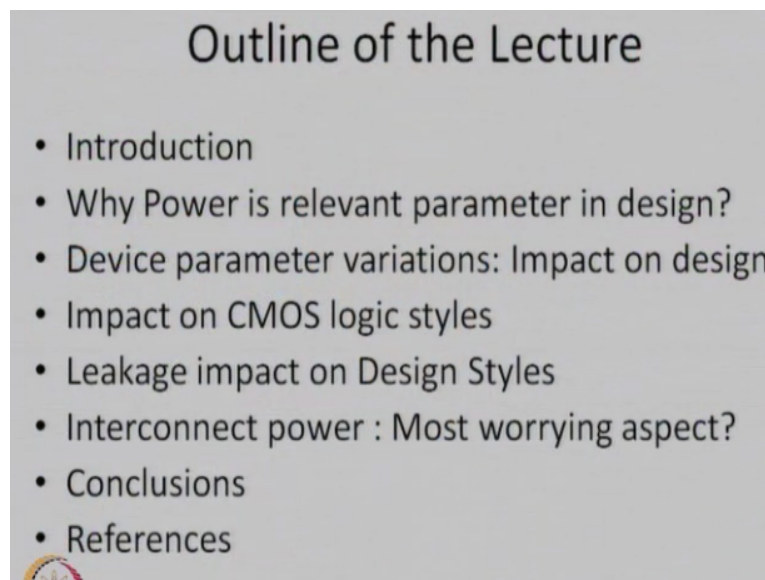


Advanced VLSI Design
Prof. A. N. Chandorkar
Department of Electrical Engineering
Indian Institute of Technology – Bombay

Lecture - 06
Power Estimation and Control in CMOS VLSI circuits

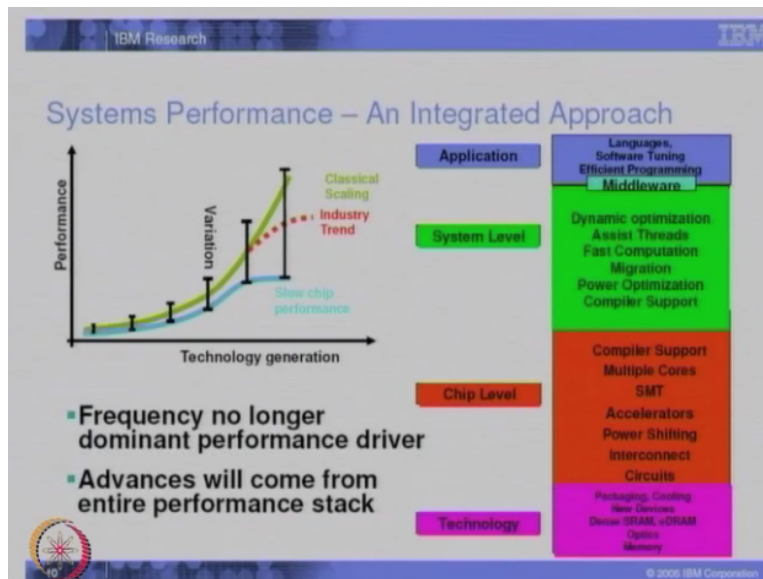
We start with a new topic today, power estimation and control in CMOS VLSI circuits. The idea behind power estimation will be very clear to you. But before that let me tell you what I am going to talk in this lecture. I have following points in which I will elaborate this, not necessarily one lecture it may be more than one lecture may be two or three lectures, but today we will start on that line.

(Refer Slide Time: 00:50)



We will start with introduction, then we will see why power, what is the relevance. Then we will talk about device parameter variation and impact on design. And then we also talk about the impact on CMOS logic style, and then the leakage impact on design styles. We will look into very important aspect which is worrying for most of the designers is the interconnect power and then we will conclude.

(Refer Slide Time: 01:14)



Now this slide is from IBM which shows a system performance of any electronic system performance and the way they said it is called an integrated approach. So if you see from the bottom side we can see there are things we are giving different colours, technology, chip level, system level, applications. The basic for those study is that if you look at the performance and we look into the technology ingestion over the years.

For example, as I said we started with technology way back in 1960's, we went for SSI, MSI then we went to LSI, VLSI, now we are continuing ultra large scalar called advanced VLSI circuits. So if you look at the technology generation nodes from the technology side and we look the performance it clearly shows, what was expected by the scaling theory, which is the green colour and that was expected as the technology start improving.

That is from nodes of 19-113 nanometer 19 nanometer down to say 28 nanometers down, this was expected performance. The industry is that the reddish one you know which is recently one sees, it is already tampering down now, it is not improving. And if you see the worst performance of the chip which is much below, which is called as low chip performance. So frequency no longer now dominant the performance of the driver, because the technology s given by something else.

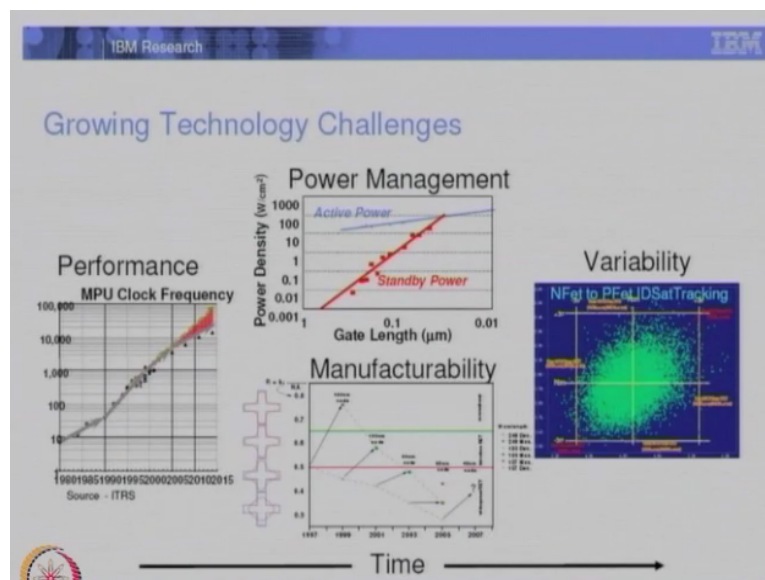
The advances will come from entire performance tag rather than only frequency of the circuit or frequency of the device. So if you look at the right side you can see that the technology essentially is one is talking about memory options, we are talking of which memories, then we are also talking packaging possibilities, we are talking of cooling, we are talking of at the

chip level compiler support, we are looking for multiple cores, SMT accelerators, power shifting, interconnect, circuits.

High level we may look for dynamic optimization, then we are looking for threads, fast like multi-thread circuits are very common these days, program is divide into multiple parts and multi-processors actually took multiple threads, and the circuit can be made much faster. Then there is a faster computation, migration, power optimization, compiler support and then finally at the application level we are looking for languages, software tuning and different technology programming or performances.

So if you see, as you go from technology level to application level, there are different ways in which optimization can be actually attained.

(Refer Slide Time: 04:07)



Now what is the challenge of today, one can see from here, this is another IBM slide which has four important problems in the VLSI design or VLSI system realisations. One is the performance which we have been talking years and years after speeds. So this is a micro processor speed clock frequency, which is as per the ITRS source the frequency against years, which is shown here to 2018 or something.

We are expecting something like 10,000 giga hertz kind of approach or more than giga hertz of approach and if you look at the other parameter is called power management and this is where this whole lecture is going to be. So I, other day also said you that there are three kinds

of circuit, which worries as in design. One of course is we say active, power related to active when the device or circuit is performing.

And the other like mobile phone we have a power which is called standby power. So if you look at the gate length versus delay, one can see from here that the active power increases as the gate length decreases from say one micron down to 100 nanometers, or if you go further below the active power will keep on increasing. However if you see the standby power, the slope of standby power with the gate length production is so steep that sooner by you know already we have crossed a level when the standby power is larger than the active power.

And around the so many watts per centimetre squares say 100 to 200 watts centimetre square at this point by typically around 0.04 micron technology one finds that this standby power is almost equal to the active power and that is more worrisome for us. Because if you reduce the device by scaling channel length to say 25-22 nanometers, one can see the active power on power may not really increase drastically.

