The architectures in a nutshell (Vectorization Edition)

Georg Zitzlsberger

georg.zitzlsberger@vsb.cz

IT4Innovations
national supercomputing
center

5th of July 2017

Agenda

- Salomon Architectures
- Intel® Xeon® processors v3 (Haswell)
- Intel® Xeon Phi™ coprocessor (KNC)
- Where to get the Information from?
- Boundness
- How to Measure Performance?

Architectures Available

- Host: Haswell
  - Intel® Xeon® E5-2680v3 (2.5 GHz)
  - 12 cores
  - 2 Sockets (24 cores total)
  - RAM: 128GB (5.3 GB per core)
  - Nodes: 576 & 432 (w/o & w/ KNC)
  - Interconnect: InfiniBand FDR56
  - SKU description on ark.intel.com
- Coprocessor: Knights Corner (KNC)
  - First generation Intel® Xeon Phi™ Coprocessor 7120P (1.24 GHz)
  - 61 cores
  - RAM: 16 GB (∼260 MB per core)
  - 2 coprocessors per node
  - Interconnect: PCIe 2.0 (via host)
  - SKU description on ark.intel.com

(Images: IT4Innovations)

Notes to Systems

Hosts:

- Hyper-Threading: off
- Turbo Boost: on, SpeedStep: off
- Cluster on Die (CoD): off (only one NUMA node per socket)
- Frequencies (base @ 2.5 GHz):

Table 2. Intel® Xeon® Processor E5-1600, E5-2600 and E5-4600 v3 Product Families Turbo Bins (Sheet 1 of 3)

S-Spec No	Stepping	Model Number	TDP (W)	# Cores	Intel® Turbo Boost Technology Maximum Core Frequency (GHz)										Notes
					Core 1 -2	Core 3	Core 4	Core 5	Core 6	Core 7	Core 8	Core 9	Core 10	Core 11+	
SR1XG	C1	E5-2695 v3	120	14	3.3	3.1	3	2.9	2.8	2.8	2.8	2.8	2.8	2.8	1,2,3,7
SR1XN	M1	E5-2690 v3	135	12	3.5	3.3	3.2	3.1	3.1	3.1	3.1	3.1	3.1	3.1	1,2,3,7
SR1XP	M1	E5-2680 v3	120	12	3.3	3.1	3	2.9	2.9	2.9	2.9	2.9	2.9	2.9	1,2,3,7

Table 3. Intel® Xeon® Processor E5-1600, E5-2600 and E5-4600 v3 Product Families Intel® AVX Turbo Bins (Sheet 1 of 3)

Model Number	Intel AVX Core Frequency (GHz)	# Cores	Cache Size (MB)	Intel® AVX Turbo Boost Technology Maximum Core Frequency (MHz)								Notes
				Cores 1-2	Cores 3	Cores 4	Cores 5	Cores 6	Cores 7	Cores 8	Cores 9+	
E5-2690 v3	2.3	12	30	3.2	3	3	3	3	3	3	3	1,2,3,7
E5-2680 v3	2.1	12	30	3.1	2.9	2.8	2.8	2.8	2.8	2.8	2.8	1,2,3,7

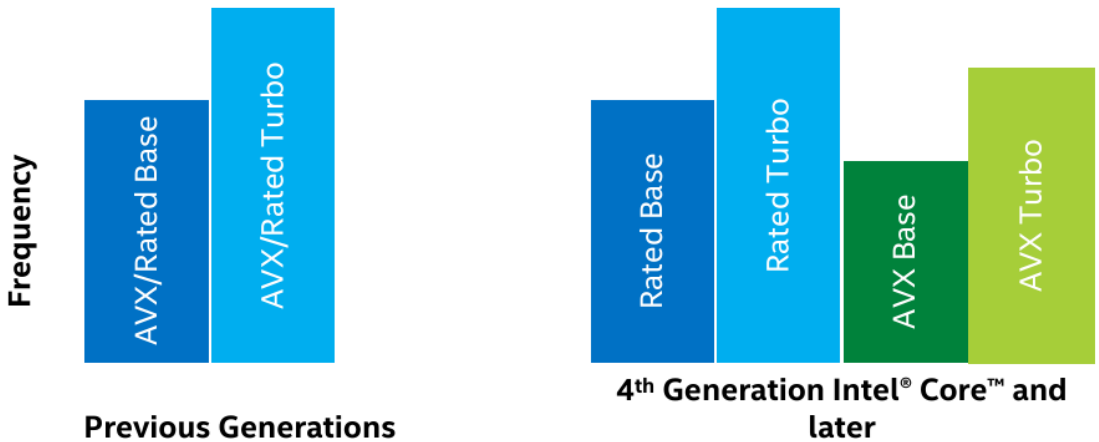
(Images: Intel)

