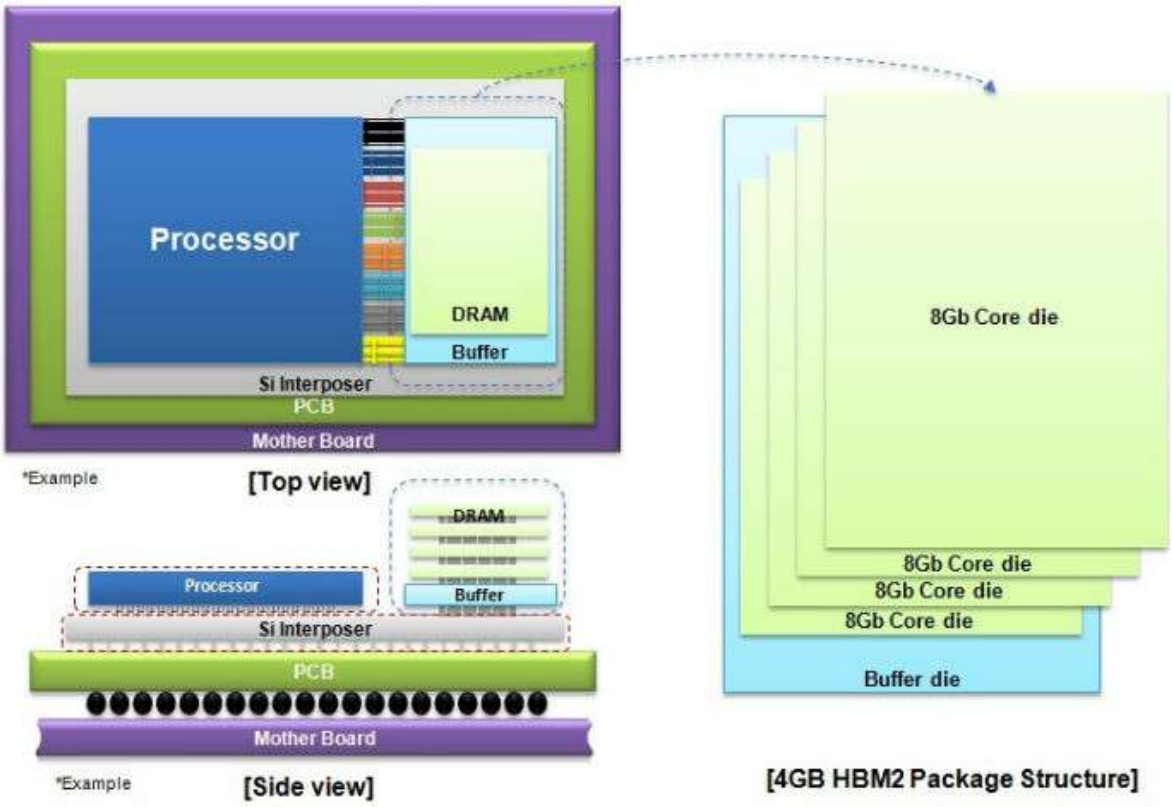

Figure 5: Layout of HBM2. Notice how the DRAM is connected directly to the processor. [B]

AMD Zen

According to a presentation from CERN, the new AMD Zen server processors will have up to 32 cores per processor (compare to Intel's 18, 28 are planned for 2017) and a 40% improvement in Instructions per Clock compared to current AMD processors. Additionally, they are said to feature eight-channel DDR4 memory, twice as much as current Intel processors offer and theoretically doubling the memory bandwidth [AZ1,AZ2].

Due to power limitations, the high number of cores might result in a lower frequency per core [AZ2]. AMD Zen processors are to be release in October 2016.

Newsflash

- IBM started selling larger versions of their enterprise POWER machines. With 2 TB per socket, the top-of-the-line E880 can now host 32 TB [NF1].

References

[MT1] http://www.lighterra.com/papers/modernmicroprocessors/
[MT2] http://www.agner.org/optimize/microarchitecture.pdf

[MT3] http://openpowerfoundation.org/wp-content/uploads/2015/03/Sadasivam-Satish_OPFS2015_IBM_031615_final.pdf
[MT4] http://www.anandtech.com/show/9193/the-xeon-e78800-v3-review/6

[HB1] http://www.theregister.co.uk/2016/01/20/ram_bam_thank_you_maam_samsung_fires_up_fastestever_memory/
[HB2] http://phys.org/news/2016-01-samsung-mass-world-fastest-dram.html
[HB3] http://insidehpc.com/2016/02/hbm/
[HB4] https://techreport.com/review/28294/amd-high-bandwidth-memory-explained
[HB5] http://www.anandtech.com/show/9794/a-few-notes-on-intels-knights-landing-and-mcdram-modes-from-sc15

[AZ1] http://www.silicon.de/41621040/amds-neuer-serverchip-zen-soll-32-kerne-bekommen/
[AZ2] http://www.extremetech.com/extreme/222921-amd-is-supposedly-planning-a-32-core-cpu-with-an-eight-channel-ddr4-interface

[NF1] http://www.nextplatform.com/2016/01/18/ibm-doubles-up-memory-adds-power8-cpus-for-big-iron/