Because the slope is not very high, whereas if you look at the standby power by then, it will actually overshoot so much, that the standby power will be larger than the active power, which essentially means that even if we are not performing, your battery is draining the power. So this is one of the major challenges of today. The third problem which we are worried about is the variability.

The way it happens that the no process or no design or no technology can control the variability in the processes. Since we are now becoming a sub 90 nanometer technologies, the variations in thresholds, due to the variation in process parameters on current, off current, transconductors, which are the basic parameters of a mass, they themselves vary. For example, in the case of Fin Fred, one sees that if the spacer thickness is different the on current will be most affected compared to any other this.

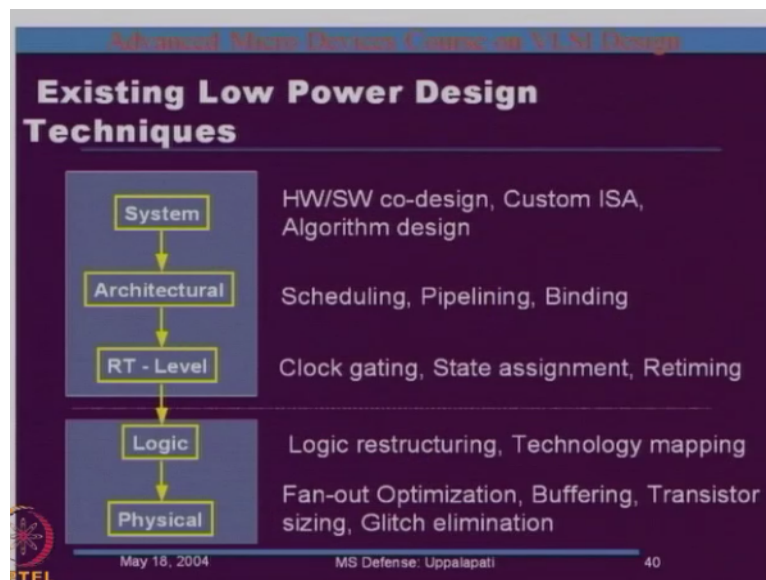
So one is worrying now that if there is a variation in the process itself, which leads to circuit parameter variations, then the chip performance will be extremely varying and if that so happen, one does not know how to design, because some devices which may run faster some may be come slower independent of what design one does. And therefore this is another issue of worry when you start designing the advanced VLSI circuit.

Now the fourth if not the, if not directly related to design area as per say, but which is worrying us most is manufacturability. So if you scale down your features from, as I said 90 nanometer down to say 22 or 16 or 11 or 7, the major worry as I see and everyone is seeing is the lithographic problems. The pattern which will become shown on the left, I mean I actually printed on a silicon number of times as the marks may increase to 30 marks actually tape manufacture.

So it is not transferred as the rectangles or the pluses or whatever symbols you make, or whatever a shapes you make and if that happens for a transistor, if the W by changes the whole performance of the circuit will change. So to transfer an image to accuracy of what in sub 90 nanometer processes has become another issue of variability. So these are the growing challenges, they are all inter connected to some extent through device and process.

However, this part of the course, which I am talking is more interested to know about power management, though as I keep saying that one cannot say each three of the other ones are not affecting the fourth one. Everyone is affecting the other one and hence the issue. Now what is existing low power design techniques.

(Refer Slide Time: 09:09)



So one can see from here if you look at the flow diagram of a chip design. We start from a system which is essentially co-design, hardware, software co-designed customize is and here we design algorithms. Then the next level of the hierarchy is architectural where essentially

we are controlling or we are trying to optimize things based on scheduling pipe-lining and binding.

And third which is the most popular hierarchy level from where most of the circuits are now-a-days designed from their IP's is RT level, register transfer level and once you know the IP, which have RTL code, then probably all that you need to know there is to start getting state assignment, real timing and things of that kind. If you go, this is essentially one group, which essentially we say more on the software side, the whole that is RTL architecture system is more on the software side control.

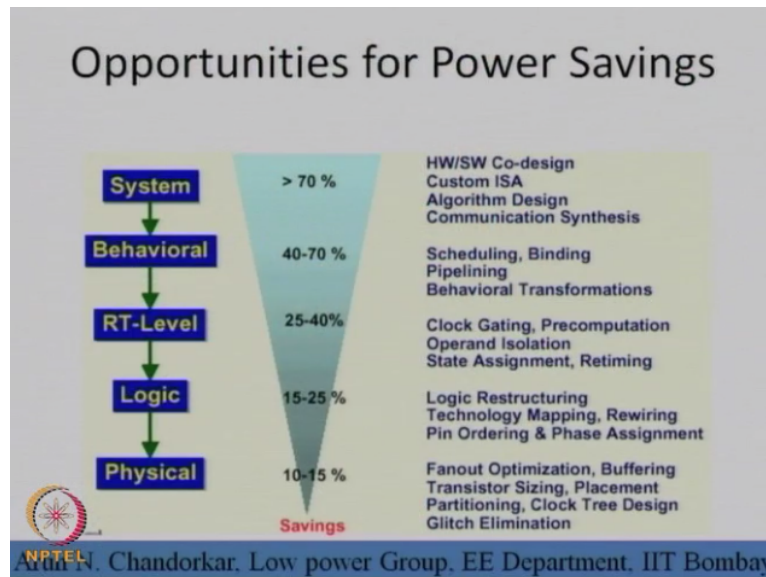
Though we have to always do hardware software co-design, but essentially still it is more software design. If you go down the second tangle below, one can see from here the RT level can then go to logic and essentially we are trying to see that any digital circuit consist of some kind of a logic blocks connection, and you can see the power probably can be affected or changed for loop or by logic restructuring or technology mapping.

And finally, the logic is consisted of the transistors and transistor design requires intruding the circuit design and then the layout and re-estimation from layout the logical performance or circuit performance. Now these essentially means you are there optimizing fan-outs, buffering, transistor sizing, and also eliminations. Now, one can see from here that the different level we are actually working differently to reduce the power.

We being more from the say circuit side, our always attempt has been to reduce power using physical and logical block restructuring or optimizing the transits themselves. However, the other part of this course surely assuredly will talk about the upper part of this from RTL to system and they will also look into the pipe-lining, binding, scheduling, clock gating, state assignments, to see how to reduce the power.

This is another standard graph available on the net, standard figure where do really can say we are power. There are just now gave you the hierarchical level from system to physical.

(Refer Slide Time: 11:57)



So of course this figure is slightly neat to be modified now because this was essentially available for 90 nanometer designs. But this numbers may slightly modify for lower than 90 nanometer nodes. One can save power 10-15% of the power by just transit sizing proper fan-outlet, that is using logical effort, putting or placing them properly, partitioning properly and how to distribute clocks and certainly how to minimize the glitches.

If you can do this, then probably one can save up to 15% of power using what we call physical design. Now that is basically transistor level design where the optimization can be tried, circuit level design. Then the next level, next 10% percent of the power can be minimized at the logical level, where you can actually can decide as we saw, you can have a branching, you can change the gates, you can logically restructure.

You can then actually map the technology and you can say some parts are critical so you need not have higher threshold or you need not have lower thresholds. And also we can have rewiring pin ordering and phase assignment can be done. If we do that another 10% power can be minimized. Then at the RT level or the RTL by properly optimizing the gating, clock gating, precomputation operation isolation and state assignment and retiming blocks.