Details can be found in ▸ Intel Xeon Processor E5 v3 Specification Update

Excursion - Frequencies

With Haswell and later (also includes KNL):

- Amount of turbo frequency achieved depends on:
  Type of workload, **number** of **active cores**, estimated **current & power consumption**, and processor **temperature**
- Due to workload dependency, separate AVX base & turbo frequencies will be defined for 4th generation Intel® Core™ and Xeon® processors and later



* Intel® AVX refers to Intel® AVX, Intel® AVX2 or Intel® AVX-512

(Image: Intel)

Notes to Systems

Coprocessors:

- Turbo Boost: off (max. 1.33 GHz with base @ 1.24 GHz)
- ECC: on (limits memory BW by ∼12%)
- Use `micsmc` to query settings, e.g.:

```
> micsmc --turbo --ecc
mic0 (turbo):
  Turbo mode is disabled
mic1 (turbo):
  Turbo mode is disabled

mic0 (ecc):
  ECC is enabled
mic1 (ecc):
  ECC is enabled
```

Use `micsmc --help` for a full list.

Theoretic Peak Performance - Hosts

- Throughput:

$$FLOPS_{SP}: \#_{cores} * frequency * SIMD * FMA$$
$$12 * 2.5GHz * 16_{AVX} * 2_{AVX2} = 0.960\,TFLOPS$$
$$FLOPS_{DP}: 12 * 2.5GHz * 8_{AVX} * 2_{AVX2} = 0.480\,TFLOPS$$

- Memory bandwidth:

$$BW_{DDR4}: \#_{channels} * frequency * byte_{cycle}$$
$$4 * 2133MT/s * 8byte \approx 68GB/s^{1}$$
$$BW_{QPI}: \#_{QPILinks} * frequency * byte_{cycle} * \#_{directions}$$
$$2 * 9.6GT/s * 2byte * 2 \approx 76.8GB/s^{2}$$

See ▸ ark.intel.com for full specifications.

[1] Note that DDR BW was just calculated for one socket!

[2] Entire system QPI BW w/o overhead

Theoretic Peak Performance - Coprocessors

- Throughput:

$$FLOPS_{SP}: \#_{cores} * frequency * SIMD * FMA$$
$$61 * 1.24GHz * 16_{MIC} * 2_{MIC} = 2.42TFLOPS$$
$$FLOPS_{DP}: 61 * 1.24GHz * 8_{MIC} * 2_{MIC} = 1.21TFLOPS$$

- Memory bandwidth:

$$BW_{GDDR5}: \#_{channels} * frequency * byte_{cycle}$$
$$8 * 2 * 5500MT/s * 4byte \approx 350GB/s^{3}$$
$$BW_{PCIe\ x16}: 8GB/s$$

