Scaling through more cores

From single to multi core

written version

by Thomas Walther
HLR Seminar on 30.11.2015
written on 25.03.2016

Index

1. Introduction

Starting with the scaling of single cores and the problems that approaches there, it will be followed by the solution through more cores with actual standards, the typical multi core usage and the barriers in parallel programming defined by Amdahl's law.

Additionally there will be an overview about hot spot and DIE size problems, leading to the current barriers for the industry, followed by the new approaching technologies for other materials in fabrication and in the end my conclusion will shown.

2. Scaling with single core until 2005

Moore's law says, that the transistors on new CPU DIEs will double every 12 to 24 month, which, in single core technology, has a direct influence on the performance of the CPU. In Theory the doubled transistors would double the performance, but in practice the physical barriers like lacking energy reduce the effective performance boost.

This is possible with new manufacturing technologies for smaller structures. With smaller structures more transistors can be placed on the DIE.

With smaller structures it is possible to raise the frequencies for more performance and achieving the goal of double the overall performance.

2.1 Scaling - Problems and barriers

A Problem with the higher frequencies were the needed higher voltage, what leads to more power consumption. More power consumption leads to higher power dissipation or in other words waste heat.

The formula for is $P = a * C * V^2 * f$ which shows the frequency is a multiplier and voltage a square multiplier.

With this strong influence for the power consumption the heat strongly increases and lead to the limits of air cooling possibilities in 2004, when the Pentium 4 with ~4 GHz marked the line.

3. Solution through more cores

In 2006 Intel announced the end of the single core age and introduced the first Dual Core Processors with 1.5 to 2.33 GHz.

In the following years they invented Hyper Threading (HT) and added more cores on the DIE for more performance.

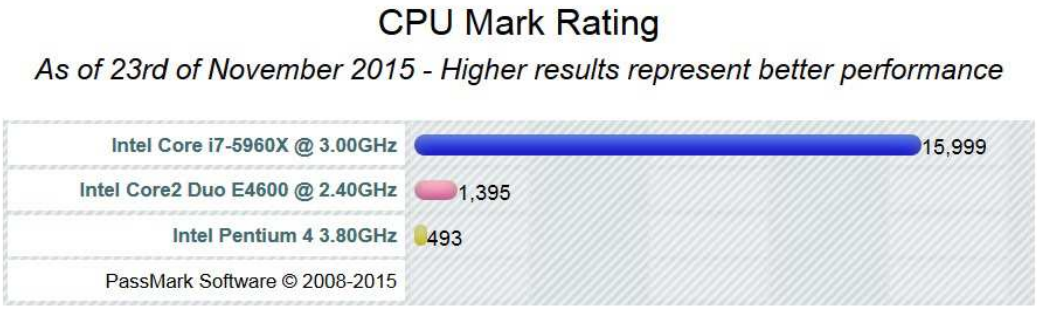
With this new way the power consumption massively decreased because of the lower clocking lower voltage was needed.

3.1 Solution through more cores - Current standard

Actual standards in consumer PCs are 2 to 8 cores with a maximum of 16 threads, if hyper threading is included and 2 to 8 cores in handheld devices.

In the Top 500 of High Performance Computing Clusters the range of cores reaches from many thousands up to 3.12 million cores in the leading HPC cluster Tianhe-2 in China.

The Graphics 1 and 2 display the performance improvements through more cores with lower frequencies. One of the latest single core processors the Pentium 4 with 3.80 GHz got out scaled by one of the first Core2 Duo processors with 2.40 GHz with a factor of 2.8.



Graphic 1, s. Attachment

One of the newest consumer processors, a Core i7 with 3.00 GHz, 8 cores and 16 threads, is 32.5 times faster than the latest P4 with a bit more than a doubled TDP.

In comparison the single core performance is only a bit more than doubled, but still far improved particularly to the lower clocking speeds.