Probably we can save another 15% of power, very large amount of power can be actually minimized at RT level. So by then the first, last three hierarchy, probably you can go up to 40%. So one can see that even with all the effort we do in the discussion we do hell of it in the design and we keep showing the slides about the physical, logical, RT level designs, the best of approach even there can get you around 40% of optimization on power saving. The

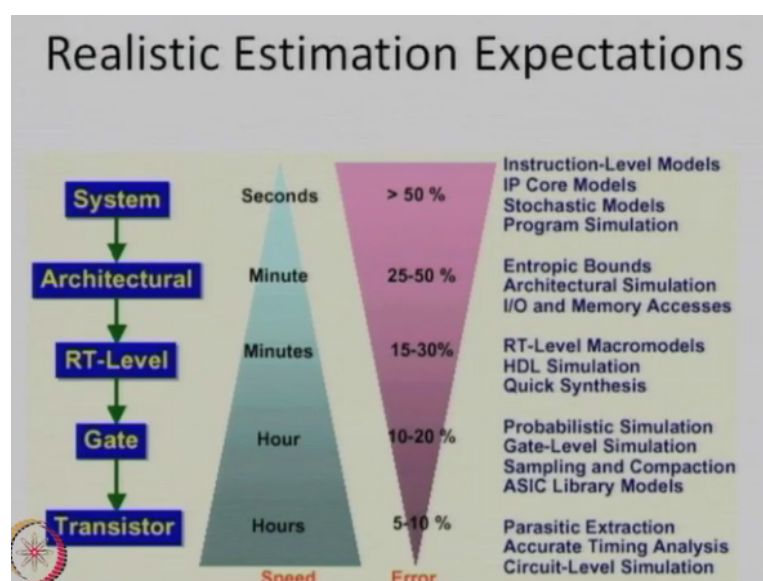
remainder power can be really worked out well by at the high levels of hierarchy in design, which is behavioural and system design.

In behavioural as I said scheduling the task, threading the task, and binding them at given time, then pipe-lining and looking for performance using behavioural, how much data to be required when and how to minimize that. So if you look at the behavioural representation of the logic you are implementing or system you are implementing, some of those standard methods can be used to reduce another 30% of the power, which is very large power reduction compared to the first, last three one which only could add up to 40%.

And to your great amusement the system level design can be particularly algorithm-based designs or used customized designs available and using more SOC or SIP kind of technique using hardware co-ware designs, hardware software co-designs, probably one can do even the remainder of the 30% or at least another 10-15% of power can be minimized right there. So when I say I have opportunity of power saving, I must realize that there are different places of in a system design where power can be minimized.

However, as being what I am, and as being most of my colleagues for the design course which is more circuit oriented from our side, we will continue to work mostly for the last three one, RT level, logic level, physical level design approach where the power can be minimized. I am not saying that the other two are not relevant, but this course probably will not, because this course is taken care by the other people in the computer science area.

(Refer Slide Time: 15:57)



So one can again see, time required to optimizations, this is called inverted matrix system. So if you look at this for example at transit level possibility the parasitic extraction, accurate timing analysis, circuit level simulations and it may take hours and it may actually add to 10% of error in calculations, in designing and it takes hours actually, actually logic circuit level or transit level design is very long, it has to run poise or spectre or models kind of equivalent of that and then it takes hell of a time if the circuit is large.

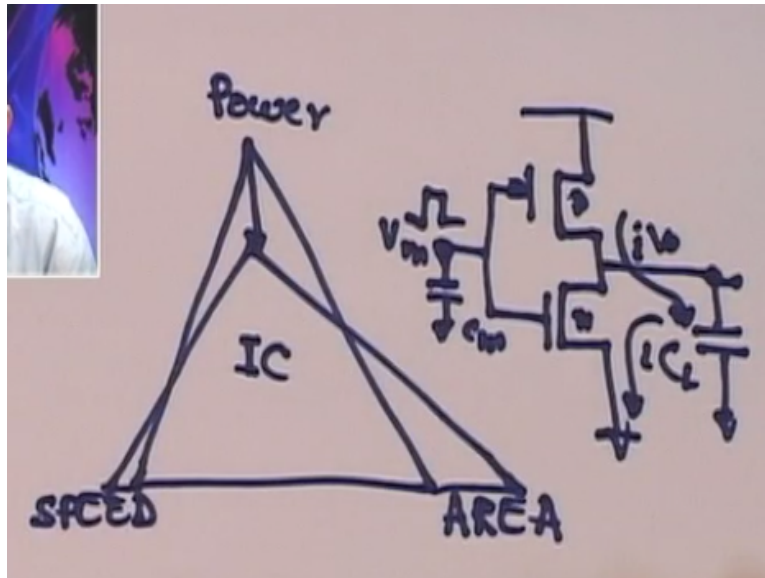
Then the next level is at the gate which is probabilistic simulation, gate-level simulation, sampling and compaction, ASIC library models. By putting these probably, you can do power estimation much faster in little hours, and the error can be around 10-20% up to 20% error with theses two together can lead to. At the RT level things can be done in minutes and you can do macro models.

You can do HDL simulation, and very quick synthesis can be done through actually do RT level design for power optimization. But you may lead to another 15% of error in all your designs. The fourth architecture you have a entropic bounds, architectural simulation, IOs and memory accesses, they are decided at architectural level. And if you particularly look for entropic minimization.

I think another 10% of it may do very fast calculations for you to reduce power. But will take 25% error bars on that, because is mostly statistical thing goes there and it is very difficult to control exact mess on that. The finally at the system level you may do very, very fast, okay, instruction level models, IP core models, stochastic models and program.

If you have all these the realistic estimation expectation can be made with but possibility since you do very fast with thinking, quick analysis, you have largest amount of error possible in your design. So having told that what are the challenges for system people and having told that power is the major criteria of design.

(Refer Slide Time: 18:14)



If you see our earlier courses, I have in my earlier VLSI design course, I used to draw a triangle, this triangle has three corners, one of them is essentially what we say power, the other is speed and the third is area. So when I am designing a simple IC, integrated circuit, I see that this form power, speed and area forms a triangle, what triangle essentially means. So if I am optimizing anything.

Let us say I want to minimize the power, so obviously what can I say if I reduce the power down, the speed means actually delay. So it may actually lead to larger area or lower speeds or area being same or on the contrary if you want to improve speed that means if you want to reduce the delay, then you may require much higher power to generate. And if you still want both power and speed be faster, power be minimal, then probably you may at least have to give larger area on the chip.

That means the number of chip per area will be smaller which means the cost of chip will increase. So essentially this optimization is also not very, very simple, in most cases we will have to trade of, is okay this much power and this much speed is possible at given package area or given chip area and that is what the best designs probably could be. And this optimization of course we have been trying.

And particularly for mass circuit this can be understood, at least the power speed can be understood very quickly from use. Let us say I take a case of a CMOS inverter, this a P channel upper device, this is N channel device and it has an input capacitance of C_n and I am

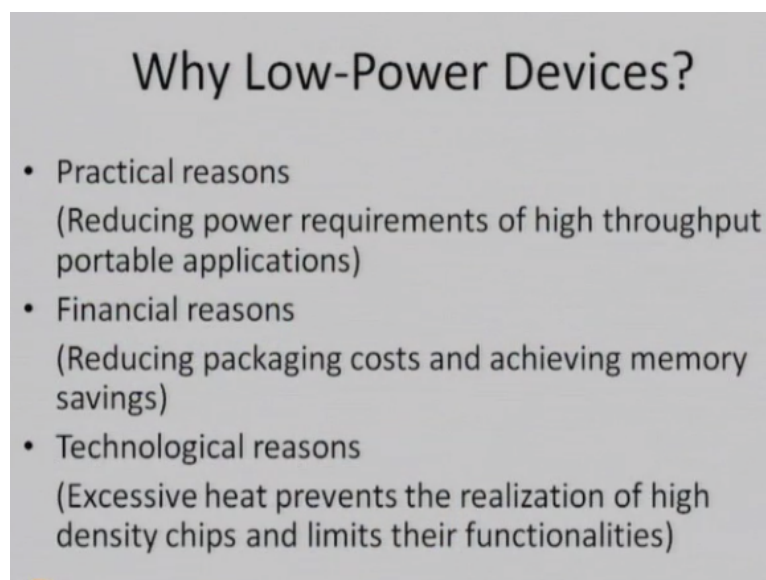
putting an input V_n which a step and I am picking up an output voltage V_0 across the net output capacitance or load capacitance here.

So basically what we are saying the speed is decided by charging and discharge transition as we looked earlier, that is time taken to charge the capacitor and time taken to discharge the capacitor, propagation delay essentially for input to go the output is $TPHL$ plus $TPLH$ by two everest delay. To improve this $TPLH$ or $TPHL$ we say actually increase the currents and if you want to increase the currents for discharge and charging transition, these devices must provide you that large currents.

