Parallel Computer Architecture
Hemangee K. Kapoor
Department of Computer Science and Engineering
Indian Institute of Technology Guwahati
Week - 01
Lecture - 02

Lec 2: Multicore Revolution

Hello everybody. We are doing the first module in this subject, Introduction to Parallel Architectures. This is lecture number 2, where we are going to see the multi-core revolution. We saw in the previous lecture, the need for a parallel architecture and we ended with saying that the hardware design is now driving towards multi-core. We have multiple cores and the software should become more parallel because of the technology constraints. So in this lecture, we are going to understand why did we go towards multi-core.

So we will start this topic with understanding Moore's law. We all know Gordon Moore, who is the co-founder of Intel. He predicted in 1965 that the transistor density of the semiconductor chips will roughly double every 18 months. It will double every 18 months and this graph is showing the trend of the Intel processor chip revolution starting from the 4004 processor up to the Pentium and so on.

Each generation is utilizing the more and more transistors which are getting packed on the same chip. So Moore's law is giving us more transistors and they are being used to design processors which are more powerful and computing fast. So if I generalize this Moore's law, instead of saying transistors double every 18 months, I am going to say that some factor X doubles every month. So what is the every 24 months? What is this X? It could be the computer performance. I can say the clock speed that is the frequency doubles, the computer performance doubles or the number of transistors on the chip doubles.

So there is a factor which doubles every 24 months and we are enjoying this improvement at a constant cost. The cost never increased. So we got more and more transistors at the same cost as the previous generation. However, this is not true for many years, but we will first concentrate on why I had a constant cost with more and more transistors. So this happened because of the concept proposed by or understood because how Dennard explained it.

So Dennard in 1974 observed that the voltage and the current are proportional to the dimension of the transistor. So if I have a transistor of this size, it is going to consume some amount of power and if I shrink this transistor to a smaller dimension, if I make a smaller transistor, it is going to consume proportionally lesser voltage and proportionally lesser current. So the voltage and current are proportional to the dimensions of the transistor. So as the transistor sizes shrank,

because if we have more transistors and the technology is improving, I am able to design smaller devices. So as the transistor shrank, the voltage and the current required to power them also reduced.

So if I quickly give you an example, the power equation is, this is the dynamic power, it is alpha multiplied by C V square f, where C is the capacitance, V is the voltage and f is the frequency. So that is the dynamic power consumed by a device. So what are we saying? As the transistor sizes shrank, the voltage and current reduced. So the V came down and the transistor sizes shrank. So I would say this impacts the capacitance because capacitance is related to the area.

So capacitance is proportional to area. So if I reduce the area, the capacitance goes down. So my C is going down, my V is going down. And if I say that I am going to provide you the same constant power, what would happen? The implication of this would be that the frequency will go up, given the constant power. If my frequency remains the same, I will save on the power.

So power is proportional to C V square f. And as we shrink the transistor sizes, C and V go down. It implies that either the power goes down or if I keep the power consumption same, I can have a higher frequency. So the conclusion is, as the transistors shrank, the voltage reduced and the circuits could operate at a higher frequency. And the transistors have been continuing to shrink generation after generation.

So you would think that, well, this is going fine and it will go on forever. But no, the end of Dennard scaling came and this is what we will see in this particular slide. So on this slide, we will see why we cannot have the same power consumption and keep on increasing the frequency. This graph shows you the device size, that is the gate length on the x axis and the power density which is watts per centimeter square on the y axis. These triangles are the active power which is the dynamic power and these diamonds are the sub threshold power density which is the leakage power.

So this is the leakage power and that is the dynamic power or the active power. So for bigger size devices, we had higher active power, but the leakage power was very low. So, we had a low leakage power and a high dynamic power. As we reduce the gate length, so as the technology shrank, we came on the side, we kept on moving here and as we move here, the gate length reduces and because of this, the leakage current increases. So, the leakage increases as we go from right to left.