	Intel Pentium 4 3.80GHz	Intel Core2 Duo E4600 @ 2.40GHz	Intel Core i7-5960X @ 3.00GHz
Socket Type	NA[2]	LGA775	LGA2011-v3
CPU Class	Desktop	Desktop	Desktop
Clockspeed	3.8 GHz	2.4 GHz	3.0 GHz
Turbo Speed	Not Supported	Not Supported	Up to 3.5 GHz
# of Physical Cores	1 (2 logical cores per physical)	2	8 (2 logical cores per physical)
Max TDP	65W	65W	140W
First Seen on Chart	Q4 2008	Q4 2008	Q2 2014
# of Samples	52	406	371
Single Thread Rating	824[3]	888	1993
CPU Mark	**493**	**1395**	**15999**

[1] - *Last seen price from our affiliates NewEgg.com & Amazon.com.*

[2] - *Information not available. Do you know? Notify Us.*

[3] - *Single thread rating may be higher than the overall rating, thread performance is just one component of the CPU Mark.*

Graphic 2, s. Attachment

3.2 Solution through more cores - Multi core usage

In the actual consumer and handheld Software only few programs using more than 2 cores. Mostly graphic programs, games or other programs with heavy workloads. Most of the time the hardware is limited by software so big improvements aren't really necessary.

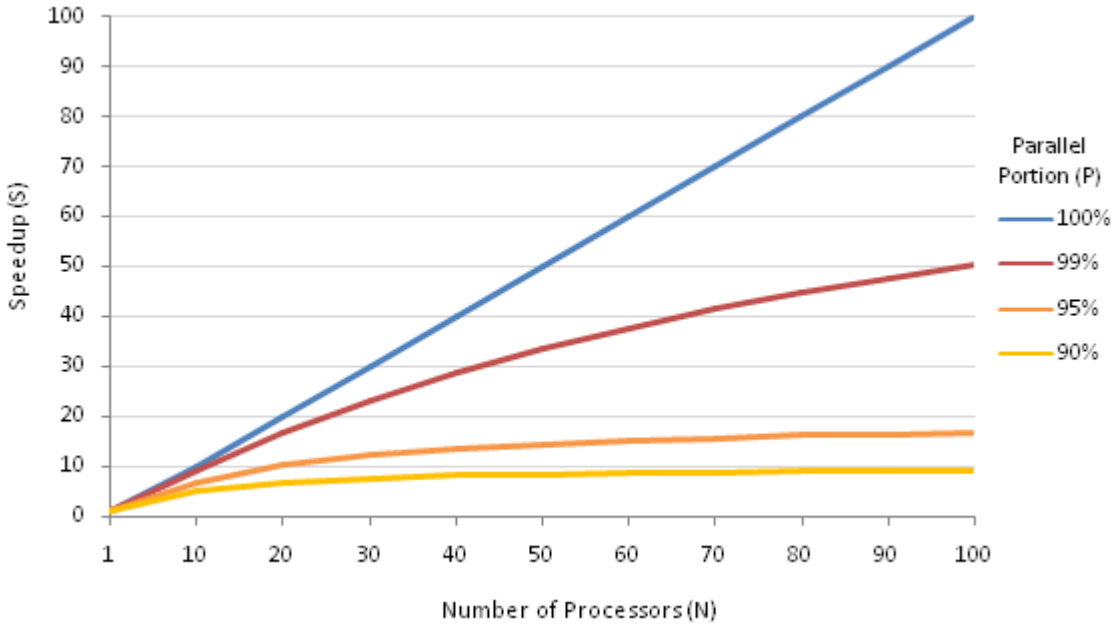
in HPC usage the programs are limited by the hardware and with more power they could calculate faster and more problems at the same time. But to reach the real hardware limits, the software has to be optimized so really every core is at the limit. this optimization is very complex and still not very common. Even with a perfect optimization there is a limit of parallelization which is named in Amdahl's law.