But the current in mass transistor is proposition to W by L for a given silicon technology and given technology node where it is fixed. So we figure it out that the only way to improve currents in design is to increase W by L . But if I increase W by L essentially I am saying larger area, and if you have a larger area the input capacitance will also increase. So we figure it out that by increasing the current, which essentially be improve the speed we actually landed up in larger area.

But apart from area if the currents are larger in a circuit and as we shall see later, because power is essentially proportional to current, so larger the current dissipation larger is the power dissipation. So attainment of this higher speed was at the cost of larger area and at times more due to larger currents and at times larger area.

(Refer Slide Time: 22:00)



Why Low-Power Devices?

- Practical reasons
(Reducing power requirements of high throughput portable applications)
- Financial reasons
(Reducing packaging costs and achieving memory savings)
- Technological reasons
(Excessive heat prevents the realization of high density chips and limits their functionalities)

So when I design a circuit or a chip I am really looking to optimize such that I do not give too much of a power, but I still improve speed and I do not have to increase my area, if I can do that, then what is I have done an optimal design. So as I say our ultimate aim to say zero watt per centimetre square power density, speeds of infinite hertz and area probably zero which is the ultimate values one probably can think. But how best we can achieve three together is the effort in all circuit design and also the system design.

So we start with the first thing we look at the power from the transistor side is coming from the device, okay, so why loop our devices will require in our designs. So the fact is there are practical reasons. What are those practical reasons, reducing power requirements of high throughput portable applications, because you know if you have a portable like PDA, if you have a mobile or any of this I-pads or I-pods or tablets you are carrying them in your hand and there is no addition power supply coming from anywhere.

So internal battery itself is essentially trying to give the power, and you also want for example these days new tablets or smart phones will show you some slides on that, they are doing multi functions too many functions including video, audio, camera and also digital normal processing, internet and of course telephone at the end of the day, and we are reading books out of it.

And we are doing all kinds of processing on a tablet or a smart phone and many other phones as well. So since your throughput is very high and it is portable and the battery is limited, we are looking for device which will reduce power, that means the battery can last longer before it is recharged. The financial reason is very important at the end of the day we all look for money so here is the problem.

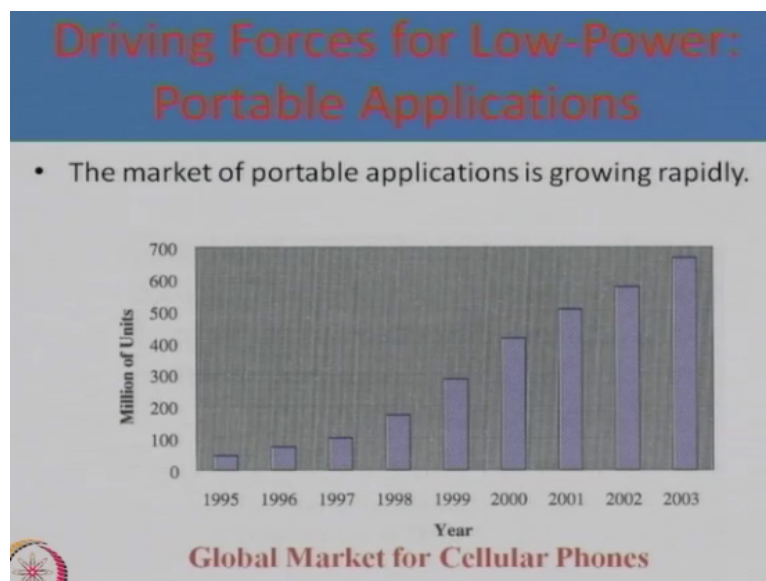
The larger the power dissipation you do the package has to actually dissipate that much power, and because of that, since the packaging cost goes higher with the number of pins and the size, because to manage it is not simple. So financially it is not very great thinking if you have a larger packaging cost for to get that extra power which you are looking or we need to reduce the power simply because the packaging cost we want to reduce.

Also if you do not have larger this, we can also achieve memory saving, because you do not have to keep in data, because something else has to be done in between. The third which what

most of us believe is that reason of thinking, for us, is the technology reasons. For example, if you have very high density appearing on any chip being used, you can see the chip has a larger thermal dissipation around.

The temperature of the silicon areas may increase and many of the functions may not remain those functions because of the ill performance of the device itself, due to heat. So if you look at the portable applications market of portable applications are growing rapidly over the years, of course this is the old graph from Intel.

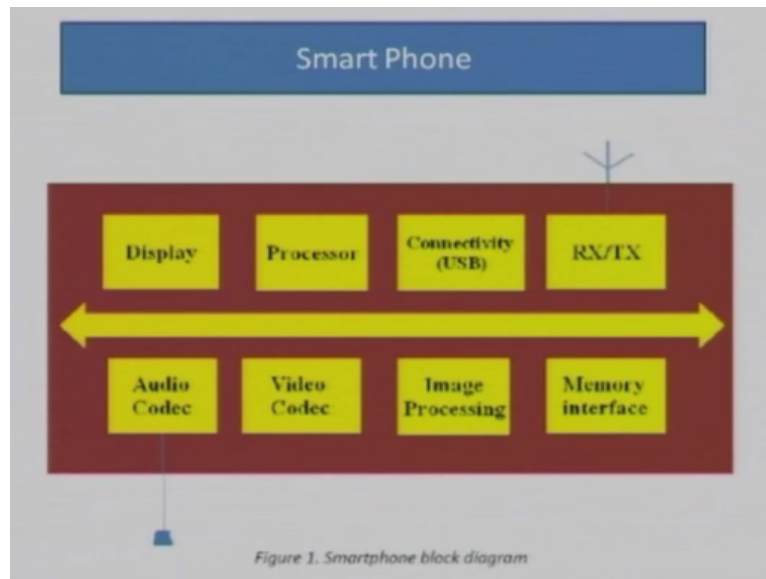
(Refer Slide Time: 25:33)



One can see 95 to around 500 millions per units, today in 2000 this is more than 2000 millions per units are required for cellular market. And India itself has a largest cellular market, of course density wise not that high even now for the population. But the number of cellular phones are really much higher than even US market as of now.

And therefore you must observed that most of these cellular companies are trying to buy, or to enter Indian market and the result of course is well known, everyday news paper is just talking about it.

(Refer Slide Time: 26:12)



So here is the smart phone, just look at the smart phone blocks inside. If you look at the system, of course there are not just eight blocks some of them are included in this but general major blocks in a smart phone are the following. There is a display which all of you see as, which the centre one is called bus structure. So we have a display, then we need a processor to do all the jobs, okay, so we have a processor.

Then you need to have connectivity externally so you are looking for USB connection, your system can be connected to any other internet or any other drives or anything which is the USB connectivity. For that is some like connectivity something different from some of the smart phones like I told between Samsung and I-pad or I-phone, that is a major difference of I-pad and Samsung tablet that is the difference, tablet has a USB connection.

Then we of course need a receiver transmitter at that frequency operation, because that is what the cellular data is coming or going. So you need from the antenna connection to receiving transmitting system. And these are of course online high frequency things going on, however there is other area where data has been to be processed, for example you have a lot of memory interface.

Because some data has to be stored, some data has to be temporarily stored before it is passed on like you have SMS messages going on, and you are receiving in between. They do have an interface between, stopping between mail for this SMS and phone together and you need data to be cached somewhere, and there are lot of memories interfaces required for the data to be stored temporarily or permanently.

Then there are blocks like image processing after all every these days have a camera and even if not camera you have internet which may have a figure most of the video pictures or video. So you need to process that data and therefore image processor has a part of any smart phones these days. Then to get this audio, video data in a proper format which can be done processed, we need an audio codec and a video codec.

So if you look at the any smart phone block diagram these are the major electronic blocks, which constitute the working of a smart phone.

(Refer Slide Time: 28:47)

Smart Phone chip Specs.