The active power is almost similar or rather it is not changing drastically, but the leakage is reducing. And as we shrink devices further, you can see this point here where the leakage power has crossed the active power. So, even if I am not using the device, the leakage current is

dissipating and I am consuming power without even using the device. So, as we shrink the transistors, the power density has started increasing and it is not coming down because the leakage is increasing. So, what happened because of this, I could not dissipate so much power.

So, if this power is dissipated on the chip, it will increase the heat and I need to dissipate the heat out to cool down the IC. So, I was not able to do this and hence the feature sizes have not scaled down further since 2004. We cannot reduce it beyond a particular limit. If I draw a window here, so beyond this, I could not reduce the feature size because the leakage current became unmanageable. So, because the voltage could not be scaled, the power could not be maintained constant.

I could not reduce the feature size, I could not shrink the transistors further, the leakage power was more and hence the power density kept on increasing. So, the power could not be maintained same to keep the same performance. What did we say in the slide related to Moore's law? We said that I was getting better and better performance at the same cost. But now this slide says that your cost is going to increase, your power consumption is going to increase if you are having more and more tinier transistors. So, I cannot have the same power budget anymore.

And if I have more power consumption, it is going to dissipate more heat and hence the heat density is now reaching unsustainable levels. And this has created the so called power wall which is limiting the growth of the system that is what is it limiting? The frequency at which I can operate my ICs, it has almost saturated to close to 4 gigahertz since 2006. Another aspect or another view of the same concept is given on this slide. So, this is showing the power consumption as a function of the threshold and the supply voltage. So, this is the x axis is the supply voltage and y axis is the power consumption.

So, if I have higher supply voltage that is I have a larger transistor, a higher supply voltage, then the dynamic power which is green in color, so it is superimposed on this blue, so the dynamic power here is very high. So higher supply voltage consumes more dynamic power, but if you see the orange line which is the leakage power, so orange line is the leakage power. So this is very low here, it is almost negligible. But if I reduce the supply voltage that is I shrink the device, as the supply voltage reduces my dynamic power that is the green line reduces, but if you see here in this window here, the orange line goes up, so my leakage current increases. So overall the blue line is the total power which is the sigma of the dynamic and the leakage power.

So, you will have more power consumption even for tinier devices and you will have more power consumption even for larger devices. So, you have to have the best design you would have to sit somewhere in this range to have a minimum power consumption. So, we need to balance the device size and variety of other factors to have a low power design. Another motivational figure which most of you must have seen is the power density in Intel CPUs. So

here, year after year the transistors kept on increasing because of Moore's law and we were able to design better and better processors.

We had starting from 4004 up to Pentium. So here we were able to manage the power staying below the limit or expected limit where the temperatures were almost little below a hot plate. But if we had continued to increase the clock frequency for such devices, the heat dissipation will be such that the chip temperatures will rise to that of a nuclear reactor and further to a rocket nozzle. Definitely unrealistic scenarios because to dissipate a cooler device at this temperature would be very difficult. Another representation of the same concept, here I have shown the power consumed by each of these ICAs.

So if you see here up to Pentium, we were below the heating plate limit but beyond this before Pentium prove we have already surpassed the heating limit of a, surpassed the heating plates temperature. And to Itanium 4, we are good enough but beyond this if we had increased the frequency, so here if the frequency goes up, my temperature will go up. So frequency increases, temperature increases. So we cannot increase the frequency further. This is another popular representation where the green line, this green line here is showing the transistors trajectory that is it is increasing every 18 months and so on.

So Moore's law is following the green line. Given these green line number of transistors, what do I do with them? I was using them to design better and better computers. The dark blue line is showing the clock speed that is the frequency was increasing as I shrank the devices, got more transistors in the same power budget, I was able to give more faster clocks. So the dark blue line was showing that the clock frequency increased. So as the clock frequency increased, the power definitely increased.