3.3 Solution through more cores - Amdahl's law

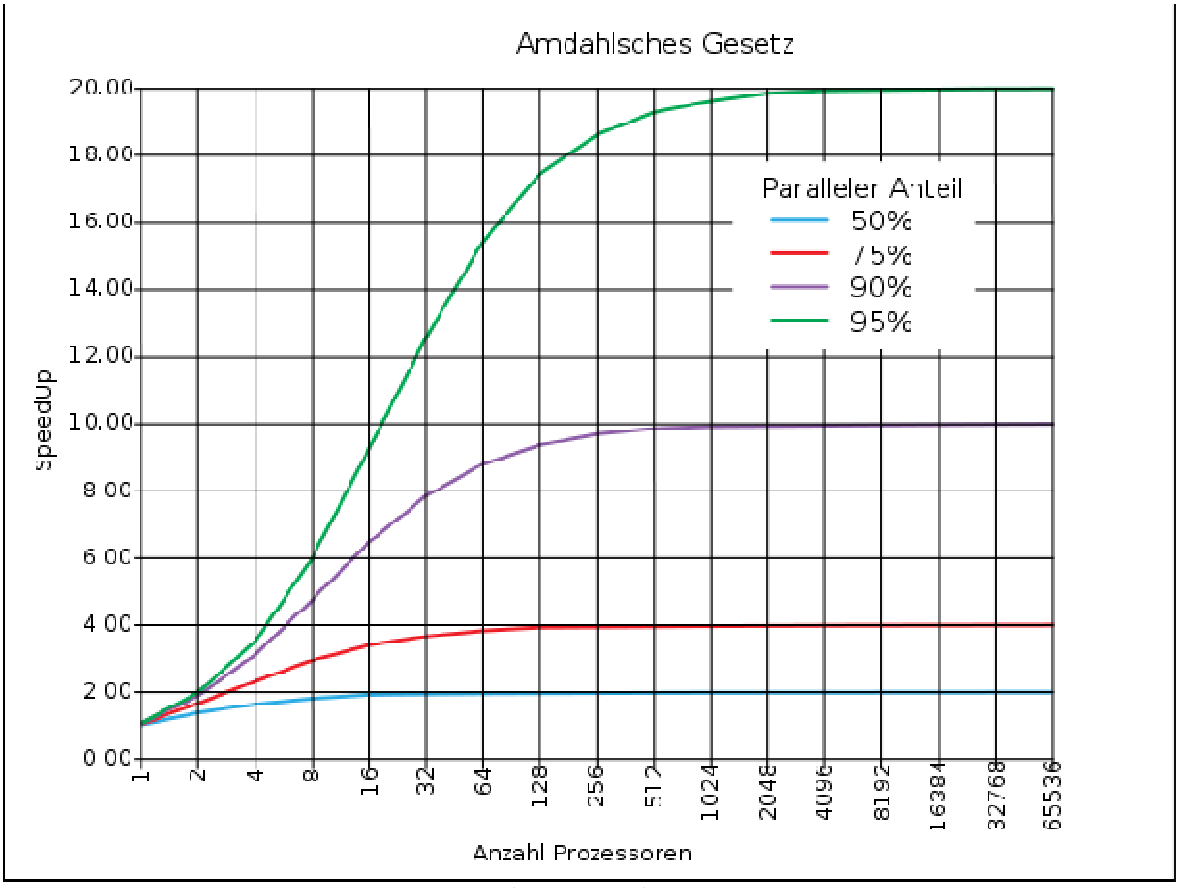
$$\eta_S = \frac{T}{t_S + t_{O(nP)} + \frac{t_P}{n_P}} \leq \frac{T}{T - t_P}$$

ns = speedup, T = running time(RT), ts = serial RT, tp = parallel RT, np = cores, to(np) = synchron time

By Amdahl's law a programs parallelization is depending on the single threaded parts for synchronization at start, end and possible areas between the parallel parts. The graphics 3 and 4 are showing the graphs how more cores would influence a program with a possible parallelization of 50 to 100 percent.