	On	Txt	Phone	PIM_acc ess	Camera	Play_b ack	Game	Standby	Off
PD_proc	1.2V	1.0V	1.0V	1.2V	1.0V	1.0V	1.2V	1.2V	OFF
PD_TX	1.2V	1.2V	1.2V	0.8V	OFF	OFF	OFF	1.0V	OFF
PD_Disp	1.2V	1.2V	0.8V	1.2V	1.2V	1.2V	1.2V	OFF	OFF
PD_img	1.2V	1.0V	0.8V	OFF	1.2V	1.2V	OFF	OFF	OFF
PD_Rest	1.2V	1.0V	0.8V	OFF	0.8V	1.2V	1.2V	OFF	OFF

Figure 2: Power domains and power modes for the Smartphone chip

These are typical specification which smart phone gives for a specialist. For example, we have processor PD processor, we have typically this on then is 1.2 volt battery is required, text is 1 volt, phone is 1 volt, PIM access of course 1.2 volts, camera requires 1 volt supply, playback require 1 volt, game requires 1.2 volts, standby requires 1.2 volts. And in off mode, we do not want any power, okay.

This is a power requirement shown here, power dissipation I am talking about. Power dissipation in the receiver, processor of course is the first one, in the case of transmitting-receiving system you have power supply requirement on when 1.2, text you need 1.2 supply and for the PIM access this, you need smaller power is 0.8, camera is not on when you are transmitting or receiving a data. Except when standby you still need 1 volt supply.

If you look at the display almost except for the case when the, you are listening phone you do not need display because you are not seeing the display. And therefore that can be stand out to 0.8 volt supply. But the all other places except for the standby and off mode you want 1.2 volt supply. Now if you look at power dissipation image processing chip chord, except for the processing access of the internet everywhere else image processing requires 1.2 volts.

There is no game when you are doing, already you have data stored so you do not need power in standby or off state. And if you are at the rest of the power dissipation, except for the PIM access and the standby, this will require constant power camera, when you are actually doing this camera requires this much lower power because you may not actually operate it, but you want to use in between any data to snap and transmit you need some camera be on.

So these are called power domains and power modes for a smart phone chip. So for any system, for example this is the system I choose to explain is smart phone, but you make any other system, you will have to first design that system level this and then break it into architectural at least sub-system blocks and then you will have to choose architecture to implement this, and from there probably you will be able to get such a power domains and power modes for a chip and then start looking that okay.

Since you require power supply sometimes you do not require power sometimes so the circuits which will operate in some cases may not have a larger power dissipation required, in some cases there are larger power and therefore corresponding power dissipations can be managed.

(Refer Slide Time: 31:51)

Power status in different operating Modes

Fully functional mode, where all the power domains are ON and working on full VDD

A texting mode where the image processing and the rest are turned to a low voltage

A phone mode when the transmit block is fully ON

A PIM access mode when the image processing is turned off

Camera mode when the image processing and the display units are ON

Playback mode when the display unit is fully ON

Game-playing mode when the processor and display units are turned ON

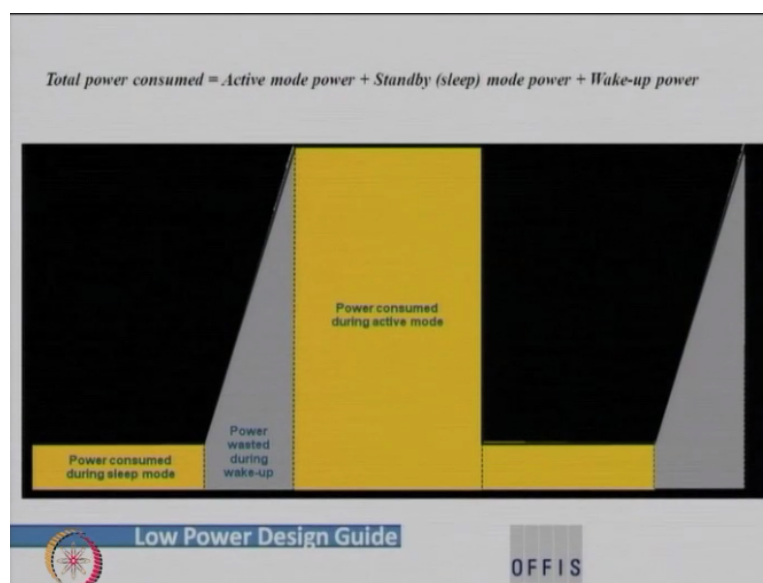
A keep-alive mode when processor is ON and the transmit unit is in low voltage

OFF mode when everything is turned OFF.

I already said I may repeat again fully functional mode of a smart phone where all power domains are on and working full VDD, a texting mode where image processing and the rest are turned to a low voltage mode. A phone mode when the transmit block is fully on. A PIM access mode when the image processing is turned off. Camera mode when the image processing and the displays are on. Playback mode when the display unit is fully on.

Game playing mode when the processor and display units are turned on and the rest can be switched off. And a keep alive mode, that is the standby mode when the processor is on transmitting is low and is still on for any time, and of course off mode means you actually switch off your mobile phone.

(Refer Slide Time: 32:41)



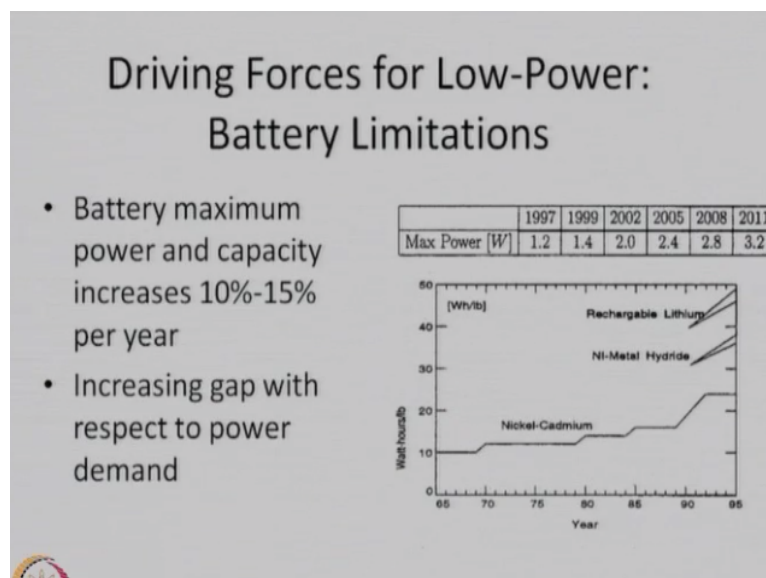
Here is another interesting picture of the same, you know many of the circuit is come back later again, when the device is in the sleep mode there are three basic parts where power is consumed. When you start a device that is called wake up power mode, when it is operating all the time we say it is active power mode, and when you are not fully switching it off but keeping it in standby mode, then it is called sleep mode power.

So if you look at typical low power design guide, this is a figure which gives you some idea where the power is going to be controlled. So for the sleep mode, the power dissipation is rather smaller and can be managed relatively simple way. The second power consumption comes from is wasted in the wake up and this is transition kind, you suddenly wake up all circuit have to turn on from their off mode.

And they actually have a much more transition power dissipation during and it does not finish very fast or something, though externally you feel you have only milliseconds or something of that kind, there are nano seconds, but actually they take milliseconds and the power dissipation is not very small. As it is a transition power and it actually takes quiet amount of power which in integration wise will be larger than the standby mode power.

And of course the largest power will be consumed when the device is fully on and among them which for a smart phone for example, not all system will be on but on average some will be on all the time except for the off mode, and in that case the power will be constantly consumed.

(Refer Slide Time: 34:31)



So what is the driving force for the driving force for low power battery limitations. So one says, driving forces for low power battery limitations come from the fact that you know in 97 we have maximum power voltage of 1.2 and in 2012, it is 3 Watts power is required, getting more right now it is 4.5 Watts of power is required currently. So if this is the power requirement, the battery maximum power and capacity increases only 10-15% per year.

Three batteries are shown here, one of course is standard nickel cadmium, people used to say it works hard battery, right from 60s to 90s, it is still working, still available, of course, but you can see the power, the capacity is not increasing very much. It is going from 13-20, 25 maximum.

Whereas the other possibility was nickel metal hydride kind of batteries and after 90s, people have started working on it and we are going from 3 per this and if you look at the rechargeable lithium ones, it is around 61 per this. So what we are trying to tell you that as the year progress, the required battery improvement is still below the power requirement and therefore, there is always a gap between the power demand and power supplied by the battery.