But what happened is beyond this limit, I had to stop increasing the frequency because I could not dissipate. So we could not dissipate the power, dissipated at that particular frequency. So the frequency capped and the to control the power dissipation. And because of these two caps, the performance per clock also almost stopped improving. Our target is given the green line which is ever increasing, can I have my purple line following the same trend? But we are not able to do this beyond these years because of the problem with heat.

So heat density is increasing because faster the clock, more the power dissipation. The clock speed is not increasing beyond that. And then I have declining benefits in instruction level parallelism. The more transistors we got were utilized for designing better pipeline processors, better out of order execution, and we were able to harness the parallelism inside a sequential program. So this was happening for several years, but it reached a limit where we were not able to further exploit the parallelism in the application because it was just not there.Ok.

So what do I do? The last generation single core ICs were probably over engineered. We had lots of ILP exploitation happening there, but beyond the limit we could not exploit it further. We had integrated lots of logic and consuming lots of power to harness the ILP parallelism, but again it was not there in the applications beyond the limit. So there were declining benefits in ILP. The heat density was saying that you cannot have faster chips.

And we also started having yield problems, that is if I have a complex IC design and if after post manufacturing and post fabrication if it does not work, we cannot sell it to the customer. So the overall yield reduced as the design complexity increased. But if you think if I had a parallel system where I had small tiny multiple cores manufactured and the yield of such an IC would also not be so good, but if n number of cores out of m are working, then we can definitely use it for marketing. So I am giving you a quick example here. For example, the IBM cell processor, it had small 8 cores.

I developed small 8 core system. After fabrication, I realized that 2, 3 of the cores are not working. So I can sell that same IC for lesser number of cores at a lesser cost instead of throwing it away. So the yield problem also can be solved if I have multiple cores. So heat density, declining benefits of ILP and yield problems have forced the designers to do a multi-core.Ok.

So the same concepts are shown here. But to this diagram, now we can see one more black line added, which is showing that the number of cores are increasing with each generation. And if you see, the trend is now going to follow the transistor increase. In the previous slide, we saw here we saw that the green line was increasing, that the transistors were going up, but the purple line which was of my interest. So the green line was increasing at the constant rate, but the purple line which is the performance per clock that was almost stagnated. It capped at a certain point, but I want the purple line also to follow the green line and you can see that I can do this provided we go multi-core.

So here the number of transistors is this top line. So if you see, the number of transistors are increasing steadily and with the multi-core evolution now we will also be able to catch up with the number of logical cores increasing at the same rate with the transistors. So we will be able to catch up with the number of transistor increase. Ok. So the Moore's law can be now reinterpreted that the number of cores of the chip will increase or double every two years instead of the number of transistors increasing, I am going to increase the number of cores per two years. The clock speeds are not going to increase, they will possibly decrease.

And when I have multiple cores, we need to deal with systems of multiple threads or concurrent several there will be several concurrent threads running on a system which need to be utilized properly by the software. And with this we need to deal with inter-chip parallelism as well as

intra-chip parallelism. Some examples of multiple cores which are now coming up. So from here we can start at the power 4 and the Pentium D.

These are two core processors. Then we have core I7 which is a four core processor. Moving up MIT-Raw and the Sunrock are 16 core processors. So these are 16 cores here, these are four cores, two cores and then the tile 64, the 64 core processors. These are examples of multi cores coming into the market. And if you see here the Moore's law is giving us transistors per chip.

So the green line in my previous slide and the multi core revolution, they are almost overlapping with each other. So the black line is the number of transistors and these small circles are showing the various multi core designs which are coming up and they are almost overlapping with each other. So we have met our target of utilizing the transistors which Moore has given us. Some real life examples of multi cores. So a quick look or peek at the history is the Intel 4004 processor which is the old Uni core system developed in 1971.