Graphic 3, s. Attachment



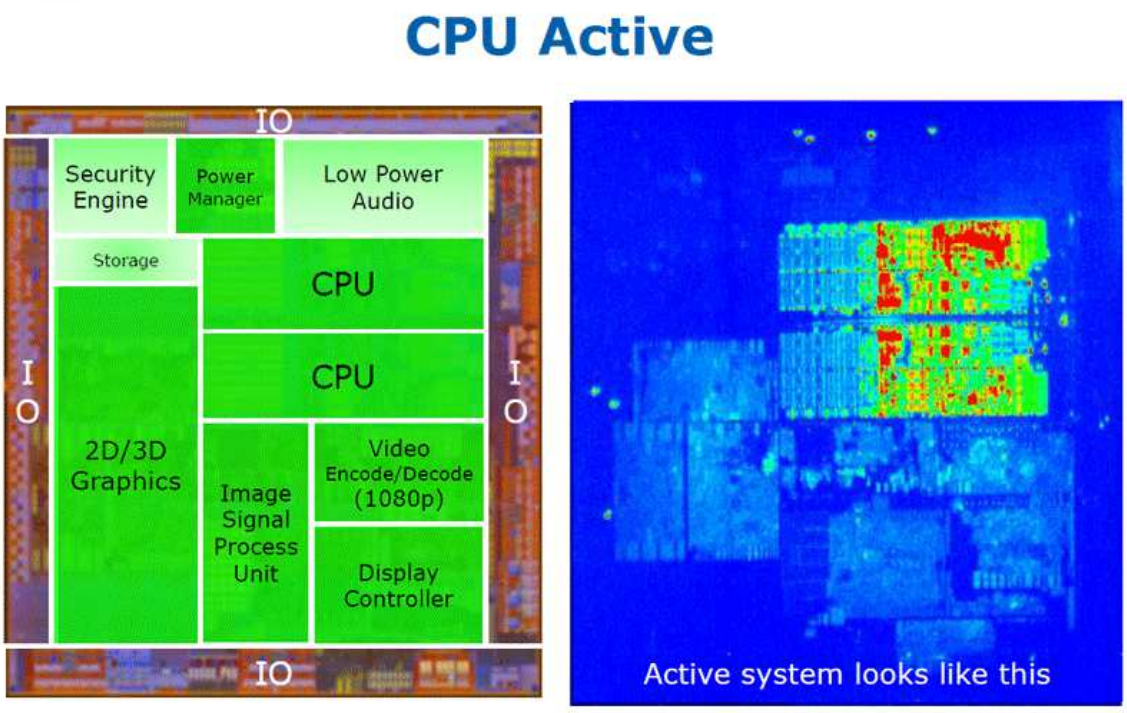
Graphic 4, s. Attachment

One possible way to improve the parallelization of a program, is to split the serial and parallel parts and achieve parts with a possible parallelization of 100 percent for a perfect scaling. But often there is a synchronization which prevents this option, so not every program can be optimized with this technique.

4.1 Problems and barriers - Hot spots

Another problem for further hardware improvements are the physical barriers, in this case the heat waste and the occurring hot spots. Because of their structure, a DIE has some parts with concentrated computational units and some with supporting units like cache, audio and graphic controller which are less used and clocking very low.

In the high clocking computational units, there are very small hot spots which are difficult to cool down, because the heat can't dispense fast enough. The smaller the structure gets, it gets more difficult to cool this spots. Multi core processors aren't different from single core CPUs, if one core is clocking very high, its small area will get exceptionally hot.

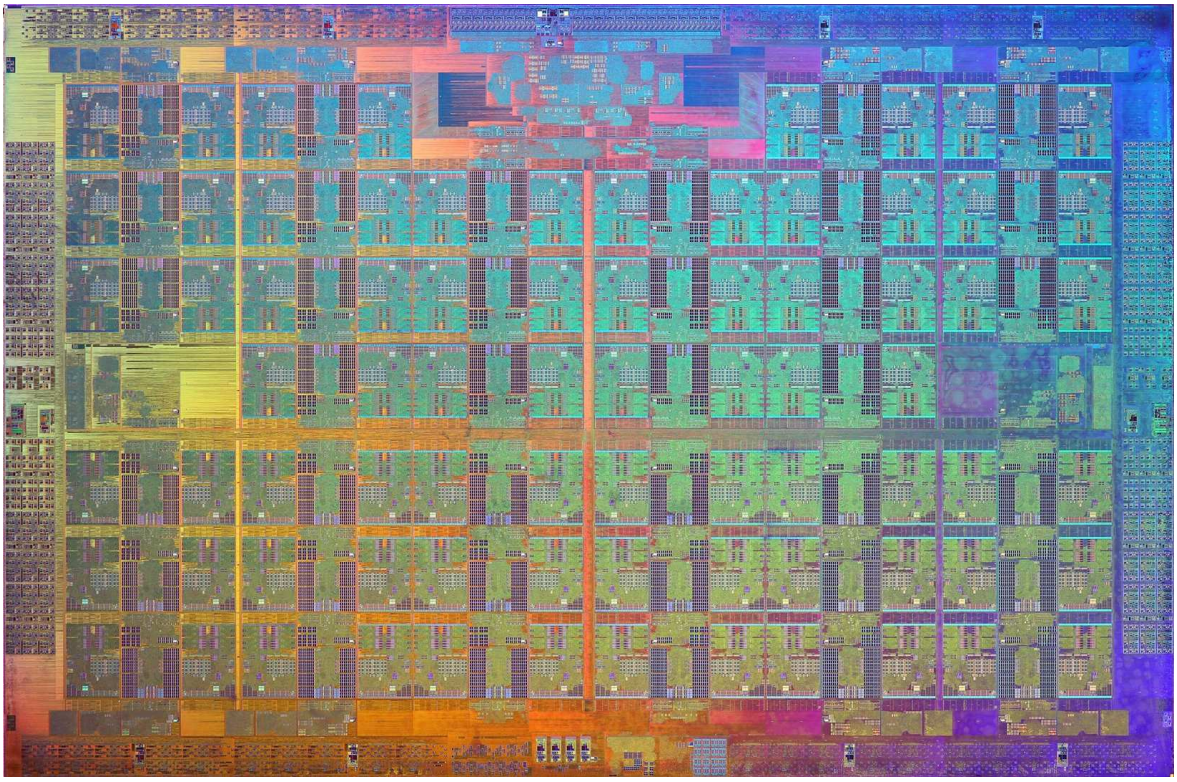


Graphic 5, s. Attachment

4.2 Problems and barriers - DIE size

The size of a DIE is one of the most important parts, if the number of cores or the size of cache should be increased. Even with a new architecture every part of the processor needs some of the limited space.