So one basic idea is that can be, if you cannot improve the battery, of course rechargeable will require quick recharging, that is one possibility, which we are anyway using, but the other method probably could be then minimize the power requirement itself and if power and capacity is not increasing with this rate per year, then probably we can have a battery, which can last longer for your work, but that is what I keep saying that the low power circuit is demanded simply.

Because there is huge limitations coming from available batteries as of now. One, of course, you can always say why work on VLSI, why not work on batteries, probably if you get a patent on a better battery, it may fetch you a lot of money and may be it will revolutionize everything ahead. Before we start really working on it, let me tell you what exactly has worked over the years on this scaling along when you start scaling from say 5 micron down to 22 nanometer process, different technology notes.

(Refer Slide Time: 37:02)

What has worked up to now?

- Voltage and process scaling
- Design methodologies
 - Power-aware design flows and tools, trade area for lower power
- Architecture Design
- Power down techniques
 - Clock gating, dynamic power management
- Dynamic voltage scaling based on workload
- Power conscious RT/ logic synthesis
- Better cell library design and resizing methods
 - Cap. reduction, threshold control, transistor layout

We are scaling voltages as well as we are scaling dimensions and therefore corresponding process has been scaled, but as we kept telling earlier in our first history perspective chapter, we said voltage is not scaling with the scale law and therefore fields are increasing even the lengths are reducing and that is our major worry. That is what the reason why power dissipation is increasing.

The second of course, as we still have reduced voltages, we went from 3.3 to 2.1 to 1.5 to 1.2, 0.8, 0.6, 0.5, I think there are some tips available now with a half a word supply. So there are effort going on and to a great extent we are successful. However, voltage scaling below 0.4 or something will be very, very difficult task, because firstly the noise margins will be so low because the thermal noise at room temperature itself is around $4 kT$ by Q or $4 kT$, which means around 100 millivolt is a typical noise even without anything.

And if the temperature starts heating on the wafer, which it will by a higher number of devices on chip. This number may actually be larger and if this total noise power is increasing then the signal power, noise power may be comparable and one will not be able to differentiate between the off state and an on state of a digital logic. Therefore all set and done, voltage scaling may not be the final option, though it is.

Because in every power dissipation, voltage term will appear I into V , so V reduction is definitely going to reduce the power and there is no denial on it. However, as I said it is not scaling as the other factors being scaled and therefore power is not getting minimized to that

new note, which by law of scaling should have. The second method, which has worked in design of a low power circuit is essentially designed methodologies.

We have started looking into power aware design flows and tools and we start reading areas for low power. Instead of speed, we are always trying to trade now on the area, because most of the low power designers do not mind giving little bigger chip, but they certainly need low power. Whereas many other high performance circuit may do otherwise they may say we do not mind too much of a power dissipation, but we want higher speeds.

So we are trading on either of them and for that we are looking for design flows. We are actually looking for everything which is right from system level down to a logic or transition level. We are actually designing what we call power of a design. For example, major work is right now power of a design is to data transfer from one point to the other through an interconnects.

And people are looking for whether this should be driven by current modes or should run by voltage modes and what kind of power can be saved by using part of analog blocks inside. So yes, there is a trig going on to reduce the power by what we call power of a designs. The other of course is change the architecture so that the data required does not require every now and then to flow from different places.

And we will see that architecture and design play a lot of this introducing the power. Then, the another circuit level method, which one is looking into reduce power is what we call power down techniques. We are trying to use clock gating, essentially which means whenever you need a data, you clock the data and do not allow it to work constantly. So that no data transfers, still it is required as if the clock is flowing down and hence at least the dynamic power is minimized.

And of course there are other techniques of dynamic power management, which is being tried using many other techniques, we will see in this part of the lecture. Then, the fifth possibility of reducing, which has worked is dynamic voltage scaling based on work load. This is called appropriation of voltage difference more than 1 voltage, which we call dual VDD or dual VTH kind of approach, in which we can assign voltages.

VDD supply value from say, let us say we are working on 1 volt, so some circuit may require 0.6, some may do 0.8 and some at the best of it may require 1 volt or 1.2 volt, whichever is available to you. So we say power management using voltage scaling can be tried at different work loads. In doing so, we may also actually scale the threshold voltage, which are dual threshold or multi-threshold circuit can be tried.

Threshold can be actually controlled using number of ways, and one we will see how to scale the V_T so that one can do lower dynamic power dissipation. There is a constant effort going on the RTL logic, which we write or do logic censes can reduce the power during RTL. RTL essentially represent the module, which does some kind of a function. So can we see that a function is one simpler function compared to a larger function.

Can we sense such a logic? Using a much different way, which will reduce the net power consumption, because the amount of logic transfers it will do, will be lower. And of course, one simplest way of doing things what people are doing right now, at least using the cell library design technique, we already have predesigned blocks with low power dissipations and design such library functions.

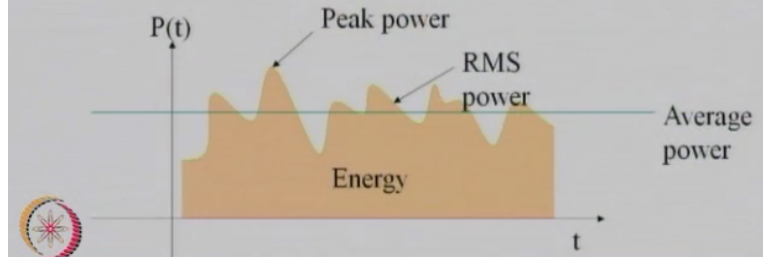
You keep resizing them if necessary, W by L may be changed, which reduce capacitance, which changes the thresholds and also lay-outs on this. Using all these, you can create logical blocks, which are slightly better optimized and you have large number of library functions, which gives you variable power dissipation and corresponding speed and areas. So this is what has been tried or this is what has worked so far.

And based on this essentially, one can say that one will be able to reduce the effort to make the low power dissipated chips and we will continue with this effort ahead.

(Refer Slide Time: 43:47)

Power Metrics

- **Average power:** Related to battery lifetime.
- **Peak power:** Related to reliability and thermal failure
- **RMS power:** Related to cycle-by-cycle power
- **Energy=power × time:** Related to power-delay product.



This is an interesting figure, which many of the designers do not look at it, but this is a system design point of view. Here we are plotting power versus time that is essentially I am plotting energy. So instead of just looking into power, we should actually start looking into energy dissipation. Because that is something, which battery has to provide. So what do we do there, we say there is a line, which say everest power.

This is my everest power and we say this is essentially governed by the battery lifetime. So let us say battery has a capacity of so many ampere hours, and for your cell kit, the net current requirement for different blocks sum up to some currents, so we say that okay, everest power ampere hours if given to you, so how many hours you want the battery to last, correspondingly that much everest power or current should be drawn from the battery.

So once we decide this, this line can be drawn. Now once you say this is everest power to be given by battery for a given time, for a given life time of a battery, then next worrisome problem is peak power. So, some part of the circuit are block at times may overshoot everest powers as shown here and this peak power related to, of course how much maximum one can allow this peak to go is decided by heating up the device.

Essentially leading to thermal failure and there are also reliability issues. I am not sure whether we will be able to talk technology here as much larger the transient is occurring, peak power is occurring, there is another worry right now going on increase of local power densities, which may lead to what we call electromigration and that may open the interconnect lines.

So there is a huge rare issue if the peak power is very large. The third point of interest is the RMS power, which is related to cycle by cycle power. Now this RMS power essentially is the power, because if there is a variation in the power requirement of a chip, then one should worry about not just the average power, but we should also worry about the RMS power for every clock cycle, how much power is really delivered is essentially very crucial.

Because that will decide the net power dissipations. So we can design a circuit, which has low RMS power requirement and you can design circuit, which requires higher RMS power requirement cycle to cycle. So some cycle may require larger power, some may not and therefore, one can probably manage low power by actually adjusting an RMS power every clock or at least number of clocks if not every cycle clock.