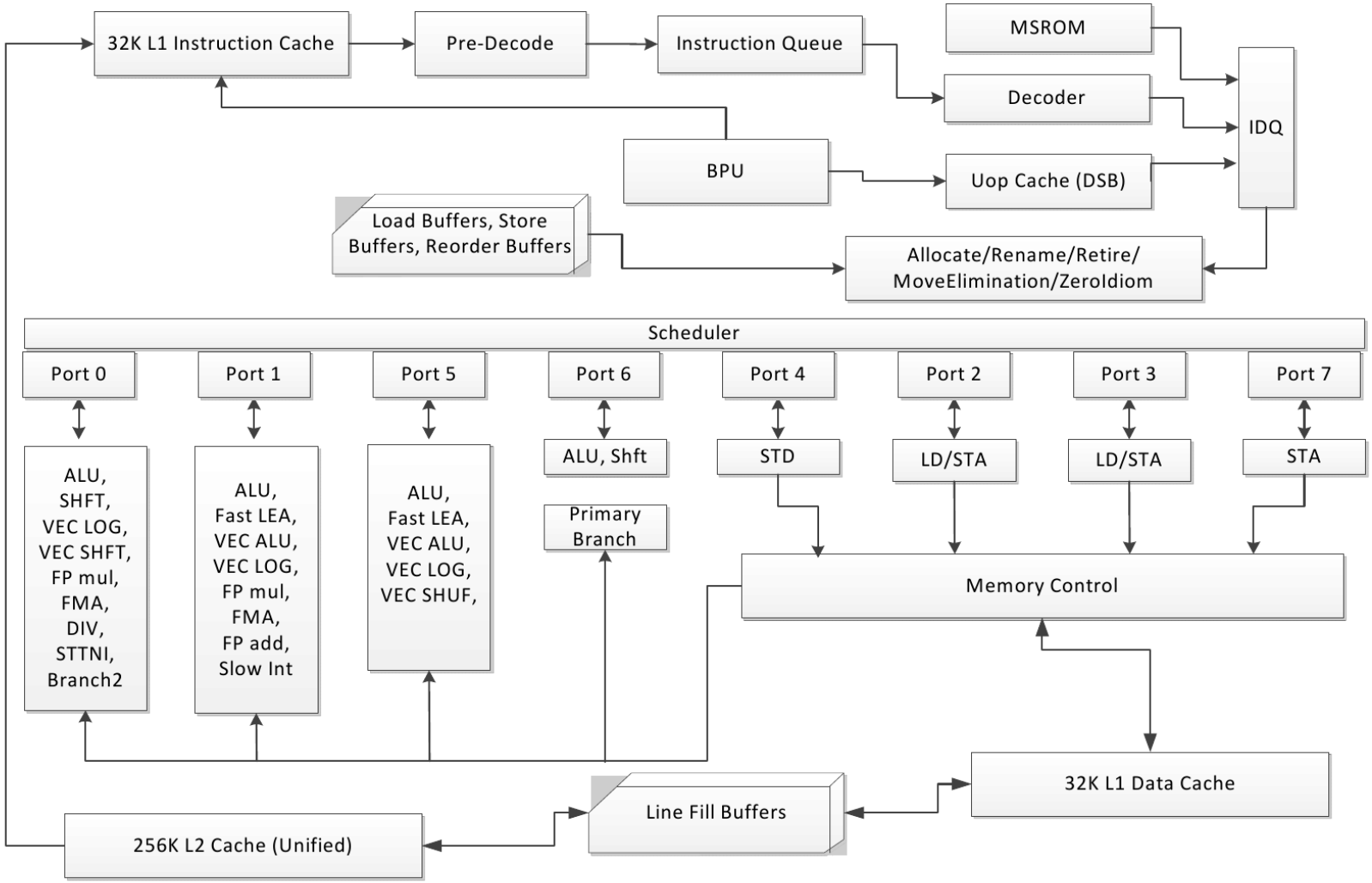
See ▸ ark.intel.com for full specifications.

[3] Note that max. ∼170 GB/s is realistic!

Intel® Xeon® processors v3 (Haswell)

- Execution:
  - Out of Order (OOO):
    The order of instructions you see is not necessarily the order they get executed!
  - Speculative execution
  - Branch predictor estimates the likely branch and speculatively executes it (due to deep pipeline)
  - FMA support with AVX-2:
    Peak performance only with 2x FMA per cycle!
- Watch out for:
  - Order of instructions can only be influenced with data flow changs in higher level (i.e. C/C++, Fortran, ...).
  - Branch prediction might be wrong - pipeline needs to be flushed
  - If FMA is not used theoretic peak performance is 50%!
- There are more but we only focus on vectorization right now...