To get more space on nearly the same size of a DIE, for example the new Xeon Phi Knights Landing with 700 m², the structure has to get smaller. The Xeon Phi is fabricated with a 14 nm Lithography and Intel achieved to place 76 cores with cache and registers on it.



Graphic 6, s. Attachment

But with smaller structures the hot spots will be even smaller and also with lower frequencies the heat generation at this hot spots will be impossible to cool. Without new cool fabrics the frequencies will have to go down more and more, which will decrease the possible performance even with hundreds of cores.

5. Fabrication and approaches for new technologies

For more performance there are many ways to go and one of it, is the specialization of cores for daily tasks so they will achieve more performance with very low frequencies because of their optimized architecture. The problem is, their can't be specialized cores for every daily task, even with hundreds of cores.

The solution for this problem would be reprogrammable simple CPU parts, which could be optimized for personal use. Beside this special CPUs there still would be the typical high clocked universal CPU, or even a dozen of them, to compute every other task.

Another possible way to improve the performance with smaller structures, is to increase the heat dissipation with new cooling materials. There are researches about viscous masses which are directly connected to the CPU and can spread the heat much faster than copper or aluminum could.

5.2 Fabrication - New materials for higher frequencies

Beside heat and small structures we are at the physical limit of silicon. The clocking can't be increased much more, without needing much more voltage, which leads us to the cooling problem again.

With new semiconductor materials like indium, germanium and gallium arsenide, it would be possible to run the same frequencies with nearly the half of voltage, what would decrease the power consumption much. This way the frequencies could be raised over the actual limit of round about 5 GHz, so we could maybe double the frequencies with air cooling.

6. Conclusion

In conclusion there are many interesting and viable ways to improve the performance of our computer.

Inventing new semiconductor materials for lower voltage and less power consumption, beside new cooling materials for faster heat dissipation, are viable solutions and both lead to an improvement of higher frequencies.

Another way would be the specialization of hardware for special tasks and with reprogrammable CPUs it would be possible to optimize it for most applications.

But the optimization of hardware can't be fully used, if we don't optimize the software. Programmers have to learn to write their programs for parallel CPUs and later they need to know, how they program the reprogrammable CPUs for their needs. This is one of the most difficult parts increasing the performance of hardware.

Sources

http://www.extremetech.com/extreme/166413-post-post-pc-the-new-materials-tech-and-cpu-designs-that-will-revive-overclocking-and-enthusiast-computing

https://de.wikipedia.org/wiki/Amdahlsches_Gesetz

https://de.wikipedia.org/wiki/Mooresches_Gesetz

https://software.intel.com/en-us/blogs/2009/06/29/why-p-scales-as-cv2f-is-so-obvious

https://en.wikipedia.org/wiki/Tianhe-2

https://en.wikipedia.org/wiki/List_of_CPU_power_dissipation_figures#Desktop_processors

http://lwn.net/Articles/483889/

http://physics.stackexchange.com/questions/34766/how-does-power-consumption-vary-with-the-processor-frequency-in-a-typical-comput

http://superuser.com/questions/988453/why-multi-core-processors-producing-less-heat

http://superuser.com/questions/543702/why-are-newer-generations-of-processors-faster-at-the-same-clock-speed/906227#906227

Graphics:
[1] , [2]:
http://www.cpubenchmark.net/compare.php?cmp[]=1081&cmp[]=937&cmp[]=2332

[3] : http://blogs.dnvgl.com/utilityofthefuture/wp-content/uploads/2011/12/Amdahls-law1.gif

[4] :
https://upload.wikimedia.org/wikipedia/commons/thumb/e/ea/AmdahlsLaw.svg/800px-AmdahlsLaw.svg.png

[5]: http://www.extremetech.com/wp-content/uploads/2013/09/big_soc-cpu2-e1379614854509-640x403.png

[6]: http://pics.computerbase.de/6/9/0/1/0/1-1260.jpg