Finally, as I keep saying at the end of the day why we always worry so much about power, one should really worry about power into time, which essentially means that we should worry more about power delay product rather than either speed or power. So if you want larger speed with lower delay and you want a low power, we are now trying to say, if you want to minimize energy, essentially you must reduce the power delay product.

And if you can do that, your battery will last longer. This is essentially because please remember battery supplies energy and this is stored by the process of battery itself ionization initially and therefore, how much energy you consume will decide the battery life and therefore, the effort in the low power design should really actually shift to low energy designs in many cases, which may be more relevant parameter.

Now if you look at, this is a Hayes paper on appearing way back an American scientist appear general.

(Refer Slide Time: 47:49)

Power Consumption at the System Level

- 500 MHz Pentium III with 17 in. Monitor: 150-200 W
- Server: 300 W
- Mainframe: 10-20 kW
- Author's home office: 2 CPUs, 2 monitors, laptop, 3 printers, scanner, plus misc. peripherals
 - Nameplate rating: 2.4 kW
 - Max power (incl. peripherals): 700 W
 - Typical usage: 150-170 W
 - Average over 10 days: 77 W (9% of total consumption)

from B. Hayes, "The Computer and the Dynamo", *American Scientist*, Sep Q 2001

This is a computer and diary, nice interesting article. If you happen to read this, September 2002. Mr. Hayes had given a very interesting comparisons there. Now he is talking about let us say, you can see since it was a old one, so it is a Pentium 3. He is talking about the 500 mHz Pentium 3. The system has a PC, which we are talking is 17-inch monitor, 150-200 Watts power it consumes.

Server requires around 300 W power. So this Pentium chip may require around 150-200, server may require 300 W of power. Mainframe may require 10-20 KW of power. The author means that he has two computers, two CPUs, two monitors, laptop, three printers, scanner, plus miscellaneous peripherals he holds and if you add all these, name plate ratings, you have around 2.4 KW maximum power including peripherals 700 W.

Typical usage is 150-171. If he averages over 10 days 77 W 10 days, 9% of total consumptions. So one can see, you seem to believe that the power even if you are using two PC at home where there is a good fan, so there is no thermal dissipation problem so much, how much power really we are consuming when we are actually using. That means, you are using 150-200 W of power equivalent of a bulb, which is constantly used by you.

And then an average, if you do not use longer time computers, hopefully these days this number is not valid, because almost 90% of the students and faculties, 90% of their time is on the computers and therefore, this Watts probably may be very, very high compared to what 77 average he is talking about, maybe as much as 770 maybe the numbers now. He is of course talking of power consumption in USA.

(Refer Slide Time: 49:58)

Power Consumption at the National Level USA

- All office equipment consumes 74 TW-hours / year (about 2% US total)
- Adding telecoms increases total to 3.2 %
- Power down and sleep modes could save 23-40 TWh
- Technology has dramatically improved power cost of computing
 - ENIAC (1940s): 18,000 vacuum tubes, 174 kW, roughly 10 W / tube
 - Today's microprocessors: 100 M transistors, 100 W, roughly 1 μ W / transistor (would draw 10 GW if no change from 1940s)

from B. Hayes, "The Computer and the Dynamo", *American Scientist*, Sep 2001

We can do similar analysis for India. Again this is 2001 paper, so please take it a little pinch of salt for 2012, but all the same gives the numbers. All office equipments consume 74 TW hours per year, about 2% of US total. Adding telecom increases to 3.2%. Power down and sleep mode could save 23-40 TW hours of power. Technology has dramatically improved power cost of computing. Eniac requires 18000 vacuum tubes, 174 KW roughly 10 W per tube.

Today's microprocessor is 100 million transistor plus 100 Watts roughly 1 microwatt per transistor. Nowadays it is not that low or that high in all processes. Since we can reduce technology dramatically this cost, one can see that even then the 3% of the net power consumption goes into only computing at homes and talking to people. So one has to worry now that when we talk about power dissipation and power when using the energy.

We talk of hell of others where no one talks about computers, no one talks about mobil, but the number right now is so high increasing day and day out that worry may start now that this 3.2% may become as high as 25% sooner and then people will suddenly wake up and say, oh, there is energy clashing, my mobile is not working. So how can we quantify quality of design. This may be my last slide for the day.

(Refer Slide Time: 51:37)

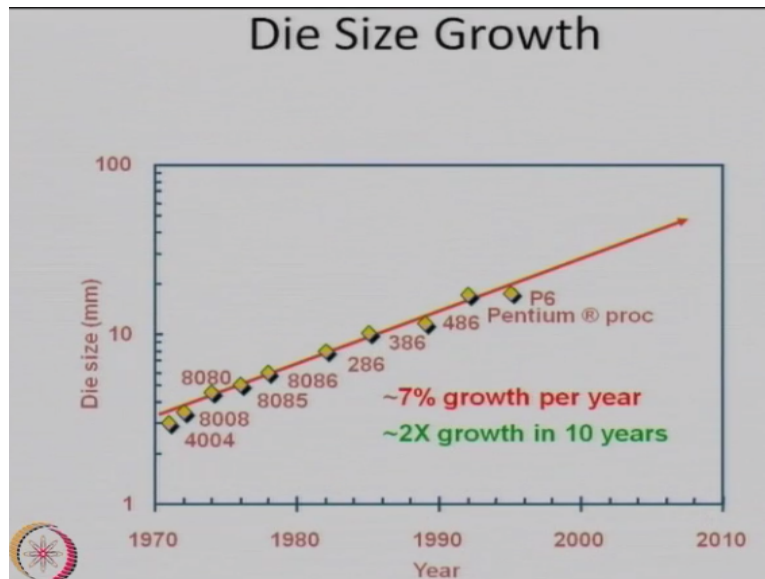
How Can We Quantify Quality of Design?

- **Want to compare designs quantitatively on a level playing field**
 - Two characteristics we care most about are delay and power consumption
 - Delay and power are related
- **CMOS logic operates by moving charge (storing energy) on and off capacitors**
 - Faster energy transfer...
 - ...implies higher power consumption...
 - means faster logic gates!

When I design a chip or when I design a system, want to compare designs, quantitatively on a level playing fields, you know you cannot have two different people designing a two different tools or two different technology and then compare. So there are two things we should have common so that we can compare. Two characteristics, we care most about delay and power consumption.

These are the two specs we must look into and we know that delay and power are related and therefore, first thing you compare in a design on these two parameters. CMOS logic operates by moving charge from one point to the other during on and off capacitors and therefore, we are looking for fast energy transfer, employs higher power consumptions means faster logic gates. So if you are looking for higher speeds, you inbuilt situation is your higher power consumption.

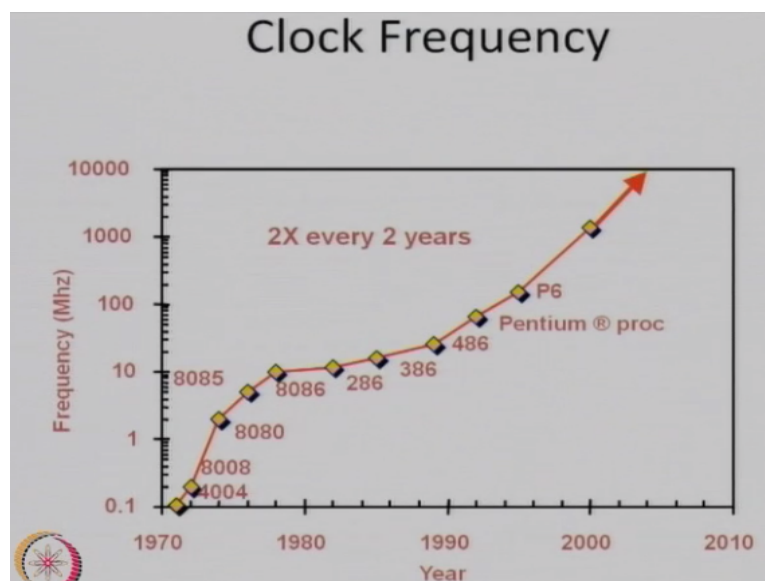
(Refer Slide Time: 52:31)



These are few slides, which we may come back again. This is a old size, the die of the microprocessor has increased size. For example, we started with 2 mm x 2 mm size of this, 2 or 3 now we are talking of 2 cm chips by now. So two or even more P6 is already 2 x 2 cm and maybe the next version quad 4 with P6 on this ethilon has something like 3 cm x 2.6 cm size chips. So there is a 7% growth per year of the size and two-X growth in last 10 years.