Haswell Pipeline Diagram

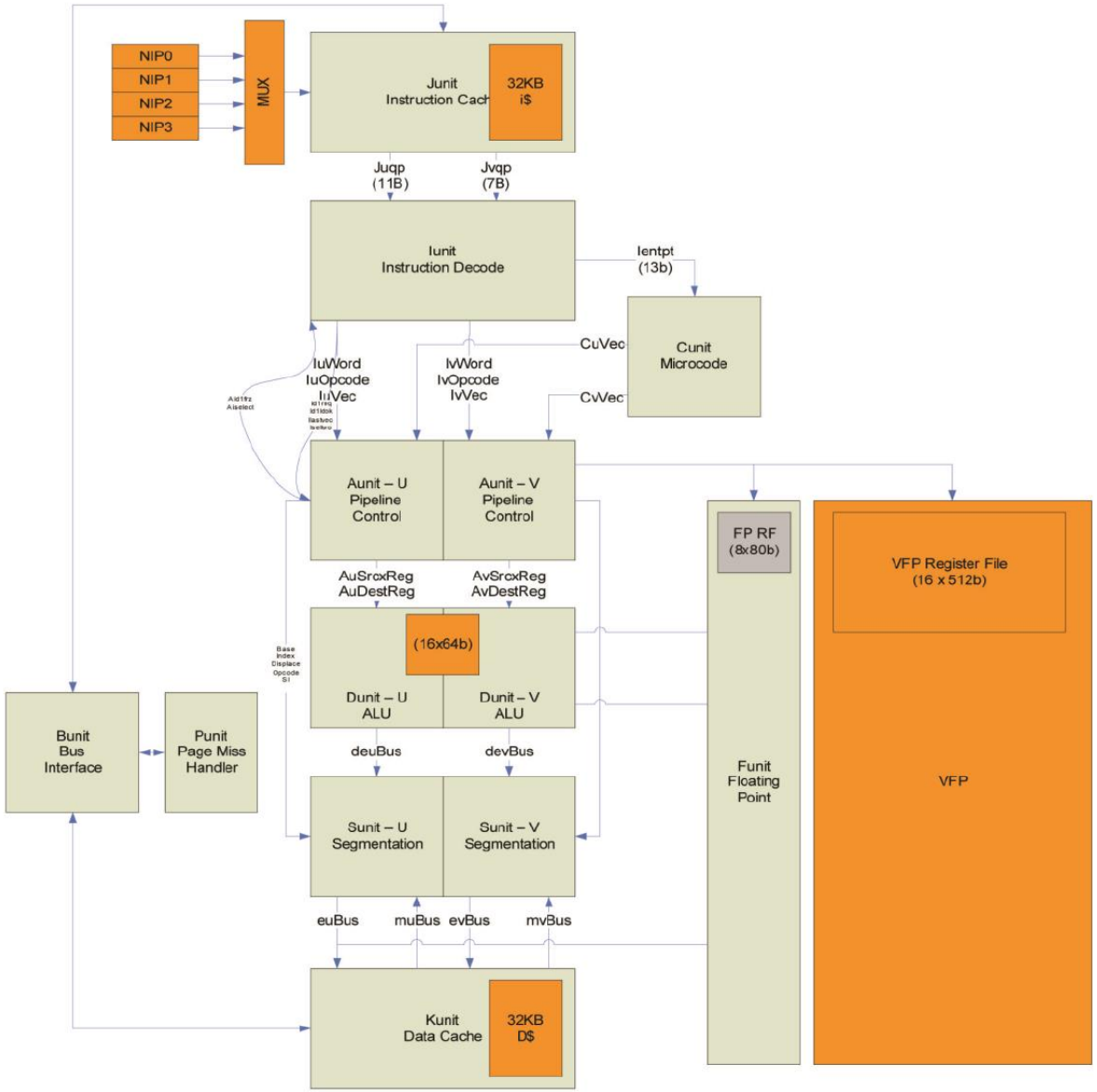


(Image: Intel)

Intel® Xeon Phi™ coprocessor (KNC)

- Execution:
  - In-order:
    The order of instructions is what is executed. Hint: Look at the assembly (Intel Compiler) to see the cycle counts when the instructions are executed.
  - FMA support
    Peak performance only with 1x FMA per cycle!
  - HW-threading is a must (2 or more threads):
    Loading (fetch & decode) instuctions takes 2 cycles per HW-thread.
- Watch out for:
  - The architecture is susceptible to the order instructions are generated
  - If FMA is not used theoretic peak performance is 50%
  - Only using one HW-thread also reduces theoretic peak performance by 50%
  - **Advantage though:** In-order makes timing for benchmarking reproducible!

KNC Pipeline Diagram



(Image: Intel)

Where to get the Information from?

- Intel® Xeon® processors v3 (Haswell) (incl. all big cores + KNL):

  ▸ Intel® 64 and IA-32 Architectures Optimization Reference Manual

  See section *2.2 The Haswell Microarchitecture*.
- Intel® Xeon Phi™ coprocessor (KNC):

  ▸ Intel® Xeon Phi™ Coprocessor System Software Developers Guide

  See section *2 Intel® Xeon Phi™ Coprocessor Architecture*.
- Very good 3rd party source (w/ unofficial but empirical numbers):

  ▸ Agner Fog's Software optimization resources

Boundness

Understand general characteristics of an application and it's phases:

- **Memory Bound:**
  Execution in the processor back-end is dominated by stalls due to memory accesses. Includes latency and bandwidth bound.
  $\Rightarrow$ Typical for HPC applications.
- **Compute Bound:**
  Means ideal case where designed processor throughput can be achieved.
- **I/O Bound:**
  Interconnects are limiting computation performance (e.g. unbalanced data flow between nodes)

Vectorization and Compute Bound

Root causes for low FLOPS:

- Vector instructions are not used efficiently (esp. FMA instructions)
- Loops have too small trip counts
- Loops are created with peeling, main and remainder loop (missing alignment guarantees)
- etc.

How to Measure Performance?

- Lock frequency (e.g. in BIOS or *cpufreq* tool) or profile it (e.g. with Intel VTune Amplifier XE)
  $\Rightarrow$ Avoid both frequency boosting (Turbo Boost) or throttling (SpeedStep).
- Measure effects of Hyper-threading: does not need to be turned off in BIOS but make sure SW-threads are pinned properly
- Beware that "high power AVX" instructions throttle the frequency to a well documented "AVX base frequency" (from Haswell onwards, incl. KNL - but not KNC).
- Deterministic threading required (ensure pinning).
- Are there other processes running and which cores handle interrupts?