It was a 4 bit microprocessor, just over 2000 transistors and developed at 10 micron PMOS technology and the size of the IC was 11 mm square. Coming to today's core i7 IC which is using the NEHLEM architecture. It is a 4 core processor instead of a single core processor. It has 48 bit virtual addresses, 40 bit physical addresses. There are 4 cores and imagine the number of transistors.

We had 2000 transistors in 4004 whereas here we have 730 million transistors. So see the big jump which has come over all these years from 1971 till 2015 or so. And the chip size also has increased to 263 mm square and the technology from 10 micron to 45 nanometer. So this big gap has, we have come a long way and these more and more transistors are not only giving me multiple cores but they are also helping me with more cache or more on chip capacity. So each core would need data and we need more cache closer to the processors to execute them.

So this IC has got an 8 megabytes of L3 cache. In addition to the L1 and L2 caches they are definitely there. This shows some details of the complex architecture of the core i7. I won't go into the details. If you are interested you can, this is covered in processor based subject but you would be, if you are interested you can look up the information. Moving on, this is the Intel Skylake picture along with the cores.

Now we have, we also have an on chip graphics processor integrator. So imagine apart from the cores, the L1, L2 and the L3 caches, we also have a graphics processor integrated with this. So the transistors which were available are being used nicely in newer features. This is a picture of the Tesla V100 GPU, 5000 CUDA cores, 640 tensor cores and see the amount of memory it can support. So this is the amount of memory and the bandwidth it can support and the number of transistors is 21 billion and it fits in a 800 plus mm square chip.

So all these multiple cores which are there made available to us because of the number of transistors, it's the job to handle them properly, be able to write softwares which can utilize all these tasks and so on. So with this advancement of having multiple cores available, what is next? So we will just take a quick look at what happens next beyond this. So beyond this is 3D integration. So as the word says 3D, we are going to have silicon ICs packed one on top of the other.

So we are going to glue the ICs together. So this is one IC, this is the second IC and this is third IC. So with this pink which is the adhesive layer, we can connect these ICs to each other. They will communicate through these through silicon vias which are the connectors buried interconnects. So we have buried interconnects which will connect these two ICs and if I manufacture a computer chip using this, how will it look like? We can have the microprocessor that is the CPU as the first layer.

The first layer could be the CPU layer. The CPU needs data from the cache. So the caches can be integrated on top of this layer as the second layer. So the SRAM which is the cache is glued as the next layer. The SRAM would need data from the main memory. So the main memory follows and you can have multiple layers of these main memory.

So DRAM 1, DRAM second layer, DRAM third layer and so on. So this could be possible setup for a 3D computer chip and if I just say that I have this here, we can have multiple such chips connected to each other with an on-chip interconnect. So we have on-chip interconnect which will connect all these stacks of ICs with each other giving us more compute facility and better performance. So another nice picture beyond computing if you want beyond basic computing facilities if you want to do more for example you want more processing that is RF, ADC, DAC units to be integrated, nano devices and MEMS to be integrated, more sensors, biochemical sensors. See all these different applications would require different fabrication technology.

So we cannot do everything on CMOS. But if I have a method of stacking one type of technology on another type of technology, so if I can glue these together and they can work more closely then we can have a better system. This is another picture a 2D layout of a system on chip translated to a 3D integration. So overall we have multi-cores, so we can have several large cores. So we can have several large cores. We can use some of these by making a heterogeneous system where some big and some small cores are there.

So these are all heterogeneous examples. We can have several small cores, an example could be a GPU or even tinier accelerator systems. Then we can have several floating point cores, then 3D stack memory and what is this giving us? It is gives us better and better computing for

our scientific, business, games and home entertainment and variety of things. So our variety of applications which we saw in the previous lecture, all of them can run if I have such a system available. So the question is not whether this is going to happen.

The question is whether we are ready to handle this in a better way. So hopefully during this course we will learn how do I handle better systems and in a course on parallel computing or parallel programming you can learn how to write better programs to harness this hardware. Thank you.