Because the die size is increasing, the number of transistors, which are going on them is increasing and therefore the power dissipation is increasing per chip. You can see I keep saying die size always increases because Moore’s Law has to be satisfied, so we are only increasing 7%, but actually one expects 14% die size increase, if we have to follow Moore’s law, that the transistor density doubles every year.

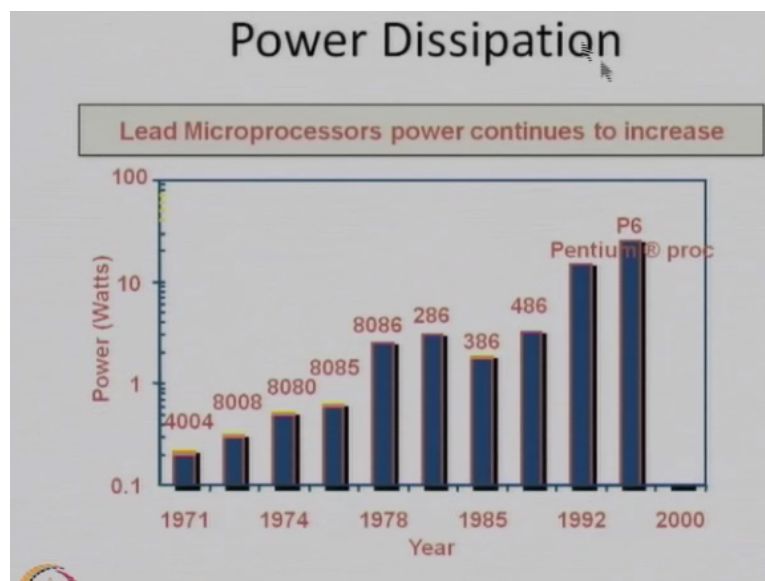
(Refer Slide Time: 53:40)



If you look at the clock frequency, then you can see by 2000, we have already crossed gigahertz, and one believes that we are looking, of course, I am not very sure, this of course is only arrow, but by 2030 or something, one is expecting that we may work on K band 60 KGHz chips. One has to keep your fingers crossed because what power dissipation, they will talk we have no idea as of now.

So lead microfrequency doubles every two years. This is again Moore's law in other form.

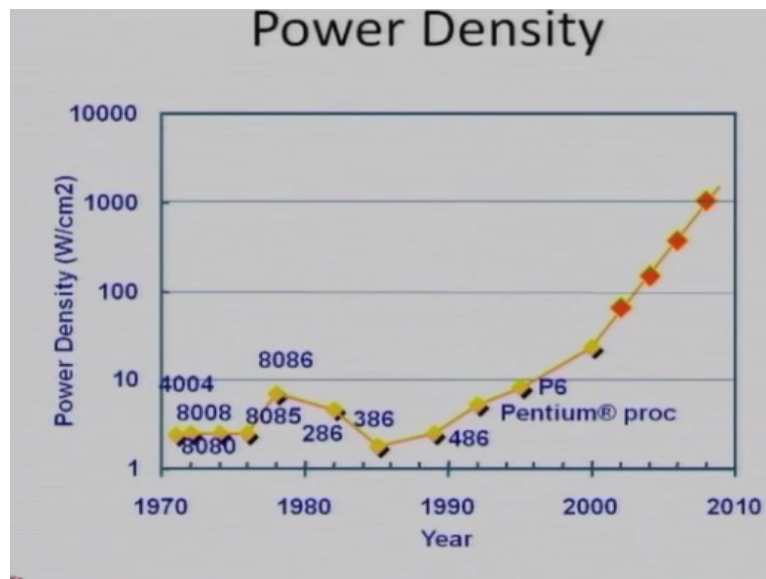
(Refer Slide Time: 54:12)



If you look at the power in microprocessors, 4004, this first microprocessor of Pentium pro P6 old slide from Intel. All that I am trying to say you that it went from say 0.2 W, now we are talking of something like 30 W to 40 W per chip and if you look at the increase in between 80 and 86 there was reduction in power, but then now it is increasing. Power deliver and dissipation will be prohibitive.

If you continue to go by this straight line as shown, then the power dissipation will be extraordinary and here is the last graph.

(Refer Slide Time: 54:51)



Showing that power density by 2000, we already have temperatures, which are equivalent of hot plate. We are reaching a place, which is 200 W per cm square, which is roughly like a nuclear reactive power density and if you continue to grow as we are looking for, it would be 1000 W per cm square, which is like a rocket nozzle temperature power density. So if you are designing a chip with anything, the first worry is how to avoid power densities of these two.


We are already here. We are trying to dissipate this power itself and if you unfortunately reach on this line, one does not know how will you really reduce or dissipate the power and how the cooling methods can be employed. And we remember, higher the temperature, the junction temperature rises, leakage current increases, the concentrations of the PNN also increases, and finally the junction breaks down.

Because once said there is no junction, both side equal concentration of carriers occur. So at higher temperature junctions, device may not operate, so we must keep temperature junctions cool. So these are the kind of power we are going to work on to reduce the device.

(Refer Slide Time: 56:13)

Components of CMOS Power Dissipation

- **Dynamic Power**
 - Charging and discharging load capacitances
- **Short Circuit (Overlap) Current**
 - Occurs when PMOS and NMOS devices on simultaneously
- **Static Current**
 - Bias circuitry in analog circuits
- **Leakage Current**
 - Reverse-biased diode leakage
 - Subthreshold leakage
 - Tunneling through gate oxide



This is my last slide for the day. We will look into CMOS alone right now. There are other technologies, but this course is strictly following as if we continue to work on CMOS chip design. So we are looking for dynamic power, which is essentially charging this charging capacitances. Then we are looking for when the transition occur, P channel turns off and channel turns on and vice versa occurs.

One is off, the other is on, but when they switch over, both are on for a while and that time, we say short circuit current. Then there is a static current particularly analog and in saturated load or unsaturated load and mass inverters are of those kinds, which are mostly popular in HEMT or Galle Watson devices. So there is a constant power supply to ground connection, we say that is the static power consumption.

And particularly like in a bias circuit in an analog, it has to be constantly on because even if you are design is not amplying signal in there, this keeping design at a particular biasing point has to be on all the times. So there is a static current dissipation in most of the chips right now and that has to be taken care. Finally, for last I would say, the major worry of 2010 or above is this power, which we call off power, which is essentially occurring because of the leakage.

We believe when the V_N is less than threshold voltage, then the device is off, but in real life, the threshold definition itself says that when the band bending is actually two times the potential $5F$, then only, but essentially that means, the inversion should occur when the

surface potential is equal to forming potential, but the actual definition says $2.5F$, which means even below V_T , there is a subthreshold leakage, subthreshold current flowing.

And that is the major leakage worry, because the slope of subthreshold slope in the new devices is too high and therefore the leakage is very high. It is not 60 volt per decade or 60 millivolt per decade. That is our major million per decade. So this is major worry for us. The other of course is the source forms a diode with the substrate, so there is a reverse biased diode leakage currents, because of the dopings we are going to use now, this reverse current leakage current should be very high.

And then, there are many other leakages current and one of the top leakage current possibility is the gate oxide tunnelling or it may be what we call girdle gate induced threshold current or the currents because of what we call change in values and things on that kind. So if you see, if I want to reduce the CMOS power in which the static power of course.

One may say is not here, at least dynamic, the short circuit power, and the leakage power essentially occurring because of the static current, leakage current and charging-discharging currents that should be minimized during the process, but when you reduce the currents, we know from our theory that the charging current discharge current is reducing, we will decrease the speed or increase the charging time or discharge time.

So circuit will become slower. So for a given speed, what else can be done to minimize power, this is what in this course ahead we will be able to talk. On all these fronts, what are the ways we can actually minimize the power. Thanks for the day.